Probability and Statistics for Data Science

Carlos Fernandez-Granda

Preface

这些笔记是为NYU数据科学中心的课程*Probability and Statistics for Data Science*开发的。目标是从基本原理出发，提供概率与统计学的基本概念概览。我想感谢Levent Sagun和Vlad Kobzar，他们是课程的教学助理，以及Brett Bernstein和David Rosenberg，他们提供了有用的建议。我还要非常感谢所有我的学生们的反馈。

在撰写这些笔记时，我得到了国家科学基金会的支持，资助编号为 NSF 奖项 DMS-1616340。

New York, August 2017

Contents

1 Basic Probability Theory **1**
1.1 概率空间 . . . . . 1 1.2 条件概率 . . . . . 4 1.3 独立性 7

2 Random Variables **11**
2.1 定义 . . . . . 11 2.2 离散随机变量 . . . . . 12 2.3 连续随机变量 . . . . . 19 2.4 以事件为条件 . . . . . 27 2.5 随机变量的函数 . . . . . 29 2.6 生成随机变量 . . . . . 30 2.7 证明 33

3 Multivariate Random Variables **35**
3.1 离散随机变量 . . . . . 35 3.2 连续随机变量 . . . . . 39 3.3 离散和连续变量的联合分布 . . . . . 47 3.4 独立性 . . . . . 51 3.5 多个随机变量的函数 . . . . . 60 3.6 生成多元随机变量 . . . . . 63 3.7 拒绝采样 64

4 Expectation **70**
4.1 期望算子 . . . . . 70 4.2 均值与方差 . . . . . 73 4.3 协方差 . . . . . 79 4.4 条件期望 . . . . . 87 4.5 证明 89

5 Random Processes **95**
5.1 定义 . . . . . 95 5.2 均值和自协方差函数 . . . . . 98 5.3 独立同分布序列 . . . . . 100 5.4 高斯过程 . . . . . 101 5.5 泊松过程 . . . . . 102 5.6 随机游走 105

5.7 证明 107

6 Convergence of Random Processes **109**
6.1 收敛类型 . . . . . 109 6.2 大数定律 . . . . . 112 6.3 中心极限定理 . . . . . 113 6.4 蒙特卡罗模拟 118

7 Markov Chains **123**
7.1 时间齐次离散时间马尔可夫链 . . . . . 123 7.2 常返性 . . . . . 127 7.3 周期性 . . . . . 131 7.4 收敛性 . . . . . 131 7.5 马尔可夫链蒙特卡洛 137

8 Descriptive statistics **142**
8.1 直方图 . . . . . 142 8.2 样本均值与方差 . . . . . 142 8.3 顺序统计量 . . . . . 145 8.4 样本协方差 . . . . . 147 8.5 样本协方差矩阵 149

9 Frequentist Statistics **154**
9.1 独立同分布抽样 . . . . . 154 9.2 均方误差 . . . . . 155 9.3 一致性 . . . . . 157 9.4 置信区间 . . . . . 160 9.5 非参数模型估计 . . . . . 163 9.6 参数模型估计 . . . . . 168 9.7 证明 176

10 Bayesian Statistics **179**
10.1 贝叶斯参数模型 . . . . . 179 10.2 共轭先验 . . . . . 181 10.3 贝叶斯估计量 183

11 Hypothesis testing **189**
11.1 假设检验框架 . . . . . 189 11.2 参数检验 . . . . . 191 11.3 非参数检验：置换检验 . . . . . 196 11.4 多重检验 200

12 Linear Regression **202**
12.1 线性模型 . . . . . 202 12.2 最小二乘估计 . . . . . 204 12.3 过拟合 . . . . . 207 12.4 全球变暖 . . . . . 208 12.5 证明 209

A Set theory **213**
A.1 基本定义 . . . . . 213 A.2 基本运算 213

B Linear Algebra **215**
B.1 向量空间 . . . . . 215 B.2 内积与范数 . . . . . 218 B.3 正交性 . . . . . 220 B.4 投影 . . . . . 222 B.5 矩阵 . . . . . 224 B.6 特征分解 227
B.7 对称矩阵的特征分解 . . . . . 229 B.8 证明 231

Chapter 1

Basic Probability Theory

在本章中，我们介绍了概率论的数学框架，它使得通过集合论以一种有原则的方式推理不确定性成为可能。附录A包含了基本集合论概念的回顾。

1.1 Probability spaces

我们的目标是建立一个数学框架，用于表示和分析不确定现象，例如掷骰子的结果、明天的天气、NBA比赛的结果等。为此，我们将感兴趣的现象建模为一个**experiment**，其中包含若干个（可能是无限的）互斥的**outcomes**。

除非在简单的情况下，当结果的数量较少时，通常会考虑结果的集合，称为 *events*。为了量化实验结果属于特定事件的可能性，我们为该事件分配一个 **probability**。更正式地，我们定义一个 **measure**（记住，度量是将集合映射到实数的函数），它为每个感兴趣的事件分配概率。

更正式地说，该实验的特征在于构建一个**probability space**。

Definition 1.1.1 (概率空间)**.** *A probability space is a triple* $(\Omega, \mathcal{F}, \mathrm{P})$ *consisting of:*

- *A* ***sample space*** Ω*, which contains all possible outcomes of the experiment.*
- *A set of events* $\mathcal{F}$*, which must be a* ***σ-algebra*** *(see Definition 1.1.2 below).*
- *A* ***probability measure*** P *that assigns probabilities to the events in* $\mathcal{F}$ *(see Definition 1.1.4 below).*

样本空间可以是**discrete**或**continuous**。离散样本空间的例子包括抛硬币的可能结果、篮球比赛的得分、出席派对的人数等。连续样本空间通常是$\mathbb{R}$或$\mathbb{R}^n$的区间，用于建模时间、位置、温度等。

σ-代数这个术语在测度理论中用来表示满足以下列出条件的集合。不要太害怕它。这只是一个复杂的表述方式，意味着如果我们给某些事件（例如*it will rain tomorrow*或*it will*）分配概率

*snow tomorrow*我们还需要为它们的补集分配概率（即 *it will not rain tomorrow* 或 *it will not snow tomorrow*）以及它们的并集（*it will rain or snow tomorrow*）。

Definition 1.1.2 (σ代数)**.** *A σ-algebra $\mathcal{F}$ is a collection of sets in Ω such that:*

1. *If a set $S \in \mathcal{F}$ then $S^c \in \mathcal{F}$.*
2. *If the sets $S_1, S_2 \in \mathcal{F}$, then $S_1 \cup S_2 \in \mathcal{F}$. This also holds for infinite sequences; if $S_1, S_2, \ldots \in \mathcal{F}$ then $\cup_{i=1}^{\infty} S_i \in \mathcal{F}$.*
3. *$\Omega \in \mathcal{F}$.*

如果我们的样本空间是离散的，σ-代数的一个可能选择是样本空间的**power set**，它包含样本空间中所有可能的元素集。如果我们抛硬币，样本空间是

$$\Omega := \{\text{heads}, \text{tails}\}, \tag{1.1}$$

然后，幂集是一个有效的 σ-代数

$$\mathcal{F} := \{\text{heads or tails}, \text{heads}, \text{tails}, \emptyset\}, \tag{1.2}$$

其中 $\emptyset$ 表示空集。然而，在许多情况下，σ-代数并不包含所有可能的结果集合。

Example 1.1.3 (胆固醇)**.** 一位医生有兴趣以概率方式建模她病人的胆固醇水平。每次病人来看她时，她都会测试他们的胆固醇水平。这里的*experiment*是胆固醇测试，结果是测得的胆固醇水平，样本空间Ω是正实数线。医生主要关心的是病人是否有低胆固醇、临界高胆固醇或高胆固醇。事件L (低胆固醇)包含所有低于200 mg/dL的结果，事件B (临界高胆固醇)包含所有在200到240 mg/dL之间的结果，事件H (高胆固醇)包含所有高于240 mg/dL的结果。因此，σ—代数$\mathcal{F}$的可能事件等于

$$\mathcal{F} := \{L \cup B \cup H, L \cup B, L \cup H, B \cup H, L, B, H, \emptyset\}. \tag{1.3}$$

事件是样本空间的划分，这简化了对应的 σ-代数的推导。△

概率测度 P 的作用是量化我们遇到 σ-代数中各个事件的可能性大小。直观地说，事件 A 的概率可以理解为当重复试验次数趋于无穷大时，实验结果落在 A 中的频率比例。因此，概率应当始终是非负的。另外，如果两个事件 A 和 B 是互不相交的（它们的交集为空），那么

$$P(A \cup B) = \frac{\text{outcomes in } A \text{ or } B}{\text{total}} \tag{1.4}$$

$$= \frac{\text{outcomes in } A + \text{outcomes in } B}{\text{total}} \tag{1.5}$$

$$= \frac{\text{outcomes in } A}{\text{total}} + \frac{\text{outcomes in } B}{\text{total}} \tag{1.6}$$

$$= P(A) + P(B). \tag{1.7}$$

不相交事件的并集的概率应该等于各个单独概率的和。此外，整个样本空间 Ω 的概率应该等于一，因为它包含所有可能的结果。

$$P(\Omega) = \frac{\text{outcomes in } \Omega}{\text{total}} \tag{1.8}$$
$$= \frac{\text{total}}{\text{total}} \tag{1.9}$$
$$= 1. \tag{1.10}$$

这些条件是一个测度成为有效概率测度所必需的。

Definition 1.1.4 (概率测度)**.** *A probability measure is a function defined over the sets in a σ-algebra $\mathcal{F}$ such that:*

1. $\mathrm{P}(S) \geq 0$ *for any event* $S \in \mathcal{F}$.
2. *If the sets* $S_1, S_2, \ldots, S_n \in \mathcal{F}$ *are disjoint (i.e.* $S_i \cap S_j = \emptyset$ *for* $i \neq j$*) then*

$$\mathrm{P}\left(\cup_{i=1}^{n} S_i\right) = \sum_{i=1}^{n} \mathrm{P}(S_i). \tag{1.11}$$

Similarly, for a countably infinite sequence of disjoint sets $S_1, S_2, \ldots \in \mathcal{F}$

$$\mathrm{P}\left(\lim_{n\to\infty} \cup_{i=1}^{n} S_i\right) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{P}(S_i). \tag{1.12}$$

3. $\mathrm{P}(\Omega) = 1$.

两个首要公理捕捉了直观的概念，即事件的概率是一种度量，例如质量（或长度或体积）：就像任何物体的质量是非负的，并且多个不同物体的总质量是它们质量的总和一样，任何事件的概率是非负的，并且多个不相交事件的并集的概率是它们概率的总和。然而，与质量不同，实验中概率的量不能是无限的。如果明天很可能下雨，那么它也不可能非常可能下雨。如果一个事件 S 的概率很大，那么它的补集 S^c 的概率必须很小。这一点由第三个公理捕捉到，第三个公理规范化了概率度量（并且意味着 $\mathrm{P}(S^c) = 1 - \mathrm{P}(S)$）。

重要的是要强调，概率度量并不是对单个结果分配概率，而是对σ代数中的事件分配概率。之所以如此，是因为当可能结果的数量是不可数无限时，就无法对所有结果分配非零概率，同时仍然满足条件$\mathrm{P}(\Omega) = 1$。这并不是一种特殊情况，例如在胆固醇的例子中，任何正实数都是可能的结果。在离散或可数样本空间的情况下，σ代数可能等于样本空间的幂集，这意味着我们确实会对只包含单个结果的事件分配概率（例如掷硬币的例子）。

Example 1.1.5 (胆固醇（续））. 示例 1.1.3 的有效概率测度为

$$\mathrm{P}(L) = 0.6, \quad \mathrm{P}(B) = 0.28, \quad \mathrm{P}(H) = 0.12. \tag{1.13}$$

利用这些性质，我们例如可以确定 $\mathrm{P}(B \cup H) = 0.6 + 0.28 = 0.88$。 △

定义 1.1.4 具有以下后果：

$$\mathrm{P}(\emptyset) = 0, \tag{1.14}$$
$$A \subseteq B \quad \text{implies} \quad \mathrm{P}(A) \leq \mathrm{P}(B), \tag{1.15}$$
$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B). \tag{1.16}$$

我们省略证明（请尝试自行证明）。

1.2 Conditional probability

条件概率是概率建模中的一个关键概念。它允许我们在揭示额外信息时更新概率模型。考虑一个概率空间 $(\Omega, \mathcal{F}, \mathrm{P})$，在其中我们发现实验的结果属于某个事件 $S \in \mathcal{F}$。显然，这会影响其他事件 $S' \in \mathcal{F}$ 发生的可能性：我们可以排除任何不属于 S 的结果。每个事件的更新概率被称为 **conditional probability** 的 S' **given** S。直观地，条件概率可以解释为 S 中也属于 S' 的结果的比例。

$$\mathrm{P}\left(S'|S\right) = \frac{\text{outcomes in } S' \text{ and } S}{\text{outcomes in } S} \tag{1.17}$$
$$= \frac{\text{outcomes in } S' \text{ and } S}{\text{总计}} \frac{\text{total}}{\text{在}S\text{中的结果}} \tag{1.18}$$
$$= \frac{\mathrm{P}(S' \cap S)}{\mathrm{P}(S)}, \tag{1.19}$$

其中我们假设 $\mathrm{P}(S) \neq 0$（稍后我们将不得不处理 S 具有零概率的情况，这在连续概率空间中经常发生）。该定义相当直观：S 现在是新的样本空间，因此如果结果在 S' 中，那么它必然属于 $S' \cap S$。然而，仅仅使用交集的概率会低估 S' 发生的可能性，因为样本空间已经缩小到 S。因此我们通过 S 的概率进行归一化。作为一致性检查，我们有 $\mathrm{P}(S|S) = 1$，并且如果 S 和 S' 不相交，那么 $\mathrm{P}(S'|S) = 0$。

条件概率 $\mathrm{P}(\cdot|S)$ 是概率空间 $(S, \mathcal{F}_S, \mathrm{P}(\cdot|S))$ 中的有效概率测度，其中 $\mathcal{F}_S$ 是一个 σ-代数，包含 S 和 $\mathcal{F}$ 中集合的交集。为了简化符号，当我们对集合的交集进行条件化时，我们将条件概率写作

$$\mathrm{P}(S|A, B, C) := \mathrm{P}(S|A \cap B \cap C), \tag{1.20}$$

对于任意事件 S、A、B、C。

Example 1.2.1 (航班与降雨). JFK机场雇用你来估计天气如何影响航班到达的准点性。你首先定义一个概率空间，其样本空间为

$$\Omega = \{\text{late and rain}, \text{late and no rain}, \text{on time and rain}, \text{on time and no rain}\} \tag{1.21}$$

并且 σ-代数是 Ω 的幂集。根据以往航班的数据，你确定该概率空间的概率测度的一个合理估计是

$$\mathrm{P}(\text{late}, \text{no rain}) = \frac{2}{20}, \qquad \mathrm{P}(\text{on time}, \text{no rain}) = \frac{14}{20}, \tag{1.22}$$

$$\mathrm{P}(\text{late}, \text{rain}) = \frac{3}{20}, \qquad \mathrm{P}(\text{on time}, \text{rain}) = \frac{1}{20}. \tag{1.23}$$

机场对下雨时航班晚点的概率感兴趣，因此你在事件*rain*上进行条件化，定义一个新的概率空间。样本空间是所有使得*rain*发生的结果的集合，σ-代数是{准点, 晚点}的幂集，概率测度为P（·|下雨）。具体而言，

$$\mathrm{P}(\text{late}|\text{rain}) = \frac{\mathrm{P}(\text{late}, \text{rain})}{\mathrm{P}(\text{rain})} = \frac{3/20}{3/20 + 1/20} = \frac{3}{4} \tag{1.24}$$

同样地，P（late|no rain）$= 1/8$。

△

条件概率可以用来以结构化的方式计算多个事件的交集。根据定义，我们可以将两个事件 $A, B \in \mathcal{F}$ 的交集的概率表示如下，

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B|A) \tag{1.25}$$

$$= \mathrm{P}(B)\,\mathrm{P}(A|B). \tag{1.26}$$

在这个公式中，P（A）被称为A的**prior**概率，因为它刻画了在揭示任何其他信息之前我们关于A所掌握的信息。类似地，P（$A|B$）被称为**posterior**概率。这些是贝叶斯模型中的基本量，在第10章中进行了讨论。将（1.25）推广到一系列事件可得到*chain rule*，它允许用条件概率来表示多个事件交集的概率。我们省略证明，因为这只是对归纳法的直接应用。

Theorem 1.2.2 (链式法则). *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space and* $S_1, S_2, \ldots$ *a collection of events in* $\mathcal{F}$*,*

$$\mathrm{P}(\cap_i S_i) = \mathrm{P}(S_1)\,\mathrm{P}(S_2|S_1)\,\mathrm{P}(S_3|S_1 \cap S_2)\cdots \tag{1.27}$$

$$= \prod_i \mathrm{P}\left(S_i| \cap_{j=1}^{i-1} S_j\right). \tag{1.28}$$

有时，直接估计某个事件的概率可能比估计其在更简单事件条件下的概率更具挑战性。一组不相交的集合 $A_1, A_2, \ldots$，使得 $\Omega = \cup_i A_i$ 被称为 Ω 的 **partition**。全概率法则允许我们将条件概率汇总在一起，通过对分区中各个事件的概率进行加权，从而计算所关注事件的概率。

Theorem 1.2.3 (总概率法则)**.** *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space and let the collection of disjoint sets* $A_1, A_2, \ldots \in \mathcal{F}$ *be any partition of* Ω*. For any set* $S \in \mathcal{F}$

$$\mathrm{P}(S) = \sum_i \mathrm{P}(S \cap A_i) \tag{1.29}$$

$$= \sum_i \mathrm{P}(A_i)\,\mathrm{P}(S|A_i)\,. \tag{1.30}$$

Proof. 这是链式法则和定义 1.1.4 中公理 2 的直接结果，因为 $S = \cup_i S \cap A_i$ 和集合 $S \cap A_i$ 是不相交的。 □

Example 1.2.4 (阿姨来访)**.** 你的阿姨明天将抵达JFK机场，你想知道她的航班准时到达的可能性。从示例1.2.1中，你记得

$$\mathrm{P}(\text{late}|\text{rain}) = 0.75, \quad \mathrm{P}(\text{late}|\text{no rain}) = 0.125. \tag{1.31}$$

在查看了一个天气网站后，你确定 P (rain) = 0.2。

现在，我们如何将所有这些信息整合起来呢? 事件 *rain* 和 *no rain* 是不相交的，且覆盖了整个样本空间，因此它们构成了一个划分。我们可以因此应用全概率法则来确定

$$\mathrm{P}(\text{late}) = \mathrm{P}(\text{late}|\text{rain})\,\mathrm{P}(\text{rain}) + \mathrm{P}(\text{late}|\text{no rain})\,\mathrm{P}(\text{no rain}) \tag{1.32}$$

$$= 0.75 \cdot 0.2 + 0.125 \cdot 0.8 = 0.25. \tag{1.33}$$

所以你姨妈的飞机晚点的概率是 $1/4$。

△

认识到这一点至关重要：通常情况下，P ($A|B$) ≠ P ($B|A$)：大多数NBA球员可能拥有篮球（P (owns ball|NBA) 很大），但大多数拥有篮球的人并不在NBA（P (NBA|owns ball) 很小）。原因在于先验概率是非常不同的：P (NBA) 比 P (owns ball) 小得多。然而，*invert* 条件概率是可能的，即从 P ($B|A$) 中找到 P ($A|B$)，只要我们考虑先验概率。这一条件概率定义的直接后果被称为贝叶斯规则。

Theorem 1.2.5 (贝叶斯法则)**.** *For any events* A *and* B *in a probability space* $(\Omega, \mathcal{F}, \mathrm{P})$

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A)\,\mathrm{P}(B|A)}{\mathrm{P}(B)}, \tag{1.34}$$

as long as P (B) > 0*.*

Example 1.2.6 (阿姨的拜访（续）)**.** 你向住在加利福尼亚的表弟马文解释了例1.2.4中描述的概率模型。一天后，你告诉他阿姨来得晚了，但你没有提到是否下雨。挂断电话后，马文想要算出下雨的概率。回想一下，雨的概率是0.2，但由于阿姨迟到，他应该更新这个估计值。应用贝叶斯定理和

全概率法则：

$$\begin{aligned}\mathrm{P}\left(\text{rain}|\text{late}\right) &= \frac{\mathrm{P}\left(\text{late}|\text{rain}\right)\mathrm{P}\left(\text{rain}\right)}{\mathrm{P}\left(\text{late}\right)} && (1.35)\\ &= \frac{\mathrm{P}\left(\text{late}|\text{rain}\right)\mathrm{P}\left(\text{rain}\right)}{\mathrm{P}\left(\text{late}|\text{rain}\right)\mathrm{P}\left(\text{rain}\right)+\mathrm{P}\left(\text{late}|\text{no rain}\right)\mathrm{P}\left(\text{no rain}\right)} && (1.36)\\ &= \frac{0.75\cdot 0.2}{0.75\cdot 0.2+0.125\cdot 0.8}=0.6. && (1.37)\end{aligned}$$

如预期那样，在假设你阿姨迟到的情况下，下雨的概率会增加。

△

1.3 Independence

如前节所述，条件概率量化了某事件发生的知识对另一个事件概率的影响程度。在某些情况下，这没有任何区别：这些事件是**independent**。更正式地说，事件A和B是独立的，当且仅当

$$\mathrm{P}\left(A|B\right)=\mathrm{P}\left(A\right). \tag{1.38}$$

如果 $\mathrm{P}\left(B\right)=0$，则该定义无效。下面的定义涵盖了这种情况，并且在其他方面是等价的。

Definition 1.3.1 (独立)**.** *Let* $(\Omega,\mathcal{F},\mathrm{P})$ *be a probability space. Two events* $A,B\in\mathcal{F}$ *are independent if and only if*

$$\mathrm{P}\left(A\cap B\right)=\mathrm{P}\left(A\right)\mathrm{P}\left(B\right). \tag{1.39}$$

Example 1.3.2 (国会)**.** 我们考虑一个数据集，汇编了1984年美国众议院成员在两个议题上的投票记录[1]。这些议题分别是水利项目的费用分担（议题1）和预算决议的通过（议题2）。我们以概率的方式对国会议员的行为进行建模，定义了一个样本空间，其中每个结果都是一系列投票。例如，一个可能的结果是 *issue 1 = yes*，*issue 2 = no*。我们选择σ代数为样本空间的幂集。为了估计与不同事件相关的概率度量，我们只需计算它们在数据中出现的频率。

$$\begin{aligned}\mathrm{P}\left(\text{issue 1 = yes}\right) &\approx \frac{\text{members voting yes on issue 1}}{\text{total votes on issue 1}} && (1.40)\\ &= 0.597, && (1.41)\\ \mathrm{P}\left(\text{issue 2 = yes}\right) &\approx \frac{\text{members voting yes on issue 2}}{\text{total votes on issue 2}} && (1.42)\\ &= 0.417, && (1.43)\\ \mathrm{P}\left(\text{issue 1 = yes}\cap\text{issue 2 = yes}\right) &\approx \frac{\text{members voting yes on issues 1 and 2}}{\text{total members voting on issues 1 and 2}} && (1.44)\\ &= 0.069. && (1.45)\end{aligned}$$

[1]The data is available here.

基于这些数据，我们可以评估在这两个议题上的投票行为是否相互依赖。换句话说，如果我们知道某位成员在议题1上的投票情况，这是否能提供其在议题2上如何投票的信息？答案是肯定的，因为

$$P\left(\text{issue } 1 = \text{yes}\right) \mathrm{P}\left(\text{issue } 2 = \text{yes}\right) = 0.249 \tag{1.46}$$

与P非常不同（议题1 = 是 ∩ 议题2 = 是）。如果一个成员在议题1上投了赞成票，他们在议题2上投赞成票的可能性较小。△

类似地，我们可以在给定第三个事件的情况下定义 **conditional independence** 发生在两个事件之间。A 和 B 在给定 C 的情况下是条件独立的，当且仅当

$$\mathrm{P}\left(A|B,C\right) = \mathrm{P}\left(A|C\right), \tag{1.47}$$

其中 $\mathrm{P}\left(A|B,C\right) := \mathrm{P}\left(A|B \cap C\right)$. 直观地说，这意味着 A 的概率不受 B 是否发生的影响，*as long as* C *occurs*。

Definition 1.3.3 (条件独立)**.** *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space. Two events* $A, B \in \mathcal{F}$ *are conditionally independent given a third event* $C \in \mathcal{F}$ *if and only if*

$$\mathrm{P}\left(A \cap B|C\right) = \mathrm{P}\left(A|C\right) \mathrm{P}\left(B|C\right). \tag{1.48}$$

Example 1.3.4 (国会（续））**.** 决定国会议员投票的主要因素是政治隶属关系。因此，我们将其纳入示例1.3.2中的概率模型。每个结果现在包括第1和第2问题的投票，以及成员的隶属关系，例如*issue 1 = yes*、*issue 2 = no*、*affiliation = republican*或*issue 1 = no*、*issue*、*2 = no*、*affiliation = democrat*。σ-代数是样本空间的幂集。我们再次使用数据估计与不同事件相关的概率度量值:

$$\mathrm{P}\left(\text{issue } 1 = \text{yes} \,\middle|\, \text{republican}\right) \approx \frac{\text{republicans voting yes on issue 1}}{\text{total republican votes on issue 1}} \tag{1.49}$$

$$= 0.134, \tag{1.50}$$

$$\mathrm{P}\left(\text{issue } 2 = \text{yes} \,\middle|\, \text{republican}\right) \approx \frac{\text{republicans voting yes on issue 2}}{\text{total republican votes on issue 2}} \tag{1.51}$$

$$= 0.988, \tag{1.52}$$

P（议题 1 = 赞成 ∩ 议题 2 = 赞成 | 共和党人）$\approx$ ~~在议题 1 和 2 上投赞成票的共和党人~~ / 在两个议题上投票的共和党人

$$= 0.134. \tag{1.53}$$

根据这些数据，我们可以评估在这两个议题上的投票行为是否相关 *conditioned on the member being a republican*。换句话说，如果我们知道某个成员在议题 1 上的投票情况，并且知道他们是共和党成员，这是否能提供有关他们在议题 2 上投票情况的信息？答案是否定的，因为

$$\mathrm{P}\left(\text{issue } 1 = \text{yes} \,\middle|\, \text{republican}\right) \mathrm{P}\left(\text{issue } 2 = \text{yes} \,\middle|\, \text{republican}\right) = 0.133 \tag{1.54}$$

非常接近P（问题1 = 是 ∩ 问题2 = 是 | 共和党）。在知道该成员是共和党的情况下，投票大致是独立的。△

如示例 1.3.2 和 1.3.4 所示，独立性并不意味着条件独立性，反之亦然。以下例子进一步说明了这一点。从现在开始，为了简化符号，我们将多个事件的交集的概率写成以下形式

$$\mathrm{P}(A, B, C) := \mathrm{P}(A \cap B \cap C). \tag{1.55}$$

Example 1.3.5 (条件独立性并不意味着独立性)**.** 你表兄Marvin在练习1.2.6中总是抱怨纽约的出租车。从他多次访问JFK机场以来，他已经计算出

$$\mathrm{P}(\text{taxi}|\text{rain}) = 0.1, \quad \mathrm{P}(\text{taxi}|\text{no rain}) = 0.6, \tag{1.56}$$

其中 *taxi* 表示在提取行李后找到空闲出租车的事件。考虑到事件 *rain* 和 *no rain*，将事件 *plane arrived late* 和 *taxi* 建模为条件独立是合理的。

$$\mathrm{P}(\text{taxi}, \text{late}|\text{rain}) = \mathrm{P}(\text{taxi}|\text{rain})\,\mathrm{P}(\text{late}|\text{rain}), \tag{1.57}$$
$$\mathrm{P}(\text{taxi}, \text{late}|\text{no rain}) = \mathrm{P}(\text{taxi}|\text{no rain})\,\mathrm{P}(\text{late}|\text{no rain}). \tag{1.58}$$

这个逻辑背后的原因是，拿到行李后，出租车的可用性取决于是否下雨，而不是飞机是否晚点（我们假设可用性在一天内是恒定的）。这个假设是否意味着这些事件是独立的?

如果它们是独立的，那么知道你的姑妈迟到这一事实并不会向马文提供任何关于出租车可用性的信息。然而，

$$\mathrm{P}(\text{taxi}) = \mathrm{P}(\text{taxi}, \text{rain}) + \mathrm{P}(\text{taxi}, \text{no rain}) \quad (\text{by the law of total probability}) \tag{1.59}$$
$$= \mathrm{P}(\text{出租车}|\text{下雨})\,\mathrm{P}(\text{下雨}) + \mathrm{P}(\text{出租车}|\text{不下雨})\,\mathrm{P}(\text{不下雨}) \tag{1.60}$$
$$= 0.1 \cdot 0.2 + 0.6 \cdot 0.8 = 0.5, \tag{1.61}$$
$$\mathrm{P}(\text{taxi}|\text{late}) = \frac{\mathrm{P}(\text{taxi}, \text{late}, \text{rain}) + \mathrm{P}(\text{taxi}, \text{late}, \text{no rain})}{\mathrm{P}(\text{late})} \quad (\text{by the law of total probability})$$
$$= \frac{\mathrm{P}(\text{taxi}|\text{rain})\,\mathrm{P}(\text{late}|\text{rain})\,\mathrm{P}(\text{rain}) + \mathrm{P}(\text{taxi}|\text{no rain})\,\mathrm{P}(\text{late}|\text{no rain})\,\mathrm{P}(\text{no rain})}{\mathrm{P}(\text{late})}$$
$$= \frac{0.1 \cdot 0.75 \cdot 0.2 + 0.6 \cdot 0.125 \cdot 0.8}{0.25} = 0.3. \tag{1.62}$$

$\mathrm{P}(\text{出租车}) \neq \mathrm{P}(\text{出租车}|\text{迟到})$ 所以这些事件是 *not* 独立的。这是有道理的，因为如果飞机迟到，那么下雨的可能性更大，这也使得出租车更难找到。

△

Example 1.3.6 (独立性并不意味着条件独立性)**.** 在查看了你在例1.2.1中的概率模型后，你在JFK的联系人指出，航班延误通常是由飞机的机械故障引起的。你查看数据并确定

$$\mathrm{P}(\text{problem}) = \mathrm{P}(\text{problem}|\text{rain}) = \mathrm{P}(\text{problem}|\text{no rain}) = 0.1, \tag{1.63}$$

因此，事件 *mechanical problem* 和 *rain in NYC* 是独立的，这在直觉上是合理的。在对数据进行进一步分析后，您估计

$$\mathrm{P}\left(\text{late}|\text{problem}\right) = 0.7, \quad \mathrm{P}\left(\text{late}|\text{no problem}\right) = 0.2, \quad \mathrm{P}\left(\text{late}|\text{no rain}, \text{problem}\right) = 0.5.$$

下次你在JFK等Marvin时，你开始想象他的飞机是否出现了机械故障的概率。在没有进一步信息的情况下，这个概率是0.1。今天是纽约的晴天，但这没有帮助，因为根据数据（和常识），事件*problem*和*rain*是独立的。

突然他们宣布马文的航班晚点了。那么，他的飞机发生机械故障的概率是多少？乍一想，你可能会像例 1.2.6 那样应用贝叶斯定理来计算 P (problem|late) = 0.28。然而，你并没有利用当天是晴天这一事实。这意味着降雨并不是造成延误的原因，因此从直觉上看，机械故障应该更有可能。确实，

$$\begin{aligned}
\mathrm{P}\left(\text{problem}|\text{late}, \text{no rain}\right) &= \frac{\mathrm{P}\left(\text{late}, \text{no rain}, \text{problem}\right)}{\mathrm{P}\left(\text{late}, \text{no rain}\right)} && (1.64)\\
&= \frac{\mathrm{P}\left(\text{late}|\text{no rain}, \text{problem}\right)\mathrm{P}\left(\text{no rain}\right)\mathrm{P}\left(\text{problem}\right)}{\mathrm{P}\left(\text{late}|\text{no rain}\right)\mathrm{P}\left(\text{no rain}\right)} && \text{(通过链式法则)}\\
&= \frac{0.5 \cdot 0.1}{0.125} = 0.4. && (1.65)
\end{aligned}$$

由于 P (problem|late, 无雨) $\neq$ P (problem|late)，事件 *mechanical problem* 和 *rain in NYC* 在给定事件 *plane is late* 的条件下 *not* 条件独立。

△

Chapter 2

Random Variables

随机变量是概率建模中的基本工具。它们使我们能够对数字量进行建模，这些数字量是*uncertain*：明天纽约的温度、航班的到达时间、卫星的位置……对这些量进行概率推理使我们能够以有原则的方式构建我们关于它们的信息。

2.1 Definition

形式化地，我们将随机变量定义为一个将概率空间中的每个结果映射到实数的函数。

Definition 2.1.1 (随机变量)**.** *Given a probability space* $(\Omega, \mathcal{F}, \mathrm{P})$*, a random variable* X *is a function from the sample space* Ω *to the real numbers* $\mathbb{R}$*. Once the outcome* $\omega \in \Omega$ *of the experiment is revealed, the corresponding* $X(\omega)$ *is known as a* ***realization*** *of the random variable.*

Remark 2.1.2 (严格定义)**.** *If we want to be completely rigorous, Definition 2.1.1 is missing some details. Consider two sample spaces* Ω_1 *and* Ω_2*, and a* σ*-algebra* $\mathcal{F}_2$ *of sets in* Ω_2*. Then, for* X *to be a random variable, there must exist a* σ*-algebra* $\mathcal{F}_1$ *in* Ω_1 *such that for any set* S *in* $\mathcal{F}_2$ *the inverse image of* S*, defined by*

$$X^{-1}(S) := \{\omega \mid X(\omega) \in S\}, \tag{2.1}$$

belongs to $\mathcal{F}_1$*. Usually, we take* Ω_2 *to be the reals* $\mathbb{R}$ *and* $\mathcal{F}_2$ *to be the Borel* σ*-algebra, which is defined as the smallest* σ*-algebra defined on the reals that contains all open intervals (amazingly, it is possible to construct sets of real numbers that do not belong to this* σ*-algebra). In any case, for the purpose of these notes, Definition 2.1.1 is sufficient (more information about the formal foundations of probability can be found in any book on measure theory and advanced probability theory).*

Remark 2.1.3 (符号)**.** *We often denote events of the form*

$$\{X(\omega) \in \mathcal{S} \,:\, \omega \in \Omega\} \tag{2.2}$$

for some random variable X *and some set* $\mathcal{S}$ *as*

$$\{X \in \mathcal{S}\} \tag{2.3}$$

to alleviate notation, since the underlying probability space is often of no significance once we have specified the random variables of interest.

随机变量量化了我们对其所表示数量的不确定性，*not* 当结果被揭示时它最终取得的那个具体取值。你*never*应当把随机变量看作具有一个固定的数值。如果结果是已知的，那么这就确定了该随机变量的一个*realization*。为了强调随机变量与其实现值之间的差别，我们用大写字母表示前者（X，Y，……），用小写字母表示后者（x，y，……）。

如果我们可以访问定义随机变量的概率空间 $(\Omega, \mathcal{F}, \mathrm{P})$，那么计算随机变量 X 属于某个集合 S:[1] 的概率就很简单：它是包含所有在 Ω 中的结果的事件的概率，这些结果由 X 映射到 S。

$$\mathrm{P}(X \in S) = \mathrm{P}(\{\omega \mid X(\omega) \in S\}). \tag{2.4}$$

然而，我们几乎从不直接对概率空间建模，因为这需要估计对应的σ-代数中每个可能事件的概率。正如我们在第2.2节和第2.3节中所解释的，有一些更实际的方法来指定随机变量，这些方法自动意味着存在一个有效的基础概率空间。这个概率空间的存在确保了整个框架在数学上是合理的，但你其实不需要太担心这个问题。

2.2 Discrete random variables

离散随机变量取值于 $\mathbb{R}$ 的一个 *finite or countably infinite* 子集，例如整数。它们用于刻画离散的数值量：如掷骰子的结果、篮球比赛的得分等。

2.2.1 Probability mass function

要指定一个离散随机变量，只需确定它可能取的每个值的概率即可。与连续随机变量的情形相比，这是可行的，因为这些取值按定义是可数的。

Definition 2.2.1 (概率质量函数)**.** *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space and* $X : \Omega \to \mathbb{Z}$ *a random variable. The probability mass function (pmf) of* X *is defined as*

$$p_X(x) := \mathrm{P}(\{\omega \mid X(\omega) = x\}). \tag{2.5}$$

In words, $p_X(x)$ *is the probability that* X *equals* x*.*

我们通常说一个随机变量根据某个 pmf 是 **distributed**。
如果 X 的离散范围用 D 表示，则三元组 $\left(D, 2^D, p_X\right)$ 是一个有效的概率空间（回忆一下，2^D 是 D 的幂集）。特别地，p_x 是一个有效的概率度量。

[1] Strictly speaking, S needs to belong to the Borel σ-algebra. Again, this comprises essentially any subset of the reals that you will ever encounter in probabilistic modeling

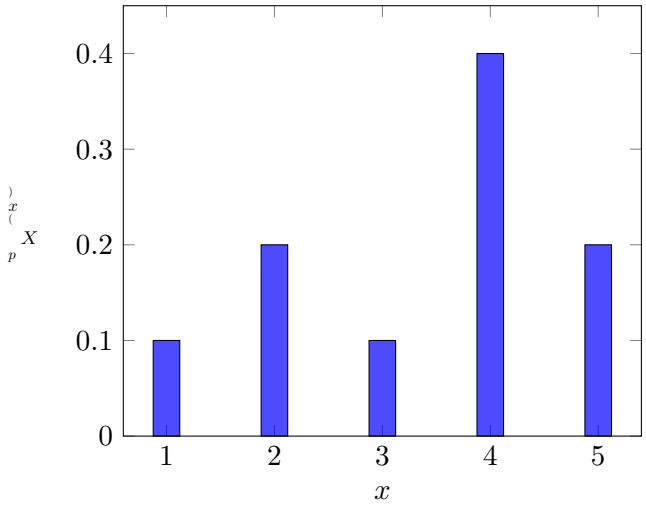


Figure 2.1: 随机变量 X 的概率质量函数，见例 2.2.2。

满足

$$p_X(x) \geq 0 \quad \text{for any } x \in D, \tag{2.6}$$

$$\sum_{x \in D} p_X(x) = 1. \tag{2.7}$$

反之亦然，如果一个定义在实数的可数子集 D 上的函数是非负的并且加起来等于 1，那么它可以被解释为一个随机变量的概率质量函数（pmf）。实际上，在实践中，我们通常通过指定其 pmf 来定义离散随机变量。

为了计算随机变量 X 属于某个集合 S 的概率，我们对 S 中包含的所有值求 pmf 的和：

$$\mathrm{P}(X \in S) = \sum_{x \in S} p_X(x). \tag{2.8}$$

Example 2.2.2 (离散随机变量)**.** 图2.1显示了离散随机变量X (的概率质量函数)，检查其总和是否为一。为了计算X属于不同集合的概率，我们应用公式(2.8)：

$$\mathrm{P}(X \in \{1, 4\}) = p_X(1) + p_X(4) = 0.5, \tag{2.9}$$

$$\mathrm{P}(X > 3) = p_X(4) + p_X(5) = 0.6. \tag{2.10}$$

△

2.2.2 Important discrete random variables

在本节中，我们描述了几种对概率建模有用的离散随机变量。

Bernoulli

伯努利随机变量用于模拟具有两个可能结果的实验。按照惯例，我们通常用0表示一个结果，用1表示另一个结果。一个经典的例子是掷一枚有偏的硬币，其中正面朝上的概率为p。如果我们将正面编码为1，反面编码为0，那么硬币掷出的结果对应一个参数为p的伯努利随机变量。

Definition 2.2.3 (贝尔努利)**.** *The pmf of a Bernoulli random variable with parameter* $p \in [0, 1]$ *is given by*

$$p_X(0) = 1 - p, \tag{2.11}$$
$$p_X(1) = p. \tag{2.12}$$

一种特殊的伯努利随机变量是事件的指示随机变量。这种随机变量在证明中特别有用。

Definition 2.2.4 (指标)**.** *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space. The indicator random variable of an event* $S \in \mathcal{F}$ *is defined as*

$$1_S(\omega) = \begin{cases} 1, & \text{if } \omega \in S, \\ 0, & \text{otherwise.} \end{cases} \tag{2.13}$$

根据定义，指示随机变量的分布是参数为 $\mathrm{P}(S)$ 的伯努利分布。

Geometric

假设我们拿一个有偏的硬币，并一直掷下去直到得到正面。如果得到正面的概率是 p，并且每次掷硬币都是独立的，那么掷 k 次的概率是

$$\begin{aligned} \mathrm{P}(k \text{ flips}) &= \mathrm{P}(1\text{st flip} = \text{tails}, \ldots, k-1\text{th flip} = \text{tails}, k\text{th flip} = \text{heads}) && (2.14)\\ &= \mathrm{P}(1\text{st flip} = \text{tails}) \cdots \mathrm{P}(k-1\text{th flip} = \text{tails})\, \mathrm{P}(k\text{th flip} = \text{heads}) && (2.15)\\ &= (1-p)^{k-1} p. && (2.16) \end{aligned}$$

这种推理可以应用于任何随机实验，其中固定概率 p 的实验会被重复进行，直到发生特定的结果，只要满足独立性假设。在这种情况下，重复次数被建模为几何随机变量。

Definition 2.2.5 (几何)**.** *The pmf of a geometric random variable with parameter* p *is given by*

$$p_X(k) = (1-p)^{k-1} p, \quad k = 1, 2, \ldots \tag{2.17}$$

图2.2展示了具有不同参数的几何随机变量的概率质量函数。p 越大，分布越集中在较小的 k 值附近。

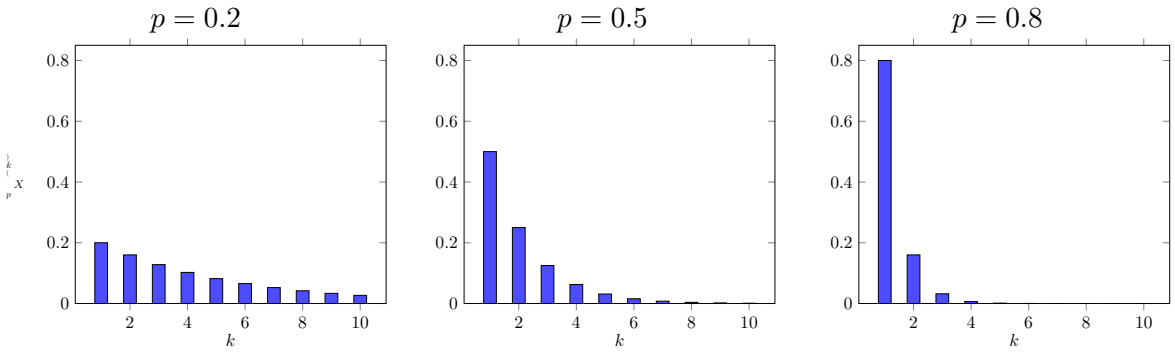


Figure 2.2: 概率 三个几何随机变量的质量函数，具有{v*} 不同的参数。

Binomial

二项随机变量在概率建模中极其有用。它们用于建模由具有相同参数、相互独立的伯努利随机变量所表示的 n 次试验中正向结果的数量。下面的例子通过抛硬币来说明这一点。

Example 2.2.6 (抛硬币). 如果我们将一枚有偏硬币抛掷 n 次，在各次抛掷相互独立且正面朝上的概率为 p 的情况下，恰好得到 k 次正面的概率是多少?

让我们先考虑一个更简单的问题：先获得 k 次正面，然后获得 $n-k$ 次反面的概率是多少？由于独立性，答案是

$$\begin{aligned} &\mathrm{P}\left(k \text{ 正面，接着 } n-k \text{ 反面}\right) \qquad (2.18)\\ &= \mathrm{P}(\text{第1次掷币为=正面，……，第}k\text{次掷币为=正面，第}k+\text{次掷币为=反面，……，第}n \\ &\text{次掷币为=反面）} = \mathrm{P}(\text{第1次掷币为=正面）} \cdots \mathrm{P}(\text{第}k\text{次掷币为=正面）}\,\mathrm{P}(\text{第}k+\text{次掷币} \\ &\text{为=反面）} \cdots \mathrm{P}(\text{第}n\text{次掷币为=反面）} = p^k\left(1-p\right)^{n-k}. \qquad (2.19)\end{aligned}$$

请注意，相同的推理意味着这也是获得恰好 k 个正面的概率 *in any fixed order*。获得恰好 k 个正面的概率是所有这些事件的并集。由于这些事件是互不相交的（我们不可能同时以两种不同的顺序获得恰好 k 个正面），我们可以将它们的各自概率相加，计算我们感兴趣事件的概率。我们只需要知道可能的排列数。根据基本的组合学原理，这由二项式系数 $\binom{n}{k}$ 给出，定义为

$$\binom{n}{k} := \frac{n!}{k!\left(n-k\right)!}. \qquad (2.20)$$

我们得出结论，

$$\mathrm{P}\left(k \text{ heads out of } n \text{ flips}\right) = \binom{n}{k} p^k\left(1-p\right)^{(n-k)}. \qquad (2.21)$$

△

在该示例中表示正面次数的随机变量称为二项随机变量。

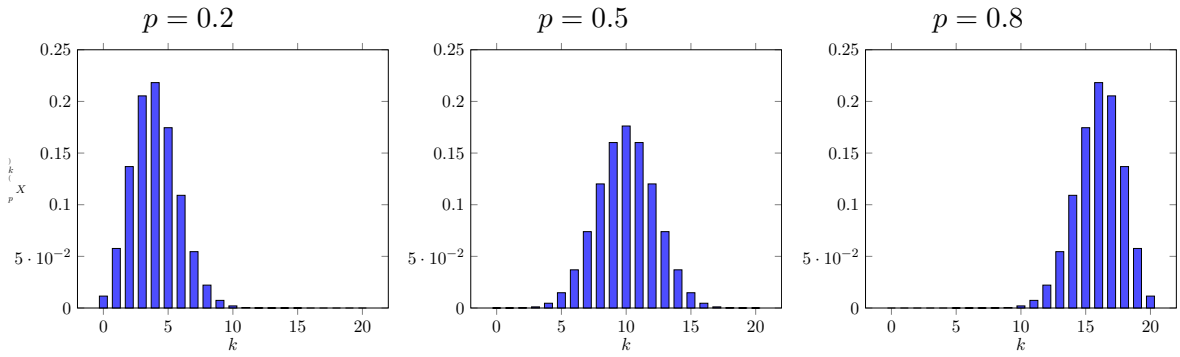


Figure 2.3: 具有不同p和n =取值的三个二项随机变量的概率质量函数 20.

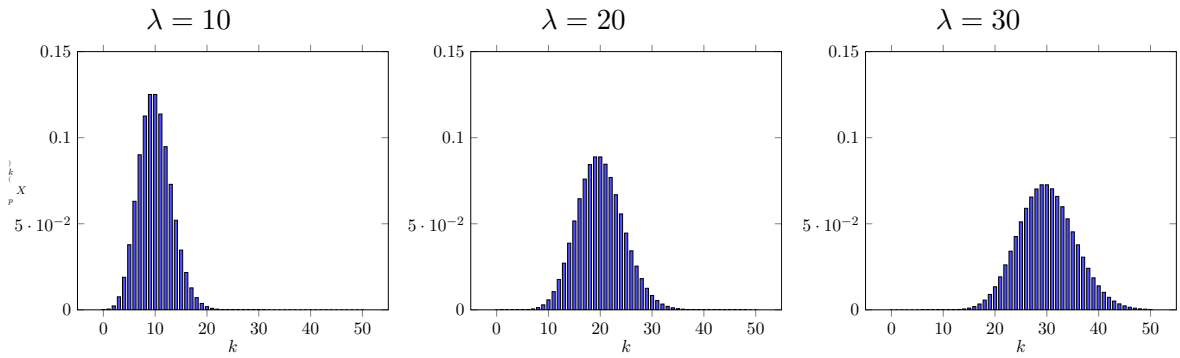


Figure 2.4: 三个具有不同参数的泊松随机变量的概率质量函数。

Definition 2.2.7 (二项式)**.** *The pmf of a binomial random variable with parameters n and p is given by*

$$p_X(k) = \binom{n}{k} p^k (1-p)^{(n-k)}, \quad k = 0, 1, 2, \ldots, n. \tag{2.22}$$

图 2.3 显示了具有不同 p 值的二项随机变量的概率质量函数。

Poisson

我们通过一个例子来引出泊松随机变量的定义。

Example 2.2.8 (呼叫中心)**.** 一个呼叫中心希望对他们一天内接收到的电话数量进行建模，以便决定雇佣多少人。他们做出以下假设：

1. 每次呼叫相互独立。 2. 每次呼叫在一天中的任何时刻发生的概率相同。 3. 呼叫发生的频率为每天 λ 次。

在第五章中，我们将看到这些假设定义了一个泊松过程。
我们的目标是计算在给定的一天内恰好接到 k 次呼叫的概率。为此，我们将一天离散化为 n 个区间，在假设每个区间都非常小的情况下计算所需的概率，然后令 $n \to \infty$。

在长度为 $1/n$ 的时间区间内发生一次呼叫的概率为 λ/n，这是根据假设 2 和假设 3 得出的。发生 $m >$ 次呼叫的概率是 $(\lambda/n)^m$。如果 n 非常大，这个概率相比于区间内接收到一次或零次呼叫的概率可以忽略不计，实际上，当我们取极限 $n \to \infty$ 时，它趋近于零。因此，整个小时内发生的总呼叫次数可以通过发生呼叫的区间数来近似，只要 n 足够大。由于每个区间内发生呼叫的概率相同，且呼叫是独立发生的，全天内的呼叫次数可以被建模为一个具有参数 n 和 p 的二项分布随机变量：$= \lambda/n$。

我们现在计算当区间任意小（即当 $n \to \infty$ 时）的呼叫分布：

$$\begin{aligned}
\mathrm{P}\left(k \text{ calls during the day }\right) &= \lim_{n\to\infty} \mathrm{P}\left(k \text{ calls in } n \text{ small intervals}\right) && (2.23)\\
&= \lim_{n\to\infty} \binom{n}{k} p^k \left(1-p\right)^{(n-k)} && (2.24)\\
&= \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{(n-k)} && (2.25)\\
&= \lim_{n\to\infty} \frac{n!\,\lambda^k}{k!\,(n-k)!\,\left(n-\lambda\right)^k} \left(1-\frac{\lambda}{n}\right)^n && (2.26)\\
&= \frac{\lambda^k\, e^{-\lambda}}{k!}. && (2.27)
\end{aligned}$$

最后一步来自于第2.7.1节中证明的以下引理。

Lemma 2.2.9.

$$\lim_{n\to\infty} \frac{n!}{(n-k)!\,\left(n-\lambda\right)^k} \left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}. \tag{2.28}$$

△

具有我们在示例中推导的概率质量函数（pmf）的随机变量称为泊松随机变量。它们用于模拟在固定速率下，某些事件不时发生的情况：互联网路由器接收数据包、地震、交通事故等。发生在固定时间间隔内的此类事件的数量服从泊松分布，只要我们在示例中列出的假设成立。

Definition 2.2.10 (泊松)**.** *The pmf of a Poisson random variable with parameter λ is given by*

$$p_X\left(k\right) = \frac{\lambda^k\, e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots \tag{2.29}$$

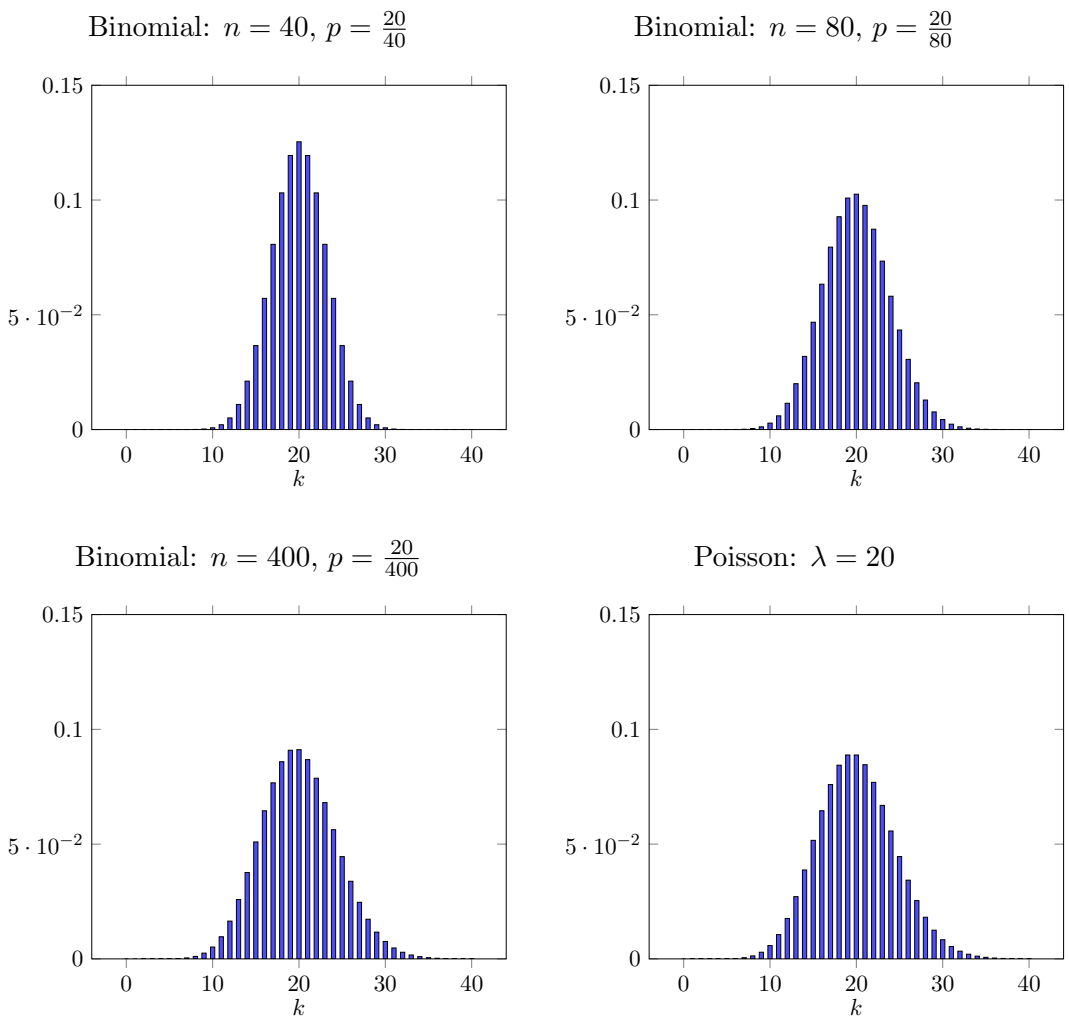


Figure 2.5: 二项分布的概率质量函数$p = \lambda/n$在n增长时收敛到参数为λ的泊松分布概率质量函数。

图 2.4 显示了不同 λ 值的泊松随机变量的概率质量函数。在例 2.2.8 中，我们证明了当 $n \to \infty$ 时，具有参数 n 和 λ/n 的二项随机变量的 pmf 会趋近于参数为 λ (的泊松分布的 pmf，正如我们将在课程中后续看到的，这是 *convergence in distribution*) 的一个例子。图 2.5 数字化地展示了这一现象；收敛速度相当快。

你可能对示例 2.2.8 感到有些怀疑：接到电话的概率肯定会随着一天中的时间变化，而且周末的概率一定不同！这确实是对的，但如果我们将注意力集中在较短的时间段内，模型实际上是非常有用的。在图 2.6 中，我们展示了使用泊松随机变量对以色列[2]一个呼叫中心在四小时（晚上 8 点到午夜）间隔内接到的电话数量进行建模的结果。我们绘制了该时间段内在 1999 年 9 月和 10 月两个月中接到的电话数量的直方图，并拟合了一个泊松概率质量函数（pmf）到数据上（稍后我们会学习如何将分布拟合到数据）。尽管我们的假设并不完全成立，模型仍然有效。

[2]The data is available here.

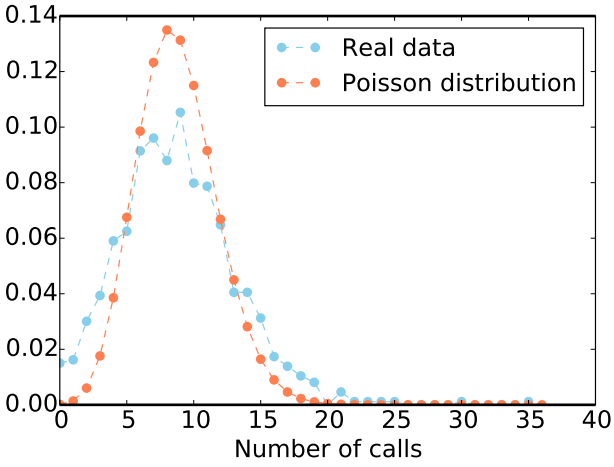


Figure 2.6: 在蓝色区域，我们可以看到以色列某呼叫中心在两个月期间每四小时接收到的电话数量的直方图。用橙色绘制的是一个泊松概率质量函数（pmf），它近似于数据的分布。

产生了一个相当好的拟合。

2.3 Continuous random variables

物理量通常最适合用连续的方式来描述：温度、持续时间、速度、重量等。为了以概率方式对这类量进行建模，我们可以将其定义域离散化，并将其表示为离散随机变量。然而，我们可能不希望结论依赖于我们如何选择离散化网格。构建连续模型使我们能够获得对 *sufficiently fine* 种网格都有效的洞见，而无需担心离散化问题。

正因为连续域模型离散结果具有任意细粒度时的极限，我们*cannot*通过设置概率值来表征连续随机变量的概率行为，就像我们对离散随机变量所做的那样，表示X等于个别结果。事实上，我们*cannot*为不确定的连续量的特定结果分配非零概率。这将导致不可数的不相交的结果，且这些结果具有非零概率。无数个正值的和是无限的，因此它们的联合概率将大于一，这是没有意义的。

更严格地说，事实证明我们无法在 $\mathbb{R}$ (的幂集上定义一个有效的概率测度；对这一点的论证需要测度论，超出了这些笔记的范围)。相反，我们考虑由 *unions of intervals of* $\mathbb{R}$ 组成的事件。这类事件形成了一个称为 Borel σ-代数的 σ-代数。这个 σ-代数足够精细，能够表示你可能感兴趣的任何集合（试着想一个不能表示为可数个区间并的集合），同时又允许在其上定义有效的概率测度。

2.3.1 Cumulative distribution function

要在Borel σ代数上指定一个随机变量，只需确定该随机变量属于任何$x \in \mathbb{R}$形式的区间（$-\infty, x$）的概率。

Definition 2.3.1 (累积分布函数). *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space and* $X : \Omega \to \mathbb{R}$ *a random variable. The cumulative distribution function (cdf) of* X *is defined as*

$$F_X(x) := \mathrm{P}(X \leq x). \tag{2.30}$$

In words, $F_X(x)$ *is the probability of* X *being smaller than* x*.*

请注意，累积分布函数可以为*both continuous and discrete*随机变量定义。

以下引理描述了累积分布函数（cdf）的基本性质。您可以在第2.7.2节找到证明。

Lemma 2.3.2 (cdf 的性质). *For any continuous random variable* X

$$\lim_{x \to -\infty} F_X(x) = 0, \tag{2.31}$$

$$\lim_{x \to \infty} F_X(x) = 1, \tag{2.32}$$

$$F_X(b) \geq F_X(a) \quad \text{if } b > a, \quad \text{i.e. } F_X \text{ is nondecreasing.} \tag{2.33}$$

为了理解为什么累积分布函数（cdf）完全决定了一个随机变量，回想一下我们只考虑可以表示为区间并集的集合。随机变量 X 属于区间 $(a, b]$ 的概率由以下公式给出

$$\begin{aligned} \mathrm{P}(a < X \leq b) &= \mathrm{P}(X \leq b) - \mathrm{P}(X \leq a) && (2.34)\\ &= F_X(b) - F_X(a). && (2.35) \end{aligned}$$

Remark 2.3.3. *Since individual points have zero probability, for any continuous random variable* X

$$\mathrm{P}(a < X \leq b) = \mathrm{P}(a \leq X \leq b) = \mathrm{P}(a < X < b) = \mathrm{P}(a \leq X < b). \tag{2.36}$$

现在，为了找到 X 属于任何特定集合的概率，我们只需要将其分解为不相交的区间，并应用公式 (2.35)，如下例所示。

Example 2.3.4 (连续随机变量). 考虑一个连续随机变量 X，其累积分布函数由以下给出

$$F_X(x) := \begin{cases} 0 & \text{for } x < 0, \\ 0.5\,x & \text{for } 0 \leq x \leq 1, \\ 0.5 & \text{for } 1 \leq x \leq 2, \\ 0.5\left(1 + (x-2)^2\right) & \text{for } 2 \leq x \leq 3, \\ 1 & \text{for } x > 3. \end{cases} \tag{2.37}$$

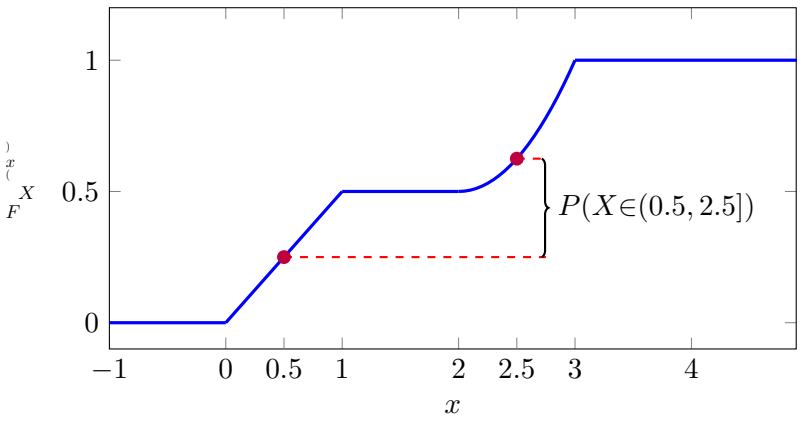


Figure 2.7: 随机变量在例 2.3.4 和 2.3.7 中的累积分布函数。

图 2.7 显示了左侧图像中的 cdf。您可以检查它是否满足引理 2.3.2 中的性质。为了确定 X 在 0.5 和 2.5 之间的概率，我们应用式 (2.35),

$$\mathrm{P}\left(0.5 < X \leq 2.5\right) = F_X\left(2.5\right) - F_X\left(0.5\right) = 0.375, \tag{2.38}$$

如图 2.7 所示。 △

2.3.2 Probability density function

如果一个连续随机变量的累积分布函数是可微的，那么它的导数可以解释为一个密度函数。随后可以对该密度进行积分，以得到该随机变量落在某个区间或区间并（从而落在任意博雷尔集）中的概率。

Definition 2.3.5 (概率密度函数). *Let $X : \Omega \to \mathbb{R}$ be a random variable with cdf F_X. If F_X is differentiable then the probability density function or pdf of X is defined as*

$$f_X\left(x\right) := \frac{dF_X\left(x\right)}{d\ x}. \tag{2.39}$$

直观地说，$f_X\left(x\right)\Delta$ 是 X 属于以 x 为中心、宽度为 Δ 的区间的概率，当 $\Delta\ \to$ 为 0 时。根据微积分基本定理，随机变量 X 属于某个区间的概率由以下公式给出

$$\mathrm{P}\left(a < X \leq b\right) = F_X\left(b\right) - F_X\left(a\right) \tag{2.40}$$

$$= \int_a^b f_X\left(x\right)\mathrm{d}x. \tag{2.41}$$

我们关注的集合属于 Borel σ-代数，因此可以分解为区间的并集，所以我们可以通过在 S 上对其概率密度函数进行积分来获得 X 属于任意此类集合 S 的概率。

$$\mathrm{P}\left(X \in S\right) = \int_S f_X\left(x\right)\mathrm{d}x. \tag{2.42}$$

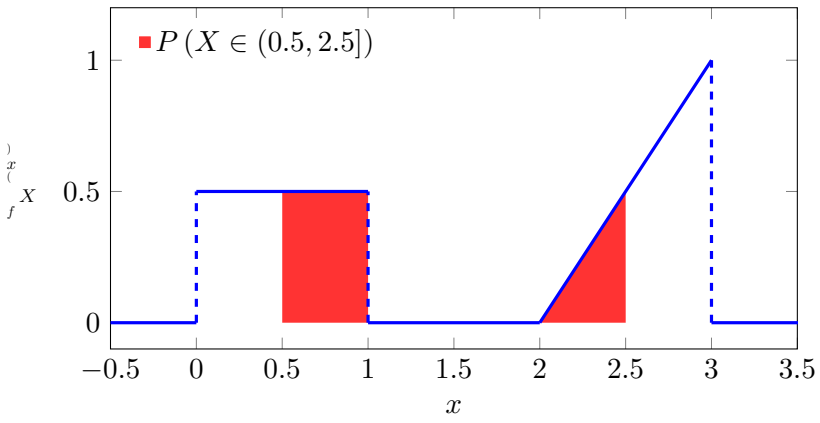


Figure 2.8: 随机变量在例 2.3.4 和 2.3.7 中的概率密度函数。

特别地，由于 X 根据定义属于 $\mathbb{R}$

$$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = \mathrm{P}(X \in \mathbb{R}) = 1. \tag{2.43}$$

由于累积分布函数（2.33）的单调性，可以得出概率密度函数是非负的。

$$f_X(x) \geq 0, \tag{2.44}$$

因为否则我们将能够找到两个点 $x_1 < x_2$，其中 $F_X(x_2) < F_X(x_1)$。

Remark 2.3.6 (PDF 不是概率测度)**.** *The pdf is a* 密度 *which must be integrated to yield a probability. In particular, it is not necessarily smaller than one (for example, take $a = 0$ and $b = 1/2$ in Definition 2.3.8 below).*

最后，正如离散随机变量的情形一样，我们常常说某个随机变量按照某个 pdf 或 cdf 是 **distributed** 的，或者说我们知道它的分布。原因在于，pmf、pdf 或 cdf 足以刻画其底层的概率空间。

Example 2.3.7 (连续随机变量（续））**.** 要计算示例 2.3.4 中随机变量的概率密度函数（pdf），我们对其累积分布函数（cdf）进行求导，得到

$$f_X(x) = \begin{cases} 0 & \text{for } x < 0, \\ 0.5 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{for } 1 \leq x \leq 2 \\ x - 2 & \text{for } 2 \leq x \leq 3 \\ 0 & \text{for } x > 3. \end{cases} \tag{2.45}$$

图 2.8 显示了概率密度函数(pdf)。你可以检查它是否积分为 1。为了确定 X 在 0.5 和 2.5 之间的概率，我们可以在该区间上进行积分，以获得与例 2.3.4 相同的答案。

$$\mathrm{P}(0.5 < X \leq 2.5) = \int_{0.5} f_X(x)\,\mathrm{d}x \tag{2.46}$$

$$= \int_{0.5}^{1} 0.5\,\mathrm{d}x + \int_{2}^{2.5} x - 2\,\mathrm{d}x = 0.375. \tag{2.47}$$

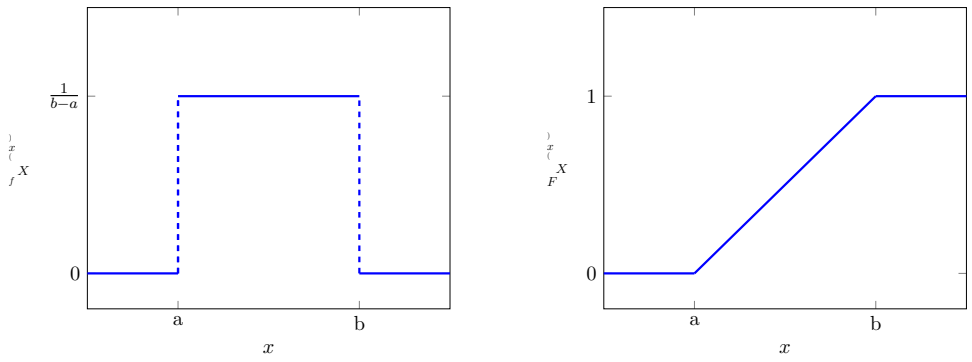


Figure 2.9: 均匀随机变量 X 的概率密度函数（左）和累积分布函数（右）。

图 2.8 说明，一旦我们将其限制到实数线的相应子集上，事件发生的概率等于 pdf 下的面积。

△

2.3.3 Important continuous random variables

在本节中，我们描述了几种在概率模型和统计中有用的连续随机变量。

Uniform

一个均匀随机变量模型表示一个实验，其中连续区间内的每个结果出现的概率相等。因此，概率密度函数在该区间内是常数。图2.9展示了均匀随机变量的概率密度函数和累积分布函数。

Definition 2.3.8 (统一). *The pdf of a uniform random variable with domain* $[a, b]$*, where* $b > a$ *are real numbers, is given by*

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases} \tag{2.48}$$

Exponential

指数随机变量通常用于模拟直到某个事件发生的时间。示例包括衰变的放射性粒子、电话通话、地震等。

Definition 2.3.9 (指数). *The pdf of an exponential random variable with parameter* λ *is given by*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2.49}$$

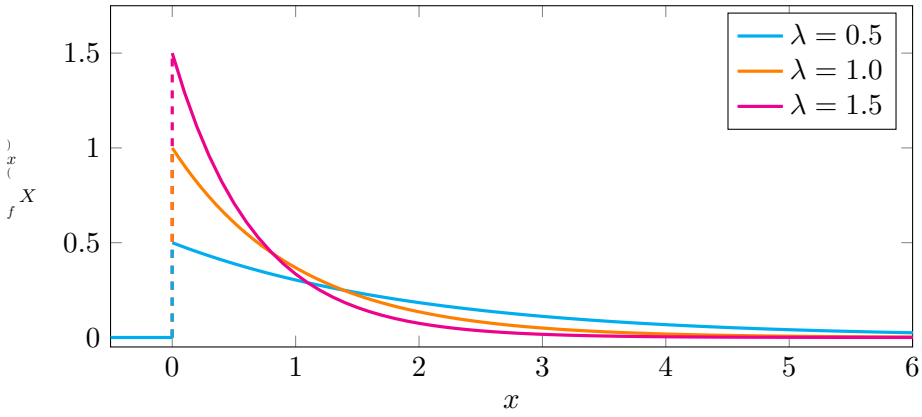


Figure 2.10: 具有不同参数的指数随机变量的概率密度函数。

图 2.10 显示了三个具有不同参数的指数随机变量的概率密度函数（pdf）。为了说明指数分布在建模实际数据中的潜力，在图 2.11 中，我们绘制了前面提到的以色列同一呼叫中心的呼叫间隔时间的直方图。具体来说，这些呼叫间隔时间是 1999 年 9 月两个晚上 8 点到午夜之间，连续呼叫之间的时间。指数模型很好地拟合了这些数据。

指数型随机变量的一个重要性质是它是 *memoryless*。我们将在第2.4节详细阐述这一性质，该性质也被几何分布所共享。

Gaussian or Normal

高斯或正态随机变量可以说是概率论和统计学中最流行的随机变量。它通常用于建模自然科学中分布未知的变量。这是因为独立随机变量的和往往会收敛到高斯分布。这个现象由中心极限定理描述，我们将在第六章讨论该定理。

Definition 2.3.10 (高斯). *The pdf of a Gaussian or normal random variable with mean μ and standard deviation σ is given by*

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{2.50}$$

A Gaussian distribution with mean μ and standard deviation σ is usually denoted by $\mathcal{N}(\mu, \sigma^2)$.

我们将在第4章中给出随机变量的均值和标准差的形式化定义。现在，你可以把它们看作是对高斯概率密度函数进行参数化的量。

高斯分布的概率密度函数的积分等于 1 并非一目了然。我们在下面的引理中证明这一点。

Lemma 2.3.11 (第2.7.3节的证明). *The pdf of a Gaussian random variable integrates to one.*

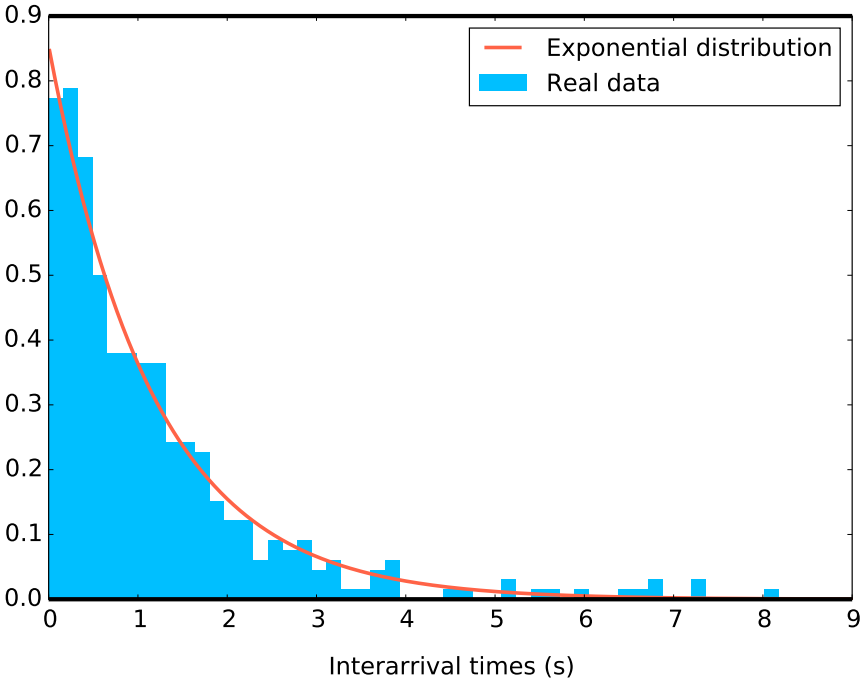


Figure 2.11: 以色列呼叫中心来电间隔时间的直方图（红色），与其由指数概率密度函数（pdf）近似的结果进行比较。

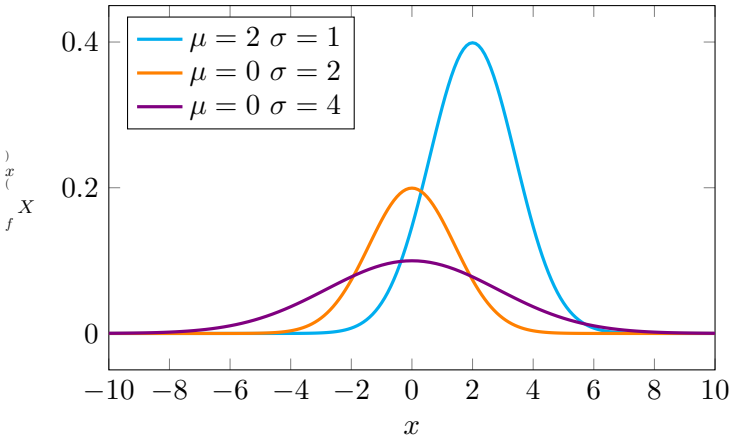


Figure 2.12: 具有不同均值和标准差的高斯随机变量。

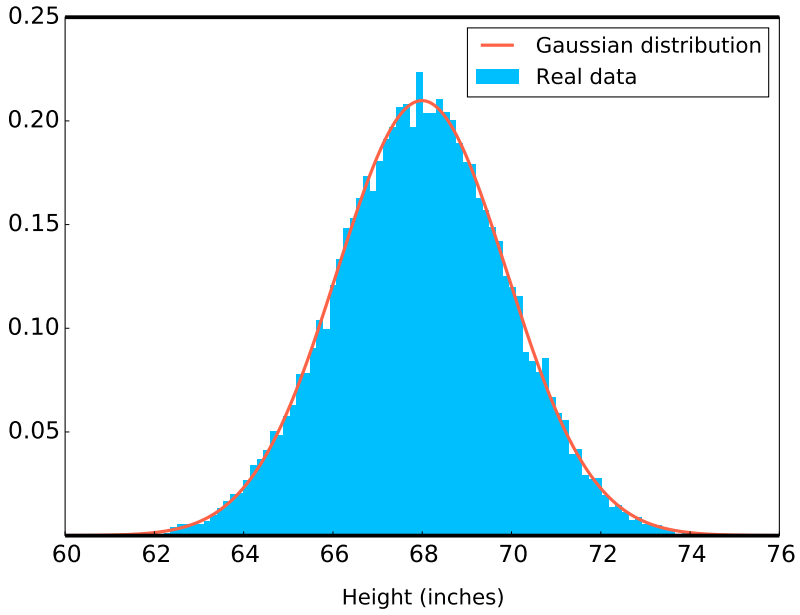


Figure 2.13: 25,000人群体的身高直方图（蓝色）及其使用高斯分布的近似（橙色）。

图 2.12 显示了两个具有不同 μ 和 σ 值的高斯随机变量的概率密度函数。图 2.13 显示了一个由 25,000 人组成的人群身高的直方图，以及它是如何被一个高斯随机变量 [3] 非常好地近似的。

高斯随机变量的一个烦人特性是其累积分布函数（cdf）没有闭式解，这与均匀和指数随机变量不同。这使得确定高斯随机变量位于某个特定区间的概率变得更加复杂。为了解决这个问题，我们利用这样一个事实：如果X是一个均值为μ、标准差为σ的高斯随机变量，那么

$$U := \frac{X - \mu}{\sigma} \tag{2.51}$$

是一个**standard**高斯随机变量，这意味着它的均值为零，标准差为一。证明见引理2.5.1。这样，我们可以通过标准高斯分布的累积分布函数（我们用Φ表示）来表示X位于区间$[a, b]$中的概率。

$$\mathrm{P}\left(X \in [a, b]\right) = \mathrm{P}\left(\frac{X - \mu}{\sigma} \in \left[\frac{a - \mu}{\sigma}, \frac{b - \mu}{\sigma}\right]\right) \tag{2.52}$$

$$= \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right). \tag{2.53}$$

只要我们能够评估 Φ，这个公式就允许我们处理任意的高斯随机变量。为了评估 Φ，人们过去常常依赖于数值计算相应积分的表格列表。如今，你可以直接使用 Matlab、WolframAlpha、SciPy 等工具。

[3] The data is available here.

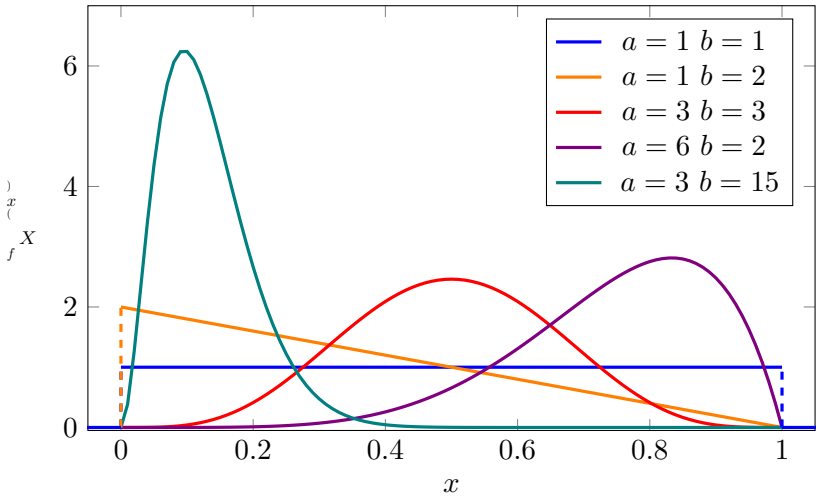


Figure 2.14: 具有不同 a 和 b 参数取值的贝塔随机变量的概率密度函数。

Beta

贝塔分布允许我们对支持单位区间的单峰连续分布进行参数化。这在贝叶斯统计中非常有用，正如我们在第10章中讨论的那样。

Definition 2.3.12 (贝塔分布). *The pdf of a beta distribution with parameters a and b is defined as*

$$f_\beta(\theta; a, b) := \begin{cases} \frac{\theta^{a-1}(1-\theta)^{b-1}}{\beta(a,b)}, & \text{if } 0 \leq \theta \leq 1, \\ 0 & \text{otherwise,} \end{cases} \tag{2.54}$$

where

$$\beta(a, b) := \int_u u^{a-1}(1-u)^{b-1}\, du. \tag{2.55}$$

$\beta(a, b)$ 是一个称为贝塔函数或第一类欧拉积分的特殊函数，必须通过数值方法计算。均匀分布是贝塔分布的一个例子（其中 $a = 1$ 且 $b = 1$）。图 2.14 显示了若干不同贝塔分布的 pdf。

2.4 Conditioning on an event

在第 1.2 节中，我们解释了如何修改概率空间的概率测度，以纳入某个事件已经发生的假设。本节中，我们回顾在涉及随机变量时的这一情形。特别地，我们考虑一个随机变量 X，其分布由 pmf、cdf 或 pdf 表示，并解释在我们假设 $X \in \mathcal{S}$ 的情况下，对于任何属于 Borel σ-代数的集合 $\mathcal{S}$，它的分布将如何变化（请记住，这几乎包括你能想到的任何有用的集合）。

如果 X 是离散的，且其概率质量函数为 p_X，那么在给定 $X \in \mathcal{S}$ 的情况下，X 的条件概率质量函数为

$$p_{X|X\in\mathcal{S}}(x) := \mathrm{P}(X = x|X \in \mathcal{S}) \tag{2.56}$$

$$= \begin{cases} \frac{p_X(x)}{\sum_{s\in\mathcal{S}} p_X(s)} & \text{if } x \in \mathcal{S} \\ 0 \quad \text{otherwise.} \end{cases} \tag{2.57}$$

这是一个有效的pmf在限制在事件 $\{X$ 上的新概率空间 $\in \mathcal{S}\}$.

类似地，如果 X 与 pdf f_X 连续，则在事件 $X \in \mathcal{S}$ 给定的条件下，X 的条件累积分布函数为

$$F_{X|X\in\mathcal{S}}(x) := \mathrm{P}(X \le x|X \in \mathcal{S}) \tag{2.58}$$

$$= \frac{\mathrm{P}(X \le x, X \in \mathcal{S})}{\mathrm{P}(X \in \mathcal{S})} \tag{2.59}$$

$$= \frac{\int_{u\le x, u\in\mathcal{S}} f_X(u)\,\mathrm{d}u}{\int_{u\in\mathcal{S}} f_X(u)\,\mathrm{d}u}, \tag{2.60}$$

根据条件概率的定义，可以验证这是在新概率空间中的有效累积分布函数（cdf）。为了得到条件概率密度函数（pdf），我们只需对这个累积分布函数（cdf）进行求导,

$$f_{X|X\in\mathcal{S}}(x) := \frac{\mathrm{d}F_{X|X\in\mathcal{S}}(x)}{\mathrm{d}x}. \tag{2.61}$$

我们现在应用这些想法来说明几何随机变量和指数随机变量是无记忆的。

Example 2.4.1 (几何随机变量是无记忆的)**.** 我们反复掷一枚硬币直到得到正面，但在掷了几次（结果都是反面）之后暂停。假设每次掷币相互独立并且具有相同的偏置p (，即每一次掷币得到正面的概率为p)。在接下来的k次掷币中得到正面的概率是多少？也许令人惊讶的是，这个概率与从一开始经过k次掷币后得到一次正面的概率完全相同。

为严格建立这一点，我们计算在事件 $\{X > k_0\}$ (条件下几何随机变量 X 的条件 pmf，也就是说，在我们的示例中前 k_0 次都是反面)。应用 (2.56) 我们有

$$p_{X|X>k_0}(k) = \frac{p_X(k)}{\sum_{m=k_0+1}^{\infty} p_X(m)} \tag{2.62}$$

$$= \frac{(1-p)^{k-1}p}{\sum_{m=k_0+1}^{\infty}(1-p)^{m-1}p} \tag{2.63}$$

$$= (1-p)^{k-k_0-1}p \tag{2.64}$$

如果是 $k > k_0$，否则为零。我们已经利用了几何级数的事实

$$\sum_{m=k_0+1}^{\infty} \alpha^m = \frac{\alpha^{k_0+1}}{1-\alpha} \tag{2.65}$$

对于任何 $\alpha < 1$。

在新的概率空间中，计数从 $k_0 + 1$ 开始，条件概率质量函数是与原始变量相同参数的几何随机变量。前 k_0 次抛掷不会影响未来，一旦揭示它们是反面。

△

Example 2.4.2 (指数随机变量是无记忆的)**.**。假设你的电子邮件到达之间的间隔时间服从指数分布（在数小时的区间内这可能是一个很好的近似，如果你验证了请告诉我们）。你收到了一封电子邮件。直到你收到下一封电子邮件的时间服从具有某个参数 λ 的指数分布。在接下来的 t_0 分钟内没有电子邮件到达。令人惊讶的是，从那时起直到你收到下一封电子邮件的时间再次服从具有相同参数的指数分布，无论 t_0 的取值如何。就像几何随机变量一样，指数随机变量是无记忆的。

让我们严谨地证明这一点。我们计算参数为 λ 的指数随机变量 T 在事件 $\{T > t_0\}$– 下的条件分布函数，对于任意的 $t_0 > 0$–，通过应用 (2.60)。

$$F_{T|T>t_0}(t) = \frac{\int_{t_0}^{t} f_T(u)\,\mathrm{d}u}{\int_{t_0}^{\infty} f_T(u)\,\mathrm{d}u} \tag{2.66}$$

$$= \frac{e^{-\lambda t} - e^{-\lambda t_0}}{-e^{-\lambda t_0}} \tag{2.67}$$

$$= 1 - e^{-\lambda(t-t_0)}. \tag{2.68}$$

对t求导会得到一个从t_0开始的指数型概率密度函数$f_{T|T>t_0}(t) = \lambda e^{-\lambda(t-t_0)}$。

△

2.5 Functions of random variables

计算随机变量函数的分布在概率建模中通常非常有用。例如，如果我们使用随机变量 X 来建模电路中的电流，我们可能对跨越具有确定性电阻 r 的电阻器所耗散的功率 $Y{:=}\ rX^2$ 感兴趣。如果我们对随机变量 X 应用确定性函数 $g{:}\mathbb{R} \to \mathbb{R}$，那么结果 $Y{:=}\ g(X)$ 就是 *not* 一个确定性量。回忆一下，随机变量是从样本空间 Ω 到 $\mathbb{R}$ 的函数。如果 X 将 Ω 的元素映射到 $\mathbb{R}$，那么 Y 也会如此，因为 $Y(\omega) = g(X(\omega))$。这意味着 Y 也是一个随机变量。在本节中，我们将解释如何在已知 X 的分布时，刻画 Y 的分布。

如果 X 是离散的，那么从 X 的 pmf 计算 $g(X)$ 的 pmf 是直接的。

$$p_Y(y) = \mathrm{P}(Y = y) \tag{2.69}$$

$$= \mathrm{P}(g(X) = y) \tag{2.70}$$

$$= \sum_{\{x \mid g(x)=y\}} p_X(x). \tag{2.71}$$

如果 X 是连续的，则该过程更加微妙。我们首先通过应用定义来计算 Y 的累积分布函数（cdf）,

$$F_Y(y) = \mathrm{P}(Y \le y) \tag{2.72}$$

$$= \mathrm{P}(g(X) \le y) \tag{2.73}$$

$$= \int_{\{x \mid g(x)\le y\}} f_X(x)\,\mathrm{d}x, \tag{2.74}$$

最后的等式显然只有在X具有概率密度函数时才成立。然后，如果Y是可微分的，我们可以从其累积分布函数中获得Y的概率密度函数。这个想法可以用来证明关于高斯随机变量的一个有用结果。

Lemma 2.5.1 (高斯随机变量). *If X is a Gaussian random variable with mean μ and standard deviation σ, then*

$$U := \frac{X-\mu}{\sigma} \tag{2.75}$$

is a standard Gaussian random variable.

Proof. 我们应用(2.74)得到

$$F_U(u) = \mathrm{P}\left(\frac{X-\mu}{\sigma} \le u\right) \tag{2.76}$$

$$= \int_{(x-\mu)/\sigma \le u} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x \tag{2.77}$$

$$= \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}}\,\mathrm{d}w \qquad \text{by the change of variables } w = \frac{x-\mu}{\sigma}. \tag{2.78}$$

对 u 求导得到

$$f_U(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}, \tag{2.79}$$

所以 U 确实是一个标准高斯随机变量 . □

2.6 Generating random variables

模拟是概率建模中的一个基本工具。模拟模型的结果需要从其中包含的随机变量中进行采样。生成来自随机变量的样本的最广泛策略将过程分解为两个步骤:

1. 从单位区间 [0, 1] 均匀生成样本。 2. 转换均匀样本，使其具有所需的分布。

这里我们关注第二步，假设我们可以使用一个随机数生成器，它能够产生在 [0, 1] 上服从均匀分布的独立样本。构造高质量的均匀随机数生成器是一个重要的问题，这超出了这些笔记的范围。

2.6.1 Sampling from a discrete distribution

让 X 是一个离散随机变量，其概率质量函数为 p_X，而 U 是一个在 [0, 1] 区间内的均匀随机变量。我们的目标是将 U 的样本转换，使其按照 p_X 分布。我们用 x_1、x_2，⋯ 来表示在 p_X 下具有非零概率的值。

对于固定的 i，假设我们将 U 的所有样本在长度为 $p_X(x_i)$ 的区间内分配给 x_i。那么，从 U 中选取的一个样本被分配给 x_i 的概率恰好是 $p_X(x_i)$!

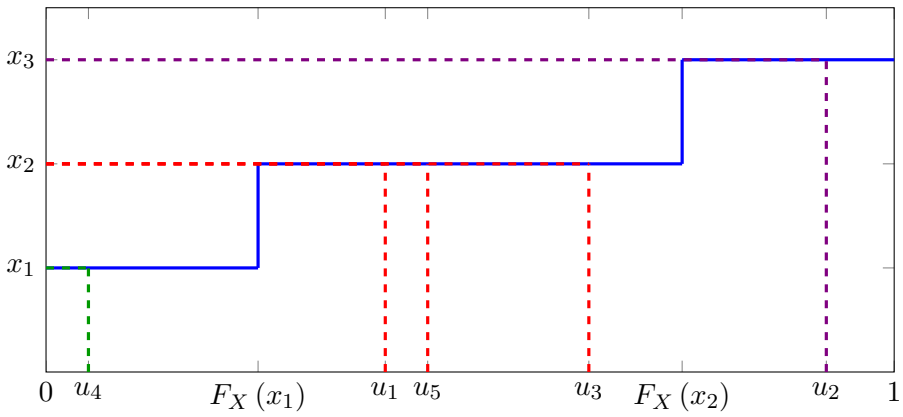


Figure 2.15: 第 2.6.1 节中所述从任意离散分布生成样本的方法示意图。离散随机变量的累积分布函数以蓝色显示。来自均匀分布的样本 u_4 和 u_2 分别映射为 x_1 和 x_3，而 u_1、u_3 和 u_5 则映射为 x_3。

非常方便地，单位区间可以划分为长度为 $p_X(x_i)$ 的区间。因此，我们可以通过从 U 中采样并设置来生成 X。

$$X = \begin{cases} x_1 & \text{if } 0 \leq U \leq p_X(x_1), \\ x_2 & \text{if } p_X(x_1) \leq U \leq p_X(x_1) + p_X(x_2), \\ \dots & \\ x_i & \text{if } \sum_{j=1}^{i-1} p_X(x_j) \leq U \leq \sum_{j=1}^{i} p_X(x_j), \\ \dots & \end{cases} \tag{2.80}$$

回忆一下，离散随机变量的累积分布函数等于

$$F_X(x) = \mathrm{P}(X \leq x) \tag{2.81}$$

$$= \sum_{x_i \leq x} p_X(x_i), \tag{2.82}$$

所以我们的算法归结为从U中获得一个样本u，然后输出x_i，使得$F_X(x_{i-1}) \leq u \leq F_X(x_i)$。如图2.15所示。

2.6.2 Inverse-transform sampling

逆变换采样通过对均匀样本应用确定性变换，使得从已知累积分布函数（cdf）的任意分布中采样成为可能。直观地，我们可以将其理解为第2.6.1节方法对连续分布的推广。

Algorithm 2.6.1 (逆变换采样). *Let X be a continuous random variable with cdf F_X and U a random variable that is uniformly distributed in* $[0, 1]$ *and independent of X.*

1. *Obtain a sample u of U.*
2. *Set $x := F_X^{-1}(u)$.*

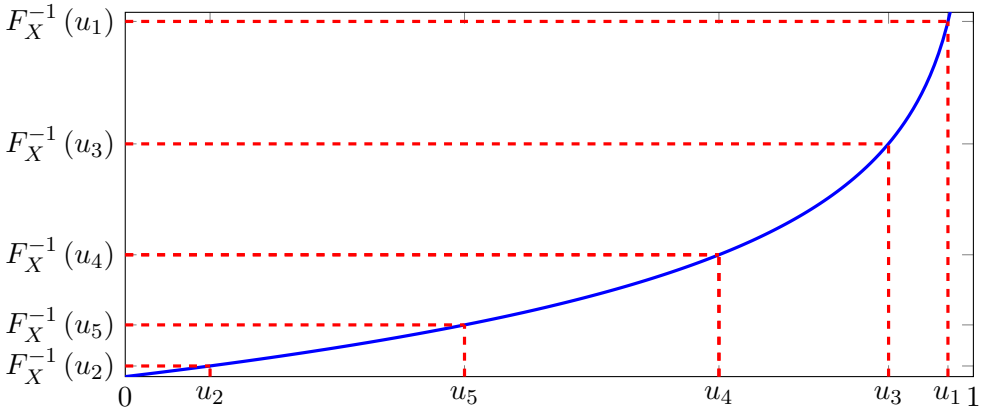


Figure 2.16: 来自参数 $\lambda = 1$ 的指数分布的样本，通过反向变换抽样获得，如示例 2.6.4 所述。样本 u_1，...，u_5 是从均匀分布中生成的。

细心的读者会指出，F_X 在每一点上可能并非可逆。为避免这一问题，我们定义 cdf 的广义逆为

$$F_X^{-1}(u) := \min_x \{F_X(x) = u\}. \tag{2.83}$$

该函数是良好定义的，因为所有的累积分布函数（cdfs）都是非递减的，因此在任何其不可逆的区间 $[x_1, x_2]$ 上，F_X 都等于一个常数 c。

我们现在证明算法2.6.1是正确的。

Theorem 2.6.2 (逆变换采样的工作原理)**.** *The distribution of* $Y = F_X^{-1}(U)$ *is the same as the distribution of* X*.*

Proof. 我们只需要证明 Y 的累积分布函数等于 F_X。我们有

$$\begin{aligned} F_Y(y) &= \mathrm{P}(Y \le y) && (2.84)\\ &= \mathrm{P}\left(F_X^{-1}(U) \le y\right) && (2.85)\\ &= \mathrm{P}(U \le F_X(y)) && (2.86)\\ &= \int_{u=0}^{F_X(y)} \mathrm{d}u && (2.87)\\ &= F_X(y), && (2.88) \end{aligned}$$

在步骤(2.86)中，我们必须考虑到我们正在使用累积分布函数的广义逆。这通过在第2.7.4节中证明的以下引理得以解决。

Lemma 2.6.3. *The events* $\left\{F_X^{-1}(U) \le y\right\}$ *and* $\{U \le F_X(y)\}$ *are equivalent.*

□

Example 2.6.4 (从指数分布中采样)**.** 设 X 为参数为 λ 的指数随机变量。其 cdf $F_X(x) := 1 - e^{-\lambda x}$ 在 $[0, \infty]$ 上是可逆的。其逆函数等于

$$F_X^{-1}(u) = \frac{1}{\lambda} \log\left(\frac{1}{1-u}\right). \tag{2.89}$$

F^{-1} 根据定理 2.6.2，$_X(U)$ 是参数为 λ 的指数随机变量。图 2.16 展示了 U 的样本如何被转换为 X 的样本。

△

2.7 Proofs

2.7.1 Proof of Lemma 2.2.9

对于任意固定常数 c_1 和 c_2

$$\lim_{n\to\infty} \frac{n - c_1}{n - c_2} = 1, \tag{2.90}$$

以便

$$\lim_{n\to\infty} \frac{n!}{(n-k)!\,(n-\lambda)^k} = \frac{n}{n-\lambda} \cdot \frac{n-1}{n-\lambda} \cdots \frac{n-k+1}{n-\lambda} = 1. \tag{2.91}$$

结果来自以下基本微积分恒等式:

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}. \tag{2.92}$$

2.7.2 Proof of Lemma 2.3.2

为建立 (2.31)

$$\begin{aligned}
\lim_{x\to-\infty} F_X(x) &= 1 - \lim_{x\to-\infty} \mathrm{P}(X > x) && (2.93)\\
&= 1 - \mathrm{P}(X > 0) - \lim_{n\to\infty} \sum_{i=0}^{n} \mathrm{P}(-i \geq X > -(i+1)) && (2.94)\\
&= 1 - \mathrm{P}\left(\lim_{n\to\infty} \{X > 0\} \cup \cup_{i=0}^{n} \{-i \geq X > -(i+1)\}\right) && (2.95)\\
&= 1 - \mathrm{P}(\Omega) = 0. && (2.96)
\end{aligned}$$

(2.32) 的证明来自这个结果。设 $Y = -X$，然后

$$\begin{aligned}
\lim_{x\to\infty} F_X(x) &= \lim_{x\to\infty} \mathrm{P}(X \leq x) && (2.97)\\
&= 1 - \lim_{x\to\infty} \mathrm{P}(X > x) && (2.98)\\
&= 1 - \lim_{x\to-\infty} \mathrm{P}(-X < x) && (2.99)\\
&= 1 - \lim_{x\to-\infty} F_Y(x) = 1 \quad \text{by (2.32)}. && (2.100)
\end{aligned}$$

最后，(2.33) 成立，因为 $\{X \leq a\} \subseteq \{X \leq b\}$。

2.7.3 Proof of Lemma 2.3.11

该结果是以下引理的推论。

Lemma 2.7.1.

$$\int_{-\infty}^{\infty} e^{-t^2}\,dt = \sqrt{\pi}. \tag{2.101}$$

Proof. 让我们定义

$$I = \int_{-\infty}^{\infty} e^{-x^2}\mathrm{d}x. \tag{2.102}$$

现在取平方并转换为极坐标,

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} e^{-x^2}\mathrm{d}x \int_{-\infty}^{\infty} e^{-y^2}\mathrm{d}y && (2.103)\\ &= \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} e^{-\left(x^2+y^2\right)}\mathrm{d}x\mathrm{d}y && (2.104)\\ &= \int_{\theta=0}^{2\pi}\int_{r=-\infty}^{\infty} re^{-\left(r^2\right)}\mathrm{d}\theta\mathrm{d}r && (2.105)\\ &= \pi e^{-\left(r^2\right)}]_0^{\infty} = \pi. && (2.106) \end{aligned}$$

□

为了完成证明，我们使用变量替换 $t = (x-\mu)/\sqrt{2}\sigma$。

2.7.4 Proof of Lemma 2.6.3

$\left\{F^{-1}{}_X\ (U) \le y\right\}$意味着
$\{U \le F_X(y)\}$

假设 $U > F_X(y)$，则对所有满足 $F_X(x) = U$ 的 x，都有 $x > y$，因为 cdf 是非递减的。特别地，$\min_x \{F_X(x) = U\} > y$。

$\{U \le F_X(y)\}$ 意味着 $\left\{F_X^{-1}(U) \le y\right\}$

假设 $\min_x \{F_X(x) = U\} > y$，则 $U > F_X(y)$，因为累积分布函数是非递减的。不等式是严格的，因为 $U = F_X(y)$ 将意味着 y 属于 $\{F_X(x) = U\}$，而这不可能发生，因为我们假设它小于该集合的最小值。

Chapter 3

Multivariate Random Variables

概率模型通常包含多个不确定的数值量。在本章中，我们将描述如何指定随机变量来表示这些量及其相互作用。在某些情况下，将这些随机变量分组为 **random vectors** 是有意义的，我们用上方带箭头的大写字母来表示：$\vec{X}$。这些随机向量的实现用小写字母表示：$\vec{x}$。

3.1 Discrete random variables

回顾一下，离散随机变量是取有限或可数无限值的数值量。在本节中，我们将解释如何操作共享同一概率空间的多个离散随机变量。

3.1.1 Joint probability mass function

如果在同一概率空间上定义了若干离散随机变量，我们通过它们的
joint probability mass function来刻画其概率行为，即每个变量取某个特定值的概率。

Definition 3.1.1 (联合概率质量函数). *Let $X : \Omega \to R_X$ and $Y : \Omega \to R_Y$ be discrete random variables (R_X and R_Y are discrete sets) on the same probability space $(\Omega, \mathcal{F}, \mathrm{P})$. The joint pmf of X and Y is defined as*

$$p_{X,Y}(x, y) := \mathrm{P}(X = x, Y = y) \quad . \tag{3.1}$$

In words, $p_{X,Y}(x, y)$ is the probability of X and Y being equal to x and y respectively.

Similarly, the joint pmf of a discrete random vector of dimension n

$$\vec{X} := \begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ X_n \end{bmatrix} \tag{3.2}$$

with entries X_i $\Omega \to R_{X_i}$ ($R_1, \ldots, R_n$ are all discrete sets) belonging to the same probability space is defined as

$$p_{\vec{X}}(\vec{x}) := \mathrm{P}(X_1 = \vec{x}_1, X_2 = \vec{x}_2, \ldots, X_n = \vec{x}_n). \tag{3.3}$$

与单个随机变量的 pmf 情形类似，如果我们考虑一个概率空间，其中样本空间在随机向量) 的情况下是 $R_X \times R_Y$[1] (or $R_{X_1} \times R_{X_2} \cdots \times R_{X_n}$，并且 σ-代数只是样本空间的幂集，那么联合 pmf 是一个有效的概率测度。这意味着联合 pmf *completely* 刻画了随机变量或随机向量，我们无需担心底层的概率空间。

根据概率测度的定义，联合概率质量函数（pmf）必须是非负的，并且其对所有可能的参数的总和必须等于一，

$$p_{X,Y}(x,y) \geq 0 \quad \text{for any } x \in R_X, y \in R_Y, \tag{3.4}$$

$$\sum_{x \in R_X} \sum_{y \in R_Y} p_{X,Y}(x,y) = 1. \tag{3.5}$$

根据全概率定律，联合 pmf 使我们能够得到 X 和 Y 属于任意集合 $\mathcal{S} \subseteq R_X \times R_Y$ 的概率，

$$\begin{aligned} \mathrm{P}((X,Y) \in \mathcal{S}) &= \mathrm{P}\left(\cup_{(x,y)\in\mathcal{S}} \{X = x, Y = y\}\right) \quad \text{(union of disjoint events)} && (3.6)\\ &= \sum_{(x,y)\in\mathcal{S}} \mathrm{P}(X = x, Y = y) && (3.7)\\ &= \sum_{(x,y)\in\mathcal{S}} p_{X,Y}(x,y). && (3.8) \end{aligned}$$

这些性质也适用于随机向量（以及多个随机变量的组）。对于任何随机向量 $\vec{X}$，

$$p_{\vec{X}}(\vec{x}) \geq 0, \tag{3.9}$$

$$\sum_{\vec{x}_1 \in R_1} \sum_{\vec{x}_2 \in R_2} \cdots \sum_{\vec{x}_n \in R_n} p_{\vec{X}}(\vec{x}) = 1. \tag{3.10}$$

$\vec{X}$ 属于离散集合 $\mathcal{S} \subseteq \mathbb{R}^n$ 的概率为

$$\mathrm{P}\left(\vec{X} \in \mathcal{S}\right) = \sum_{\vec{x} \in \mathcal{S}} p_{\vec{X}}(\vec{x}). \tag{3.11}$$

3.1.2 Marginalization

假设我们可以访问某一概率空间中多个随机变量的联合概率质量函数（pmf），但我们只对其中一个随机变量的行为感兴趣。为了计算该随机变量在特定值下的概率质量函数，我们固定该值并对其余随机变量求和。实际上，根据全概率法则

$$\begin{aligned} p_X(x) &= \mathrm{P}(X = x) && (3.12)\\ &= \mathrm{P}\left(\cup_{y\in R_Y} \{X = x, Y = y\}\right) \quad \text{(union of disjoint events)} && (3.13)\\ &= \sum_{y \in R_Y} \mathrm{P}(X = x, Y = y) && (3.14)\\ &= \sum_{y \in R_Y} p_{X,Y}(x,y). && (3.15) \end{aligned}$$

[1] This is the Cartesian product of the two sets, defined in Section A.2, which contains all possible pairs (x, y) where $x \in R_X$ and $y \in R_y$.

当联合pmf涉及两个以上的随机变量时，论点完全相同。这被称为**marginalizing**对其他随机变量的情况。在这种情况下，单个随机变量的pmf称为其**marginal pmf**。表3.1显示了一个联合pmf的示例及相应的边际pmf。

如果我们对计算随机向量中多个条目的联合概率质量函数（pmf）感兴趣，而不仅仅是一个条目，那么边际化过程本质上是相同的。pmf 仍然通过对其余条目求和获得。设 $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$ 是 $m < n$ 条目子集，属于一个 n 维随机向量 $\vec{X}$，$\vec{X}_{\mathcal{I}}$ 为对应的随机子向量。为了计算 $\vec{X}_{\mathcal{I}}$ 的联合pmf，我们对所有不在 $\mathcal{I}$ 中的条目求和，记作 $\{j_1, j_2, \ldots, j_{n-m}\} := \{1, 2, \ldots, n\} / \mathcal{I}$

$$p_{\vec{X}_{\mathcal{I}}}(\vec{x}_{\mathcal{I}}) = \sum_{\vec{x}_{j_1} \in R_{j_1}} \sum_{\vec{x}_{j_2} \in R_{j_2}} \cdots \sum_{\vec{x}_{j_{n-m}} \in R_{j_{n-m}}} p_{\vec{X}}(\vec{x}). \tag{3.16}$$

3.1.3 Conditional distributions

条件概率使我们能够在新信息揭示时更新我们对概率模型中量的未知性。随机变量的条件分布指定了当我们假设概率空间中的其他随机变量取固定值时，随机变量的行为。

Definition 3.1.2 (条件概率质量函数)**.** *The conditional probability mass function of Y given X, where X and Y are discrete random variables defined on the same probability space, is given by*

$$p_{Y|X}(y|x) = \mathrm{P}(Y = y | X = x) \tag{3.17}$$

$$= \frac{p_{X,Y}(x, y)}{p_X(x)} \qquad \text{if } p_X(x) > 0 \tag{3.18}$$

and is undefined otherwise.

条件概率质量函数 $p_{X|Y}(\cdot|y)$ 表征了我们对 X 在事件 $\{Y = y\}$ 条件下的不确定性。这个对象是 X 的有效概率质量函数，因此，如果 R_X 是 X 的取值范围

$$\sum_{x \in R_X} p_{X|Y}(x|y) = 1 \tag{3.19}$$

对于任何 y 都是良好定义的。然而，它是 *not* 一个 Y 的概率质量函数（pmf）。特别地，$\sum_{y \in R_Y} p_{X|Y}(x|y)$ 没有理由加起来等于一！
我们现在定义给定其他随机变量（或随机向量的条目）条件下多个随机变量（等价地为随机向量的子向量）的联合条件概率质量函数（pmf）。

Definition 3.1.3 (条件pmf)**.** *The conditional pmf of a discrete random subvector $\vec{X}_{\mathcal{I}}, \mathcal{I} \subseteq \{1, 2, \ldots, n\}$, given another subvector $\vec{X}_{\mathcal{J}}$ is*

$$p_{\vec{X}_{\mathcal{I}}|\vec{X}_{\mathcal{J}}}(\vec{x}_{\mathcal{I}}|\vec{x}_{\mathcal{J}}) := \frac{p_{\vec{X}}(\vec{x})}{p_{\vec{X}_{\mathcal{J}}}(\vec{x}_{\mathcal{J}})}, \tag{3.20}$$

where $\{j_1, j2, \ldots, j_{n-m}\} := \{1, 2, \ldots, n\} / \mathcal{I}$.

R

L \ $p_{L,R}$	0	1
0	$\frac{14}{20}$	$\frac{1}{20}$
1	$\frac{2}{20}$	$\frac{3}{20}$

p_L
$\frac{15}{20}$
$\frac{5}{20}$

$p_{L\|R}(\cdot\|0)$
$\frac{7}{8}$
$\frac{1}{8}$

$p_{L\|R}(\cdot\|1)$
$\frac{1}{4}$
$\frac{3}{4}$

p_R	$\frac{16}{20}$	$\frac{4}{20}$

$p_{R\|L}(\cdot\|0)$	$\frac{14}{15}$	$\frac{1}{15}$

$p_{R\|L}(\cdot\|1)$	$\frac{2}{5}$	$\frac{3}{5}$

Table 3.1: 联合、边际和条件概率质量函数（pmf）定义在例 3.1.5 中的随机变量 L 和 R。

条件概率质量函数 $p_{Y|X}(\cdot|x)$ 和 $p_{\vec{X}_{\mathcal{I}}|\vec{X}_{\mathcal{J}}}(\cdot|\vec{x}_{\mathcal{J}})$ 分别是在以 $X = x$ 或$\vec{X}_{\mathcal{J}} = \vec{x}_{\mathcal{J}}$为条件的概率空间中的有效概率质量函数。例如，它们必须非负且加和为一。

由条件概率质量函数（pmf）的定义，可以推导出离散随机变量和向量的链式法则。

Lemma 3.1.4 (离散随机变量和向量的链式法则)**.**

$$p_{X,Y}(x,y) = p_X(x)\, p_{Y|X}(y|x), \tag{3.21}$$

$$p_{\vec{X}}(\vec{x}) = p_{X_1}(\vec{x}_1)\, p_{X_2|X_1}(\vec{x}_2|\vec{x}_1) \dots p_{X_n|X_1,\dots,X_{n-1}}(\vec{x}_n|\vec{x}_1,\dots,\vec{x}_{n-1}) \tag{3.22}$$

$$= \prod_{i=1}^{n} p_{X_i|\vec{X}_{\{1,\dots,i-1\}}}\left(\vec{x}_i|\vec{x}_{\{1,\dots,i-1\}}\right), \tag{3.23}$$

where the order of indices in the random vector is arbitrary (any order works).

以下示例说明了边际和条件概率质量函数（pmf）的定义。

Example 3.1.5 (航班和降雨（续）)**.** 在例1.2.1中描述的概率空间内，我们定义了一个随机变量

$$L = \begin{cases} 1 & \text{if plane is late,} \\ 0 & \text{otherwise,} \end{cases} \tag{3.24}$$

用于表示飞机是否晚点。同样地,

$$R = \begin{cases} 1 & \text{it rains,} \\ 0 & \text{otherwise,} \end{cases} \tag{3.25}$$

表示是否下雨。等效地，这些随机变量只是指示器 $R = 1_{\text{rain}}$ 和 $L = 1_{\text{late}}$。表 3.1 显示了 L 和 R 的联合、边际和条件 pmf。

△

3.2 Continuous random variables

连续随机变量使我们能够建模连续量，而无需担心离散化。作为交换，用于操作它们的数学工具比离散情况稍微复杂一些。

3.2.1 Joint cdf and joint pdf

如同单变量连续随机变量的情况，我们通过它们属于Borel集（或等效地，区间的并集）的概率来描述定义在同一概率空间上的多个连续随机变量的行为。在这种情况下，我们考虑多维Borel集，它是单维Borel集的笛卡尔积。多维Borel集可以表示为多维区间或超矩形的并集（定义为单维区间的笛卡尔积）。**joint cdf** 汇总了随机变量属于形式为$(-\infty, r]$的区间笛卡尔积的概率，对于每一个$r \in \mathbb{R}$。

Definition 3.2.1 (联合累积分布函数)**.** *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space and* $X, Y \colon \Omega \to \mathbb{R}$ *random variables. The* ***joint cdf*** *of* X *and* Y *is defined as*

$$F_{X,Y}(x, y) := \mathrm{P}(X \le x, Y \le y). \tag{3.26}$$

In words, $F_{X,Y}(x, y)$ *is the probability of* X *and* Y *being smaller than* x *and* y *respectively.*

Let $\vec{X}\ \Omega \to \mathbb{R}^n$ *be a random vector of dimension* n *on a probability space* $(\Omega, \mathcal{F}, \mathrm{P})$*. The joint cdf of* $\vec{X}$ *is defined as*

$$F_{\vec{X}}(\vec{x}) := \mathrm{P}\left(\vec{X}_1 \le \vec{x}_1, \vec{X}_2 \le \vec{x}_2, \ldots, \vec{X}_n \le \vec{x}_n\right). \tag{3.27}$$

In words, $F_{\vec{X}}(\vec{x})$ *is the probability that* $\vec{X}_i \le \vec{x}_i$ *for all* $i = 1, 2, \ldots, n$.

我们现在记录联合累积分布函数的一些性质。

Lemma 3.2.2 (联合累积分布函数)**.**的性质

$$\lim_{x \to -\infty} F_{X,Y}(x, y) = 0, \tag{3.28}$$

$$\lim_{y \to -\infty} F_{X,Y}(x, y) = 0, \tag{3.29}$$

$$\lim_{x \to \infty, y \to \infty} F_{X,Y}(x, y) = 1, \tag{3.30}$$

$$F_{X,Y}(x_1, y_1) \le F_{X,Y}(x_2, y_2) \quad \text{if } x_2 \ge x_1, y_2 \ge y_1, \quad \text{i.e. } F_{X,Y} \text{ is nondecreasing.} \tag{3.31}$$

Proof. 证明沿用与引理2.3.2相同的思路。 □

联合累积分布函数完全指定了相应随机变量的行为。事实上，我们可以将任何Borel集合分解为若干个不相交的n维区间，并通过计算联合累积分布函数来求得它们的概率。我们以二元变量的情况为例说明这一点：

$$\begin{aligned}\mathrm{P}\left(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2\right) &= \mathrm{P}\left(\{X \leq x_2, Y \leq y_2\} \cap \{X > x_1\} \cap \{Y > y_1\}\right) && (3.32)\\ &= \mathrm{P}\left(X \leq x_2, Y \leq y_2\right) - \mathrm{P}\left(X \leq x_1, Y \leq y_2\right) && (3.33)\\ &\quad - \mathrm{P}\left(X \leq x_2, Y \leq y_1\right) + \mathrm{P}\left(X \leq x_1, Y \leq y_1\right) && (3.34)\\ &= F_{X,Y}\left(x_2, y_2\right) - F_{X,Y}\left(x_1, y_2\right) - F_{X,Y}\left(x_2, y_1\right) + F_{X,Y}\left(x_1, y_1\right).\end{aligned}$$

这意味着，与一维情形一样，要定义一个随机向量或一组随机变量，我们所需要做的只是定义它们的联合累积分布函数。我们不必担心底层的概率空间。

如果联合累积分布函数是可微的，我们可以对其求导，以获得**joint probability density function** 的 X 和 Y。与单变量随机变量的情况类似，这通常是一种更方便的指定联合分布的方法。

Definition 3.2.3 (联合概率密度函数)**.** *If the joint cdf of two random variables X, Y is differentiable, then their joint pdf is defined as*

$$f_{X,Y}\left(x, y\right) := \frac{\partial^2 F_{X,Y}\left(x, y\right)}{\partial x \partial y}. \tag{3.35}$$

If the joint cdf of a random vector $\vec{X}$ is differentiable, then its joint pdf is defined as

$$f_{\vec{X}}\left(\vec{x}\right) := \frac{\partial^n F_{\vec{X}}\left(\vec{x}\right)}{\partial \vec{x}_1 \, \partial \vec{x}_2 \cdots \partial \vec{x}_n}. \tag{3.36}$$

联合pdf应理解为一个n维密度，*not*作为一个概率（例如，它可以大于一）。在二维情况下，

$$\lim_{\Delta_x \to 0, \Delta_y \to 0} \mathrm{P}\left(x \leq X \leq x + \Delta_x, y \leq Y \leq y + \Delta_y\right) = f_{X,Y}\left(x, y\right) \Delta_x \Delta_y. \tag{3.37}$$

由于联合累积分布函数在每个变量中的单调性，联合概率质量函数总是非负的。

X 和 Y的联合概率密度函数允许我们通过对 $\mathcal{S}$ 积分来计算任何 Borell 集 $\mathcal{S} \subseteq \mathbb{R}^2$ 的概率。

$$\mathrm{P}\left((X, Y) \in \mathcal{S}\right) = \int_{(x,y) \in \mathcal{S}} f_{X,Y}\left(x, y\right) \, \mathrm{d}x \, \mathrm{d}y. \tag{3.38}$$

同样地，一个 n 维随机向量$\vec{X}$ 的联合概率密度函数可以计算$\vec{X}$ 属于某个 Borel 集 $\mathcal{S} \subseteq \mathbb{R}^n$ 的概率，

$$\mathrm{P}\left(\vec{X} \in \mathcal{S}\right) = \int_{\vec{x} \in \mathcal{S}} f_{\vec{X}}\left(\vec{x}\right) \, \mathrm{d}\vec{x}. \tag{3.39}$$

特别地，如果我们对整个空间 $\mathbb{R}^n$ 积分一个联合概率密度函数，那么根据全概率法则，它必须积分为 1。

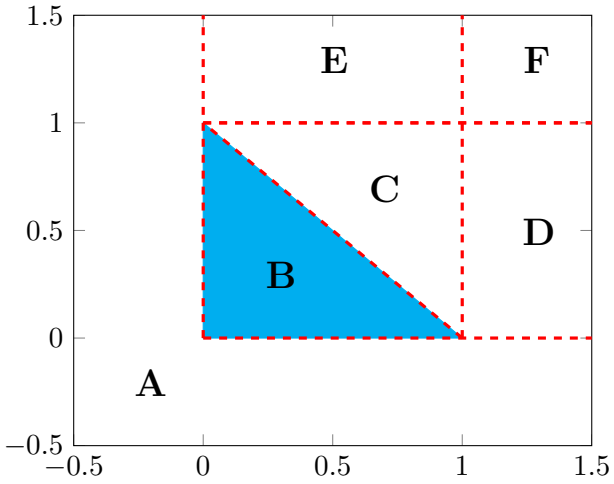


Figure 3.1: 示例 3.2.12 中的三角湖。

Example 3.2.4 (三角湖)**.** 一位生物学家正在追踪一只生活在湖中的水獭。她决定以概率方式建模水獭的位置。该湖恰好是三角形，如图 3.1 所示，因此我们可以通过集合表示它。

$$\text{Lake} := \{\vec{x} \mid \vec{x}_1 \geq 0, \vec{x}_2 \geq 0, \vec{x}_1 + \vec{x}_2 \leq 1\}. \tag{3.40}$$

这位生物学家不知道水獭在哪里，因此她将其位置建模为一个在湖面上均匀分布的随机向量$\vec{X}$。换言之，$\vec{X}$ 的联合概率密度函数是常数，

$$f_{\vec{X}}(\vec{x}) = \begin{cases} c & \text{if } \vec{x} \in \text{Lake}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.41}$$

为了找到归一化常数 c，我们利用这样一个事实：作为有效的联合概率密度函数 $f_{\vec{X}}$，它应该积分为 1。

$$\int_{x_1=-\infty}^{\infty}\int_{x_2=-\infty}^{\infty} c\,\mathrm{d}x_1\,\mathrm{d}x_2 = \int_{x_2=0}^{1}\int_{x_1=0}^{1-x_2} c\,\mathrm{d}x_1\,\mathrm{d}x_2 \tag{3.42}$$

$$= c\int_{x_2=0}^{1}(1-x_2)\,\mathrm{d}x_2 \tag{3.43}$$

$$= \frac{c}{2} = 1, \tag{3.44}$$

所以 $c = 2$。

我们现在计算$\vec{X}$ 的累积分布函数（cdf）。$F_{\vec{X}}(\vec{x})$ 表示水獭位于 $\vec{x}$ 点的西南方的概率。计算联合累积分布函数需要将范围划分为图 3.1 中所示的集合，并对联合概率密度函数（pdf）进行积分。如果 $\vec{x} \in A$，则 $F_{\vec{X}}(\vec{x}) = 0$，因为 $\mathrm{P}\left(\vec{X} \leq \vec{x}\right) = 0$。如果 $(\vec{x}) \in B$，

$$F_{\vec{X}}(\vec{x}) = \int_{u=0}^{\vec{x}_2}\int_{v=0}^{\vec{x}_1} 2\,\mathrm{d}v\,\mathrm{d}u = 2\vec{x}_1\vec{x}_2. \tag{3.45}$$

如果 $\vec{x} \in C$,

$$F_{\vec{X}}(\vec{x}) = \int_{u=0}^{1-\vec{x}_1} \int_{v=0}^{\vec{x}_1} 2\,\mathrm{d}v\,\mathrm{d}u + \int_{u=1-\vec{x}_1}^{\vec{x}_2} \int_{v=0}^{1-u} 2\,\mathrm{d}v\,\mathrm{d}u = 2\vec{x}_1 + 2\vec{x}_2 - \vec{x}_2^2 - \vec{x}_1^2 - 1. \tag{3.46}$$

如果 $\vec{x} \in D$,

$$F_{\vec{X}}(\vec{x}) = \mathrm{P}\left(\vec{X}_1 \le \vec{x}_1, \vec{X}_2 \le \vec{x}_2\right) = \mathrm{P}\left(\vec{X}_1 \le 1, \vec{X}_2 \le \vec{x}_2\right) = F_{\vec{X}}(1, \vec{x}_2) = 2\vec{x}_2 - \vec{x}_2^2, \tag{3.47}$$

最后一步由(3.46)得出。交换$\vec{x}_1$和$\vec{x}_2$，我们得到$F_{\vec{X}}(\vec{x}) = 2\vec{x}_1 - \vec{x}_1^2$对于$\vec{x} \in E$，理由相同。最后，$\vec{x} \in F$ $F_{\vec{X}}(\vec{x}) = 1$因为$\mathrm{P}\left(\vec{X}_1 \le x_1, \vec{X}_2 \le x_2\right) = 1$。将所有内容合并，

$$F_{\vec{X}}(\vec{x}) = \begin{cases} 0 & \text{if } \vec{x}_1 < 0 \text{ or } \vec{x}_2 < 0, \\ 2\vec{x}_1\vec{x}_2, & \text{if } \vec{x}_1 \ge 0, \vec{x}_2 \ge 0, \vec{x}_1 + \vec{x}_2 \le 1, \\ 2\vec{x}_1 + 2\vec{x}_2 - \vec{x}_2^2 - \vec{x}_1^2 - 1, & \text{if } \vec{x}_1 \le 1, \vec{x}_2 \le 1, \vec{x}_1 + \vec{x}_2 \ge 1, \\ 2\vec{x}_2 - \vec{x}_2^2, & \text{if } \vec{x}_1 \ge 1, 0 \le \vec{x}_2 \le 1, \\ 2\vec{x}_1 - \vec{x}_1^2, & \text{if } 0 \le \vec{x}_1 \le 1, \vec{x}_2 \ge 1, \\ 1, & \text{if } \vec{x}_1 \ge 1, \vec{x}_2 \ge 1. \end{cases} \tag{3.48}$$

△

3.2.2 Marginalization

我们现在讨论如何从联合 cdf 或联合 pdf 中刻画单个随机变量的边缘分布。考虑联合 cdf $F_{X,Y}(x,y)$。当 $x \to \infty$ 时，$F_{X,Y}(x,y)$ 的极限按定义等于 Y 小于 y 的概率，这恰好就是 Y 的边缘 cdf。更正式地说，

$$\lim_{x\to\infty} F_{X,Y}(x,y) = \lim_{n\to\infty} \mathrm{P}\left(\cup_{i=1}^{n} \{X \le i, Y \le y\}\right) \tag{3.49}$$

$$= \mathrm{P}\left(\lim_{n\to\infty} \{X \le n, Y \le y\}\right) \tag{3.50}$$

$$= \mathrm{P}(Y \le y) \tag{3.51}$$

$$= F_Y(y). \tag{3.52}$$

如果随机变量具有联合概率密度函数（pdf），我们也可以通过对 x 积分来计算边缘累积分布函数（cdf）。

$$F_Y(y) = \mathrm{P}(Y \le y) \tag{3.53}$$

$$= \int_{u=-\infty}^{y} \int_{x=-\infty}^{\infty} f_{X,Y}(x,u)\,\mathrm{d}x\,\mathrm{d}y. \tag{3.54}$$

对后者方程关于y求导，我们得到Y的边际概率密度函数。

$$f_Y(y) = \int_{x=-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x. \tag{3.55}$$

类似地，随机向量$\vec{X}$的一个由$\mathcal{I}:=\{i_1,i_2,\ldots,i_m\}$索引的子向量$\vec{X}_{\mathcal{I}}$的边缘 pdf 是通过对其余分量$\{j_1,j_2,\ldots,j_{n-m}\}:=\{1,2,\ldots,n\}/\mathcal{I}$进行积分得到的，

$$f_{\vec{X}_{\mathcal{I}}}\left(\vec{x}_{\mathcal{I}}\right)=\int_{\vec{x}_{j_1}}\int_{\vec{x}_{j_2}}\cdots\int_{\vec{x}_{j_{n-m}}}f_{\vec{X}}\left(\vec{x}\right)\,\mathrm{d}\vec{x}_{j_1}\,\mathrm{d}\vec{x}_{j_2}\cdots\mathrm{d}\vec{x}_{j_{n-m}}. \tag{3.56}$$

Example 3.2.5 (三角湖（续）)**.** 生物学家对水獭位于x_1以南的概率感兴趣。这个信息编码在随机向量的累积分布函数中，我们只需要在$x_2\to\infty$时取极限，以对x_2进行边缘化处理。

$$F_{X_1}\left(x_1\right)=\begin{cases}0 & \text{if } x_1<0,\\ 2x_1-x_1^2 & \text{if } 0\le x_1\le 1,\\ 1 & \text{if } x_1\ge 1.\end{cases} \tag{3.57}$$

为了获得 X_1 的边缘概率密度函数（pdf），它表示海獭位置的纬度，我们对边缘累积分布函数（cdf）求导

$$f_{X_1}\left(x_1\right)=\frac{\mathrm{d}F_{X_1}\left(x_1\right)}{\mathrm{d}x_1}=\begin{cases}2\left(1-x_1\right) & \text{if } 0\le x_1\le 1,\\ 0, & \text{otherwise.}\end{cases}. \tag{3.58}$$

或者，我们也可以在x_2 (上对联合均匀概率密度函数进行积分，我们鼓励你检查结果是相同的)。

△

3.2.3 Conditional distributions

在本节中，我们讨论如何在概率空间中获得给定其他随机变量信息的条件分布。首先，我们考虑两个随机变量的情况。与单变量分布的情况一样，我们可以通过应用条件概率的定义，定义给定事件形式 $\{(X,Y)\in\mathcal{S}\}$ 的两个随机变量的联合累积分布函数（cdf）和概率密度函数（pdf），其中 $\mathbb{R}^2$ 为任意Borel集合。

Definition 3.2.6 (联合条件累积分布函数和概率密度函数给定事件)**.** *Let X,Y be random variables with joint pdf $f_{X,Y}$ and let $\mathcal{S}\subseteq\mathbb{R}^2$ be any Borel set with nonzero probability, the conditional cdf and pdf of X and Y given the event $(X,Y)\in\mathcal{S}$ is defined as*

$$F_{X,Y|(X,Y)\in\mathcal{S}}\left(x,y\right):=\mathrm{P}\left(X\le x,Y\le y|\left(X,Y\right)\in\mathcal{S}\right) \tag{3.59}$$

$$=\frac{\mathrm{P}\left(X\le x,Y\le y,\left(X,Y\right)\in\mathcal{S}\right)}{\mathrm{P}\left(\left(X,Y\right)\in\mathcal{S}\right)} \tag{3.60}$$

$$=\frac{\int_{u\le x,v\le y,(u,v)\in\mathcal{S}}f_{X,Y}\left(u,v\right)\,du\,dv}{\int_{(u,v)\in\mathcal{S}}f_{X,Y}\left(u,v\right)\,du\,dv}, \tag{3.61}$$

$$f_{X,Y|(X,Y)\in\mathcal{S}}\left(x,y\right):=\frac{\partial^2F_{X,Y|(X,Y)\in\mathcal{S}}\left(x,y\right)}{\partial x\partial y}. \tag{3.62}$$

这个定义只对概率非零的事件成立。然而，形如 $\{X=x\}$ 的事件的概率等于零，因为该随机变量是连续的。事实上，

X的范围是不可数的，因此几乎每个事件$\{X = x\}$的概率必须为零，否则它们的并集的概率将是无界的。

那么，我们如何刻画在给定 $X = x$ 的情况下对 Y的不确定性呢？我们定义一个 **conditional pdf**，它刻画了我们在极限情况下试图做的事情，然后对其进行积分以获得条件 cdf。

Definition 3.2.7 (条件概率密度函数和累积分布函数
)**.** *If $F_{X,Y}$ is differentiable, then the conditional pdf of Y given X is defined as*

$$f_{Y|X}(y|x) := \frac{f_{X,Y}(x,y)}{f_X(x)} \quad \text{if } f_X(x) > 0 \tag{3.63}$$

and is undefined otherwise.

The conditional cdf of Y given X is defined as

$$F_{Y|X}(y|x) := \int_{u=-\infty}^{y} f_{Y|X}(u|x)\ du \quad \text{if } f_X(x) > 0 \tag{3.64}$$

and is undefined otherwise.

我们现在论证这一定义的合理性，而不仅仅是与(3.18)的类比。假设$f_X(x) > 0$。让我们用极限的形式来写条件概率密度函数的定义。我们有

$$f_X(x) = \lim_{\Delta_x \to 0} \frac{\mathrm{P}(x \le X \le x + \Delta_x)}{\Delta_x}, \tag{3.65}$$

$$f_{X,Y}(x,y) = \lim_{\Delta_x \to 0} \frac{1}{\Delta_x} \frac{\partial \mathrm{P}(x \le X \le x + \Delta_x, Y \le y)}{\partial y}. \tag{3.66}$$

这意味着

$$\frac{f_{X,Y}(x,y)}{f_X(x)} = \lim_{\Delta_x \to 0, \Delta_y \to 0} \frac{1}{\mathrm{P}(x \le X \le x + \Delta_x)} \frac{\partial \mathrm{P}(x \le X \le x + \Delta_x, Y \le y)}{\partial y}. \tag{3.67}$$

我们现在可以将条件累积分布函数写为

$$F_{Y|X}(y|x) = \int_{u=-\infty}^{y} \lim_{\Delta_x \to 0, \Delta_y \to 0} \frac{1}{\mathrm{P}(x \le X \le x + \Delta_x)} \frac{\partial \mathrm{P}(x \le X \le x + \Delta_x, Y \le u)}{\partial y}\, \mathrm{d}u \tag{3.68}$$

$$= \lim_{\Delta_x \to 0} \frac{1}{\mathrm{P}(x \le X \le x + \Delta_x)} \int_{u=-\infty}^{y} \frac{\partial \mathrm{P}(x \le X \le x + \Delta_x, Y \le u)}{\partial y}\, \mathrm{d}u \tag{3.69}$$

$$= \lim_{\Delta_x \to 0} \frac{\mathrm{P}(x \le X \le x + \Delta_x, Y \le y)}{\mathrm{P}(x \le X \le x + \Delta_x)} \tag{3.70}$$

$$= \text{极限}_{\Delta_x \to 0} \mathrm{P}(Y \le y | x \le X \le x + \Delta_x). \tag{3.71}$$

因此，我们可以将条件 cdf 解释为：当区间的宽度趋于零时，在 X 属于 x 附近的一个区间这一条件下，Y在 y 处的 cdf 的极限。

Remark 3.2.8. *Interchanging limits and integrals as in* (3.69) *is not necessarily justified in general. In this case it is, as long as the integral converges and the quantities involved are bounded.*

定义 3.2.7 的一个直接推论是连续随机变量的链式法则。

Lemma 3.2.9 (连续随机变量的链式法则)**.**

$$f_{X,Y}(x,y) = f_X(x) f_{Y|X}(y|x). \tag{3.72}$$

沿用二元情形中的相同思想，我们定义在给定随机向量其余部分时子向量的条件分布。

Definition 3.2.10 (条件概率密度函数
)**.** *The conditional pdf of a random subvector* $\vec{X}_{\mathcal{I}}$*,* $\mathcal{I} \subseteq \{1, 2$
$, \ldots, n\}$*, given the subvector* $\vec{X}_{\{1,\ldots,n\}/\mathcal{I}}$ *is*

$$f_{\vec{X}_{\mathcal{I}}|\vec{X}_{\{1,\ldots,n\}/\mathcal{I}}}\left(\vec{x}_{\mathcal{I}}|\vec{x}_{\{1,\ldots,n\}/\mathcal{I}}\right) := \frac{f_{\vec{X}}(\vec{x})}{f_{\vec{X}_{\{1,\ldots,n\}/\mathcal{I}}}\left(\vec{x}_{\{1,\ldots,n\}/\mathcal{I}}\right)}. \tag{3.73}$$

通常，利用随机向量的链式法则将随机向量的联合概率密度函数分解为条件概率密度函数来表示它是很有用的。

Lemma 3.2.11 (随机向量链式法则)**.** *The joint pdf of a random vector* $\vec{X}$ *can be decomposed into*

$$f_{\vec{X}}(\vec{x}) = f_{\vec{X}_1}(\vec{x}_1) f_{\vec{X}_2|\vec{X}_1}(\vec{x}_2|\vec{x}_1) \ldots f_{\vec{X}_n|\vec{X}_1,\ldots,\vec{X}_{n-1}}(\vec{x}_n|\vec{x}_1,\ldots,\vec{x}_{n-1}) \tag{3.74}$$

$$= \prod_{i=1}^{n} f_{\vec{X}_i|\vec{X}_{\{1,\ldots,i-1\}}}\left(\vec{x}_i|\vec{x}_{\{1,\ldots,i-1\}}\right). \tag{3.75}$$

Note that the order is arbitrary, you can reorder the components of the vector in any way you like.

Proof. 该结果源于递归地应用条件概率密度函数的定义。 □

Example 3.2.12 (三角湖（续)）**.** 生物学家在湖岸上发现了水獭。她站在湖的西侧，位于纬度 $x_1 = 0.75$，向东望去，水獭正好在她的正前方。因此，水獭也位于纬度 $x_1 = 0.75$，但她无法判断其距离。给定水獭的纬度 X_1 时，其位置分布由在给定 X_1 条件下经度 X_2 的条件概率密度函数来表征。

$$f_{X_2|X_1}(x_2|x_1) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)} \tag{3.76}$$

$$= \frac{1}{1-x_1}, \quad 0 \le x_2 \le 1 - x_1. \tag{3.77}$$

这位生物学家对水獭距离她小于 x_2 的概率感兴趣。该概率由条件累积分布函数给出。

$$F_{X_2|X_1}(x_2|x_1) = \int_{-\infty}^{x_2} f_{X_2|X_1}(u|x_1)\, \mathrm{d}u \tag{3.78}$$

$$= \frac{x_2}{1-x_1}. \tag{3.79}$$

概率 水獭距离小于 x_2 的概率为 $4x_2$，对于 $0 \le x$ $2 \le 1/4$.

△

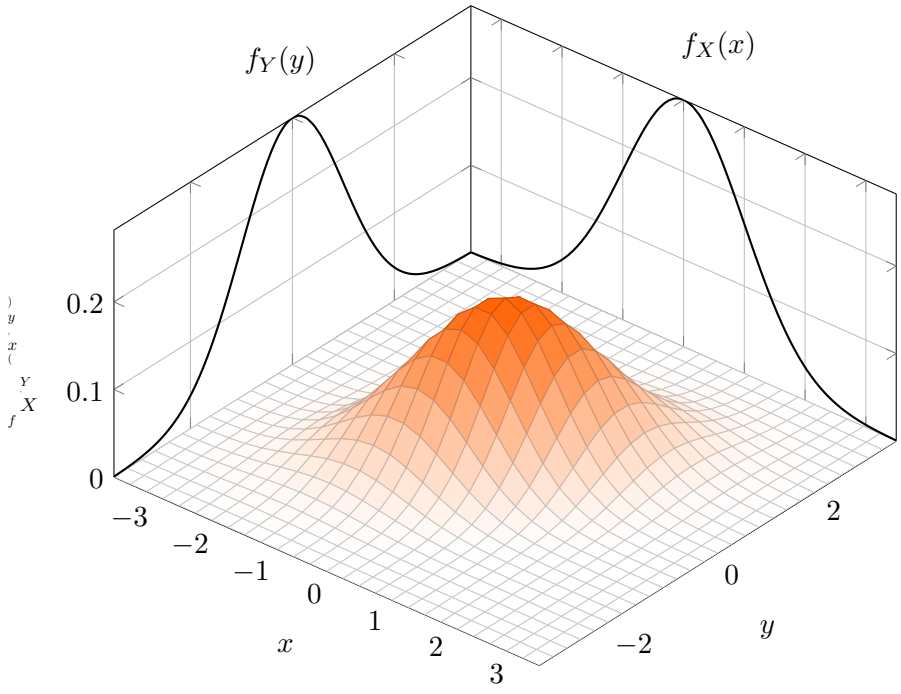


Figure 3.2: 二元高斯随机变量（X, Y）的联合概率密度函数，以及 X 和 Y 的边缘概率密度函数。

3.2.4 Gaussian random vectors

高斯随机向量是高斯随机变量的多维推广。它们由一个向量和一个矩阵来参数化，分别对应其均值和协方差矩阵（我们在第4章中为一般的多元随机变量定义这些量）。

Definition 3.2.13 (高斯随机向量)**.** *A Gaussian random vector $\vec{X}$ is a random vector with joint pdf*

$$f_{\vec{X}}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})\right) \tag{3.80}$$

where the mean vector $\vec{\mu} \in \mathbb{R}^n$ and the covariance matrix Σ, which is symmetric and positive definite, parametrize the distribution. A Gaussian distribution with mean $\vec{\mu}$ and covariance matrix Σ is usually denoted by $\mathcal{N}(\vec{\mu}, \Sigma)$.

高斯随机向量的一个基本性质是，对它们进行线性变换总会得到联合分布仍然为高斯分布的向量。我们不会对这一结果进行形式化证明，但其证明与引理 2.5.1 类似（事实上，这是该结果的一个多维推广）。

Theorem 3.2.14 (高斯随机向量的线性变换是高斯的)**.** *Let $\vec{X}$ be a Gaussian random vector of dimension n with mean $\vec{\mu}$ and covariance matrix Σ. For any matrix $A \in \mathbb{R}^{m\times n}$ and $\vec{b} \in \mathbb{R}^m$ $\vec{Y} = A\vec{X} + \vec{b}$ is a Gaussian random vector with mean $A\vec{\mu} + \vec{b}$ and covariance matrix $A\Sigma A^T$.*

该结果的一个推论是，高斯随机向量的一个子向量的联合概率密度函数仍为高斯分布。

Corollary 3.2.15 (高斯随机向量的边缘分布是高斯分布)**.** *The joint pdf of any subvector of a Gaussian random vector is Gaussian. Without loss of generality, assume that the subvector $\vec{X}$ consists of the first m entries of the Gaussian random vector,*

$$\vec{Z} := \begin{bmatrix} \vec{X} \\ \vec{Y} \end{bmatrix}, \qquad \text{with mean} \quad \vec{\mu} := \begin{bmatrix} \mu_{\vec{X}} \\ \mu_{\vec{Y}} \end{bmatrix} \tag{3.81}$$

and covariance matrix

$$\Sigma_{\vec{Z}} = \begin{bmatrix} \Sigma_{\vec{X}} & \Sigma_{\vec{X}\vec{Y}} \\ \Sigma_{\vec{X}\vec{Y}}^T & \Sigma_{\vec{Y}} \end{bmatrix}. \tag{3.82}$$

Then $\vec{X}$ is a Gaussian random vector with mean $\mu_{\vec{X}}$ and covariance matrix $\Sigma_{\vec{X}}$.

Proof. 请注意

$$\vec{X} = \begin{bmatrix} I_m & 0_{m\times n-m} \\ 0_{n-m\times m} & 0_{n-m\times n-m} \end{bmatrix} \begin{bmatrix} \vec{X} \\ \vec{Y} \end{bmatrix} = \begin{bmatrix} I_m & 0_{m\times n-m} \\ 0_{n-m\times m} & 0_{n-m\times n-m} \end{bmatrix} \vec{Z}, \tag{3.83}$$

其中 $I \in \mathbb{R}^{m\times m}$ 是单位矩阵，$0_{c\times d}$ 表示维度为 $c \times d$ 的零矩阵。该结果随后由定理 3.2.14 得出。□

图3.2展示了一个二维高斯随机变量的联合概率密度函数及其边缘概率密度函数。

3.3 Joint distributions of discrete and continuous variables

概率模型通常包含离散和连续随机变量。然而，离散和连续随机变量的联合概率质量函数（pmf）或概率密度函数（pdf）并不明确。在这种情况下，为了指定联合分布，我们使用它们的边际和条件概率质量函数（pmf）和概率密度函数（pdf）。

假设我们有一个连续随机变量 C 和一个取值范围为 R_D 的离散随机变量 D。我们将给定 D 时 C 的条件 cdf 和 pdf 定义如下。

Definition 3.3.1 (给定离散随机变量
)**.** *Let C and D be a continuous and a discrete random variable defined on the same probability space. Then, the conditional cdf and pdf of C given D are of the form* 的连续随机变量的条件累积分布函数和概率密度函数

$$F_{C|D}(c|d) := \mathrm{P}(C \le c|d), \tag{3.84}$$

$$f_{C|D}(c|d) := \frac{\mathrm{d}F_{C|D}(c|d)}{\mathrm{d}c}. \tag{3.85}$$

我们通过计算加权和，从条件 cdf 和 pdf 得到 C 的边缘 cdf 和 pdf。

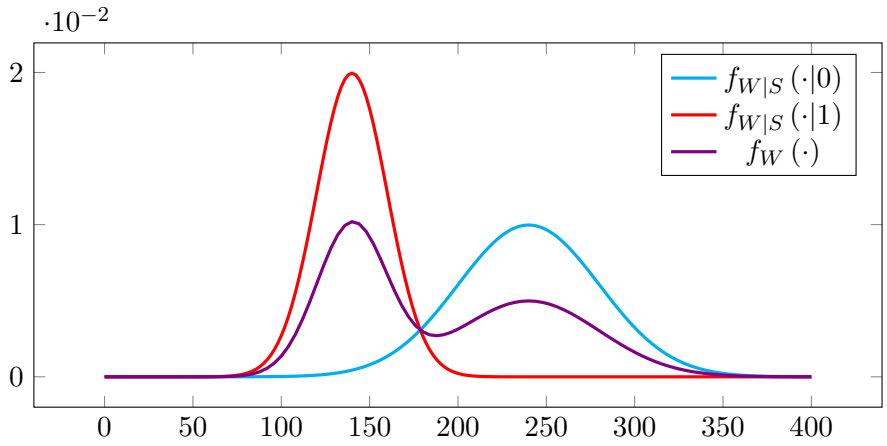


Figure 3.3: 例3.3.3中熊的体重W的条件分布和边缘分布。

Lemma 3.3.2. *Let $F_{C|D}$ and $f_{C|D}$ be the conditional cdf and pdf of a continuous random variable C given a discrete random variable D. Then,*

$$F_C(c) = \sum_{d \in R_D} p_D(d) F_{C|D}(c|d), \tag{3.86}$$

$$f_C(c) = \sum_{d \in R_D} p_D(d) f_{C|D}(c|d). \tag{3.87}$$

Proof. 事件 $\{D = d\}$ 是整个概率空间的一个划分（它们必须发生其中之一，并且它们是互不相交的），因此

$$F_C(c) = \mathrm{P}(C \le c) \tag{3.88}$$

$$= \sum_{d \in R_D} \mathrm{P}(D = d)\,\mathrm{P}(C \le c|d) \quad \text{by the Law of Total Probability} \tag{3.89}$$

$$= \sum_{d \in R_D} p_D(d) F_{C|D}(c|d). \tag{3.90}$$

现在，通过求导即可得到（3.87）。 □

将离散的边缘概率质量函数（pmf）与连续的条件分布相结合，使我们能够定义 **mixture models**，其中数据来自一个连续分布，而其参数是从一个离散集合中选择的。如果将高斯分布用作连续分布，这将得到一个高斯混合模型。拟合高斯混合模型是一种用于数据聚类的流行技术。

Example 3.3.3 (黄石公园的灰熊)**.** 一位科学家正在收集黄石公园灰熊的数据。结果表明，雄性灰熊的体重可以很好地用一个均值为 240 千克、标准差为 40 千克的高斯随机变量来建模，而雌性灰熊的体重可以很好地用一个均值为 140 千克、标准差为 20 千克的高斯随机变量来建模。雌性和雄性的数量大致相同。

所有灰熊体重的分布因此可以通过一个高斯混合模型来表示，该模型包括一个连续随机变量 W 来表示体重，以及一个离散随机变量 S 来表示熊的性别。S 是参数为 1/2 的伯努利分布，W 在 $S=$ 为 0（雄性）时为 $\mathcal{N}(240, 1600)$，而 W 在 $S=$ 为 1（雌性）时为 $\mathcal{N}(140, 400)$。根据 (3.87)，W 的概率密度函数因此具有以下形式

$$f_W(w) = \sum_{s=0}^{1} p_S(s) f_{W|S}(w|s) \tag{3.91}$$

$$= \frac{1}{2\sqrt{2\pi}} \left(\frac{e^{-\frac{(w-240)^2}{3200}}}{40} + \frac{e^{-\frac{(w-140)^2}{800}}}{20} \right). \tag{3.92}$$

图3.3显示了W的条件分布和边际分布。

△

在给定连续随机变量 C 的条件下，定义离散随机变量 D 的条件概率质量函数是具有挑战性的，因为事件 $\{C=c\}$ 的概率为零。我们遵循定义 3.2.7 中的相同方法，并将条件概率质量函数定义为一个极限。

Definition 3.3.4 (给定连续随机变量条件下离散随机变量的条件概率质量函数)**.** *Let C and D be a continuous and a discrete random variable defined on the same probability space. Then, the conditional pmf of D given C is defined as*

$$p_{D|C}(d|c) := \lim_{\Delta \to 0} \frac{\mathrm{P}(D=d, c \le C \le c+\Delta)}{\mathrm{P}(c \le C \le c+\Delta)}. \tag{3.93}$$

类似于引理3.3.2，通过计算加权和，我们从条件概率质量函数得到 D 的边缘概率质量函数。

Lemma 3.3.5. *Let $p_{D|C}$ be the conditional pmf of a discrete random variable D given a continuous random variable C. Then,*

$$p_D(d) = \int_{c=-\infty}^{\infty} f_C(c)\, p_{D|C}(d|c)\ dc. \tag{3.94}$$

Proof. 我们不会给出形式化的证明，而是提供一种直观的论证，这种论证可以被严格化。如果我们取一组关于 c 的网格值，这些值位于宽度为 Δ 的网格 $\ldots, c_{-1}, c_0, c_1, \ldots$ 上，那么

$$p_D(d) = \sum_{i=-\infty}^{\infty} \mathrm{P}(D=d, c_i \le C \le c_i+\Delta) \tag{3.95}$$

根据全概率定律。取$\Delta \to 0$ 的极限，求和变为积分，于是我们有

$$p_D(d) = \int_{c=-\infty}^{\infty} \lim_{\Delta \to 0} \frac{\mathrm{P}(D=d, c \le C \le c+\Delta)}{\Delta} \mathrm{d}c \tag{3.96}$$

$$= \int_{c=-\infty}^{\infty} \lim_{\Delta \to 0} \frac{\mathrm{P}(c \le C \le c+\Delta)}{\Delta} \cdot \frac{\mathrm{P}(D=d, c \le C \le c+\Delta)}{\mathrm{P}(c \le C \le c+\Delta)} \mathrm{d}c \tag{3.97}$$

$$= \int_{c=-\infty}^{\infty} f_C(c)\, p_{D|C}(d|c)\ \mathrm{d}c. \tag{3.98}$$

由于 $f_C(c) = \lim_{\Delta \to 0} \frac{\mathrm{P}(c \le C \le c+\Delta)}{\Delta}$。 □

将连续边际分布与离散条件分布相结合，在贝叶斯统计模型中特别有用，如以下示例所示（有关更多信息，请参见第10章）。连续分布用于量化我们对离散分布参数的不确定性。

Example 3.3.6 (贝叶斯硬币抛掷)**.** 你叔叔打赌十美元，认为硬币抛掷会是正面。你怀疑硬币有偏，但不确定偏差的程度。为了建模这种不确定性，你将偏差表示为一个连续随机变量 B，其概率密度函数如下：

$$f_B(b) = 2b \quad \text{for } b \in [0,1] \ . \tag{3.99}$$

现在，您可以使用引理 3.3.5 计算硬币正面朝上的概率，记作 X。在偏差 B 的条件下，硬币投掷的结果是参数为 B 的伯努利分布。

$$\begin{aligned} p_X(1) &= \int_{b=-\infty}^{\infty} f_B(b)\, p_{X|B}(1|b)\ \mathrm{d}b & (3.100)\\ &= \int_{b=0}^{1} 2b^2\, \mathrm{d}b & (3.101)\\ &= \frac{2}{3}. & (3.102) \end{aligned}$$

根据您的模型，硬币正面朝上的概率是2/3。 △

以下引理提供了联合分布的连续和离散随机变量的链式法则的类比。

Lemma 3.3.7 (联合分布的连续和离散随机变量的链式法则)**.** *Let C be a continuous random variable with conditional pdf $f_{C|D}$ and D a discrete random variable with conditional pmf $p_{D|C}$. Then,*

$$p_D(d)\, f_{C|D}(c|d) = f_C(c)\, p_{D|C}(d|c)\,. \tag{3.103}$$

Proof. 应用这些定义，

$$\begin{aligned} p_D(d)\, f_{C|D}(c|d) &= \lim_{\Delta\to 0} \mathrm{P}(D=d)\, \frac{\mathrm{P}(c \le C \le c+\Delta | D=d)}{\Delta} & (3.104)\\ &= \lim_{\Delta\to 0} \frac{\mathrm{P}(D=d, c \le C \le c+\Delta)}{\Delta} & (3.105)\\ &= \lim_{\Delta\to 0} \frac{\mathrm{P}(c \le C \le c+\Delta)}{\Delta} \cdot \frac{\mathrm{P}(D=d, c \le C \le c+\Delta)}{\mathrm{P}(c \le C \le c+\Delta)} & (3.106)\\ &= f_C(c)\, p_{D|C}(d|c)\,. & (3.107) \end{aligned}$$

□

Example 3.3.8 (黄石公园的灰熊（续））**.** 这位科学家用她的双筒望远镜观察到一只熊。根据其体型，她估计其体重为180千克。这只熊是雄性的概率是多少?

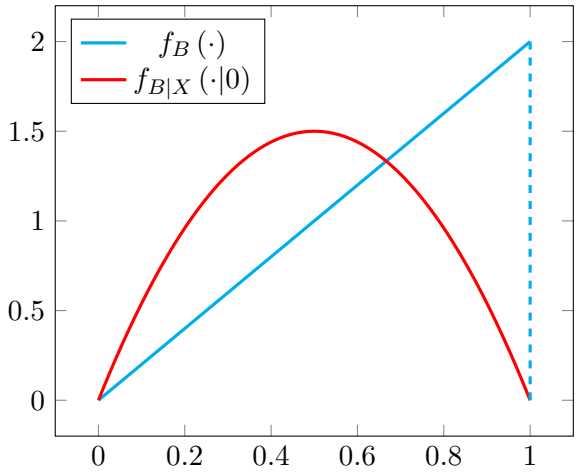


Figure 3.4: 例 3.3.9 中硬币抛掷偏置的条件分布和边缘分布。

我们应用引理 3.3.7 来计算

$$p_{S|W}(0|180) = \frac{p_S(0) f_{W|S}(180|0)}{f_W(180)} \tag{3.108}$$

$$= \frac{\frac{1}{40}\exp\left(-\frac{60^2}{3200}\right)}{\frac{1}{40}\exp\left(-\frac{60^2}{3200}\right) + \frac{1}{20}\exp\left(-\frac{40^2}{800}\right)} \tag{3.109}$$

$$= 0.545. \tag{3.110}$$

根据概率模型，它是男性的概率是 0.545。

△

Example 3.3.9 (贝叶斯硬币投掷（续））. 硬币落在反面。你决定在此信息的条件下重新计算偏倚的分布。根据引理 3.3.7

$$f_{B|X}(b|0) = \frac{f_B(b) p_{X|B}(0|b)}{p_X(0)} \tag{3.111}$$

$$= \frac{2b(1-b)}{1/3} \tag{3.112}$$

$$= 6b(1-b). \tag{3.113}$$

条件在结果的基础上，偏差的概率密度函数现在是集中在中间，而不是像以前那样集中在接近1的位置，如图3.4所示。

△

3.4 Independence

在本节中，我们定义随机变量和随机向量的独立性与条件独立性。

3.4.1 Definition

当关于随机变量X的知识不会影响我们对另一个随机变量Y的不确定性时，我们称X和Y是**independent**。形式上，这体现在边缘和条件累积分布函数（cdf）以及条件概率质量函数（pmf）或概率密度函数（pdf）必须相等，即。

$$F_Y(y) = F_{Y|X}(y|x) \tag{3.114}$$

和

$$p_Y(y) = p_{Y|X}(y|x) \quad \text{or} \quad f_Y(y) = f_{Y|X}(y|x), \tag{3.115}$$

取决于变量是离散还是连续，对于任意 x 以及任意 y，只要条件分布是良好定义的。等价地，联合 cdf 与条件 pmf 或 pdf 可分解为各自的边缘分布。

Definition 3.4.1 (独立随机变量)**.** *Two random variables X and Y are independent if and only if*

$$F_{X,Y}(x,y) = F_X(x) F_Y(y), \quad \text{for all } (x,y) \in \mathbb{R}^2. \tag{3.116}$$

If the variables are discrete, the following condition is equivalent

$$p_{X,Y}(x,y) = p_X(x) p_Y(y), \quad \text{for all } x \in R_X, y \in R_Y. \tag{3.117}$$

If the variables are continuous have joint and marginal pdfs, the following condition is equivalent

$$f_{X,Y}(x,y) = f_X(x) f_Y(y), \quad \text{for all } (x,y) \in \mathbb{R}^2. \tag{3.118}$$

我们现在将该定义扩展，以涵盖彼此不提供信息的多个随机变量（或等价地，随机向量中的多个分量）。

Definition 3.4.2 (独立随机变量)**.** *The n entries X_1, X_2, ..., X_n in a random vector $\vec{X}$ are independent if and only if*

$$F_{\vec{X}}(\vec{x}) = \prod_{i=1}^{n} F_{X_i}(\vec{x}_i), \tag{3.119}$$

which is equivalent to

$$p_{\vec{X}}(\vec{x}) = \prod_{i=1}^{n} p_{X_i}(\vec{x}_i) \tag{3.120}$$

for discrete vectors and

$$f_{\vec{X}}(\vec{x}) = \prod_{i=1}^{n} f_{X_i}(\vec{x}_i) \tag{3.121}$$

for continuous vectors, if the joint pdf exists.

下面的例子表明，两两独立性*not imply*独立性。

Example 3.4.3 (两两独立并不意味着联合独立). 设X_1和X_2为相互独立且无偏的硬币抛掷结果。令X_3为事件{X_1与X_2具有相同结果}的指示变量,

$$X_3 = \begin{cases} 1 & \text{if } X_1 = X_2, \\ 0 & \text{if } X_1 \neq X_2. \end{cases} \tag{3.122}$$

X_3的概率质量函数是

$$p_{X_3}(1) = p_{X_1,X_2}(1,1) + p_{X_1,X_2}(0,0) = \frac{1}{2}, \tag{3.123}$$

$$p_{X_3}(0) = p_{X_1,X_2}(0,1) + p_{X_1,X_2}(1,0) = \frac{1}{2}. \tag{3.124}$$

X_1 和 X_2 假设上是独立的。X_1 和 X_3 是独立的，因为

$$p_{X_1,X_3}(0,0) = p_{X_1,X_2}(0,1) = \frac{1}{4} = p_{X_1}(0)\, p_{X_3}(0), \tag{3.125}$$

$$p_{X_1,X_3}(1,0) = p_{X_1,X_2}(1,0) = \frac{1}{4} = p_{X_1}(1)\, p_{X_3}(0), \tag{3.126}$$

$$p_{X_1,X_3}(0,1) = p_{X_1,X_2}(0,0) = \frac{1}{4} = p_{X_1}(0)\, p_{X_3}(1), \tag{3.127}$$

$$p_{X_1,X_3}(1,1) = p_{X_1,X_2}(1,1) = \frac{1}{4} = p_{X_1}(1)\, p_{X_3}(1). \tag{3.128}$$

X_2 和 X_3 也是独立的（推理过程相同）。
然而，X_1、X_2 和 X_3 是否都是独立的?

$$p_{X_1,X_2,X_3}(1,1,1) = \mathrm{P}(X_1 = 1, X_2 = 1) = \frac{1}{4} \neq p_{X_1}(1)\, p_{X_2}(1)\, p_{X_3}(1) = \frac{1}{8}. \tag{3.129}$$

它们不是，这很有道理，因为 X_3 是 X_1 和 X_2 的函数。 △

Conditional independence 表示在已知另一个随机变量的情况下，两个随机变量相互独立。

Definition 3.4.4 (条件独立的随机变量). *Two random variables X and Y are independent with respect to another random variable Z if and only if*

$$F_{X,Y\,|\,Z}(x,y\,|\,z) = F_{X\,|\,Z}(x\,|\,z)\, F_{Y\,|\,Z}(y\,|\,z), \quad \text{for all } (x,y) \in \mathbb{R}^2, \tag{3.130}$$

and any z for which the conditional cdfs are well defined. If the variables are discrete, the following condition is equivalent

$$p_{X,Y\,|\,Z}(x,y\,|\,z) = p_{X\,|\,Z}(x\,|\,z)\, p_{Y\,|\,Z}(y\,|\,z), \quad \text{for all } x \in R_X, y \in R_Y, \tag{3.131}$$

and any z for which the conditional pmfs are well defined. If the variables are continuous have joint and marginal pdfs, the following condition is equivalent

$$f_{X,Y\,|\,Z}(x,y\,|\,z) = f_{X\,|\,Z}(x\,|\,z)\, f_{Y\,|\,Z}(y\,|\,z), \quad \text{for all } (x,y) \in \mathbb{R}^2, \tag{3.132}$$

and any z for which the conditional pmfs are well defined.

定义可以扩展到依赖于多个随机变量。

Definition 3.4.5 (条件独立的随机变量). *The components of a sub- vector* $\vec{X}_{\mathcal{I}}$, $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$ *are conditionally independent given another subvector* $\vec{X}_{\mathcal{J}}$, $\mathcal{J} \subseteq \{1, 2, \ldots, n\}$, *if and only if*

$$F_{\vec{X}_{\mathcal{I}}|\vec{X}_{\mathcal{J}}}\left(\vec{x}_{\mathcal{I}}|\vec{x}_{\mathcal{J}}\right) = \prod_{i\in\mathcal{I}} F_{X_i|\vec{X}_{\mathcal{J}}}\left(\vec{x}_i|\vec{x}_{\mathcal{J}}\right), \tag{3.133}$$

which is equivalent to

$$p_{\vec{X}_{\mathcal{I}}|\vec{X}_{\mathcal{J}}}\left(\vec{x}_{\mathcal{I}}|\vec{x}_{\mathcal{J}}\right) = \prod_{i\in\mathcal{I}} p_{X_i|\vec{X}_{\mathcal{J}}}\left(\vec{x}_i|\vec{x}_{\mathcal{J}}\right) \tag{3.134}$$

for discrete vectors and

$$f_{\vec{X}_{\mathcal{I}}|\vec{X}_{\mathcal{J}}}\left(\vec{x}_{\mathcal{I}}|\vec{x}_{\mathcal{J}}\right) = \prod_{i\in\mathcal{I}} f_{X_i|\vec{X}_{\mathcal{J}}}\left(\vec{x}_i|\vec{x}_{\mathcal{J}}\right) \tag{3.135}$$

for continuous vectors if the conditional joint pdf exists.

如示例 1.3.5 和 1.3.6 所示，独立性 **not** 并不意味着条件独立性，反之亦然。

3.4.2 Variable dependence in probabilistic modeling

设计概率模型时，一个基本的考虑因素是不同变量之间的依赖关系，即哪些变量是独立的或条件独立的。尽管这可能听起来令人惊讶，但如果变量数量很大，引入一些独立性假设可能是使模型可处理的必要条件，即使我们知道所有变量都是相关的。为了解释这一点，考虑一个关于美国总统选举的模型，其中有50个随机变量，每个代表一个州。如果这些变量只有两个可能的值（表示哪个候选人赢得该州），它们分布的联合概率质量函数（pmf）有$2^{50} - 1 \geq 10^{15}$个自由度。即使拥有全世界所有的计算机内存，我们也无法存储这个pmf！相比之下，如果我们假设所有变量都是独立的，那么该分布只有50个自由参数。当然，这不一定是一个好主意，因为未能表示依赖关系可能会严重影响模型的预测准确性，如下面的例子3.5.1所示。在可处理性和准确性之间取得平衡是概率建模中的一个关键挑战。

我们现在说明如何利用概率模型中随机变量的依赖结构，通过适当的联合概率质量函数（pmf）或概率密度函数（pdf）分解，减少描述分布的参数数量。考虑三个伯努利随机变量 A、B 和 C。一般来说，我们需要 $7 = 2^3 - 1$ 个参数来描述 pmf。然而，如果 B 和 C 在给定 A 的条件下是条件独立的，我们可以进行如下分解

$$p_{A,B,C} = p_A \, p_{B\,|\,A} \, p_{C\,|\,A} \tag{3.136}$$

仅依赖于五个参数（一个用于p_A，两个分别用于$p_{B\,|\,A}$和$p_{C\,|\,B}$）。需要注意的是，还有许多其他可能的分解方式，这些方式并未利用依赖假设，例如例如

$$p_{A,B,C} = p_B \, p_{A\,|\,B} \, p_{C\,|\,A,B}. \tag{3.137}$$

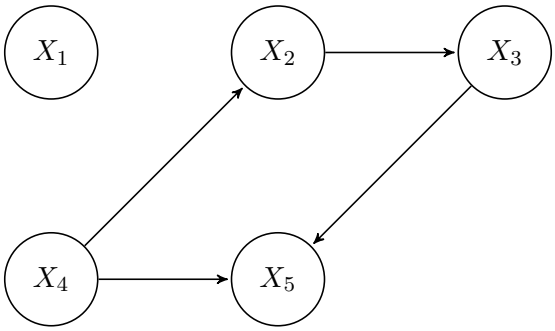


Figure 3.5: 示例：表示概率模型的有向无环图。

对于大型概率模型，找到能够尽可能减少参数数量的分解是至关重要的。

3.4.3 Graphical models

图形模型是表征概率模型依赖结构的一种工具。在本节中，我们简要描述了 **directed** 图形模型，也称为贝叶斯网络。无向图形模型，通常称为马尔可夫随机场，超出了本文的讨论范围。我们建议对概率建模和机器学习有兴趣的读者参考更高级的文本，以深入了解图形模型的内容。

有向无环图（Directed Acyclic Graphs，简称 DAG）可以被解释为表示概率模型的联合 pmf 或 pdf 的一种因子分解的图示。为了给出一个有效的分解，这些图被约束为不包含任何环（因此称为“无环”）。DAG 中的每个节点表示一个随机变量。节点之间的边表示变量之间的依赖关系。与一个 DAG 相对应的分解包含:

- 与所有没有输入边的节点对应的变量的边际pmf或pdf。
- 在给定其 *parents* 的条件下，其余随机变量的条件 pmf 或 pdf。 如果从（分配给）A 的节点到（分配给）B 的节点存在一条有向边，则 A 是 B 的父节点。

为具体起见，考虑图 3.5 中的 DAG。为简便起见，我们用相应的随机变量表示每个节点，并假设它们都是离散的。节点 X_1 和 X_4 没有父节点，因此联合 pmf 的因子分解中包含它们的边缘 pmf。节点 X_2 仅由 X_4 派生，因此我们包含 $p_{X_2\,|\,X_4}$。 节点 X_3 由 X_2 派生，因此我们包含 $p_{X_3\,|\,X_2}$。最后，节点 X_5 由 X_3 和 X_4 派生，因此我们包含 $p_{X_5\,|\,X_3,X_4}$。 该因子分解的形式为

$$p_{X_1,X_2,X_3,X_4,X_5} = p_{X_1}\; p_{X_4}\; p_{X_2\,|\,X_4}\; p_{X_3\,|\,X_2}\; p_{X_5\,|\,X_3,X_4}. \tag{3.138}$$

这个因式分解揭示了一些依赖假设。通过链式法则，联合概率质量函数的另一个有效因式分解是

$$p_{X_1,X_2,X_3,X_4,X_5} = p_{X_1}\; p_{X_4\,|\,X_1}\; p_{X_2\,|\,X_1,X_4}\; p_{X_3\,|\,X_1,X_2,X_4}\; p_{X_5\,|\,X_1,X_2,X_3,X_4}. \tag{3.139}$$

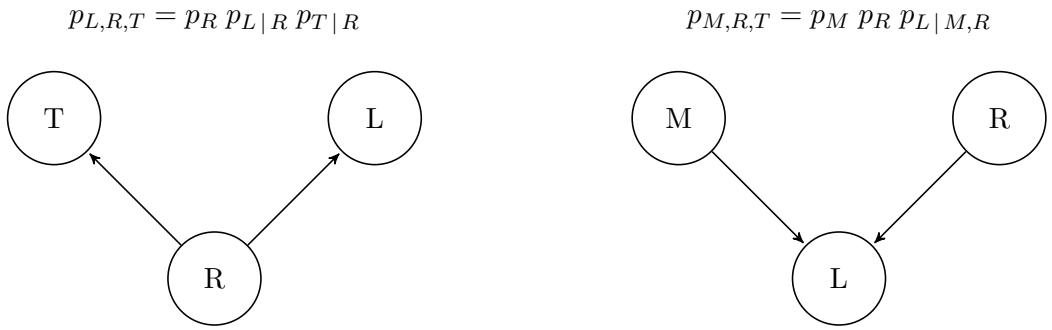


Figure 3.6: 与例 1.3.5 和 1.3.6 中的变量对应的有向图模型。

比较这两个表达式，我们可以看到 X_1 和所有其他变量是独立的，因为 $p_{X_4\,|\,X_1} = p_{X_4}$、$p_{X_2\,|\,X_1,X_4} = p_{X_2\,|\,X_4}$ 等等。此外，X_3 在给定 X_2 的条件下与 X_4 条件独立，因为 $p_{X_3\,|\,X_2,X_4} = p_{X_3\,|\,X_2}$。这些依赖假设可以直接从图中读取，使用以下性质。

Theorem 3.4.6 (局部马尔可夫性质)**.** *The factorization of the joint pmf or pdf represented by a DAG satisfies the local Markov property: each variable is conditionally independent of its non-descendants given all its parent variables. In particular, if it has no parents, it is independent of its non-descendants. To be clear, B is a non-descendant of A if there is no directed path from A to B.*

Proof. 令 X_i 为一个任意变量。我们用 X_N 表示 X_i 的非后代集合，用 X_P 表示父节点集合，用 X_D 表示后代集合。该图模型所表示的因子分解具有如下形式

$$p_{X_1,\ldots,X_n} = p_{X_N}\, p_{X_P|X_N}\, p_{X_i|X_P}\, p_{X_D|X_i}. \tag{3.140}$$

通过链式法则，另一个有效的因式分解是

$$p_{X_1,\ldots,X_n} = p_{X_N}\, p_{X_P|X_N}\, p_{X_i|X_P,X_N}\, p_{X_D|X_i,X_P,X_N}. \tag{3.141}$$

比较这两个表达式，我们得出结论：$p_{X_i|X_P,X_N} = p_{X_i|X_P}$，因此在给定 X_P 的条件下，X_i 与 X_N 条件独立。 □

我们通过展示示例 1.3.5 和 1.3.6 的 DAG 来阐释这些思想。

Example 3.4.7 (图示模型示例 1.3.5)**.** 我们使用指示随机变量建模示例 1.3.5 中的不同事件。T 表示出租车是否可用（$T = 1$）或不可用（$T = 0$），L 表示飞机是否延误（$L = 1$）或不延误（$L = 0$），R 表示是否下雨（$R = 1$）或不下雨（$R = 0$）。在该示例中，T 和 L 在给定 R 的条件下是条件独立的。我们可以使用图 3.6 左侧的图表示相应的因式分解。

△

Example 3.4.8 (示例 1.3.6 的图模型)**.** 我们使用指示随机变量来对示例 1.3.6 中的不同事件进行建模。M 表示是否发生机械故障

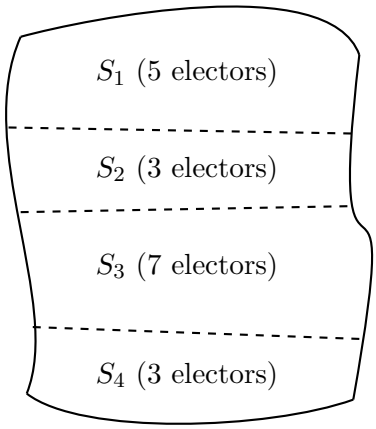


Figure 3.7: 虚构国家，见例 3.4.9。

（$M = 1$）或 非（$M = 0$），并且 L 和 R 与例 3.4.7 中的相同。在该示例中，M 和 R 相互独立，但 L 依赖于它们二者。我们可以使用图 3.6 右侧的图来表示相应的因子分解。

△

以下示例介绍了一类重要的图形模型，称为马尔可夫链，我们将在第七章中详细讨论。

Example 3.4.9 (选举)**.** 在图3.7所示的国家中，总统选举遵循与美国相同的制度。公民为选举团中的*electors*投票。每个州都有一定数量的选举人（在美国，这通常与国会议员人数相同）。在每个州，选举人会承诺支持赢得该州的候选人。我们的目标是对选举进行概率建模。我们假设只有两位候选人A和B。每个州由一个随机变量S_i表示，$1 \leq i \leq 4$,

$$S_i = \begin{cases} 1 & \text{if candidate A wins state } i, \\ -1 & \text{if candidate B wins state } i. \end{cases} \tag{3.142}$$

一个重要的决策是对模型应作出何种独立性假设。图 3.8 展示了三种不同的选项。如果我们将每个州建模为相互独立，那么只需要为每个州估计一个参数。然而，该模型可能并不准确，因为人口统计特征相似的州，其结果必然是相关的。另一种选择是估计完整的联合概率质量函数。问题在于，计算这些参数可能相当具有挑战性。我们可以利用民调数据来估计各个州的边缘概率质量函数，但条件概率则更难估计。此外，对于规模较大的模型，考虑完全依赖的模型是不可处理的（例如前文提到的美国大选情形）。一个合理的折中方案是：在给定其间州的条件下，将不相邻的州建模为条件独立。例如，我们假设州 1 和州 3 的结果仅通过州 2 相关。相应的图模型如图 3.8 右侧所示，称为马尔可夫链。它对应

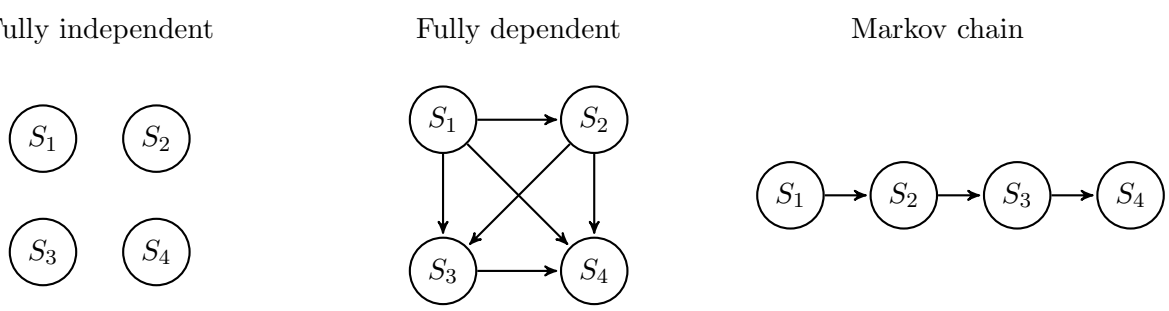


Figure 3.8: 图形模型捕捉了关于例子 3.4.9 中考虑的随机变量分布的不同假设。

到如下形式的因式分解

$$p_{S_1,S_2,S_3,S_4} = p_{S_1}\, p_{S_2\,|\,S_1}\, p_{S_3\,|\,S_2}\, p_{S_4\,|\,S_3}. \tag{3.143}$$

在这个模型下，我们只需要关注估计成对条件概率，而不是完整的联合概率质量函数（pmf）。我们在第七章详细讨论了马尔可夫链。

△

我们通过一个涉及连续变量的例子来结束这一节。

Example 3.4.10 (沙漠)**.** 丹尼和费利克斯正在亚利桑那州的沙漠中旅行。他们开始担心他们的车可能会抛锚，决定建立一个概率模型来评估风险。他们将车抛锚的时间建模为一个指数随机变量T，其参数依赖于发动机的状态M和道路的状态R。这三个量在同一概率空间中由随机变量表示。

不幸的是，他们不知道电动机的状态如何，因此他们假设其状态在0（电动机没有问题）和1（电动机几乎坏掉）之间是均匀分布的。同样，他们没有关于道路的信息，因此他们也假设道路的状态是一个在0（道路没有问题）和1（道路很糟糕）之间的均匀随机变量。此外，他们假设道路和汽车的状态是独立的，并且表示直到发生故障的时间（单位：小时）的指数随机变量的参数等于$M + R$。相应的图形模型如图3.9所示。

要找到随机变量的联合分布，我们应用链式法则得到，

$$\begin{aligned} f_{M,R,T}\,(m,r,t) &= f_M\,(m)\, f_{R|M}\,(r|m)\, f_{T|M,R}\,(t|m,r) && (3.144)\\ &= f_M\,(m)\, f_R\,(r)\, f_{T|M,R}\,(t|m,r) \quad \text{(by independence of } M \text{ and } R) && (3.145)\\ &= \begin{cases} (m+r)\, e^{-(m+r)t} & \text{for } t \geq 0,\ 0 \leq m \leq 1,\ 0 \leq r \leq 1,\\ 0 & \text{otherwise.} \end{cases} && (3.146) \end{aligned}$$

请注意，我们从 M 和 R 开始，因为我们知道它们的边际分布，而我们只知道 T 在给定 M 和 R 条件下的条件分布。

15分钟后，汽车发生故障。道路看起来还行，按照他们为R定义的标准，约为0.2，因此他们自然会想知道发动机的状况。

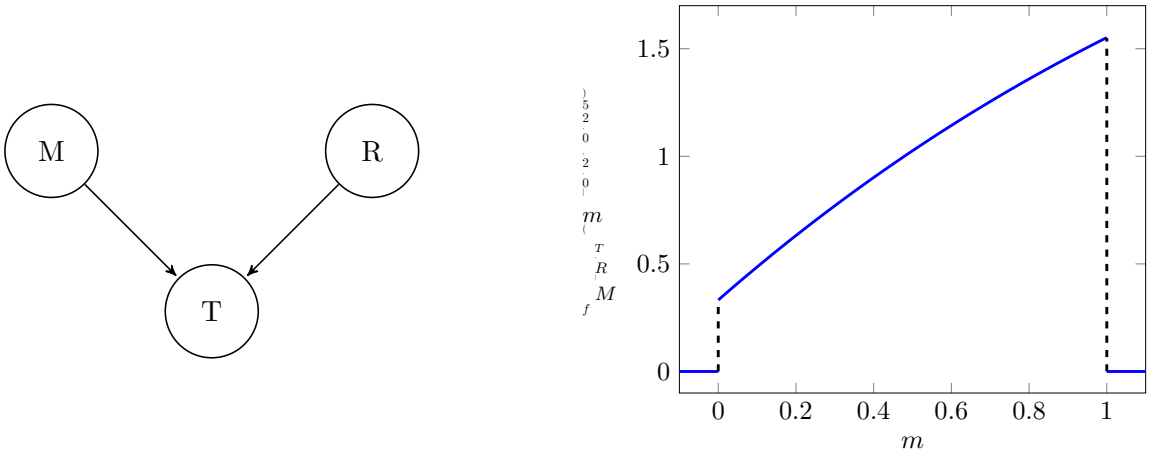


Figure 3.9: 左图是表示例 3.4.10 中随机变量的图形模型。右图显示了在给定 $T = 0.25$ 和 $R = 0.2$ 的条件下，M 的条件概率密度函数。

在概率模型中，鉴于所有这些信息，他们对电机的不确定性由在给定 T 和 R 条件下 M 的条件分布来刻画。

要计算条件概率密度函数，我们首先需要通过对M进行边际化来计算T和R的联合边际分布。为了简化计算，我们使用以下简单引理。

Lemma 3.4.11. *For any constant $c > 0$,*

$$\int_0^1 e^{-cx}\,dx = \frac{1-e^{-c}}{c}, \tag{3.147}$$

$$\int_0^1 xe^{-cx}\,dx = \frac{1-(1+c)\,e^{-c}}{c^2}. \tag{3.148}$$

Proof. 式(3.147)是通过对指数函数（其本身）求原函数得到的，而分部积分则得到(3.148)。 △

我们有

$$f_{R,T}(r,t) = \int_{m=0}^{1} f_{M,R,T}(m,r,t)\,\mathrm{d}m \tag{3.149}$$

$$= e^{-tr}\left(\int_{m=0}^{1} me^{-tm}\,\mathrm{d}m + r\int_{m=0}^{1} e^{-tm}\,\mathrm{d}m\right) \tag{3.150}$$

$$= e^{-tr}\left(\frac{1-(1+t)\,e^{-t}}{t^2} + \frac{r\left(1-e^{-t}\right)}{t}\right) \quad \text{by (3.147) and (3.148)} \tag{3.151}$$

$$= \frac{e^{-tr}}{t^2}\left(1+tr-e^{-t}\left(1+t+tr\right)\right), \tag{3.152}$$

对于 $t \geq 0, 0 \leq r \leq 1$。

M 在给定 T 和 R 的条件下的条件概率密度函数是

$$f_{M|R,T}(m|r,t) = \frac{f_{M,R,T}(m,r,t)}{f_{R,T}(r,t)} \tag{3.153}$$

$$= \frac{(m+r)\,e^{-(m+r)t}}{\frac{e^{-tr}}{t^2}(1+tr-e^{-t}(1+t+tr))} \tag{3.154}$$

$$= \frac{(m+r)\,t^2 e^{-tm}}{1+tr-e^{-t}(1+t+tr)}, \tag{3.155}$$

对于 $t \geq 0, 0 \leq m \leq 1, 0 \leq r \leq 1$。代入观测到的数值，条件 pdf 等于

$$f_{M|R,T}(m|0.2,0.25) = \frac{(m+0.2)\,0.25^2 e^{-0.25m}}{1+0.25\cdot 0.2-e^{-0.25}(1+0.25+0.25\cdot 0.2)} \tag{3.156}$$

$$= 1.66\,(m+0.2)\,e^{-0.25m}. \tag{3.157}$$

对于 $0 \leq m \leq 1$，否则为零。概率密度函数在图 3.9 中绘制。根据模型，似乎电机的状态不好。△

3.5 Functions of several random variables

随机变量 $Y := g(X_1,\ldots,X_n)$ 的 pmf 被定义为若干个离散随机变量 $X_1,\ldots,X_n$ 的函数 $g: \mathbb{R}^n \to \mathbb{R}$，其由下式给出

$$p_Y(y) = \sum_{y=g(x_1,\ldots,x_n)} p_{X_1,\ldots,X_n}(x_1,\ldots,x_n). \tag{3.158}$$

这直接由（3.11）推出。用文字表述，$g(X_1,\ldots,X_n)=y$ 的概率等于在所有可能取值上、使得 $y=g(x_1,\ldots,x_n)$ 的联合 pmf 之和。

Example 3.5.1 (选举)**.** 在例 3.4.9 中，我们讨论了一个由四个州组成的国家进行总统选举的几种可能模型。设想你正试图利用来自各个州的民意调查数据来预测选举结果。目标是预测选举的结果，该结果由随机变量表示

$$O := \begin{cases} 1 & \text{if } \sum_{i=1}^4 n_i S_i > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{3.159}$$

其中 n_i 表示州 i (中选举人的数量，注意总和永远不可能为零)。

通过分析民意调查数据，你得出结论：候选人A在每个州获胜的概率为0.15。如果你假设所有州相互独立，这就足以刻画联合概率质量函数（pmf）。表3.2列出了该模型下所有可能结果的概率。根据式(3.158)，我们只需要把满足 $O=1$ 的结果加总。在完全独立的假设下，候选人A获胜的概率为6%。

你对结果不满意，因为你怀疑不同状态下的结果高度依赖。从过去的选举中，你确定了a的条件概率{v*}

S_1	S_2	S_3	S_4	O	Prob. (indep.)	Prob. (Markov)
-1	-1	-1	-1	0	0.5220	0.6203
-1	-1	-1	1	0	0.0921	0.0687
-1	-1	1	-1	0	0.0921	0.0431
-1	-1	1	1	1	0.0163	0.0332
-1	1	-1	-1	0	0.0921	0.0431
-1	1	-1	1	0	0.0163	0.0048
-1	1	1	-1	1	0.0163	0.0208
-1	1	1	1	1	0.0029	0.0160
1	-1	-1	-1	0	0.0921	0.0687
1	-1	-1	1	0	0.0163	0.0077
1	-1	1	-1	1	0.0163	0.0048
1	-1	1	1	1	0.0029	0.0037
1	1	-1	-1	0	0.0163	0.0332
1	1	-1	1	1	0.0029	0.0037
1	1	1	-1	1	0.0029	0.0160
1	1	1	1	1	0.0005	0.0123

Table 3.2: 示例 3.5.1 的辅助值表。

如果某位候选人在赢得一个相邻州的情况下赢得某个州，其概率确实非常高。你将对条件概率的估计纳入由（3.143）所描述的马尔可夫链模型中：

$$p_{S_1}(1) = 0.15, \tag{3.160}$$

$$p_{S_{i+1}\,|\,S_i}(1\,|\,1) = 0.435, \quad 2 \le i \le 4, \tag{3.161}$$

$$p_{S_{i+1}\,|\,S_i}(-1\,|\,-1) = 0.900 \quad 2 \le i \le 4. \tag{3.162}$$

这意味着如果候选人B赢得一个州，他们很可能也会赢得相邻的州。如果候选人A赢得一个州，其赢得相邻州的概率显著高于未赢得该州时（但仍低于候选人B）。在该模型下，候选人A赢得每个州的边际概率仍然是0.15。表3.2列出了所有可能结果的概率。候选人A获胜的概率现在为11%，几乎是完全独立模型下所得概率的两倍。这说明了未考虑各州之间依赖关系的危险性，例如这可能是许多预测在2016年选举中严重低估唐纳德·特朗普胜算的原因之一。

△

第2.5节解释了如何通过先计算单变量随机变量的累积分布函数（cdf），再对其求导以获得其概率密度函数（pdf），从而推导其函数的分布。这一方法可以直接推广到多变量随机函数。设 X, Y 为定义在同一概率

空间，并令 $U = g(X,Y)$ 和 $V = h(X,Y)$ 为两个任意函数 $g, h : \mathbb{R}^2 \to \mathbb{R}$。然后，

$$F_{U,V}(u,v) = \mathrm{P}(U \le u, V \le v) \tag{3.163}$$
$$= \mathrm{P}(g(X,Y) \le u, h(X,Y) \le v) \tag{3.164}$$
$$= \int_{\{(x,y) \mid g(x,y)\le u, h(x,y)\le v\}} f_{X,Y}(x,y)\, \mathrm{d}x\, \mathrm{d}y, \tag{3.165}$$

最后的等式仅在X和Y的联合概率密度函数存在时成立。然后可以通过微分得到联合概率密度函数。

Theorem 3.5.2 (两个独立随机变量)**.** *The pdf of $Z = X + Y$, where X and Y are independent random variables is equal to the* ***convolution*** *of their respective pdfs f_X and f_Y,*的和的概率密度函数

$$f_Z(z) = \int_{u=-\infty}^{\infty} f_X(z-u)\, f_Y(u)\ du. \tag{3.166}$$

Proof. 首先我们推导 Z 的累积分布函数

$$F_Z(z) = \mathrm{P}(X + Y \le z) \tag{3.167}$$
$$= \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{z-y} f_X(x)\, f_Y(y)\, \mathrm{d}x\, \mathrm{d}y \tag{3.168}$$
$$= \int_{y=-\infty}^{\infty} F_X(z-y)\, f_Y(y)\, \mathrm{d}y. \tag{3.169}$$

请注意，X 和 Y的联合概率密度函数是边际概率密度函数的乘积，因为这些随机变量是独立的。我们现在对累积分布函数进行求导，以获得概率密度函数。请注意，这需要将极限算符与求导算符互换，并且将积分算符与求导算符互换，这些操作是合理的，因为所涉及的函数是有界且可积的。

$$f_Z(z) = \frac{\mathrm{d}}{\mathrm{d}z} \lim_{u\to\infty} \int_{y=-u}^{u} F_X(z-y)\, f_Y(y)\, \mathrm{d}y \tag{3.170}$$
$$= \lim_{u\to\infty} \frac{\mathrm{d}}{\mathrm{d}z} \int_{y=-u}^{u} F_X(z-y)\, f_Y(y)\, \mathrm{d}y \tag{3.171}$$
$$= \lim_{u\to\infty} \int_{y=-u}^{u} \frac{\mathrm{d}}{\mathrm{d}z} F_X(z-y)\, f_Y(y)\, \mathrm{d}y \tag{3.172}$$
$$= \lim_{u\to\infty} \int_{y=-u}^{u} f_X(z-y)\, f_Y(y)\, \mathrm{d}y. \tag{3.173}$$

□

Example 3.5.3 (咖啡豆)**.** 一家生产咖啡的公司从哥伦比亚和越南的两个小型当地生产商购买咖啡豆。它们可以从每个生产商购买的豆子数量取决于天气。这些数量C 和 V 被建模为独立的随机变量（假设哥伦比亚的天气与越南的天气独立），并且分别在[0, 1]和[0, 2]的范围内服从均匀分布（单位是吨）。

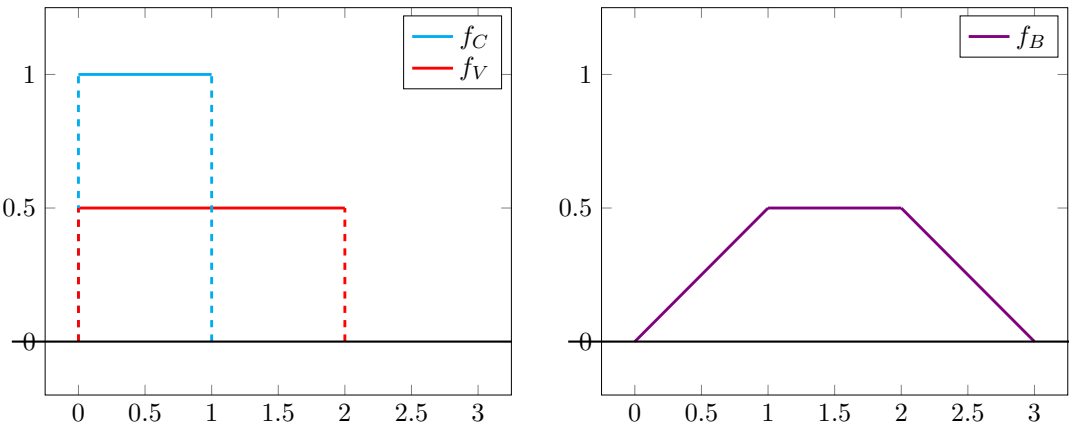


Figure 3.10: 例 3.5 中的概率密度函数

.3. 我们现在计算总量咖啡豆 $B := E+V$的概率密度函数，应用定理 3.5.2,

$$f_B(b) = \int_{u=-\infty}^{\infty} f_C(b-u)\, f_V(u)\, \mathrm{d}u \tag{3.174}$$

$$= \frac{1}{2}\int_{u=0}^{2} f_C(b-u)\, \mathrm{d}u \tag{3.175}$$

$$= \begin{cases} \frac{1}{2}\int_{u=0}^{b} \mathrm{d}u = \frac{b}{2} & \text{if } b \le 1 \\ \frac{1}{2}\int_{u=b-1}^{b} \mathrm{d}u = \frac{1}{2} & \text{if } 1 \le b \le 2 \\ \frac{1}{2}\int_{u=b-1}^{2} \mathrm{d}u = \frac{3-b}{2} & \text{if } 2 \le b \le 3. \end{cases} \tag{3.176}$$

B 的概率密度函数如图3.10所示。

△

3.6 Generating multivariate random variables

在第2.6节中，我们考虑从任意一元分布生成独立样本的问题。假设存在实现这一目标的过程，我们可以通过从相应的条件分布中生成样本，来对任意多元分布进行抽样。

Algorithm 3.6.1 (从多变量分布中采样)**.** *Let X_1, X_2, ..., X_n be random variables belonging to the same probability space. To generate samples from their joint distribution we sequentially sample from their conditional distributions:*

1. *Obtain a sample x_1 of X_1*
2. *For $i = 2, 3, \ldots, n$, obtain a sample x_i of X_i given the event $\{X_1 = x_1, \ldots, X_{i-1} = x_{i-1}\}$ by sampling from $F_{X_i|X_1,\ldots,X_{i-1}}(\cdot|x_1, \ldots, x_{i-1})$.*

链式法则意味着该过程的输出 $x_1, \ldots, x_n$ 是随机变量联合分布的样本。以下示例考虑从指数型随机变量混合分布中进行采样的问题。

Example 3.6.2 (指数混合). 设B为一个伯努利随机变量，其参数为p，X为一个指数随机变量，当B =为0时其参数为1，当B =为1时其参数为2。假设我们可以获取来自[0, 1]区间的两个独立样本u_1和u_2，它们来自均匀分布。为了获得来自B和X的样本:

1. 我们设置 $b :=$ 为 1 如果 $u_1 \leq p$ 和 $b :=$ 为 0，否则为 0。这确保了 b 是一个具有正确参数的伯努利样本。

2. 然后，我们设置

$$x := \frac{1}{\lambda} \log\left(\frac{1}{1-u_2}\right) \tag{3.177}$$

其中 λ: = 当 $b = 0$ 时取 1，且 λ: = 当 $b = 1$ 时取 2。由例 2.6.4，x 服从参数为 λ 的指数分布。

△

3.7 Rejection sampling

本章最后我们介绍拒绝采样（rejection sampling），也称为接受–拒绝法（accept-reject method），这是一种从一维分布中进行抽样的替代性方法。之所以将其推迟到本章，是因为对该技术的分析需要理解多元随机变量。在介绍该方法之前，我们先用离散随机变量来对其进行动机说明。

3.7.1 Rejection sampling for discrete random variables

我们的目标是通过从另一个随机变量 X 的样本中模拟一个随机变量 Y。为了简化阐述，我们假设它们的概率质量函数 p_X 和 p_Y 在集合 $\{1, 2, \ldots, n\}$ (中具有非零值，推广到其他离散集合是直接的)。拒绝采样的思想是，我们可以以某种方式选择 X 的样本子集，从而重新塑造其分布。当我们获得一个 X 的样本时，我们决定是否接受它或以某个概率拒绝它。这个概率依赖于样本 x 的值，如果 $p_X(x)$ 比 $p_Y(x)$ 大得多，我们通常应该拒绝它（但不是总是！）。对于每个 $x \in \{1, 2, \ldots, n\}$，我们通过 a_x 来定义接受样本的概率。

我们关注的只是被接受样本的分布。从数学上讲，被接受样本的 pmf 等于在样本被接受这一事件条件下 X 的条件 pmf,

$$p_{X \mid \text{Accepted}}(x \mid \text{Accepted}) = \frac{p_X(x)\,\mathrm{P}(\text{Accepted} \mid X = x)}{\sum_{i=1}^{n} p_X(i)\,\mathrm{P}(\text{Accepted} \mid X = i)} \quad \text{by Bayes' rule} \tag{3.178}$$

$$= \frac{p_X(x)\,a_x}{\sum_{i=1}^{n} p_X(i)\,a_i}. \tag{3.179}$$

我们希望固定接受概率，使得对于所有 $x \in \{1, 2, \ldots, n\}$

$$p_{X \mid \text{Accepted}}(x \mid \text{Accepted}) = p_Y(x). \tag{3.180}$$

这可以通过固定 {v*} 来实现

$$a_x := \frac{p_Y(x)}{c p_X(x)}, \qquad x \in \{1, \ldots, n\}, \tag{3.181}$$

对于任意常数 c。然而，这对于任意的 c 都不会产生有效的概率，因为 a_i 可能大于一！为避免这个问题，我们需要

$$c \geq \max_{x \in \{1,\ldots,n\}} \frac{p_Y(x)}{p_X(x)}, \quad \text{for all } x \in \{1, \ldots, n\}. \tag{3.182}$$

最后，我们可以使用一个介于 0 和 1 之间的均匀随机变量 U 来进行接受或拒绝，当 $U \leq a_x$ 时接受每个样本 x。你可能会想，为什么我们不能直接从 U 中生成 Y。那样确实可行，而且要简单得多；这里只是将离散情形作为连续情形的教学性引入来介绍。

Algorithm 3.7.1 (拒绝采样). *Let X and Y be random variables with pmfs p_X and p_Y such that*

$$c \geq \max_{x \in \{1,\ldots,n\}} \frac{p_Y(x)}{p_X(x)} \tag{3.183}$$

for all x such that $p_Y(x)$ is nonzero, and U a random variable that is uniformly distributed in [0, 1] *and independent of X.*

1. *Obtain a sample y of X.*
2. *Obtain a sample u of U.*
3. *Declare y to be a sample of Y if*

$$u \leq \frac{p_Y(y)}{c\, p_X(y)}. \tag{3.184}$$

3.7.2 Rejection sampling for continuous random variables

在这里，我们展示了前一节中提出的想法可以应用于连续情况。目标是通过选择根据不同的概率密度函数 f_X 获得的样本来获取符合目标概率密度函数 f_Y 的样本。与离散情况类似，我们需要

$$f_Y(y) \leq c\, f_X(y) \tag{3.185}$$

对于所有 y，其中 c 是一个固定的正的常数。换言之，Y 的 pdf 必须被 X 的 pdf 的一个按比例缩放的版本所界定。

Algorithm 3.7.2 (拒绝采样). *Let X be a random variable with pdf f_X and U a random variable that is uniformly distributed in* [0, 1] *and independent of X. We assume that* (3.185) *holds.*

1. *Obtain a sample y of X.*
2. *Obtain a sample u of U.*

3. Declare y to be a sample of Y if

$$u \leq \frac{f_Y(y)}{c\, f_X(y)}. \tag{3.186}$$

以下定理确立了通过拒绝采样获得的样本具有所需的分布。

Theorem 3.7.3 (拒绝采样有效)**.** *If assumption* (3.185) *holds, then the samples produced by rejection sampling are distributed according to f_Y.*

Proof. 让 Z 表示由拒绝采样生成的随机变量。Z 的累积分布函数等于

$$F_Z(y) = \mathrm{P}\left(X \leq y \,|\, U \leq \frac{f_Y(X)}{c\, f_X(X)}\right) \tag{3.187}$$

$$= \frac{\mathrm{P}\left(X \leq y,\, U \leq \frac{f_Y(X)}{c\, f_X(X)}\right)}{\mathrm{P}\left(U \leq \frac{f_Y(X)}{c\, f_X(X)}\right)}. \tag{3.188}$$

为了计算分子，我们在感兴趣的区域上对 U 和 X 的联合概率密度函数进行积分

$$\mathrm{P}\left(X \leq y,\, U \leq \frac{f_Y(X)}{c\, f_X(X)}\right) = \int_{x=-\infty}^{y} \int_{u=0}^{\frac{f_Y(x)}{c\, f_X(x)}} f_X(x)\, \mathrm{d}u\, \mathrm{d}x \tag{3.189}$$

$$= \int_{x=-\infty}^{y} \frac{f_Y(x)}{c\, f_X(x)} f_X(x)\, \mathrm{d}x \tag{3.190}$$

$$= \frac{1}{c} \int_{x=-\infty}^{y} f_Y(x)\, \mathrm{d}x \tag{3.191}$$

$$= \frac{1}{c} F_Y(y). \tag{3.192}$$

分母也是以类似的方式得到的

$$\mathrm{P}\left(U \leq \frac{f_Y(X)}{c\, f_X(X)}\right) = \int_{x=-\infty}^{\infty} \int_{u=0}^{\frac{f_Y(x)}{c\, f_X(x)}} f_X(x)\, \mathrm{d}u\, \mathrm{d}x \tag{3.193}$$

$$= \int_{x=-\infty}^{\infty} \frac{f_Y(x)}{c\, f_X(x)} f_X(x)\, \mathrm{d}x \tag{3.194}$$

$$= \frac{1}{c} \int_{x=-\infty}^{\infty} f_Y(x)\, \mathrm{d}x \tag{3.195}$$

$$= \frac{1}{c}. \tag{3.196}$$

我们得出结论，

$$F_Z(y) = F_Y(y), \tag{3.197}$$

因此，该方法从 Y 的分布中生成样本。 □

我们现在通过将该方法应用于从指数分布和均匀分布随机变量生成高斯随机变量来进行说明。

Example 3.7.4 (生成高斯随机变量)**.** 在例子2.6.4中，我们学习了如何使用均匀分布的样本生成指数随机变量。在本例中，我们将使用来自指数分布的样本，通过拒绝采样法生成标准高斯随机变量。

下面的引理表明，我们可以通过以下方式生成一个标准高斯随机变量 Y:

1. 生成具有概率密度函数(pdf)的随机变量 H

$$f_H(h) := \begin{cases} \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{h^2}{2}\right) & \text{if } h \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.198}$$

2. 生成一个随机变量 S，其值为 1 或 -1，概率为 1/2，例如通过应用第 2.6.1 节中描述的方法。

3. 设置 $Y := SH$.

Lemma 3.7.5. *Let H be a continuous random variable with pdf given by* (3.198) *and S a discrete random variable which equals* 1 *with probability 1/2 and* -1 *with probability 1/2. The random variable of $Y := SH$ is a standard Gaussian.*

Proof. Y在给定 S 的条件下的概率密度函数为

$$f_{Y|S}(y|1) = \begin{cases} f_H(y) & \text{if } y \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{3.199}$$

$$f_{Y|S}(y|-1) = \begin{cases} f_H(-y) & \text{if } y < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.200}$$

根据引理 3.3.5，我们有

$$f_Y(y) = p_S(1) f_{Y|S}(y|1) + p_S(-1) f_{Y|S}(y|-1) \tag{3.201}$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right). \tag{3.202}$$

△

我们之所以将问题化简为生成 H，是因为它的 pdf 只在正轴上非零，这使我们能够用参数为 1 的指数随机变量 X 的指数 pdf 对其进行上界约束。若我们设 $c := \sqrt{2e/\pi}$，则对所有 x 都有 $f_H(x) \leq c f_X(x)$，如图 3.11 所示。事实上，

$$\frac{f_H(x)}{f_X(x)} = \frac{\frac{2}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)}{\exp(-x)} \tag{3.203}$$

$$= \sqrt{\frac{2e}{\pi}} \exp\left(\frac{-(x-1)^2}{2}\right) \tag{3.204}$$

$$\leq \sqrt{\frac{2e}{\pi}}. \tag{3.205}$$

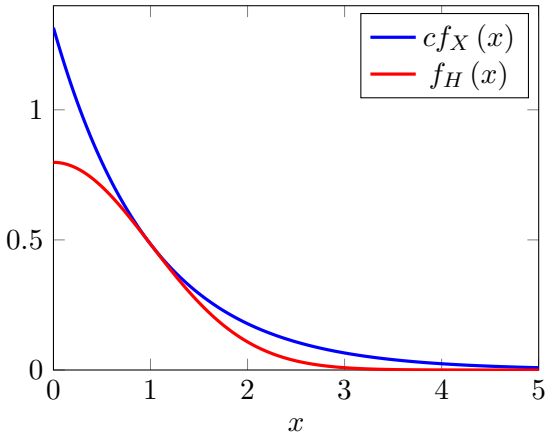


Figure 3.11: 例3.7.4中目标分布的概率密度函数的界。

我们现在可以应用拒绝采样来生成 H。步骤如下:

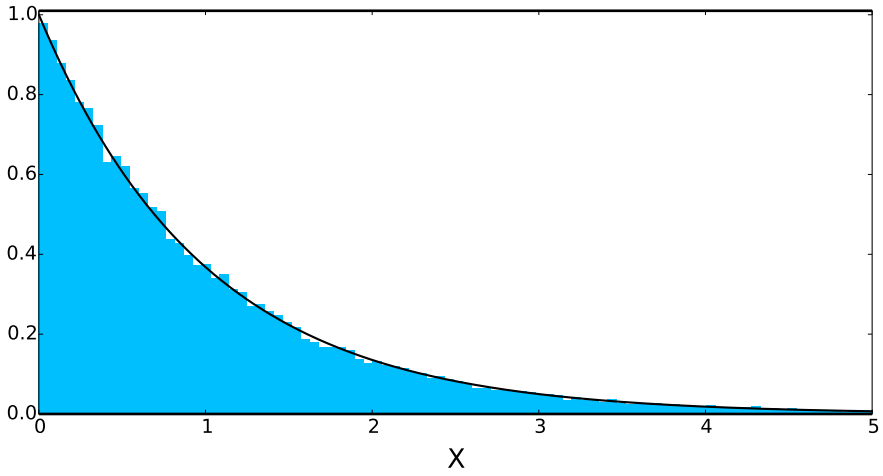
1. 从具有参数为一的指数随机变量 X 中获取一个样本 x

2. 从 U 中获取一个样本 u，该样本在 [0, 1] 内均匀分布。 3. 如果 x 作为 H 的样本被接受

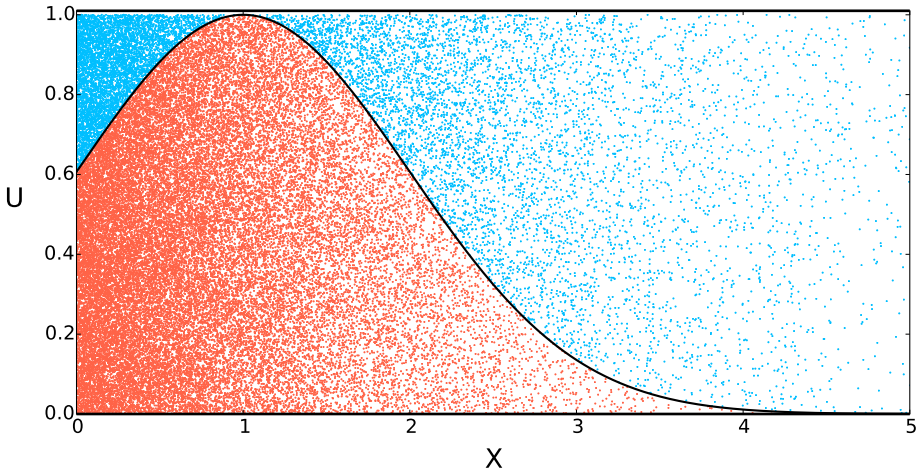
$$u \leq \exp\left(\frac{-(x-1)^2}{2}\right). \tag{3.206}$$

此过程如图 3.12 所示。拒绝机制确保接受的样本具有正确的分布。△

Histogram of 50,000 iid samples from X (f_X is shown in black)



Scatterplot of samples from X and samples from U (accepted samples are colored red, $\exp\left(-(x-1)^2/2\right)$ is shown in black)



Histogram of accepted samples (f_H is shown in black)

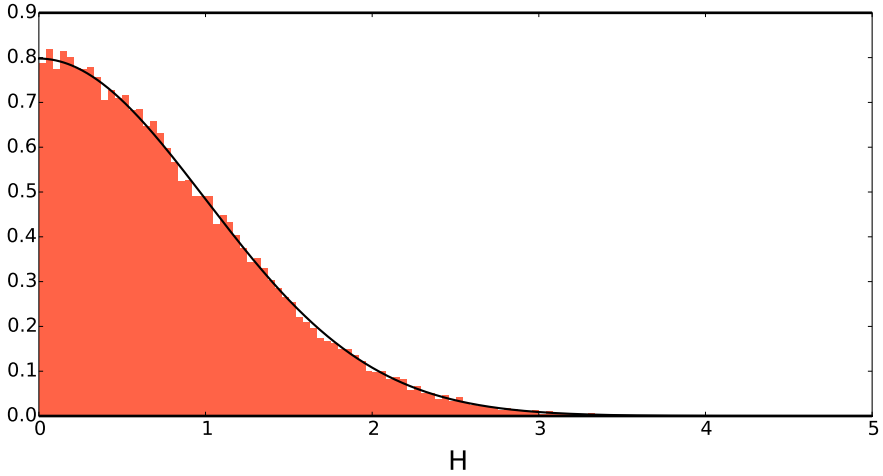


Figure 3.12: 通过拒绝采样从例3.7.4中定义的随机变量H生成50,000个样本的方法示意图。

Chapter 4

Expectation

在本节中，我们介绍了一些简洁地描述随机变量行为的量。均值是随机变量分布的中心值。方差量化了随机变量围绕均值波动的程度。两个随机变量的协方差指示它们是否倾向于以相似的方式偏离各自的均值。在多维空间中，随机向量的协方差矩阵编码了它在各个可能方向上的方差。这些量并不能完全表征随机变量或向量的分布，但它们通过少量的数字提供了关于其行为的有用总结。

4.1 Expectation operator

期望算子使我们能够严格地定义均值、方差和协方差。它将一个随机变量或多个随机变量的函数映射到一个由相应的概率质量函数（pmf）或概率密度函数（pdf）加权的平均值。

Definition 4.1.1 (离散随机变量的期望)**.** *Let X be a discrete random variable with range R. The expected value of a function $g(X)$, $g : \mathbb{R} \to \mathbb{R}$, of X is*

$$\mathrm{E}\left(g\left(X\right)\right) := \sum_{x \in R} g\left(x\right) p_X\left(x\right). \tag{4.1}$$

Similarly, if X, Y are both discrete random variables with ranges R_X and R_Y then the expected value of a function $g(X, Y)$, $g\ \mathbb{R}2 \to \mathbb{R}$, of X and Y is

$$\mathrm{E}\left(g\left(X, Y\right)\right) := \sum_{x \in R_X} \sum_{x \in R_Y} g\left(x, y\right) p_{X,Y}\left(x, y\right). \tag{4.2}$$

If $\vec{X}$ is an n-dimensional discrete random vector, the expected value of a function $g(\vec{X})$, g $\mathbb{R}^n \to \mathbb{R}$, of $\vec{X}$ is

$$\mathrm{E}\left(g(\vec{X})\right) := \sum_{\vec{x}_1} \sum_{\vec{x}_2} \cdots \sum_{\vec{x}_n} g\left(\vec{x}\right) p_{\vec{X}}\left(\vec{x}\right). \tag{4.3}$$

Definition 4.1.2 (连续随机变量的期望)**.** *Let X be a continuous random variable. The expected value of a function $g(X)$, $g:\mathbb{R}\to\mathbb{R}$, of X is*

$$\mathrm{E}\left(g\left(X\right)\right) := \int_{x=-\infty}^{\infty} g\left(x\right) f_X\left(x\right)\, dx. \tag{4.4}$$

Similarly, if X, Y are both continuous random variables then the expected value of a function $g(X,Y)$, $g\,\mathbb{R}2\to\mathbb{R}$, of X and Y is

$$\mathrm{E}\left(g\left(X,Y\right)\right) := \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} g\left(x,y\right) f_{X,Y}\left(x,y\right)\, dx\, dy. \tag{4.5}$$

If $\vec{X}$ is an n-dimensional random vector, the expected value of a function $g(X)$, $g:\mathbb{R}^n\to\mathbb{R}$, of $\vec{X}$ is

$$\mathrm{E}\left(g(\vec{X})\right) := \int_{x_1=-\infty}^{\infty}\int_{x_2=-\infty}^{\infty}\cdots\int_{x_n=-\infty}^{\infty} g\left(\vec{x}\right) f_{\vec{X}}\left(\vec{x}\right)\, dx_1\, dx_2 \ldots\, dx_n \tag{4.6}$$

在同时依赖连续型和离散型随机变量的量的情形下，边缘分布与条件分布的乘积起到联合 pdf 或 pmf 的作用。

Definition 4.1.3 (关于连续型和离散型随机变量的期望)**.** *If C is a continuous random variable and D a discrete random variable with range R_D defined on the same probability space, the expected value of a function $g(C,D)$ of C and D is*

$$\mathrm{E}\left(g\left(C,D\right)\right) := \int_{c=-\infty}^{\infty}\sum_{d\in R_D} g\left(c,d\right) f_C\left(c\right) p_{D|C}\left(d|c\right)\, dc \tag{4.7}$$

$$= \sum_{d\in R_D}\int_{c=-\infty}^{\infty} g\left(c,d\right) p_D\left(d\right) f_{C|D}\left(c|d\right)\, dc. \tag{4.8}$$

某个量的期望值在相应的求和或积分趋于无穷或具有未定义值时，可能是无穷大，甚至根本不存在。下面的例4.1.4和4.2.2对此作了说明。

Example 4.1.4 (圣彼得堡悖论)**.** 一家赌场向你提供以下游戏。你将投掷一枚公平的硬币，直到它正面朝上为止，赌场将支付你 2^k 美元，其中 k 是投掷次数。你愿意为此支付多少费用?

让我们计算预期收益。如果翻转是独立的，则翻转的总次数 X 是一个几何随机变量，因此 $p_X(k) = 1/2^k$。收益是 2^X，这意味着

$$\mathrm{E}\left(\text{Gain}\right) = \sum_{k=1}^{\infty} 2^k\cdot\frac{1}{2^k} = \infty. \tag{4.9}$$

预期收益是无限的，但由于你只能玩一次，你愿意支付的金额可能是有限的。这被称为圣彼得堡悖论。

△

期望算子的一个基本属性是它是线性的。

Theorem 4.1.5 (期望的线性)**.** *For any constant $a \in \mathbb{R}$, any function $g : \mathbb{R} \to \mathbb{R}$ and any continuous or discrete random variable X*

$$\mathrm{E}\left(a\, g\left(X\right)\right) = a\, \mathrm{E}\left(g\left(X\right)\right). \tag{4.10}$$

For any constants $a, b \in \mathbb{R}$, any functions $g_1, g2 : \mathbb{R}^n \to \mathbb{R}$ and any continuous or discrete random variables X and Y

$$\mathrm{E}\left(a\, g_1\left(X,Y\right) + b\, g_2\left(X,Y\right)\right) = a\, \mathrm{E}\left(g_1\left(X,Y\right)\right) + b\, \mathrm{E}\left(g_2\left(X,Y\right)\right). \tag{4.11}$$

Proof. 定理立即从和与积分的线性性质中得出。 □

期望的线性性质使得计算随机变量线性函数的期望变得非常简单。相比之下，计算联合概率密度函数或概率质量函数通常复杂得多。

Example 4.1.6 (咖啡豆（接续自例子 3.5.3））**.** 让我们计算能够购买的咖啡豆的预期总量。C 在 [0, 1] 内均匀分布，因此 $\mathrm{E}\left(C\right) = 1/2$。$V$ 在 [0, 2] 内均匀分布，因此 $\mathrm{E}\left(V\right) = 1$。根据期望的线性性质

$$\begin{aligned} \mathrm{E}\left(C + V\right) &= \mathrm{E}\left(C\right) + \mathrm{E}\left(V\right) && (4.12)\\ &= 1.5 \text{ tons.} && (4.13) \end{aligned}$$

请注意，即使两个量是 *not* 独立的，这一点仍然成立。

△

如果两个随机变量是独立的，则它们的乘积的期望等于期望的乘积。

Theorem 4.1.7 (独立随机变量的函数的期望)**.** *If X, Y are independent random variables defined on the same probability space, and $g, h : \mathbb{R} \to \mathbb{R}$ are univariate real-valued functions, then*

$$\mathrm{E}\left(g\left(X\right) h\left(Y\right)\right) = \mathrm{E}\left(g\left(X\right)\right) \mathrm{E}\left(h\left(Y\right)\right). \tag{4.14}$$

Proof. 我们证明了连续随机变量的结果，但离散随机变量的证明本质上是相同的。

$$\begin{aligned} \mathrm{E}\left(g\left(X\right) h\left(Y\right)\right) &= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} g\left(x\right) h\left(y\right) f_{X,Y}\left(x,y\right)\, \mathrm{d}x\, \mathrm{d}y && (4.15)\\ &= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} g\left(x\right) h\left(y\right) f_X\left(x\right) f_Y\left(y\right)\, \mathrm{d}x\, \mathrm{d}y \quad \text{by independence} && (4.16)\\ &= \mathrm{E}\left(g\left(X\right)\right) \mathrm{E}\left(h\left(Y\right)\right). && (4.17) \end{aligned}$$

□

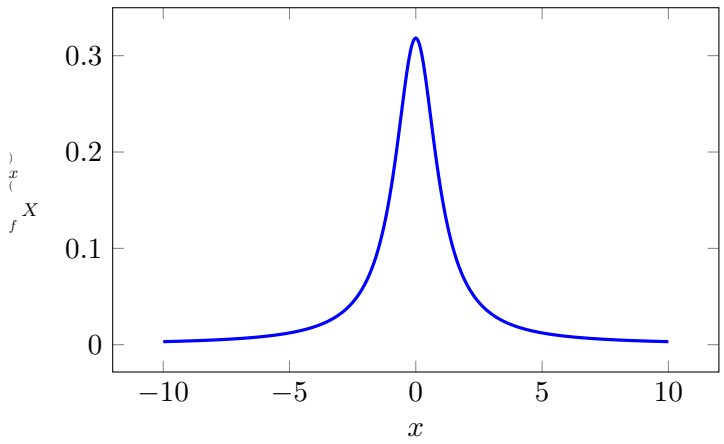


Figure 4.1: 柯西随机变量的概率密度函数。

4.2 Mean and variance

4.2.1 Mean

随机变量的均值等于其期望值。

Definition 4.2.1 (均值)**.** *The mean or first moment of X is the expected value of X:* $\mathrm{E}(X)$.

表4.1列出了一些重要随机变量的均值。推导可见第4.5.1节。如图4.3所示，均值是相应随机变量的 pmf 或 pdf 的质心。

如果一个随机变量的分布非常 *heavy tailed*，这意味着随机变量取大值的概率衰减得很慢，那么它的均值可能是无限的。这就是例子4.1.4中表示收益的随机变量的情况。以下例子显示了如果相应的和或积分值没有明确定义，均值可能不存在。

Example 4.2.2 (柯西随机变量)**.** 柯西随机变量的概率密度函数（pdf）如图 4.1 所示，其表达式为

$$f_X(x) = \frac{1}{\pi(1+x^2)}. \tag{4.18}$$

根据期望值的定义，

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)}\,\mathrm{d}x = \int_{0}^{\infty} \frac{x}{\pi(1+x^2)}\,\mathrm{d}x - \int_{0}^{\infty} \frac{x}{\pi(1+x^2)}\,\mathrm{d}x. \tag{4.19}$$

现在，通过变量替换 $t = x^2$,

$$\int_0^\infty \frac{x}{\pi(1+x^2)}\,\mathrm{d}x = \int_0^\infty \frac{1}{2\pi(1+t)}\mathrm{d}t = \lim_{t\to\infty} \frac{\log(1+t)}{2\pi} = \infty, \tag{4.20}$$

所以 E(X) 并不　存在，因为它是两个趋于 t 的极限之差　　o 无穷大。

△

随机向量的均值被定义为由其各个分量的均值组成的向量。

Definition 4.2.3 (随机向量的均值)**.** *The mean of a random vector* $\vec{X}$ *is*

$$\mathrm{E}(\vec{X}) := \begin{bmatrix} \mathrm{E}\left(\vec{X}_1\right) \\ \mathrm{E}\left(\vec{X}_2\right) \\ \cdots \\ \mathrm{E}\left(\vec{X}_n\right) \end{bmatrix}. \tag{4.21}$$

如同单变量情况，均值可以被解释为随机向量分布围绕其中心的值。

由一维情形下期望算子的线性性可以立即推出，均值算子是线性的。

Theorem 4.2.4 (随机向量线性变换的均值)**.** *For any random vector* $\vec{X}$ *of dimension* n*, any matrix* $A \in \mathbb{R}^{m\times n}$ *and* $\vec{b} \in \mathbb{R}^m$

$$\mathrm{E}\left(A\vec{X} + \vec{b}\right) = A\,\mathrm{E}(\vec{X}) + \vec{b}. \tag{4.22}$$

Proof.

$$\mathrm{E}\left(A\vec{X} + \vec{b}\right) = \begin{bmatrix} \mathrm{E}\left(\sum_{i=1}^n A_{1i}\vec{X}_i + b_1\right) \\ \mathrm{E}\left(\sum_{i=1}^n A_{2i}\vec{X}_i + b_2\right) \\ \cdots \\ \mathrm{E}\left(\sum_{i=1}^n A_{mi}\vec{X}_i + b_n\right) \end{bmatrix} \tag{4.23}$$

$$= \begin{bmatrix} \sum_{i=1}^n A_{1i}\mathrm{E}\left(\vec{X}_i\right) + b_1 \\ \sum_{i=1}^n A_{2i}\mathrm{E}\left(\vec{X}_i\right) + b_2 \\ \cdots \\ \sum_{i=1}^n A_{mi}\mathrm{E}\left(\vec{X}_i\right) + b_n \end{bmatrix} \quad \text{by linearity of expectation} \tag{4.24}$$

$$= A\,\mathrm{E}(\vec{X}) + \vec{b}. \tag{4.25}$$

□

4.2.2 Median

均值通常被解释为随机变量所取的一个 *typical* 值。然而，随机变量等于其均值的概率可能为零！例如，伯努利随机变量不可能等于 0.5。此外，均值可能会被一小部分极端值严重扭曲，如下文示例 4.2.6 所示。中位数是对随机变量所取 *typical* 值的另一种刻画方式，旨在在这种情况下更加稳健。它被定义为随机变量的 pmf 或 pdf 的中点。如果随机变量是连续的，则其大于或小于中位数的概率均等，均为 1/2。

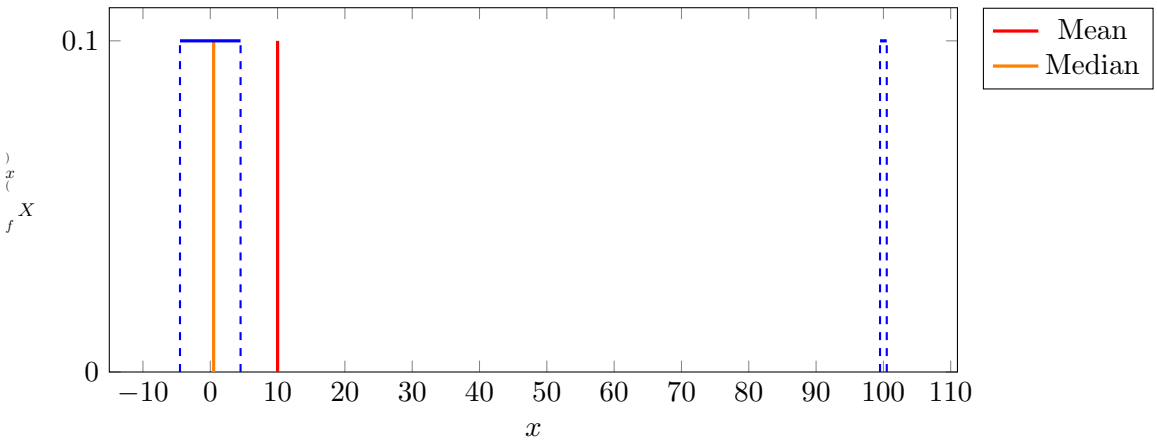


Figure 4.2: 在[−4.5, 4.5] ∪ [99.5, 100.5]上的均匀概率密度函数。均值为10，中位数为0.5。

Definition 4.2.5 (中位数)**.** *The median of a discrete random variable X is a number m such that*

$$P(X \leq m) \geq \frac{1}{2} \quad \textit{and} \quad P(X \geq m) \geq \frac{1}{2}. \tag{4.26}$$

The median of a continuous random variable X is a number m such that

$$F_X(m) = \int_{-\infty}^{m} f_X(x)\, dx = \frac{1}{2}. \tag{4.27}$$

以下示例说明了中位数在存在具有非零概率的小部分极端值时的稳健性。

Example 4.2.6 (均值与中位数)**.** 考虑一个均匀随机变量 X，其定义域为 [−4.5, 4.5] ∪ [99.5, 100.5]。X 的均值为

$$\begin{aligned} \mathrm{E}(X) &= \int_{x=-4.5}^{4.5} x f_X(x)\, \mathrm{d}x + \int_{x=99.5}^{100.5} x f_X(x)\, \mathrm{d}x && (4.28)\\ &= \frac{1}{10}\frac{100.5^2 - 99.5^2}{2} && (4.29)\\ &= 10. && (4.30) \end{aligned}$$

X 在 -4.5 到 4.5 之间的累积分布函数 (cdf) 等于

$$\begin{aligned} F_X(m) &= \int_{-4.5}^{m} f_X(x)\, \mathrm{d}x && (4.31)\\ &= \frac{m + 4.5}{10}. && (4.32) \end{aligned}$$

将此设置为1/2可以计算中位数，结果为0.5。图4.2显示了X的概率密度函数以及中位数和均值的位置。中位数提供了分布中心的更实际的度量。

△

Random variable	Parameters	Mean	Variance
Bernoulli	p	p	$p(1-p)$
Geometric	p	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Binomial	n, p	np	$np(1-p)$
Poisson	λ	λ	λ
Uniform	a, b	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential	λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gaussian	μ, σ	μ	σ^2

Table 4.1: 常见随机变量的均值和方差，推导见附录第 4.5.1 节。

4.2.3 Variance and standard deviation

随机变量平方的期望值有时用于量化随机变量的*energy*。

Definition 4.2.7 (二阶矩)**.** *The mean square or second moment of a random variable X is the expected value of X^2:* $\mathrm{E}\left(X^2\right)$.

该定义可推广到更高阶矩，对大于二的整数定义为 $\mathrm{E}\left(X^p\right)$。随机变量与其均值之差的均方称为随机变量的方差。它刻画了随机变量围绕其均值的波动，也称为该分布的第二 *centered* 阶矩。该量的平方根称为随机变量的标准差。

Definition 4.2.8 (方差和标准差)**.** *The variance of X is the mean square deviation from the mean*

$$\mathrm{Var}(X) := \mathrm{E}\left(\left(X - \mathrm{E}(X)\right)^2\right) \tag{4.33}$$

$$= \mathrm{E}\left(X^2\right) - \mathrm{E}^2(X). \tag{4.34}$$

The standard deviation σ_X of X is

$$\sigma_X := \sqrt{\mathrm{Var}(X)}. \tag{4.35}$$

我们在表 4.1 中汇总了一些重要随机变量的方差。其推导过程可见第 4.5.1 节。在图 4.3 中，我们绘制了这些随机变量的 pmf 和 pdf，并展示了落在均值一个标准差之内的取值范围。

方差算子不是线性的，但确定随机变量线性函数的方差是直接的。

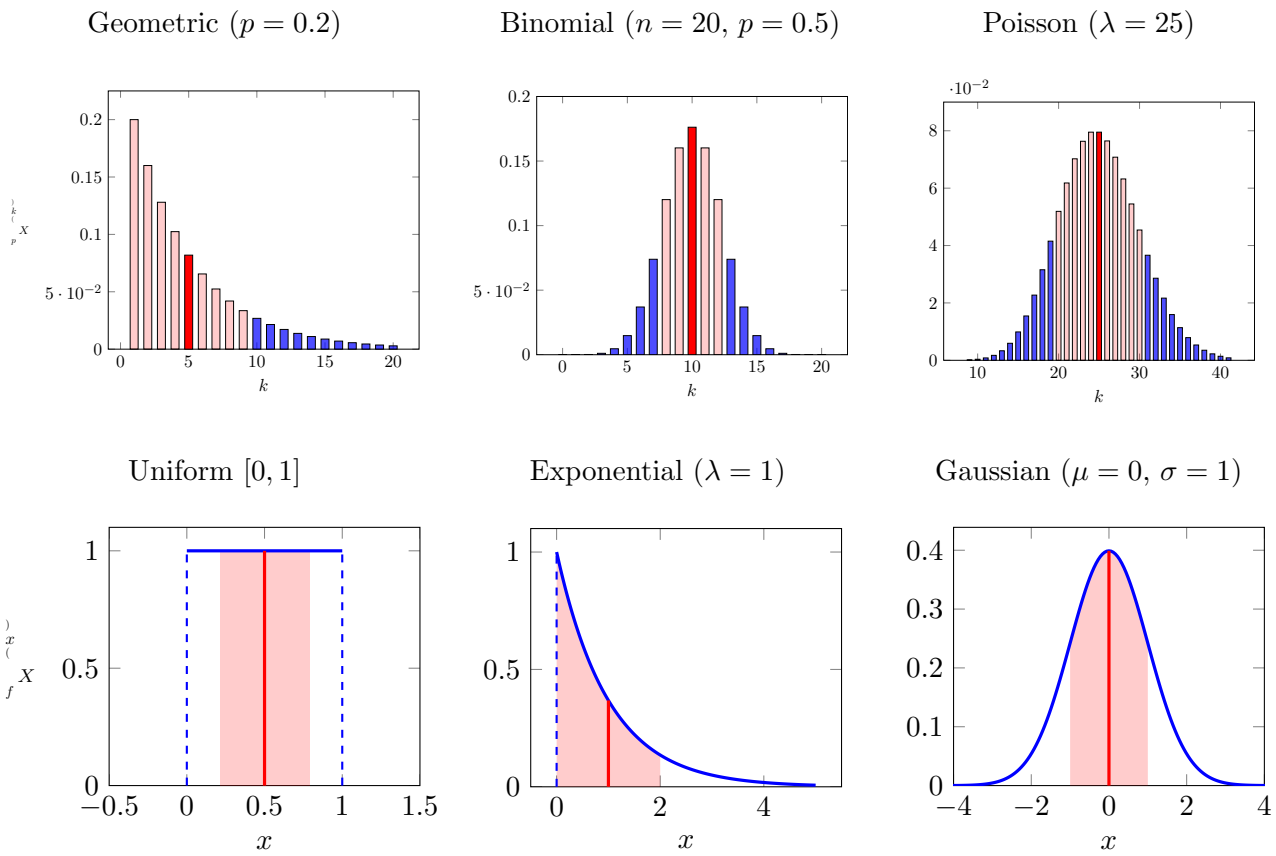


Figure 4.3: 离散随机变量的pmf（上排）和连续随机变量的pdf（下排）。随机变量的均值用红色标记。均值的一个标准差范围内的值用粉色标记。

Lemma 4.2.9 (线性函数的方差)**.** *For any constants a and b*

$$\operatorname{Var}(a\,X + b) = a^2 \operatorname{Var}(X)\,. \tag{4.36}$$

Proof.

$$\begin{aligned}
\operatorname{Var}(a\,X + b) &= \mathrm{E}\left((a\,X + b - \mathrm{E}(a\,X + b))^2\right) && (4.37)\\
&= \mathrm{E}\left((a\,X + b - a\mathrm{E}(X) - b)^2\right) && (4.38)\\
&= a^2\,\mathrm{E}\left((X - \mathrm{E}(X))^2\right) && (4.39)\\
&= a^2 \operatorname{Var}(X)\,. && (4.40)
\end{aligned}$$

□

这个结果是有道理的：如果我们通过加上一个常数来改变随机变量的中心，那么方差不会受到影响，因为方差仅衡量与均值的偏差。如果我们将随机变量乘以一个常数，标准差将按相同的比例缩放。

4.2.4 Bounding probabilities using the mean and variance

在本节中，我们介绍两个不等式，它们允许仅通过已知随机变量的均值和方差，在一定程度上刻画其行为。第一个是不等式是马尔可夫不等式，它量化了这样一种直观想法：如果一个随机变量是非负且通常较小，那么它取到较大值的概率必然很小。

Theorem 4.2.10 (马尔可夫不等式)**.** *Let X be a nonnegative random variable. For any positive constant $a > 0$,*

$$\mathrm{P}(X \geq a) \leq \frac{\mathrm{E}(X)}{a}. \tag{4.41}$$

Proof. 考虑指示变量 $1_{X \geq a}$。我们有

$$X - a\,1_{X \geq a} \geq 0. \tag{4.42}$$

特别地，它的期望是非负的（因为它是在正实轴上对一个非负量的求和或积分）。由期望的线性性以及 $1_{X \geq a}$ 是一个期望为 $\mathrm{P}(X \geq a)$ 的伯努利随机变量这一事实，我们有

$$\mathrm{E}(X) \geq a\,\mathrm{E}(1_{X \geq a}) = a\,\mathrm{P}(X \geq a). \tag{4.43}$$

□

Example 4.2.11 (学生的年龄)**.** 你听说纽约大学学生的平均年龄是20岁，但你知道有一些学生已经超过30岁。你决定使用马尔可夫不等式来界定年龄超过30岁的学生比例，并将年龄建模为一个非负随机变量A。

$$\mathrm{P}(A \geq 30) \leq \frac{\mathrm{E}(A)}{30} = \frac{2}{3}. \tag{4.44}$$

至多三分之二的学生超过30岁。

△

如示例 4.2.11 所示，马尔可夫不等式可能相当宽松。原因在于它几乎没有利用关于随机变量分布的任何信息。

切比雪夫不等式控制随机变量偏离其均值的程度。直观地说，如果方差（因此标准差）很小，那么随机变量远离其均值的概率必然很低。

Theorem 4.2.12 (切比雪夫不等式)**.** *For any positive constant $a > 0$ and any random variable X with bounded variance,*

$$\mathrm{P}(|X - \mathrm{E}(X)| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}. \tag{4.45}$$

Proof. 将马尔可夫不等式应用于随机变量 $Y = (X - \mathrm{E}(X))^2$，即可得到该结果。

□

切比雪夫不等式的一个有趣推论表明，如果一个随机变量的方差为零，那么该随机变量是一个常数；更准确地说，它偏离其均值的概率为零。

Corollary 4.2.13. *If* 变量 $(X) = 0$ *then* $\mathrm{P}(X \neq \mathrm{E}(X)) = 0$.

Proof. 取任何 $\epsilon > 0$，根據切比雪夫不等式

$$\mathrm{P}(|X - \mathrm{E}(X)| \geq \epsilon) \leq \frac{\mathrm{Var}(X)}{\epsilon^2} = 0. \tag{4.46}$$

□

Example 4.2.14 (学生年龄（续））**.** 你对30岁以上学生人数的上界并不太满意。你发现学生年龄的标准差实际上只有3年。应用切比雪夫不等式，这意味着

$$\mathrm{P}(A \geq 30) \leq \mathrm{P}(|A - \mathrm{E}(A)| \geq 10) \tag{4.47}$$

$$\leq \frac{\mathrm{Var}(A)}{100} = \frac{9}{100}. \tag{4.48}$$

所以实际上至少有91%的学生年龄在30岁以下（并且在10岁以上）。

△

4.3 Covariance

4.3.1 Covariance of two random variables

两个随机变量的协方差描述了它们的联合行为。它是随机变量与各自均值之差的乘积的期望值。直观地说，它衡量了随机变量一起波动的程度。

Definition 4.3.1 (协方差)**.** *The covariance of* X *and* Y *is*

$$\mathrm{Cov}(X, Y) := \mathrm{E}((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))) \tag{4.49}$$

$$= \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y). \tag{4.50}$$

If 协方差 $(X, Y) = 0$*,* X *and* Y *are* ***uncorrelated****.*

图 4.4 展示了来自具有不同协方差的二元高斯分布的样本。如果协方差为零，则联合概率密度函数呈球形。如果协方差为正且较大，则联合概率密度函数会发生偏斜，使得两个变量倾向于具有相似的取值。如果协方差较大且为负，则两个变量将倾向于具有数值相近但符号相反的取值。

两个随机变量之和的方差可以用它们各自的方差以及它们的协方差来表示。因此，如果协方差为正，它们的波动会相互加强；如果协方差为负，它们的波动会相互抵消。

Theorem 4.3.2 (两个随机变量)**.**和的方差

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y). \tag{4.51}$$

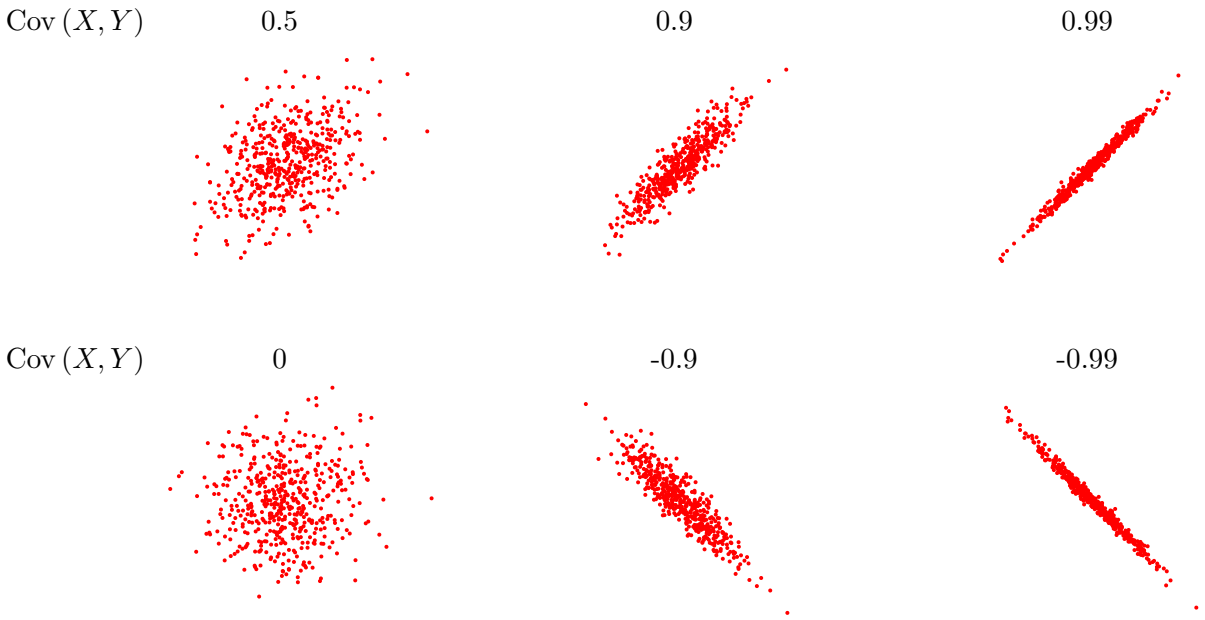


Figure 4.4: 来自二维高斯向量 (X, Y) 的样本，其中 X 和 Y 是均值为零、方差为一的标准高斯随机变量，针对 X 与 Y 之间协方差的不同取值。

Proof.

$$\begin{aligned}\operatorname{Var}(X+Y) &= \mathrm{E}\left((X+Y-\mathrm{E}(X+Y))^2\right) &(4.52)\\ &= \mathrm{E}\left((X-\mathrm{E}(X))^2\right)+\mathrm{E}\left((Y-\mathrm{E}(Y))^2\right)+2\mathrm{E}\left((X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\right)\\ &= \operatorname{Var}(X)+\operatorname{Var}(Y)+2\operatorname{Cov}(X,Y). &(4.53)\end{aligned}$$

□

一个直接的结果是，如果两个随机变量不相关，那么它们之和的方差等于它们方差之和。

Corollary 4.3.3. *If X and Y are uncorrelated, then*

$$\operatorname{Var}(X+Y)=\operatorname{Var}(X)+\operatorname{Var}(Y). \tag{4.54}$$

下面的引理和例子表明，独立性蕴含不相关性，但不相关性并不总是蕴含独立性。

Lemma 4.3.4 (独立性蕴含不相关性)**.** *If two random variables are independent, then they are uncorrelated.*

Proof. 根据定理 4.1.7，如果 X 和 Y 是独立的

$$\operatorname{Cov}(X,Y)=\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)=\mathrm{E}(X)\mathrm{E}(Y)-\mathrm{E}(X)\mathrm{E}(Y)=0. \tag{4.55}$$

□

Example 4.3.5 (不相关并不意味着独立). 设X和Y为两个独立的伯努利随机变量，参数为1/2。考虑这些随机变量

$$U = X + Y, \tag{4.56}$$
$$V = X - Y. \tag{4.57}$$

请注意

$$p_U(0) = \mathrm{P}(X = 0, Y = 0) = \frac{1}{4}, \tag{4.58}$$
$$p_V(0) = \mathrm{P}(X = 1, Y = 1) + \mathrm{P}(X = 0, Y = 0) = \frac{1}{2}, \tag{4.59}$$
$$p_{U,V}(0,0) = \mathrm{P}(X = 0, Y = 0) = \frac{1}{4} \neq p_U(0)\, p_V(0) = \frac{1}{8}, \tag{4.60}$$

因此，U 和 V 不是独立的。然而，它们是无相关的，公式为：

$$\mathrm{Cov}(U,V) = \mathrm{E}(UV) - \mathrm{E}(U)\,\mathrm{E}(V) \tag{4.61}$$
$$= \mathrm{E}((X+Y)(X-Y)) - \mathrm{E}(X+Y)\,\mathrm{E}(X-Y) \tag{4.62}$$
$$= \mathrm{E}\left(X^2\right) - \mathrm{E}\left(Y^2\right) - \mathrm{E}^2(X) + \mathrm{E}^2(Y) = 0. \tag{4.63}$$

最终的等式成立，因为 X 和 Y 具有相同的分布。

△

4.3.2 Correlation coefficient

协方差并未考虑所涉及随机变量的方差大小。皮尔逊相关系数是通过使用两个变量的标准差对协方差进行标准化得到的。

Definition 4.3.6 (皮尔逊相关系数). *The Pearson correlation coefficient of two random variables X and Y is*

$$\rho_{X,Y} := \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}. \tag{4.64}$$

X 与 Y 之间的相关系数等于 X/σ_X 与 Y/σ_Y 之间的协方差。图 4.5 比较了具有相同相关系数但不同协方差的双变量高斯随机变量样本，反之亦然。

尽管这可能并不立刻显而易见，相关系数的绝对值被限制在 1 之内，因为两个随机变量的协方差不能超过它们标准差的乘积。相关系数的一个有用解释是，它量化了 X 与 Y 之间线性相关的程度。事实上，如果它等于 1 或 −1，那么其中一个变量就是另一个变量的线性函数！所有这些都源自柯西—施瓦茨不等式。证明见第 4.5.3 节。

Theorem 4.3.7 (柯西-施瓦茨不等式). *For any random variables X and Y defined on the same probability space*

$$|\mathrm{E}(XY)| \leq \sqrt{\mathrm{E}(X^2)\,\mathrm{E}(Y^2)}. \tag{4.65}$$

$\sigma_Y = 1$, $\mathrm{Cov}(X,Y) = 0.9$, $\rho_{X,Y} = 0.9$ | $\sigma_Y = 3$, $\mathrm{Cov}(X,Y) = 0.9$, $\rho_{X,Y} = 0.3$ | $\sigma_Y = 3$, $\mathrm{Cov}(X,Y) = 2.7$, $\rho_{X,Y} = 0.9$

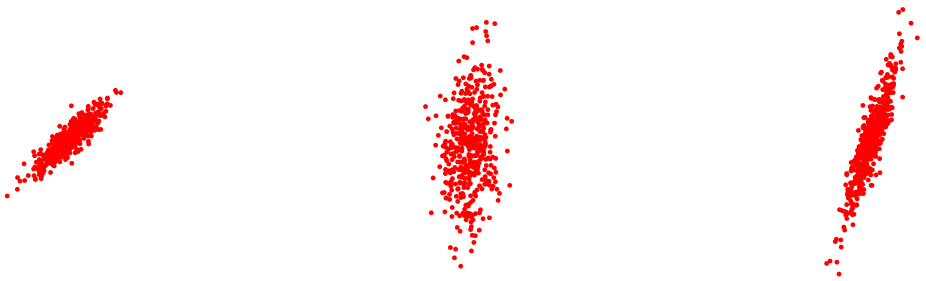


Figure 4.5: 来自二维高斯向量（X, Y）的样本，其中 X 是均值为零、方差为一的标准高斯随机变量；对应于 Y（（其均值为零））的标准差 σ_Y 的不同取值，以及 X 与 Y 之间的协方差。

Assume $\mathrm{E}\left(X^2\right) \neq 0$,

$$\mathrm{E}(XY) = \sqrt{\mathrm{E}(X^2)\,\mathrm{E}(Y^2)} \iff Y = \sqrt{\frac{\mathrm{E}(Y^2)}{\mathrm{E}(X^2)}}X, \tag{4.66}$$

$$\mathrm{E}(XY) = -\sqrt{\mathrm{E}(X^2)\,\mathrm{E}(Y^2)} \iff Y = -\sqrt{\frac{\mathrm{E}(Y^2)}{\mathrm{E}(X^2)}}X. \tag{4.67}$$

Corollary 4.3.8. *For any random variables* X *and* Y,

$$\mathrm{Cov}(X,Y) \leq \sigma_X \sigma_Y. \tag{4.68}$$

Equivalently, the Pearson correlation coefficient satisfies

$$|\rho_{X,Y}| \leq 1, \tag{4.69}$$

with equality if and only if there is a linear relationship between X *and* Y

$$|\rho_{X,Y}| = 1 \iff Y = cX + d. \tag{4.70}$$

where

$$c := \begin{cases} \frac{\sigma_Y}{\sigma_X} & \text{if } \rho_{X,Y} = 1, \\ -\frac{\sigma_Y}{\sigma_X} & \text{if } \rho_{X,Y} = -1, \end{cases} \qquad d := \mathrm{E}(Y) - c\,\mathrm{E}(X). \tag{4.71}$$

Proof. 让

$$U := X - \mathrm{E}(X), \tag{4.72}$$

$$V := Y - \mathrm{E}(Y). \tag{4.73}$$

根据方差和相关系数的定义,

$$\mathrm{E}\left(U^2\right) = \mathrm{Var}(X), \tag{4.74}$$

$$\mathrm{E}\left(V^2\right) = \mathrm{Var}(Y) \tag{4.75}$$

$$\rho_{X,Y} = \frac{\mathrm{E}(UV)}{\sqrt{\mathrm{E}(U^2)\,\mathrm{E}(V^2)}}. \tag{4.76}$$

该结果现可通过将定理4.3.7应用于U和V而得到。 □

4.3.3 Covariance matrix of a random vector

随机向量的协方差矩阵刻画了向量各个分量之间的相互作用。它在对角线上包含每个分量的方差，在非对角线上包含不同分量之间的协方差。

Definition 4.3.9. *The covariance matrix of a random vector $\vec{X}$ is defined as*

$$\Sigma_{\vec{X}} := \begin{bmatrix} \operatorname{Var}\left(\vec{X}_1\right) & \operatorname{Cov}\left(\vec{X}_1, \vec{X}_2\right) & \cdots & \operatorname{Cov}\left(\vec{X}_1, \vec{X}_n\right) \\ \operatorname{Cov}\left(\vec{X}_2, \vec{X}_1\right) & \operatorname{Var}\left(\vec{X}_2\right) & \cdots & \operatorname{Cov}\left(\vec{X}_2, \vec{X}_n\right) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}\left(\vec{X}_n, \vec{X}_1\right) & \operatorname{Cov}\left(\vec{X}_n, \vec{X}_2\right) & \cdots & \operatorname{Var}\left(\vec{X}_n\right) \end{bmatrix} \tag{4.77}$$

$$= \operatorname{E}\left(\vec{X}\vec{X}^T\right) - \operatorname{E}(\vec{X})\operatorname{E}(\vec{X})^T. \tag{4.78}$$

注意，如果一个向量的所有分量彼此不相关，那么它的协方差矩阵是对角矩阵。
来自定理 4.2.4，我们得到随机向量线性变换的协方差矩阵的简单表达式。

Theorem 4.3.10 (线性变换后的协方差矩阵)**.** *Let $\vec{X}$ be a random vector of dimension n with covariance matrix Σ. For any matrix $A \in \mathbb{R}^{m\times n}$ and $\vec{b} \in \mathbb{R}^m$,*

$$\Sigma_{A\vec{X}+\vec{b}} = A\Sigma_{\vec{X}}A^T. \tag{4.79}$$

Proof.

$$\Sigma_{A\vec{X}+\vec{b}} = \operatorname{E}\left(\left(A\vec{X}+\vec{b}\right)\left(A\vec{X}+\vec{b}\right)^T\right) - \operatorname{E}\left(A\vec{X}+\vec{b}\right)\operatorname{E}\left(A\vec{X}+\vec{b}\right)^T \tag{4.80}$$

$$= A\operatorname{E}\left(\vec{X}\vec{X}^T\right)A^T + \vec{b}\operatorname{E}\left(\vec{X}\right)^T A^T + A\operatorname{E}(\vec{X})\vec{b}^T + \vec{b}\vec{b}^T$$
$$- A\operatorname{E}(\vec{X})\operatorname{E}(\vec{X})^T A^T - A\operatorname{E}(\vec{X})\vec{b}^T - \vec{b}\operatorname{E}(\vec{X})^T A^T - \vec{b}\vec{b}^T \tag{4.81}$$

$$= A\left(\operatorname{E}\left(\vec{X}\vec{X}^T\right) - \operatorname{E}(\vec{X})\operatorname{E}(\vec{X})^T\right)A^T \tag{4.82}$$

$$= A\Sigma_{\vec{X}}A^T. \tag{4.83}$$

□

这个结果的直接推论是，我们可以轻松地从协方差矩阵中解码随机向量 *in any direction* 的方差。从数学上讲，随机向量在单位向量 $\vec{v}$ 方向上的方差等于其在 $\vec{v}$ 上投影的方差。

Corollary 4.3.11. *Let $\vec{v}$ be a unit vector,*

$$\operatorname{Var}\left(\vec{v}^T\vec{X}\right) = \vec{v}^T\Sigma_{\vec{X}}\vec{v}. \tag{4.84}$$

考虑n维随机向量$\vec{X}$的协方差矩阵的特征分解

$$\Sigma_{\vec{X}} = U\Lambda U^T \tag{4.85}$$

$$= \begin{bmatrix}\vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n\end{bmatrix} \begin{bmatrix}\lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_n\end{bmatrix} \begin{bmatrix}\vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n\end{bmatrix}^T, \tag{4.86}$$

其中特征值按 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ 排列。协方差矩阵按定义是对称的，因此根据定理 B.7.1，特征向量 $\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_n$ 可以选择为正交的。这些特征向量及其特征值完全刻画了随机向量在不同方向上的方差。该定理是推论 4.3.11 和定理 B.7.2 的直接结果。

Theorem 4.3.12. *Let $\vec{X}$ be a random vector of dimension n with covariance matrix $\Sigma_{\vec{X}}$. The eigendecomposition of $\Sigma_{\vec{X}}$ given by* (4.86) *satisfies*

$$\lambda_1 = \max_{||\vec{v}||_2=1} \mathrm{Var}\left(\vec{v}^T\vec{X}\right), \tag{4.87}$$

$$\vec{u}_1 = \arg\max_{||\vec{v}||_2=1} \mathrm{Var}\left(\vec{v}^T\vec{X}\right), \tag{4.88}$$

$$\lambda_k = \max_{||\vec{v}||_2=1, \vec{v}\perp\vec{u}_1,\ldots,\vec{u}_{k-1}} \mathrm{Var}\left(\vec{v}^T\vec{X}\right), \tag{4.89}$$

$$\vec{u}_k = \arg\max_{||\vec{v}||_2=1, \vec{v}\perp\vec{u}_1,\ldots,\vec{u}_{k-1}} \mathrm{Var}\left(\vec{v}^T\vec{X}\right). \tag{4.90}$$

用文字来说，$\vec{u}_1$ 是 *direction of maximum variance*。与第二大特征值 λ_2 对应的特征向量 $\vec{u}_2$ 是与 $\vec{u}_1$ 正交的最大变化方向。一般而言，与第 k 大特征值 λ_k 对应的特征向量 $\vec{u}_k$ 揭示了与 $\vec{u}_1$、$\vec{u}_2$、……、$\vec{u}_{k-1}$ 正交的最大变化方向。最后，$\vec{u}_n$ 是最小方差的方向。图 4.6 通过一个示例说明了这一点，其中 $n = 2$。正如我们在第 8 章中讨论的那样，主成分分析——一种流行的无监督学习与降维方法——应用同样的原理来确定数据集的变化方向。

为了总结本节内容，我们描述了一种算法，用于将来自无相关随机向量的样本转换为具有指定协方差矩阵的样本。为了这个目的，将无相关样本进行转换的过程称为 *coloring*，因为无相关样本通常被描述为 *white* 噪声。正如我们将在下一节看到的，着色使得模拟高斯随机向量成为可能。

Algorithm 4.3.13 (对不相关样本进行着色)**.** *Let $\vec{x}$ be a realization from an n-dimensional random vector with covariance matrix I. To generate samples with covariance matrix Σ, we:*

1. *Compute the eigendecomposition $\Sigma = U\Lambda U^T$.*

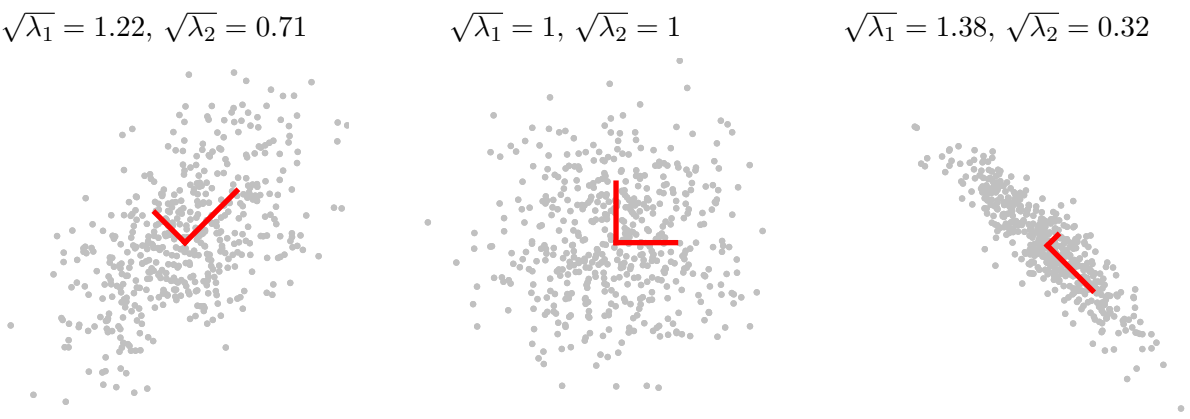


Figure 4.6: 来自具有不同协方差矩阵的二元高斯随机向量的样本以灰色显示。协方差矩阵的特征向量以红色绘制。每个特征向量都按相应特征值的平方根 λ_1 或 λ_2 进行了缩放。

2. *Set* $\vec{y} := U\sqrt{\Lambda}\vec{x}$*, where* $\sqrt{\Lambda}$ *is a diagonal matrix containing the square roots of the eigenvalues of* Σ*,*

$$\sqrt{\Lambda} := \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \sqrt{\lambda_n} \end{bmatrix}. \tag{4.91}$$

由定理4.3 .10 Y的协方差矩阵:$= U\sqrt{\Lambda}\vec{x}$确实等 als Σ.

$$\Sigma_{\vec{Y}} = U\sqrt{\Lambda}\Sigma_{\vec{X}}\sqrt{\Lambda}^T U^T \tag{4.92}$$
$$= U\sqrt{\Lambda}I\sqrt{\Lambda}^T U^T \tag{4.93}$$
$$= \Sigma. \tag{4.94}$$

图 4.7 说明了二维着色的两个步骤：首先，样本根据 Σ 的特征值进行拉伸，然后它们被旋转以与相应的特征向量对齐。

4.3.4 Gaussian random vectors

我们主要使用高斯向量来可视化协方差算子的不同性质。与其他随机向量不同，高斯随机向量完全由其均值和协方差矩阵决定。一个重要的结果是，如果一个高斯随机向量的各个分量不相关，那么它们也相互独立。

Lemma 4.3.14 (对于高斯随机向量而言，不相关意味着相互独立)**.** *If all the components of a Gaussian random vector* $\vec{X}$ *are uncorrelated, this implies that they are mutually independent.*

Proof. 高斯随机向量的联合概率密度函数的参数 Σ 是其协方差矩阵（可以通过应用协方差的定义并进行积分来验证）。如果所有分量

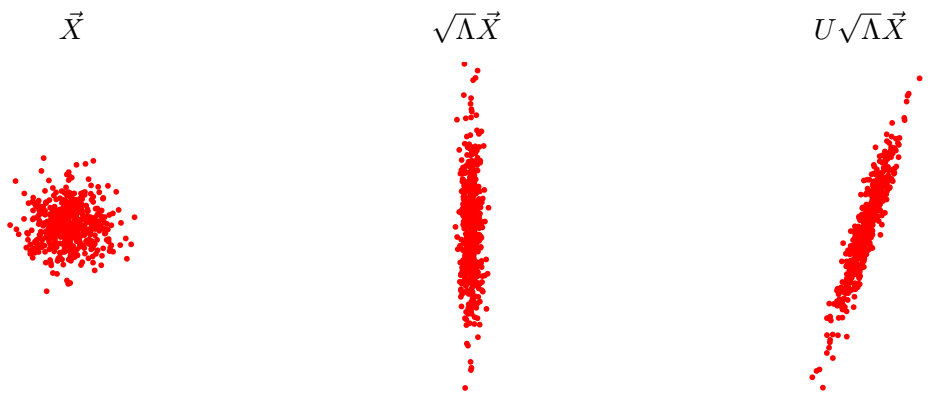


Figure 4.7: 当我们对二维无关样本进行着色（左图）时，首先通过对角矩阵 $\sqrt{\Lambda}$ *stretches* 按照期望协方差矩阵的特征值沿不同方向对其进行不同的变换（中图），然后 U 将其旋转，使其与对应的特征向量对齐（右图）。

若无相关性，则

$$\Sigma_{\vec{X}} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}, \tag{4.95}$$

其中 σ_i 是 i 维分量的标准差。现在，这个对角矩阵的逆就是

$$\Sigma_{\vec{X}}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{bmatrix}, \tag{4.96}$$

并且它的行列式是 $|\Sigma| = \prod_{i=1}^{n} \sigma_i^2$，从而

$$f_{\vec{X}}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})\right) \tag{4.97}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)\sigma_i}} \exp\left(-\frac{(\vec{x}_i - \mu_i)^2}{2\sigma_i^2}\right) \tag{4.98}$$

$$= \prod_{i=1}^{n} f_{\vec{X}_i}(\vec{x}_i)\,. \tag{4.99}$$

由于联合概率密度函数可以分解为各边际的乘积，因此各组件都是相互独立的。

□

以下算法通过对来自标准高斯分布的独立样本向量进行着色（并居中），生成具有任意均值和协方差矩阵的高斯随机向量的样本。

Algorithm 4.3.15 (生成高斯随机向量)**.** *To sample from an n-dimensional Gaussian random vector with mean $\vec{\mu}$ and covariance matrix Σ, we:*

1. *Generate a vector $\vec{x}$ containing n independent standard Gaussian samples.*
2. *Compute the eigendecomposition $\Sigma = U\Lambda U^T$.*
3. *Set $\vec{y} := U\sqrt{\Lambda}\vec{x} + \vec{\mu}$, where $\sqrt{\Lambda}$ is defined by* (8.20)*.*

算法仅对随机向量 $\vec{Y}$:= $U\sqrt{\Lambda}\vec{X} + \vec{\mu}$ 进行中心化和着色。根据期望的线性性质，它的均值是

$$\mathrm{E}(\vec{Y}) = U\sqrt{\Lambda}\mathrm{E}(\vec{X}) + \vec{\mu} \tag{4.100}$$
$$= \vec{\mu} \tag{4.101}$$

由于$\vec{X}$ 的均值为零。与方程 (4.94) 中使用的相同论证表明，$\vec{X}$ 的协方差矩阵为 Σ。由于着色和中心化是线性操作，根据定理 3.2.14，$\vec{Y}$是具有期望均值和协方差矩阵的高斯分布。例如，在图 4.7 中，生成的样本是高斯分布的。对于非高斯随机向量，着色会修改协方差矩阵，但不一定保持分布不变。

4.4 Conditional expectation

条件期望是操作随机变量的有用工具。不幸的是，它可能有些令人困惑（正如我们下面看到的，它是一个随机变量，而不是期望！）。考虑一个由两个随机变量 X 和 Y组成的函数 g。对于任何固定值 x，可以使用给定 X 的 Y的条件 pmf 或 pdf 来计算条件于事件 $X = x$ 下 g 的期望。

$$\mathrm{E}\left(g\left(X,Y\right)|X=x\right) = \sum_{y\in R} g\left(x,y\right) p_{Y|X}\left(y|x\right), \tag{4.102}$$

如果 Y是离散的并且具有范围 R，而

$$\mathrm{E}\left(g\left(X,Y\right)|X=x\right) = \int_{y=-\infty}^{\infty} g\left(x,y\right) f_{Y|X}\left(y|x\right)\,\mathrm{d}y, \tag{4.103}$$

如果 Y是连续的。

请注意，$\mathrm{E}\left(g\left(X,Y\right)|X=x\right)$ 实际上可以解释为 *function of* x，因为它将 x 的每个值映射到一个实数。这使得可以定义 $g\left(X,Y\right)$ 在给定 X 条件下的条件期望，如下所示。

Definition 4.4.1 (条件期望)**.** *The conditional expectation of $g\left(X,Y\right)$ given X is*

$$\mathrm{E}\left(g\left(X,Y\right)|X\right) := h\left(X\right), \tag{4.104}$$

where

$$h\left(x\right) := \mathrm{E}\left(g\left(X,Y\right)|X=x\right). \tag{4.105}$$

当心这个容易混淆的定义：条件期望实际上是一个随机变量！
条件期望的主要用途之一是利用迭代期望来计算期望值。其思想是，某个量的期望值可以表示为该量的条件期望的期望。

Theorem 4.4.2 (迭代期望)**.** *For any random variables X and Y and any function g: $\mathbb{R}^2 \to \mathbb{R}$*

$$\mathrm{E}\left(g\left(X,Y\right)\right) = \mathrm{E}\left(\mathrm{E}\left(g\left(X,Y\right)|X\right)\right). \tag{4.106}$$

Proof. 我们对连续随机变量证明该结果；对离散随机变量的证明，以及对同时依赖连续和离散随机变量的量的证明，几乎是完全相同的。为使说明更加清晰，我们定义

$$\begin{aligned} h\left(x\right) &:= \mathrm{E}\left(g\left(X,Y\right)|X=x\right) && (4.107)\\ &= \int_{y=-\infty}^{\infty} g\left(x,y\right) f_{Y|X}\left(y|x\right)\,\mathrm{d}y. && (4.108) \end{aligned}$$

现在，

$$\begin{aligned} \mathrm{E}\left(\mathrm{E}\left(g\left(X,Y\right)|X\right)\right) &= \mathrm{E}\left(h\left(X\right)\right) && (4.109)\\ &= \int_{x=-\infty}^{\infty} h\left(x\right) f_X\left(x\right)\,\mathrm{d}x && (4.110)\\ &= \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} f_X\left(x\right) f_{Y|X}\left(y|x\right) g\left(x,y\right)\,\mathrm{d}y\,\mathrm{d}x && (4.111)\\ &= \mathrm{E}\left(g\left(X,Y\right)\right). && (4.112) \end{aligned}$$

□

如果我们能够获得边缘分布和条件分布，迭代期望就可以非常容易地求出依赖于多个量的量的期望。我们通过前几章中的若干示例来说明这一点。

Example 4.4.3 (沙漠（续自例 3.4.10））**.** 让我们计算汽车发生故障的平均时间，即 T 的均值。通过迭代期望

$$\begin{aligned} \mathrm{E}\left(T\right) &= \mathrm{E}\left(\mathrm{E}\left(T|M,R\right)\right) && (4.113)\\ &= \mathrm{E}\left(\frac{1}{M+R}\right) \quad \text{because } T \text{ is exponential when conditioned on } M \text{ and } R && (4.114)\\ &= \int_m^{1\,0} \mathrm{d}r \int^{1\,0\,1} \frac{}{m+r}\,\mathrm{d} && (4.115)\\ &= \int_0^1 \log\left(r+1\right) - \log\left(r\right)\,\mathrm{d}r && (4.116)\\ &= \log 4 \approx 1.39 \qquad \text{integrating by parts.} && (4.117) \end{aligned}$$

△

Example 4.4.4 (黄石公园的灰熊（续自示例3.3.3））. 让我们计算约塞米蒂公园一只熊的平均体重。通过迭代期望

$$\begin{aligned} \mathrm{E}(W) &= \mathrm{E}(\mathrm{E}(W|S)) && (4.118)\\ &= \frac{\mathrm{E}(W|S=0)+\mathrm{E}(W|S=1)}{2} && (4.119)\\ &= 180 \text{ kg}. && (4.120) \end{aligned}$$

△

Example 4.4.5 (贝叶斯抛硬币（续自例 3.3.6).）让我们计算抛硬币结果 X 的均值。根据迭代期望

$$\begin{aligned} \mathrm{E}(X) &= \mathrm{E}(\mathrm{E}(X|B)) && (4.121)\\ &= \mathrm{E}(B) \quad \text{because } X \text{ is Bernoulli when conditioned on } B && (4.122)\\ &= \int_0^1 2b^2 \,\mathrm{d}b && (4.123)\\ &= \frac{2}{3}. && (4.124) \end{aligned}$$

△

4.5 Proofs

4.5.1 Derivation of means and variances in Table 4.1

Bernoulli

$$\begin{aligned} \mathrm{E}(X) &= p_X(1) = p, && (4.125)\\ \mathrm{E}\left(X^2\right) &= p_X(1), && (4.126)\\ \mathrm{Var}(X) &= \mathrm{E}\left(X^2\right) - \mathrm{E}^2(X) = p(1-p). && (4.127) \end{aligned}$$

Geometric

为了计算几何随机变量的均值，我们需要处理一个几何级数。根据下面第4.5.2节中的引理4.5.3，我们有：

$$\begin{aligned} \mathrm{E}(X) &= \sum_{k=1}^{\infty} k\, p_X(k) && (4.128)\\ &= \sum_{k=1}^{\infty} k\, p(1-p)^{k-1} && (4.129)\\ &= \frac{p}{1-p}\sum_{k=1}^{\infty} k(1-p)^k = \frac{1}{p}. && (4.130) \end{aligned}$$

为了计算均方值，我们应用同一节中的引理 4.5.4：

$$\mathrm{E}\left(X^2\right) = \sum_{k=1}^{\infty} k^2 \, p_X(k) \tag{4.131}$$

$$= \sum_{k=1}^{\infty} k^2 \, p \, (1-p)^{k-1} \tag{4.132}$$

$$= \frac{p}{1-p} \sum_{k=1}^{\infty} k^2 \, (1-p)^k \tag{4.133}$$

$$= \frac{2-p}{p^2}. \tag{4.134}$$

Binomial

如例 2.2.6 所示，我们可以将参数为 n 和 p 的二项随机变量表示为 n 个相互独立、参数为 p 的伯努利随机变量 $B_1, B_2, \ldots$ 的和。

$$X = \sum_{i=1}^{n} B_i. \tag{4.135}$$

由于伯努利随机变量的均值为p，根据期望的线性性

$$\mathrm{E}(X) = \sum_{i=1}^{n} \mathrm{E}(B_i) = np. \tag{4.136}$$

注意，由于独立性，$\mathrm{E}\left(B_i^2\right) = p$ 和 $\mathrm{E}\left(B_i B_j\right) = p^2$

$$\mathrm{E}\left(X^2\right) = \mathrm{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} B_i B_j\right) \tag{4.137}$$

$$= \sum_{i=1}^{n} \mathrm{E}\left(B_i^2\right) + 2\sum_{i=1}^{n-1}\sum_{i=j+1}^{n} \mathrm{E}\left(B_i B_j\right) = np + n\,(n-1)\,p^2. \tag{4.138}$$

Poisson

从微积分我们知道

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}, \tag{4.139}$$

即泰勒 指数函数的级数展开。这 i mplies

$$\mathrm{E}(X) = \sum_{k=1}^{\infty} k p_X(k) \tag{4.140}$$

$$= \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-1)!} \tag{4.141}$$

$$= e^{-\lambda} \sum_{m=0}^{\infty} \frac{\lambda^{m+1}}{m!} = \lambda, \tag{4.142}$$

和

$$\mathrm{E}\left(X^2\right) = \sum_{k=1}^{\infty} k^2 p_X(k) \tag{4.143}$$

$$= \sum_{k=1}^{\infty} \frac{k\lambda^k e^{-\lambda}}{(k-1)!} \tag{4.144}$$

$$= e^{-\lambda}\left(\sum_{k=1}^{\infty} \frac{(k-1)\lambda^k}{(k-1)!} + \frac{k\lambda^k}{(k-1)!}\right) \tag{4.145}$$

$$= e^{-\lambda}\left(\sum_{m=1}^{\infty} \frac{\lambda^{m+2}}{m!} + \sum_{m=1}^{\infty} \frac{\lambda^{m+1}}{m!}\right) = \lambda^2 + \lambda. \tag{4.146}$$

Uniform

我们应用连续随机变量的期望值定义得到

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x = \int_a^b \frac{x}{b-a}\mathrm{d}x \tag{4.147}$$

$$= \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}. \tag{4.148}$$

同样地,

$$\mathrm{E}\left(X^2\right) = \int_a^b \frac{x^2}{b-a}\mathrm{d}x \tag{4.149}$$

$$= \frac{b^3 - a^3}{3(b-a)} \tag{4.150}$$

$$= \frac{a^2 + ab + b^2}{3}. \tag{4.151}$$

Exponential

应用分部积分,

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x \tag{4.152}$$

$$= \int_0^{\infty} x\lambda e^{-\lambda x}\mathrm{d}x \tag{4.153}$$

$$= xe^{-\lambda x}]_0^{\infty} + \int_0^{\infty} e^{-\lambda x}\mathrm{d}x = \frac{1}{\lambda}. \tag{4.154}$$

同样地,

$$\mathrm{E}\left(X^2\right) = \int_0^{\infty} x^2\lambda e^{-\lambda x}\mathrm{d}x \tag{4.155}$$

$$= x^2 e^{-\lambda x}]_0^{\infty} + 2\int_0^{\infty} xe^{-\lambda x}\mathrm{d}x = \frac{2}{\lambda^2}. \tag{4.156}$$

Gaussian

我们应用变量变换 $t = (x-\mu)/\sigma$。

$$\begin{aligned}
\mathrm{E}(X) &= \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x && (4.157)\\
&= \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x && (4.158)\\
&= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t e^{-\frac{t^2}{2}}\,\mathrm{d}t + \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}}\,\mathrm{d}t && (4.159)\\
&= \mu, && (4.160)
\end{aligned}$$

最后一步源自于这样一个事实：在对称区间上，有限的奇函数的积分为零。

应用变量代换 $t = (x-\mu)/\sigma$ 并进行分部积分，我们得到

$$\begin{aligned}
\mathrm{E}\left(X^2\right) &= \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x && (4.161)\\
&= \int_{-\infty}^{\infty} \frac{x^2}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x && (4.162)\\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 e^{-\frac{t^2}{2}}\,\mathrm{d}t + \frac{2\mu\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t e^{-\frac{t^2}{2}}\,\mathrm{d}t + \frac{\mu^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}}\,\mathrm{d}t && (4.163)\\
&= \frac{\sigma^2}{\sqrt{2\pi}} \left(t^2 e^{-\frac{t^2}{2}}]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}}\,\mathrm{d}t \right) + \mu^2 && (4.164)\\
&= \sigma^2 + \mu^2. && (4.165)
\end{aligned}$$

4.5.2 Geometric series

Lemma 4.5.1. *For any $\alpha \neq 0$ and any integers n_1 and n_2*

$$\sum_{k=n_1}^{n_2} \alpha^k = \frac{\alpha^{n_1} - \alpha^{n_2+1}}{1-\alpha}. \tag{4.166}$$

Corollary 4.5.2. *If $0 < \alpha < 1$*

$$\sum_{k=0}^{\infty} \alpha^k = \frac{\alpha}{1-\alpha}. \tag{4.167}$$

Proof. 我们只是将该和乘以因子$(1-\alpha)/(1-\alpha)$，它显然等于一，

$$\begin{aligned}
\alpha^{n_1} + \alpha^{n_1+1} + \cdots + \alpha^{n_2-1} + \alpha^{n_2} &= \frac{1-\alpha}{1-\alpha}\left(\alpha^{n_1} + \alpha^{n_1+1} + \cdots + \alpha^{n_2-1} + \alpha^{n_2}\right) && (4.168)\\
&= \frac{\alpha^{n_1} - \alpha^{n_1+1} + \alpha^{n_1+1} + \cdots - \alpha^{n_2} + \alpha^{n_2} - \alpha^{n_2+1}}{1-\alpha}\\
&= \frac{\alpha^{n_1} - \alpha^{n_2+1}}{1-\alpha} && (4.169)
\end{aligned}$$

□

Lemma 4.5.3. *For* $0 < \alpha < 1$

$$\sum_{k=1}^{\infty} k\,\alpha^k = \frac{\alpha}{(1-\alpha)^2}. \tag{4.170}$$

Proof. 由推论 4.5.2,

$$\sum_{k=0}^{\infty} \alpha^k = \frac{1}{1-\alpha}. \tag{4.171}$$

由于左极限收敛，我们可以对两边进行微分得到

$$\sum_{k=0}^{\infty} k\alpha^{k-1} = \frac{1}{(1-\alpha)^2} \tag{4.172}$$

□

Lemma 4.5.4. *For* $0 < \alpha < 1$

$$\sum_{k=1}^{\infty} k^2\,\alpha^k = \frac{\alpha\,(1+\alpha)}{(1-\alpha)^3}. \tag{4.173}$$

Proof. 由引理 4.5.3,

$$\sum_{k=1}^{\infty} k\,\alpha^k = \frac{\alpha}{(1-\alpha)^2}. \tag{4.174}$$

由于左极限收敛，我们可以对两边求导，从而得到

$$\sum_{k=1}^{\infty} k^2\alpha^{k-1} = \frac{1+\alpha}{(1-\alpha)^3}. \tag{4.175}$$

□

4.5.3 Proof of Theorem 4.3.7

如果 $\mathrm{E}\left(X^2\right) = 0$，则由推论 4.2.13 可得 $X = 0$，且 $X =$ 以概率 1 为 0，这意味着 $\mathrm{E}\,(XY) = 0$ ，从而 (4.65) 中等号成立。若 $\mathrm{E}\left(Y^2\right) = 0$，情况亦同。

现在假设 $\mathrm{E}\left(X^2\right) \neq 0$ 且 $\mathrm{E}\left(Y^2\right) \neq 0$。我们定义常数 $a = \sqrt{\mathrm{E}\,(Y^2)}$ 和 $b = \sqrt{\mathrm{E}\,(X^2)}$。根据期望的线性性,

$$\mathrm{E}\left((aX+bY)^2\right) = a^2\mathrm{E}\left(X^2\right) + b^2\mathrm{E}\left(Y^2\right) + 2\,a\,b\,\mathrm{E}\,(XY) \tag{4.176}$$

$$= 2\left(\mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right) + \sqrt{\mathrm{E}\,(X^2)\,\mathrm{E}\,(Y^2)}\mathrm{E}\,(XY)\right), \tag{4.177}$$

$$\mathrm{E}\left((aX-bY)^2\right) = a^2\mathrm{E}\left(X^2\right) + b^2\mathrm{E}\left(Y^2\right) - 2\,a\,b\,\mathrm{E}\,(XY) \tag{4.178}$$

$$= 2\left(\mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right) - \sqrt{\mathrm{E}\,(X^2)\,\mathrm{E}\,(Y^2)}\mathrm{E}\,(XY)\right). \tag{4.179}$$

非负量的期望值是非零的，因为非负量的积分或求和是非负的。因此，(4.176) 和 (4.178) 的左侧是非负的，因此 (B.117) 和 (B.118) 都是非负的，这意味着 (4.65)。

让我们通过证明两个推论来证明 (B.21)。

($\Rightarrow$)。假设 $\mathrm{E}(XY) = -\sqrt{\mathrm{E}(X^2)\,\mathrm{E}(Y^2)}$。那么 (B.117) 等于零，因此

$$\mathrm{E}\left(\left(\sqrt{\mathrm{E}(X^2)}X + \sqrt{\mathrm{E}(X^2)}Y\right)^2\right) = 0, \tag{4.180}$$

根据推论 4.2.13，这意味着 $\sqrt{\mathrm{E}(Y^2)}X = -\sqrt{\mathrm{E}(X^2)}Y$的概率为 1。

($\Leftarrow$)。假设 $Y = -\frac{\mathrm{E}(Y^2)}{\mathrm{E}(X^2)}X$。然后可以轻松验证 (B.117) 等于零，这意味着 $\mathrm{E}(XY) = -\sqrt{\mathrm{E}(X^2)\,\mathrm{E}(Y^2)}$。

(B.22) 的证明几乎是相同的（使用 (4.176) 代替 (B.117)）。

Chapter 5

Random Processes

随机过程，也称为随机过程，用于建模随时间演变的不确定量：粒子的轨迹、石油价格、纽约的温度、美国的国债等。在这些笔记中，我们引入了一个数学框架，使得能够对这些量进行概率性推理。

5.1 Definition

我们用大写字母上方的波浪号来表示随机过程$\widetilde{X}$。这是**not**的标准记号，但我们希望强调其与随机变量和随机向量之间的区别。形式化地，随机过程$\widetilde{X}$是一个函数，它将样本空间Ω中的元素映射为实值函数。

Definition 5.1.1 (随机过程)**.** *Given a probability space* $(\Omega, \mathcal{F}, \mathsf{P})$*, a random process* $\widetilde{X}$ *is a function that maps each element* ω *in the sample space* Ω *to a function* $\widetilde{X}(\omega, \cdot) : \mathcal{T} \to \mathbb{R}$*, where* $\mathcal{T}$ *is a discrete or continuous set.*

对于$\widetilde{X}(\omega, t)$，有两种可能的解释：

- 如果我们固定 ω，那么$\widetilde{X}(\omega, t)$ 是 t 的一个 *deterministic function*，称为该随机过程的一个 **realization**。
- 如果我们固定 t，那么$\widetilde{X}(\omega, t)$ 是一个 *random variable*，我们通常就将其记为$\widetilde{X}(t)$。

因此，我们可以将$\widetilde{X}$解释为由t索引的无限个随机变量的集合。对于固定的t，随机变量$\widetilde{X}(t)$可能取值的集合称为该随机过程的**statespace**。随机过程可以根据索引变量或其状态空间进行分类。

- 如果索引变量 t 在 $\mathbb{R}$ 上定义，或者在某个 $t_0 \in \mathbb{R}$ 的半无限区间 (t_0, ∞) 上定义，那么$\widetilde{X}$ 是一个 **continuous-time** 随机过程。
- 如果索引变量 t 定义在离散集上，通常是整数或自然数，则$\widetilde{X}$ 是一个 **discrete-time** 随机过程。在这种情况下，我们通常使用与 t 不同的字母，如 i，作为索引变量。

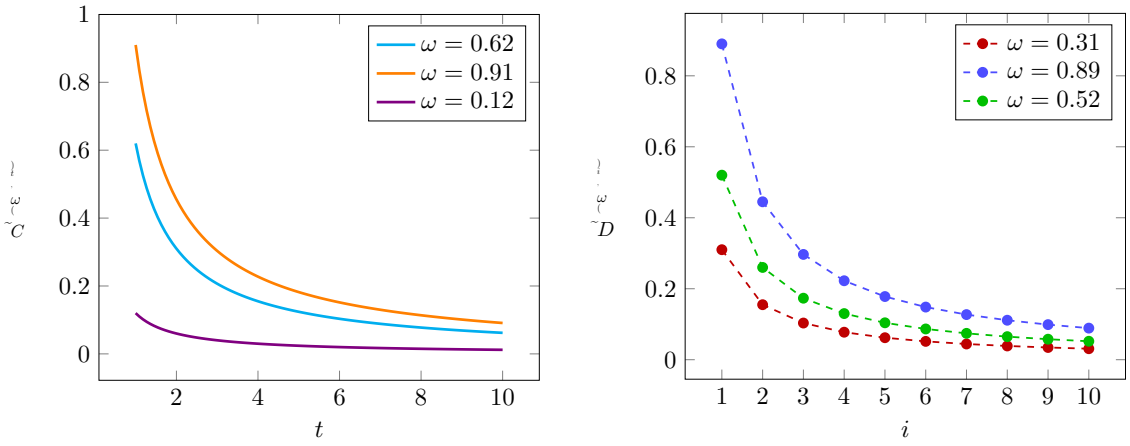


Figure 5.1: 例 5.1.2 中定义的连续时间（左）和离散时间（右）随机过程的实现。

- 如果$\widetilde{X}(t)$是一个离散随机变量，对于所有 t，则$\widetilde{X}$是一个 **discrete-state** 随机过程。如果离散随机变量取有限个值，并且对所有 t 相同，则$\widetilde{X}$是一个 **finite-state** 随机过程。

- 如果$\widetilde{X}(t)$是对所有 t 的连续随机变量，那么$\widetilde{X}$是一个 **continuous-state** 随机过程。

请注意，存在连续状态的离散时间随机过程，以及离散状态的连续时间随机过程。任何组合都是可能的。

底层概率空间 $(\Omega, \mathcal{F}, \mathrm{P})$ 在定义中提到，完全决定了随机过程的随机行为。原则上，我们可以通过定义 (1) 一个概率空间 $(\Omega, \mathcal{F}, \mathrm{P})$ 和 (2) 一个将函数分配给 Ω 中每个元素的映射来指定随机过程，如下例所示。这种指定随机过程的方式仅适用于非常简单的情况。

Example 5.1.2 (水坑)**.** Bob让Mary以概率方式对水坑进行建模。当水坑形成时，它包含的水量在0到1加仑之间均匀分布。随着时间的推移，水分蒸发。在时间间隔t之后，剩余的水量是初始数量的t倍。

玛丽将水注中的水建模为一个连续状态、连续时间的随机过程$\widetilde{C}$。其基础样本空间为 $(0, 1)$，σ 代数是对应的 Borel σ 代数（在 $(0, 1)$ 中所有可能的可数区间并集），概率测度是 $(0, 1)$ 上的均匀概率测度。对于样本空间中的一个特定元素 $\omega \in (0, 1)$。

$$\widetilde{C}(\omega, t) := \frac{\omega}{t}, \quad t \in [1, \infty), \tag{5.1}$$

在此示例中，t 的单位是天。图 6.1 显示了随机过程的不同实现。每个实现是 $[1, \infty)$ 上的一个确定性函数。

鲍勃指出，他只关心每天水坑的状态，而不是任何时间的状态 t。玛丽决定通过使用连续状态离散时间随机模型来简化该模型。

过程$\widetilde{D}$。底层的概率空间与之前完全相同，但时间索引现在是离散的。对于样本空间中的某个特定元素 $\omega \in (0,1)$

$$\widetilde{D}(\omega, i) := \frac{\omega}{i}, \quad i = 1, 2, \ldots \tag{5.2}$$

图6.1展示了连续随机过程的不同实现。注意，每一个实现都只是一个确定性的离散序列。

△

回顾一下，随机过程在某一特定时刻的取值是一个随机变量。因此，我们可以通过计算相应随机变量的分布来刻画该时刻过程的行为。同样地，我们可以考虑在 n 个固定时刻对该过程取样所得的联合分布。这由该随机过程的第 n 阶分布给出。

Definition 5.1.3 ($n$th阶分布)**.** *The nth-order distribution of a random process* $\widetilde{X}$ *is the joint distribution of the random variables* $\widetilde{X}(t_1)$, $\widetilde{X}(t_2)$, ..., $\widetilde{X}(t_n)$ *for any* n *samples* $\{t_1, t_2, \ldots, t_n\}$ *of the time index* t.

Example 5.1.4 (水洼（续））**.** 例 5.1.2 中$\widetilde{C}(t)$ 的一阶累积分布函数为

$$F_{\widetilde{C}(t)}(x) := \mathrm{P}\left(\widetilde{C}(t) \leq x\right) \tag{5.3}$$

$$= \mathrm{P}(\omega \leq t\, x) \tag{5.4}$$

$$= \begin{cases} \int_{u=0}^{t\,x} \mathrm{d}u = t\, x & \text{if } 0 \leq x \leq \frac{1}{t}\ , \\ 1 & \text{if } x > \frac{1}{t}\ , \\ 0 & \text{if } x < 0. \end{cases} \tag{5.5}$$

我们通过求导得到一阶概率密度函数。

$$f_{\widetilde{C}(t)}(x) = \begin{cases} t & \text{if } 0 \leq x \leq \frac{1}{t}\ , \\ 0 & \text{otherwise.} \end{cases} \tag{5.6}$$

△

如果一个随机过程的n阶分布是平移不变的，那么该过程被称为严格平稳或强平稳。

Definition 5.1.5 (严格/强平稳过程)**.** *A process is stationary in a strict or strong sense if for any* $n \geq 0$ *if we select* n *samples* t_1, t_2, ..., t_n *and any displacement* τ *the random variables* $\widetilde{X}(t_1)$, $\widetilde{X}(t_2)$, ..., $\widetilde{X}(t_n)$ *have the same joint distribution as* $\widetilde{X}(t_1 + \tau)$, $\widetilde{X}(t_2 + \tau)$, ..., $\widetilde{X}(t_n + \tau)$.

例 5.1.2 中的随机过程显然不是严格平稳的，因为它们的一阶概率密度函数和概率质量函数在各点并不相同。严格平稳过程的一个重要例子是独立同分布序列，将在第 5.3 节中介绍。

如同随机变量和随机向量的情况一样，定义基础的概率空间以指定随机过程通常并不非常实用，除非是非常简单的情况。

像例子5.1.2中的情况。原因在于，提出一个概率空间来产生给定的n-阶分布是具有挑战性的。幸运的是，我们也可以通过直接指定其n-阶分布*for all values of* $n = 1, 2, \ldots$ 来指定一个随机过程。这完全表征了随机过程。本章中描述的大多数随机过程，例如独立同分布序列、马尔可夫链、泊松过程和高斯过程，都是通过这种方式来指定的。

最后，随机过程也可以通过将其表示为其他随机过程的函数来加以刻画。随机过程$\widetilde{X}$ 的一个函数$\widetilde{Y} := g(\widetilde{X})$ 也是一个随机过程，因为它将样本空间 Ω 中的任意元素 ω 映射为一个函数 $\widetilde{Y}(\omega, \cdot) := g(\widetilde{X}(\omega, \cdot))$。在第 5.6 节中，我们以这种方式定义随机游走。

5.2 Mean and autocovariance functions

与随机变量和随机向量的情形一样，期望算子可以推导出用于概括随机过程行为的量。随机向量的均值是在任意固定时间 t 时$\widetilde{X}(t)$ 的均值。

Definition 5.2.1 (平均值)**.** *The mean of a random process is the function*

$$\mu_{\widetilde{X}}(t) := \mathrm{E}\left(\widetilde{X}(t)\right). \tag{5.7}$$

注意，均值是 t 的一个 *deterministic* 函数。随机过程的自协方差是另一个确定性函数，对于任意两点 t_1 和 t_2，它等于$\widetilde{X}(t_1)$ 与$\widetilde{X}(t_2)$ 的协方差。如果我们令 $t_1 := t_2$，那么自协方差等于 t_1 处的方差。

Definition 5.2.2 (自协方差)**.** *The autocovariance of a random process is the function*

$$R_{\widetilde{X}}(t_1, t_2) := \mathrm{Cov}\left(\widetilde{X}(t_1), \widetilde{X}(t_2)\right). \tag{5.8}$$

In particular,

$$R_{\widetilde{X}}(t, t) := \mathrm{Var}\left(\widetilde{X}(t)\right). \tag{5.9}$$

直观地说，自协方差量化了过程在两个不同时间点之间的相关性。如果这种相关性仅依赖于两点之间的间隔，则该过程被称为宽义平稳过程。

Definition 5.2.3 (宽平稳/弱平稳过程)**.** *A process is stationary in a wide or weak sense if its mean is constant*

$$\mu_{\widetilde{X}}(t) := \mu \tag{5.10}$$

and its autocovariance function is 平移不变*, i.e.*

$$R_{\widetilde{X}}(t_1, t_2) := R_{\widetilde{X}}(t_1 + \tau, t_2 + \tau) \tag{5.11}$$

for any t_1 and t_2 and any shift τ. For weakly stationary processes, the autocovariance is usually expressed as a function of the difference between the two time points,

$$R_{\widetilde{X}}(s) := R_{\widetilde{X}}(t, t + s) \quad \text{for any } t. \tag{5.12}$$

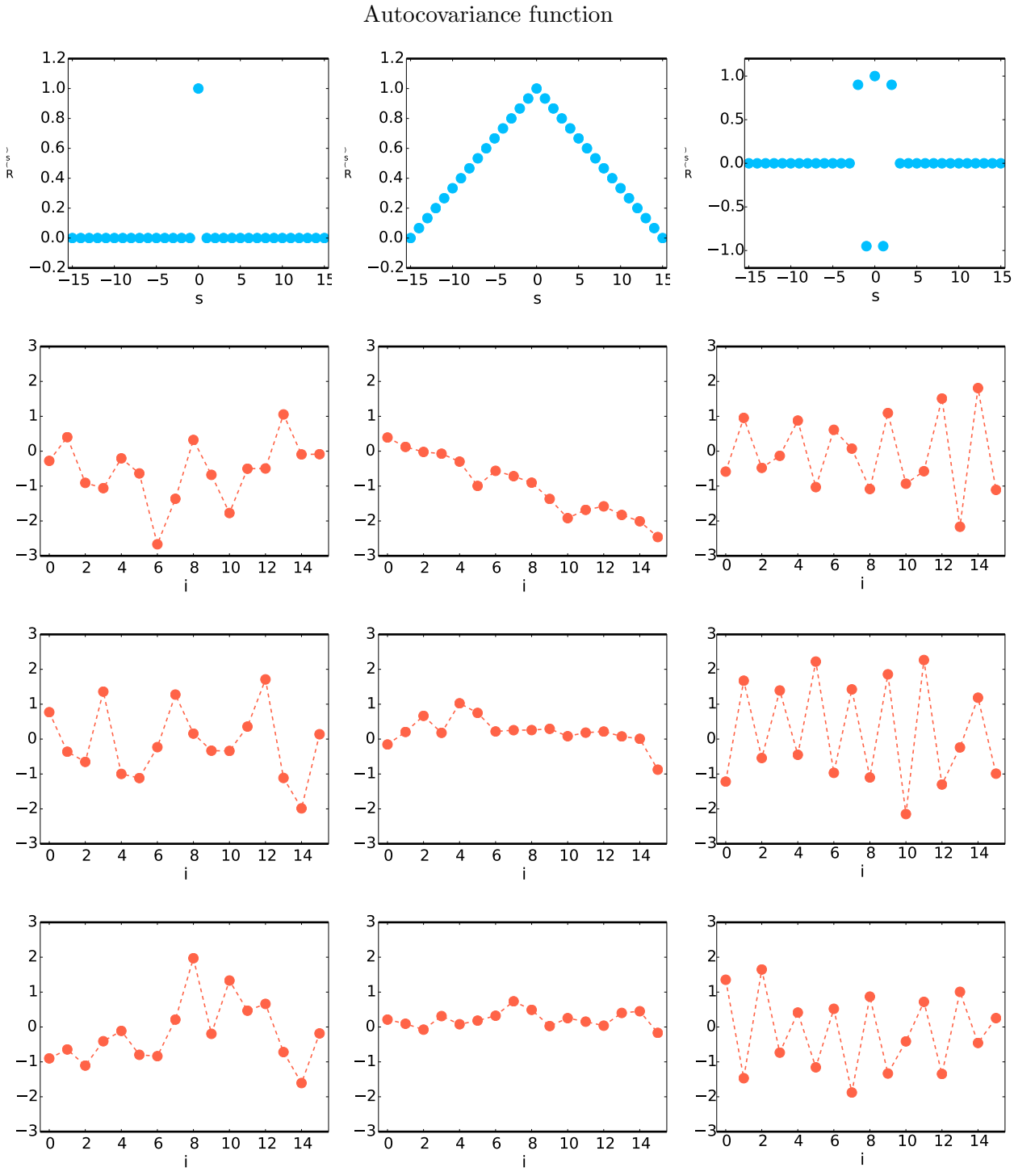


Figure 5.2: 零均值高斯过程的实现（底部三行），其自协方差函数显示在顶部一行。

注意，任何严格平稳过程必然是弱平稳的，因为其一阶和二阶分布具有移位不变性。

图 5.2 显示了具有不同自协方差函数的几个平稳随机过程。如果自协方差函数在除原点外的所有地方都为零，那么随机过程在不同点的值是无相关的。这会导致不规则的波动。当相邻时间的自协方差较高时，轨迹随机过程变得更加平滑。自相关还可以引起更有结构的行为，如图中右列所示。在该示例中，$\widetilde{X}(i)$ 与其两个邻居$\widetilde{X}(i-1)$ 和$\widetilde{X}(i+1)$ 负相关，但与$\widetilde{X}(i-2)$ 和 $\widetilde{X}(i+2)$ 正相关。这导致了快速的周期性波动。

5.3 Independent identically-distributed sequences

一个独立同分布（iid）序列$\widetilde{X}$ 是一个离散时间随机过程，其中$\widetilde{X}(i)$ 对于任何固定的 i 和$\widetilde{X}(i_1)$, $\widetilde{X}(i_2)$，...，$\widetilde{X}(i_n)$ 在任何 n 固定的索引和任何 $n \geq 2$ 下都是相互独立的。如果$\widetilde{X}(i)$ 是一个离散随机变量（或者等价地，随机过程的状态空间是离散的），则我们用 $p_{\widetilde{X}}$ 表示与每个条目的分布相关的概率质量函数（pmf）。这个概率密度函数（pdf）完全表征了随机过程，因为对于任何 n 索引 $i_1, i_2, \ldots, i_n$ 和任何 n:

$$p_{\widetilde{X}(i_1),\widetilde{X}(i_2),\ldots,\widetilde{X}(i_n)}(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) = \prod_{i=1}^{n} p_{\widetilde{X}}(x_i). \tag{5.13}$$

注意，分布在我们将每个索引平移相同的量时不会变化，因此该过程是严格平稳的。

类似地，如果$\widetilde{X}(i)$ 是一个连续随机变量，则我们用 $f_{\widetilde{X}}$ 表示与该分布相关的概率密度函数(pdf)。对于任何 n 索引 $i_1, i_2, \ldots, i_n$ 和任何 n，我们有

$$f_{\widetilde{X}(i_1),\widetilde{X}(i_2),\ldots,\widetilde{X}(i_n)}(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) = \prod_{i=1}^{n} f_{\widetilde{X}}(x_i). \tag{5.14}$$

图5.3展示了来自服从均匀分布和几何分布的独立同分布（iid）序列的若干实现。

独立同分布随机序列的均值是恒定的，且等于其相关分布的均值，我们用 μ 表示。

$$\mu_{\widetilde{X}}(i) := \mathrm{E}\left(\widetilde{X}(i)\right) \tag{5.15}$$

$$= \mu. \tag{5.16}$$

我们将与iid序列相关的分布的方差记为 σ^2。自协方差函数由以下公式给出:

$$R_{\widetilde{X}}(i,j) := \mathrm{E}\left(\widetilde{X}(i)\,\widetilde{X}(j)\right) - \mathrm{E}\left(\widetilde{X}(i)\right)\mathrm{E}\left(\widetilde{X}(j)\right) \tag{5.17}$$

$$= \begin{cases} \sigma^2, \\ 0. \end{cases} \tag{5.18}$$

这并不奇怪，$\widetilde{X}(i)$ 和$\widetilde{X}(j)$ 对于所有 $i \neq j$ 都是独立的，因此它们也是不相关的。

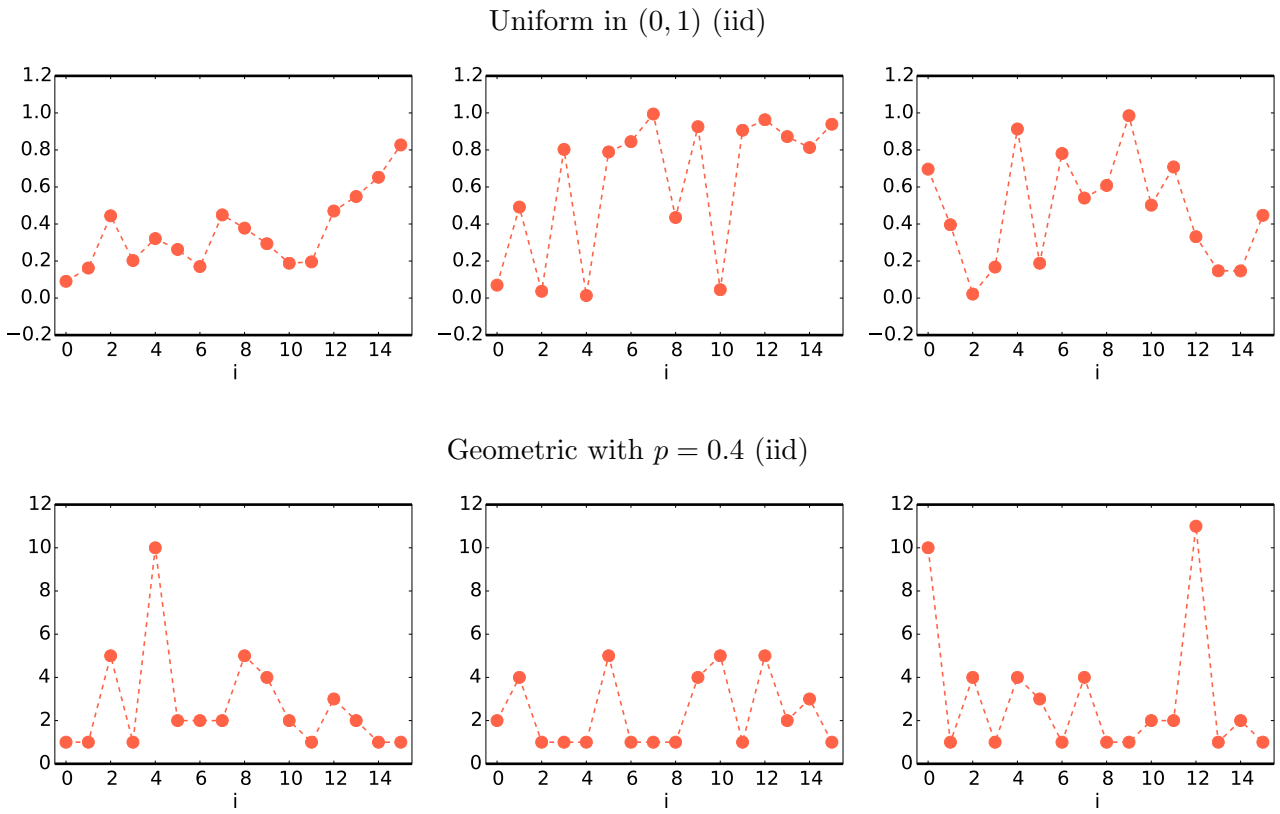


Figure 5.3: （0, 1）上的 iid 均匀序列的实现（第一行），以及参数为 $p = 0.4$ 的 iid 几何序列的实现（第二行）。

5.4 Gaussian process

一个随机过程 $\widetilde{X}$ 是高斯的，如果任何一组样本都是高斯随机向量。一个高斯过程 $\widetilde{X}$ 完全由其均值函数 $\mu_{\widetilde{X}}$ 和自协方差函数 $R_{\widetilde{X}}$ 特征化。对于所有 t_1，t_2，...，t_n 和任何 $n \geq 1$，随机向量

$$\vec{X} := \begin{bmatrix} \widetilde{X}(t_1) \\ \widetilde{X}(t_2) \\ \cdots \\ \widetilde{X}(t_n) \end{bmatrix} \tag{5.19}$$

是一个均值为的高斯随机向量

$$\vec{\mu}_{\vec{X}} := \begin{bmatrix} \mu_{\widetilde{X}}(t_1) \\ \mu_{\widetilde{X}}(t_2) \\ \cdots \\ \mu_{\widetilde{X}}(t_n) \end{bmatrix} \tag{5.20}$$

和协方差矩阵

$$\Sigma_{\vec{X}} := \begin{bmatrix} R_{\widetilde{X}}(t_1,t_1) & R_{\widetilde{X}}(t_1,t_2) & \cdots & R_{\widetilde{X}}(t_1,t_n) \\ R_{\widetilde{X}}(t_1,t_2) & R_{\widetilde{X}}(t_2,t_2) & \cdots & R_{\widetilde{X}}(t_2,t_n) \\ \vdots & \vdots & \ddots & \vdots \\ R_{\widetilde{X}}(t_2,t_n) & R_{\widetilde{X}}(t_2,t_n) & \cdots & R_{\widetilde{X}}(t_n,t_n) \end{bmatrix} \tag{5.21}$$

图 5.2 展示了若干具有不同自协方差函数的离散高斯过程的实现。从高斯随机过程中采样归结为对具有适当均值和协方差矩阵的高斯随机向量进行采样。

Algorithm 5.4.1 (生成高斯随机过程). *To sample from an Gaussian random process with mean function $\mu_{\widetilde{X}}$ and autocovariance function $\Sigma_{\widetilde{X}}$ at n points $t_1, \ldots, t_n$ we:*

1. *Compute the mean vector $\vec{\mu}_{\vec{X}}$ given by* (5.20) *and the covariance matrix $\Sigma_{\vec{X}}$ given by* (5.21).
2. *Generate n independent samples from a standard Gaussian.*
3. *Color the samples according to $\Sigma_{\vec{X}}$ and center them around $\vec{\mu}_{\vec{X}}$, as described in Algorithm 4.3.15.*

5.5 Poisson process

在例 2.2.8 中，我们通过推导在以下条件下固定时间区间内发生的事件数量的分布，来引出泊松随机变量的定义：

1. 每个事件都独立于其他所有事件发生。 2. 事件均匀发生。
3. 事件以每个时间间隔 λ 个事件的速率发生。

我们现在假设这些条件在半无限区间$[0,\infty)$内成立，并定义一个随机过程$\widetilde{N}$来计数事件。为了明确，$\widetilde{N}(t)$是发生在0和t之间的事件数量。

基于与例 2.2.8 相同的推理，随机变量$\widetilde{N}(t_2) - \widetilde{N}(t_1)$ 的分布（它表示在 t_1 与 t_2 之间发生的事件数）是参数为 $\lambda(t_2 - t_1)$ 的泊松随机变量。这对任意 t_1 和 t_2 都成立。此外，只要区间 $[t_1, t_2]$ 和 (t_3, t_4) 按条件 1 不重叠，随机变量$\widetilde{N}(t_2) - \widetilde{N}(t_1)$ 和$\widetilde{N}(t_4) - \widetilde{N}(t_3)$ 就是相互独立的。泊松过程是一种满足这两个性质的离散状态连续随机过程。

泊松过程通常用于模拟事件，如地震、电话呼叫、放射性粒子的衰变、神经脉冲等。图2.6展示了一个实际场景的例子，其中在一个呼叫中心接收到的电话数量可以很好地近似为泊松过程（只要我们考虑的时间范围为几小时）。请注意，在这里我们使用词语*event*来表示*something that happens*，例如电子邮件的到达，而不是在样本空间内的一个集合，这是在这些笔记中其他地方通常的含义。

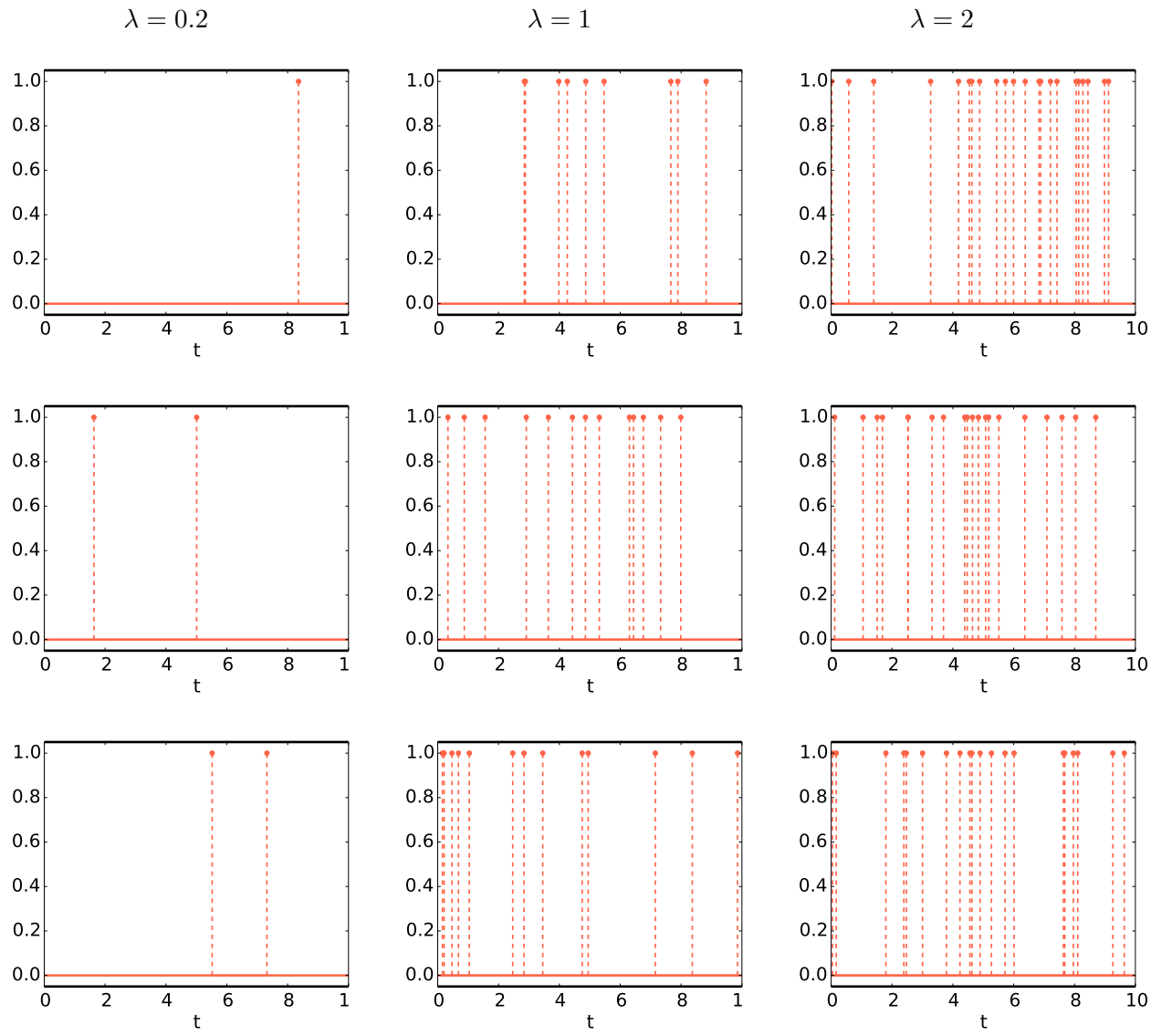


Figure 5.4: 与不同参数λ值的泊松过程$\widetilde{N}$的实现对应的事件。$\widetilde{N}(t)$等于直到时间t为止的事件数量。

Definition 5.5.1 (泊松过程). *A Poisson process with parameter λ is a discrete-state continuous random process $\widetilde{N}$ such that*

1. $\widetilde{N}(0) = 0$.
2. *For any $t_1 < t_2 < t_3 < t_4$ $\widetilde{N}(t_2) - \widetilde{N}(t_1)$ is a Poisson random variable with parameter $\lambda(t_2 - t_1)$.*
3. *For any $t_1 < t_2 < t_3 < t_4$ the random variables $\widetilde{N}(t_2) - \widetilde{N}(t_1)$ and $\widetilde{N}(t_4) - \widetilde{N}(t_3)$ are independent.*

我们现在检查随机过程是否定义良好，通过证明我们可以在任何n点$t_1 < t_2 < \ldots < t_n$的任何$n \geq 0$下推导出$\widetilde{N}$的联合pmf。为了简化符号，令$p\left(\tilde{\lambda}, x\right)$为参数$\tilde{\lambda}$下泊松随机变量在$x$处的pmf值，即。

$$p\left(\tilde{\lambda}, x\right) := \frac{\tilde{\lambda}^x \, e^{-\tilde{\lambda}}}{x!}. \tag{5.22}$$

我们有

$$\begin{aligned}
&p_{\widetilde{N}(t_1),\ldots,\widetilde{N}(t_n)}\left(x_1, \ldots, x_n\right) && (5.23)\\
&= \mathrm{P}\left(\widetilde{N}\left(t_1\right) = x_1, \ldots, \widetilde{N}\left(t_n\right) = x_n\right) && (5.24)\\
&= \mathrm{P}\left(\widetilde{N}\left(t_1\right) = x_1, \widetilde{N}\left(t_2\right) - \widetilde{N}\left(t_1\right) = x_2 - x_1, \ldots, \widetilde{N}\left(t_n\right) - \widetilde{N}\left(t_{n-1}\right) = x_n - x_{n-1}\right) && (5.25)\\
&= \mathrm{P}\left(\widetilde{N}\left(t_1\right) = x_1\right) \mathrm{P}\left(\widetilde{N}\left(t_2\right) - \widetilde{N}\left(t_1\right) = x_2 - x_1\right) \ldots \mathrm{P}\left(\widetilde{N}\left(t_n\right) - \widetilde{N}\left(t_{n-1}\right) = x_n - x_{n-1}\right)\\
&= p\left(\lambda t_1, x_1\right) p\left(\lambda\left(t_2 - t_1\right), x_2 - x_1\right) \ldots p\left(\lambda\left(t_n - t_{n-1}\right), x_n - x_{n-1}\right). && (5.26)
\end{aligned}$$

用文字表达，我们已经将事件$\widetilde{N}\left(t_i\right) = x_i$对于$1 \leq i \leq n$通过随机变量$\widetilde{N}\left(t_1\right)$和$\widetilde{N}\left(t_i\right) - \widetilde{N}\left(t_{i-1}\right)$表示，其中$\leq i \leq n$是独立的泊松随机变量，参数分别为$\lambda t_1$和$\lambda\left(t_i - t_{i-1}\right)$。

图 5.4 显示了对应于泊松过程$\widetilde{N}$实现的多个事件序列，参数λ（$\widetilde{N}\left(t\right)$等于到时间$t$）之前的事件数量。值得注意的是，事件之间的间隔时间，即连续事件之间的时间，始终具有相同的分布：它是一个指数随机变量。

Lemma 5.5.2 (泊松过程的到达间隔时间服从指数分布)**.** *Let T denote the time between two contiguous events in a Poisson process with parameter λ. T is an exponential random variable with parameter λ.*

证明见附录第5.7.1节。图2.11显示，呼叫中心的电话呼叫到达时间确实可以很好地用指数分布来建模。

引理 5.5.2 表明，为了模拟泊松过程，我们所需要做的就是从指数分布中抽样。

Algorithm 5.5.3 (生成泊松随机过程)**.** *To sample from a Poisson random process with parameter λ we:*

1. *Generate independent samples from an exponential random variable with parameter λ t_1, t_2, t_3, ...*
2. *Set the events of the Poisson process to occur at t_1, $t_1 + t_2$, $t_1 + t_2 + t_3$, ...*

图 5.4 是以这种方式生成的。为了确认该算法能够从泊松过程中进行采样，我们必须证明所得过程满足定义 5.5.1 中的条件。事实上确实如此，但我们省略证明。

以下引理推导了泊松过程的均值和自协方差函数，证明见第5.7.2节。

Lemma 5.5.4 (泊松过程)**.** *The mean and autocovariance of a Poisson process equal* 的均值与自协方差

$$\mathrm{E}\left(\widetilde{X}(t)\right) = \lambda t, \tag{5.27}$$

$$R_{\widetilde{X}}(t_1, t_2) = \lambda \min\{t_1, t_2\}. \tag{5.28}$$

泊松过程的均值不是常数，其自协方差也不是平移不变的，因此该过程既不是严格平稳的，也不是宽意义上的平稳过程。

Example 5.5.5 (地震)**.** 在旧金山半岛上，震中强度至少为里氏3级的地震数量可以通过参数为0.3次地震/年 的泊松过程来建模。接下来十年内没有地震，接下来的二十年内至少发生一次地震的概率是多少?

我们定义一个参数为 0.3 的泊松过程$\widetilde{X}$ 来建模该问题。未来 10 年的地震次数，即$\widetilde{X}(10)$，是一个泊松随机变量，参数为 $0.3 \cdot 10 = 3$。接下来的 20 年中的地震，即$\widetilde{X}(30) - \widetilde{X}(10)$，是一个泊松随机变量，参数为 $0.3 \cdot 20 = 6$。由于时间间隔不重叠，两个随机变量是独立的。

$$\mathrm{P}\left(\widetilde{X}(10) = 0, \widetilde{X}(30) \geq 1\right) = \mathrm{P}\left(\widetilde{X}(10) = 0, \widetilde{X}(30) - \widetilde{X}(10) \geq 1\right) \tag{5.29}$$

$$= \mathrm{P}\left(\widetilde{X}(10) = 0\right) \mathrm{P}\left(\widetilde{X}(30) - \widetilde{X}(10) \geq 1\right) \tag{5.30}$$

$$= \mathrm{P}\left(\widetilde{X}(10) = 0\right) \left(1 - \mathrm{P}\left(\widetilde{X}(30) - \widetilde{X}(10) = 0\right)\right) \tag{5.31}$$

$$= e^{-3}\left(1 - e^{-6}\right) = 4.97\,10^{-2}. \tag{5.32}$$

概率是4.97%。

△

5.6 Random walk

随机游走是一个离散时间随机过程，模拟沿随机方向进行的一系列步骤。为了正式地定义一个随机游走，我们首先定义一个独立同分布的步骤序列$\widetilde{S}$，使得

$$\widetilde{S}(i) = \begin{cases} +1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2}. \end{cases} \tag{5.33}$$

我们将随机游走$\widetilde{X}$ 定义为离散状态离散时间随机过程

$$\widetilde{X}(i) := \begin{cases} 0 & \text{for } i = 0, \\ \sum_{j=1}^{i} \widetilde{S}(j) & \text{for } i = 1, 2, \ldots \end{cases} \tag{5.34}$$

我们已将$\widetilde{X}$ 定义为独立同分布序列的函数，因此它是良好定义的。图5.5显示了随机游走的若干实现。

$\widetilde{X}$ 是对称的（采取正步和负步的概率相同），并且从原点开始。很容易定义走法在非对称的情况下并且从

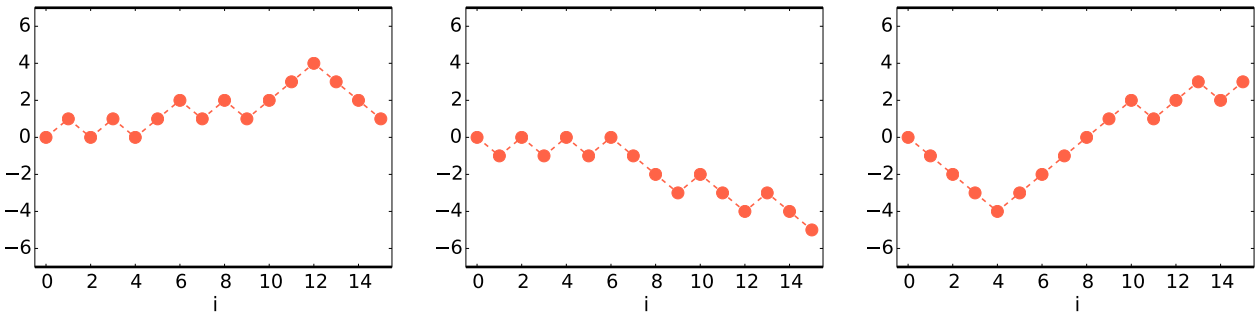


Figure 5.5: 随机游走在第5.5节中定义的实现。

在另一个点上。对更高维空间的泛化——例如在二维表面上建模随机过程——也是可能的。

我们在以下引理中推导了随机游走的一级概率质量函数（pmf），该引理在附录的第5.7.3节中证明。

Lemma 5.6.1 (随机游走)**.** *The first-order pmf of the random walk* $\widetilde{X}$ *is*的一阶pmf

$$p_{\widetilde{X}(i)}(x) = \begin{cases} \binom{i}{\frac{i+x}{2}} \frac{1}{2^i} & \text{if } i+x \text{ is even and } -i \le x \le i \\ 0 & \text{otherwise.} \end{cases} \tag{5.35}$$

随机游走的一阶分布显然随时间变化，因此该随机过程不是严格平稳的。根据下述引理，随机游走的均值是常数（其值为零）。然而，自协方差并不具有平移不变性，所以该过程也不是弱平稳的。

Lemma 5.6.2 (随机游走)**.** *The mean and autocovariance of the random walk* $\widetilde{X}$ *are* 的均值和自协方差

$$\mu_{\widetilde{X}}(i) = 0, \tag{5.36}$$
$$R_{\widetilde{X}}(i,j) = \min\{i,j\}. \tag{5.37}$$

Proof.

$$\mu_{\widetilde{X}}(i) := \mathrm{E}\left(\widetilde{X}(i)\right) \tag{5.38}$$
$$= \mathrm{E}\left(\sum_{j=1}^{i} \widetilde{S}(j)\right) \tag{5.39}$$
$$= \sum_{j=1}^{i} \mathrm{E}\left(\widetilde{S}(j)\right) \qquad \text{by linearity of expectation} \tag{5.40}$$
$$= 0. \tag{5.41}$$

$$R_{\widetilde{X}}(i,j) := \mathrm{E}\left(\widetilde{X}(i)\,\widetilde{X}(j)\right) - \mathrm{E}\left(\widetilde{X}(i)\right)\mathrm{E}\left(\widetilde{X}(j)\right) \tag{5.42}$$

$$= \mathrm{E}\left(\sum_{k=1}^{i}\sum_{l=1}^{j}\widetilde{S}(k)\,\widetilde{S}(l)\right) \tag{5.43}$$

$$= \mathrm{E}\left(\sum_{k=1}^{\min\{i,j\}}\widetilde{S}(k)^2 + \sum_{k=1}^{i}\sum_{\substack{l=1\\ l\neq k}}^{j}\widetilde{S}(k)\,\widetilde{S}(l)\right) \tag{5.44}$$

$$= \sum_{k=1}^{\min\{i,j\}} 1 + \sum_{k=1}^{i}\sum_{\substack{l=1\\ l\neq k}}^{j}\mathrm{E}\left(\widetilde{S}(k)\right)\mathrm{E}\left(\widetilde{S}(l)\right) \tag{5.45}$$

$$= \text{最小值}\,\{i,j\}, \tag{5.46}$$

其中(5.45)来自期望的线性性和独立性。 □

$\widetilde{X}$ 在 i 的方差等于 $R_{\widetilde{X}}(i,i) = i$，这意味着随机漫步的标准差按 $\sqrt{i}$ 变化。

Example 5.6.3 (赌博者)**.** 一名赌博者正在玩以下游戏。一个公平的硬币被连续掷出。每次结果是正面时，赌博者赢得一美元，每次结果是反面时，她失去一美元。我们可以将赌博者赚取（或失去）的钱数建模为一个随机漫步，只要掷硬币的结果是独立的。这使得我们可以估算出期望收益为零，或者估算出赌博者在前10次掷硬币后赚取6美元或更多的概率是

$$\mathrm{P}\,(\text{gambler is up \$6 or more}) = p_{\widetilde{X}(10)}(6) + p_{\widetilde{X}(10)}(8) + p_{\widetilde{X}(10)}(10) \tag{5.47}$$

$$= \binom{10}{8}\frac{1}{2^{10}} + \binom{10}{9}\frac{1}{2^{10}} + \frac{1}{2^{10}} \tag{5.48}$$

$$= 5.47\,10^{-2}. \tag{5.49}$$

△

5.7 Proofs

5.7.1 Proof of Lemma 5.5.2

我们首先推导 T 的累积分布函数（cdf），

$$F_T(t) := \mathrm{P}\,(T \le t) \tag{5.50}$$

$$= 1 - \mathrm{P}\,(T > t) \tag{5.51}$$

$$= 1 - \mathrm{P}\,(\text{no events in an interval of length } t) \tag{5.52}$$

$$= 1 - e^{-\lambda t} \tag{5.53}$$

因为一个长度为 t 的区间内的点数服从参数为 λt 的泊松分布。对其求导，我们得出结论：

$$f_T(t) = \lambda e^{-\lambda t}. \tag{5.54}$$

5.7.2 Proof of Lemma 5.5.4

根据定义，介于 0 和 t 之间的事件数服从参数为 λt 的泊松随机变量分布，因此其均值等于 λt 。

自协方差等于

$$R_{\widetilde{X}}(t_1,t_2) := \mathrm{E}\left(\widetilde{X}(t_1)\,\widetilde{X}(t_2)\right) - \mathrm{E}\left(\widetilde{X}(t_1)\right)\mathrm{E}\left(\widetilde{X}(t_2)\right) \tag{5.55}$$

$$= \mathrm{E}\left(\widetilde{X}(t_1)\,\widetilde{X}(t_2)\right) - \lambda^2 t_1 t_2. \tag{5.56}$$

根据假设，$\widetilde{X}(t_1)$ 和$\widetilde{X}(t_2) - \widetilde{X}(t_1)$ 相互独立，因此

$$\mathrm{E}\left(\widetilde{X}(t_1)\,\widetilde{X}(t_2)\right) = \mathrm{E}\left(\widetilde{X}(t_1)\left(\widetilde{X}(t_2) - \widetilde{X}(t_1)\right) + \widetilde{X}(t_1)^2\right) \tag{5.57}$$

$$= \mathrm{E}\left(\widetilde{X}(t_1)\right)\mathrm{E}\left(\widetilde{X}(t_2) - \widetilde{X}(t_1)\right) + \mathrm{E}\left(\widetilde{X}(t_1)^2\right) \tag{5.58}$$

$$= \lambda^2 t_1 (t_2 - t_1) + \lambda t_1 + \lambda^2 t_1^2 \tag{5.59}$$

$$= \lambda^2 t_1 t_2 + \lambda t_1. \tag{5.60}$$

5.7.3 Proof of Lemma 5.6.1

让我们定义随机游走所采取的正步数 S_+。鉴于对 $\widetilde{S}$ 的假设，这是一个参数为 i 和 $1/2$ 的二项随机变量。负步数为 $S_- := i - S_+$。为了使$\widetilde{X}(i)$ 等于 x，我们需要净步数等于 x，这意味着

$$x = S_+ - S_- \tag{5.61}$$

$$= 2S_+ - i. \tag{5.62}$$

这意味着 S_+ 必须等于 $\frac{i+x}{2}$。我们得出结论，

$$p_{\widetilde{X}(i)}(i) = \mathrm{P}\left(\sum_{j=0}^{i} \widetilde{S}(i) = x\right) \tag{5.63}$$

$$= \binom{i}{\frac{i+x}{2}} \frac{1}{2^i} \qquad \text{if } \frac{i+x}{2} \text{ is an integer between 0 and } i. \tag{5.64}$$

Chapter 6

Convergence of Random Processes

在本章中，我们研究离散随机过程的收敛性。这使我们能够刻画统计估计和概率建模中两个基本现象：大数定律和中心极限定理。

6.1 Types of convergence

让我们快速回顾一个确定性的实数序列 x_1、x_2、... 的收敛性概念。我们有

$$\lim_{i\to\infty} x_i = x \tag{6.1}$$

如果 x_i 在索引 i 增长时任意接近 x。更正式地说，序列收敛到 x，如果对于任何 $\epsilon > 0$，都存在一个索引 i_0，使得对于所有大于 i_0 的索引 i，我们都有 $|x_i - x| < \epsilon$。回想一下，任何离散时间随机过程$\widetilde{X}(\omega, i)$ 的实现，当我们固定结果 ω 时，是一个确定性序列。因此，通过计算相应的极限，可以实现这样的实现的收敛到一个固定数值。然而，如果我们考虑的是随机过程本身而不是一个实现，并且我们想要确定它是否最终收敛到一个随机变量 X，那么确定性收敛就不再有意义。在本节中，我们描述了几种收敛的替代定义，允许将这一概念扩展到随机量。

6.1.1 Convergence with probability one

考虑一个离散随机过程$\widetilde{X}$ 和一个在同一概率空间上定义的随机变量 X。如果我们固定样本空间 Ω 中的一个元素 ω，那么$\widetilde{X}(i, \omega)$ 就是一个确定性序列，而 $X(\omega)$ 是一个常数。因此，我们可以验证$\widetilde{X}(i, \omega)$ 是否会确定性地收敛到 $X(\omega)$，当 $i \to \infty$ *for that particular value of* ω 时。事实上，我们可以问：发生这种情况的概率是多少？更准确地说，这将是如果我们抽取 ω，我们得到的概率。

$$\lim_{i\to\infty} \widetilde{X}(i, \omega) = X(\omega). \tag{6.2}$$

如果这个概率等于一，那么我们称$\widetilde{X}(i)$ 以概率一收敛到 X。

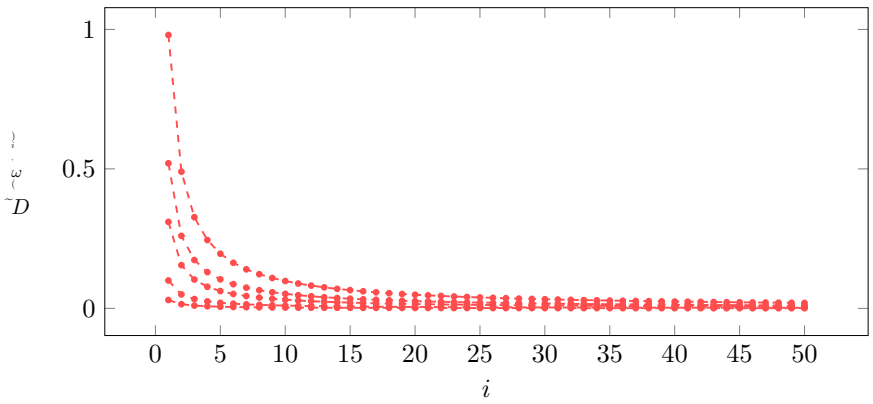


Figure 6.1: 例5.1.2中定义的离散随机过程$\widetilde{D}$ 的零收敛性。

Definition 6.1.1 (几乎必然收敛). *A discrete random vector* $\widetilde{X}$ *converges with probability one to a random variable* X *belonging to the same probability space* $(\Omega, \mathcal{F}, \mathrm{P})$ *if*

$$\mathrm{P}\left(\left\{\omega \mid \omega \in \Omega, \quad \lim_{i\to\infty} \widetilde{X}(\omega, i) = X(\omega)\right\}\right) = 1. \tag{6.3}$$

回顾一下，一般而言，除非是非常简单的情况，否则样本空间 Ω 很难被明确定义并进行显式操作。

Example 6.1.2 (水洼（接续自例 5.1.2））**.** 让我们考虑在例 5.1.2 中定义的离散随机过程$\widetilde{D}$。如果我们固定 $\omega \in (0, 1)$

$$\lim_{i\to\infty} \widetilde{D}(\omega, i) = \lim_{i\to\infty} \frac{\omega}{i} \tag{6.4}$$

$$= 0. \tag{6.5}$$

事实证明，在样本空间中 ω 的所有可能取值下，其实现值都趋于零。这意味着$\widetilde{D}$ 以概率一收敛到零。

△

6.1.2 Convergence in mean square and in probability

为验证以概率为一的收敛性，我们固定结果 ω，并检查随机过程的相应实现是否以确定性的方式收敛。另一种观点是固定索引变量 i，并在我们增大 i 时，考察随机变量$\widetilde{X}(i)$ 与另一随机变量 X 的接近程度。

两个随机变量之间的可能距离度量是它们差异的均方值。如果 $\mathrm{E}((X-Y)^2) = 0$，那么根据切比雪夫不等式，$X = Y$的概率为一。$\widetilde{X}(i)$ 和 X 之间的均方偏差是一个确定性量（一个数字），因此我们可以评估其收敛性为 $i \to \infty$。如果它收敛到零，那么我们说该随机序列在均方意义上收敛。

Definition 6.1.3 (均方收敛). *A discrete random process $\widetilde{X}$ converges in mean square to a random variable X belonging to the same probability space if*

$$\lim_{i\to\infty} \mathrm{E}\left(\left(X-\widetilde{X}\left(i\right)\right)^2\right)=0. \tag{6.6}$$

或者，我们可以考虑$\widetilde{X}\left(i\right)$与 X 相距某个给定的固定 $\epsilon>0$ 的概率。如果对任意 ϵ，无论多小，当 $i\to\infty$ 时该概率收敛到零，那么我们称该随机序列以概率收敛。

Definition 6.1.4 (依概率收敛). *A discrete random process $\widetilde{X}$ converges in probability to another random variable X belonging to the same probability space if for any $\epsilon>0$*

$$\lim_{i\to\infty} \mathrm{P}\left(\left|X-\widetilde{X}\left(i\right)\right|>\epsilon\right)=0. \tag{6.7}$$

注意，与均方收敛的情形一样，该定义中的极限是确定性的，因为它是概率的极限，而概率只是实数。

作为马尔可夫不等式的直接推论，均方收敛蕴含依概率收敛。

Theorem 6.1.5. *Convergence in mean square implies convergence in probability.*

Proof. 我们有

$$\begin{aligned}
\lim_{i\to\infty} \mathrm{P}\left(\left|X-\widetilde{X}\left(i\right)\right|>\epsilon\right) &= \lim_{i\to\infty} \mathrm{P}\left(\left(X-\widetilde{X}\left(i\right)\right)^2>\epsilon^2\right) && (6.8)\\
&\le \lim_{i\to\infty} \frac{\mathrm{E}\left(\left(X-\widetilde{X}\left(i\right)\right)^2\right)}{\epsilon^2} \quad \text{by Markov's inequality} && (6.9)\\
&= 0, && (6.10)
\end{aligned}$$

如果该序列在均方意义下收敛。 □

事实证明，几乎必然收敛也蕴含依概率收敛。几乎必然收敛并不蕴含均方收敛，反之亦然。对于本课程的目的而言，这三种收敛方式之间的差别并不十分重要。

6.1.3 Convergence in distribution

在某些情况下，随机过程$\tilde{X}$ 并不收敛到任何随机变量的取值，但$\tilde{X}\left(i\right)$ 的累积分布函数（cdf）逐点收敛到另一个随机变量 X 的累积分布函数。在这种情况下，$\tilde{X}\left(i\right)$ 和 X 的实际取值是 *not* 必然接近的，但在极限中它们具有相同的 *distribution*。在这种情况下，我们称$\widetilde{X}$ 以分布收敛到 X。

Definition 6.1.6 (依分布收敛). *A random process $\widetilde{X}$ converges in distribution to a random variable X belonging to the same probability space if*

$$\lim_{i\to\infty} F_{\widetilde{X}(i)}\left(x\right)=F_X\left(x\right) \tag{6.11}$$

for all $x\in\mathbb{R}$ where F_X is continuous.

注意，分布收敛是一个比以概率1收敛、均方收敛或概率收敛弱得多的概念。如果离散随机过程 $\widetilde{X}$ 在分布意义下收敛到随机变量 X，这仅仅意味着当 i 变大时，$\widetilde{X}(i)$ 的分布趋向于 X 的分布，而并不意味着这两个随机变量的 *the values* 很接近。然而，概率收敛（因此也包括以概率1收敛或均方收敛）确实蕴含分布收敛。

Example 6.1.7 (二项分布收敛到泊松)**.** 让我们定义一个离散随机过程$\widetilde{X}(i)$，使得$\widetilde{X}(i)$ 的分布是参数为 i 和 p 的二项分布：$= \lambda/i$。在 $i \neq j$ 下，$\widetilde{X}(i)$ 和$\widetilde{X}(j)$ 相互独立，这完全刻画了该过程对所有 $n > 1$ 的 n 阶分布。考虑一个参数为 λ 的泊松随机变量 X，它对所有 i 与$\tilde{X}(i)$ 相互独立。当 $i \to \infty$ 时，你是否期望 X 和$\tilde{X}(i)$ 的取值接近?

不！事实上，即使是$\tilde{X}(i)$ 和$\tilde{X}(i+1)$ 通常也不会接近。然而，$\tilde{X}$ 在分布上收敛到 X，正如示例 2.2.8 中所建立的那样：

$$\lim_{i\to\infty} p_{\widetilde{X}(i)}(x) = \lim_{i\to\infty} \binom{i}{x} p^x (1-p)^{(i-x)} \tag{6.12}$$

$$= \frac{\lambda^x e^{-\lambda}}{x!} \tag{6.13}$$

$$= p_X(x). \tag{6.14}$$

△

6.2 Law of large numbers

让我们定义离散随机过程的平均值。

Definition 6.2.1 (移动平均)**.** *The moving or running average $\widetilde{A}$ of a discrete random process $\widetilde{X}$, defined for $i = 1, 2, \ldots$ (i.e. 1 is the starting point), is equal to*

$$\widetilde{A}(i) := \frac{1}{i}\sum_{j=1}^{i} \widetilde{X}(j). \tag{6.15}$$

考虑一个 iid 序列。对移动平均的一种非常自然的解释是，它是对均值的实时估计。事实上，从统计学角度看，移动平均是该过程截至时间 i (的样本均值；样本均值在第 8 章中定义)。大数定律表明，该平均值确实收敛到该 iid 序列的均值。

Theorem 6.2.2 (弱大数定律)**.** *Let $\widetilde{X}$ be an iid discrete random process with mean $\mu_{\widetilde{X}} := \mu$ such that the variance of $\widetilde{X}(i)$ σ^2 is bounded. Then the average $\widetilde{A}$ of $\widetilde{X}$ converges in mean square to μ.*

Proof. 首先，我们确定$\widetilde{A}(i)$的均值是恒定的且等于μ，

$$\mathrm{E}\left(\widetilde{A}(i)\right) = \mathrm{E}\left(\frac{1}{i}\sum_{j=1}^{i}\widetilde{X}(j)\right) \tag{6.16}$$

$$= \frac{1}{i}\sum_{j=1}^{i}\mathrm{E}\left(\widetilde{X}(j)\right) \tag{6.17}$$

$$= \mu. \tag{6.18}$$

由于独立性假设，方差在i中线性缩放。回想一下，对于独立随机变量，和的方差等于方差的和，

$$\mathrm{Var}\left(\widetilde{A}(i)\right) = \mathrm{Var}\left(\frac{1}{i}\sum_{j=1}^{i}\widetilde{X}(j)\right) \tag{6.19}$$

$$= \frac{1}{i^2}\sum_{j=1}^{i}\mathrm{Var}\left(\widetilde{X}(j)\right) \tag{6.20}$$

$$= \frac{\sigma^2}{i}. \tag{6.21}$$

我们得出结论，

$$\lim_{i\to\infty}\mathrm{E}\left(\left(\widetilde{A}(i)-\mu\right)^2\right) = \lim_{i\to\infty}\mathrm{E}\left(\left(\widetilde{A}(i)-\mathrm{E}\left(\widetilde{A}(i)\right)\right)^2\right) \quad \text{by (6.18)} \tag{6.22}$$

$$= \lim_{i\to\infty}\mathrm{Var}\left(\widetilde{A}(i)\right) \tag{6.23}$$

$$= \lim_{i\to\infty}\frac{\sigma^2}{i} \quad \text{by (6.21)} \tag{6.24}$$

$$= 0. \tag{6.25}$$

□

根据定理6.1.5，平均值也以概率收敛到独立同分布序列的均值。实际上，在相同假设下，还可以证明以概率1收敛。这个结果被称为大数法则的强法则，但证明超出了本笔记的范围。我们建议有兴趣的读者查阅更高级的概率论教材。

图 6.2 显示了多个独立同分布（iid）序列的实现的平均值。当 iid 序列是高斯分布或几何分布时，我们观察到它们趋向于分布的均值；然而，当序列是柯西分布时，移动平均值发散。原因是，如示例 4.2.2 所示，柯西分布没有明确的均值！直观地说，柯西分布下极端值具有非忽略的概率，因此 iid 序列时不时会取到非常大的数值，这阻止了移动平均值的收敛。

6.3 Central limit theorem

在上一节中，我们已经确定，一系列独立同分布（iid）随机变量的移动平均数会收敛到它们分布的均值（前提是均值是良好定义的并且

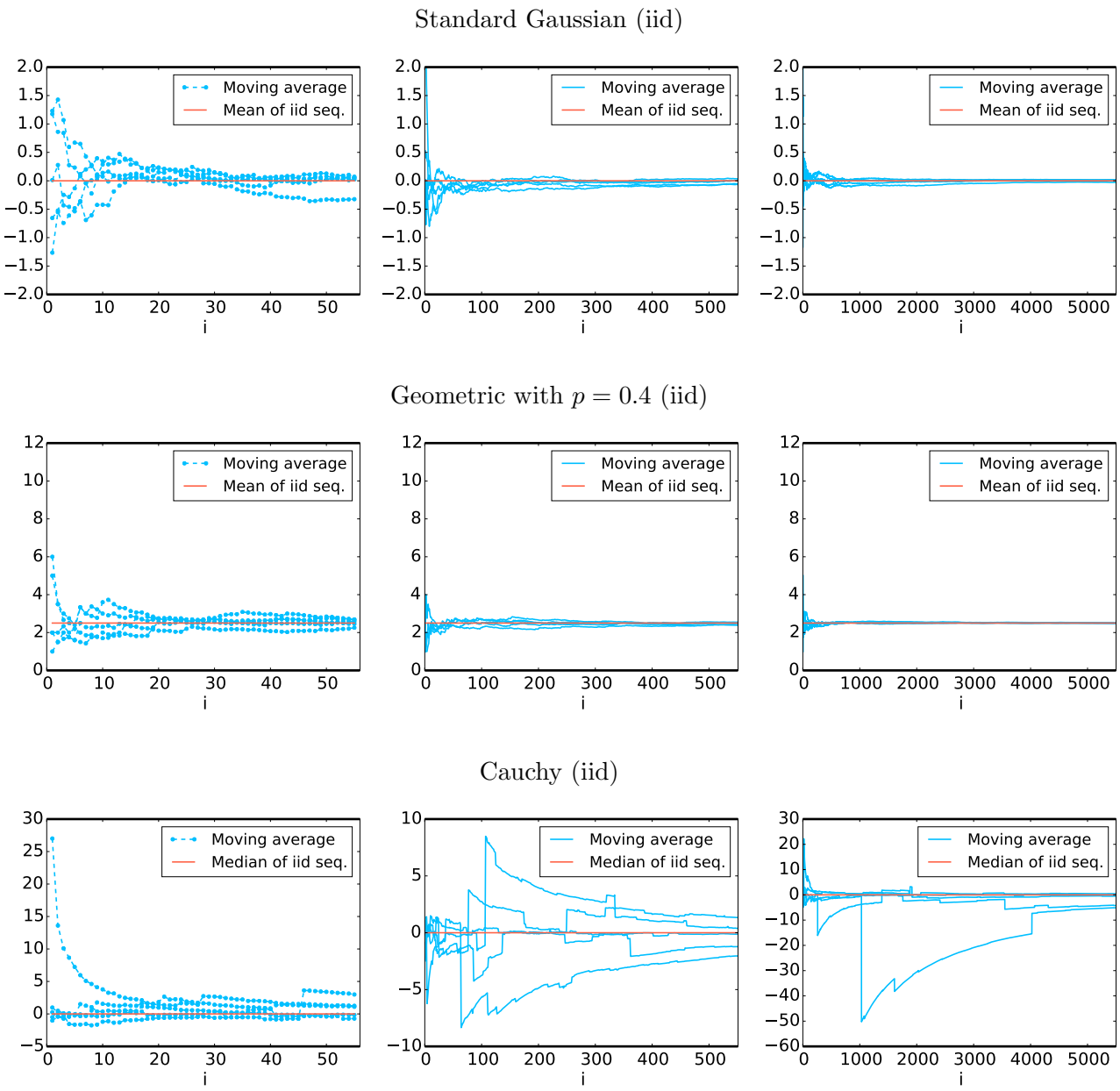


Figure 6.2: 独立同分布标准高斯序列的移动平均的一个实现（上），参数为 $p = 0.4$ 的独立同分布几何分布序列（中）以及独立同分布柯西分布序列（下）。

方差是有限的)。在本节中，我们描述了当i增加时，平均$\widetilde{A}(i)$的*distribution*。事实证明，$\widetilde{A}$在分布上收敛为高斯随机变量，这在统计学中非常有用，正如我们稍后将看到的。

这个结果，被称为中心极限定理，证明了使用高斯分布来建模由许多不同独立因素导致的数据的合理性。例如，某一人群的身高或体重分布通常呈高斯形状——如图 2.13 所示——因为一个人的身高和体重依赖于许多大致独立的不同因素。在许多信号处理应用中，噪声也因同样的原因被很好地建模为高斯分布。

Theorem 6.3.1 (中心极限定理)**.** *Let $\widetilde{X}$ be an iid discrete random process with mean $\mu_{\widetilde{X}} := \mu$ such that the variance of $\widetilde{X}(i)$ σ^2 is bounded. The random process $\sqrt{n}(\widetilde{A}-\mu)$, which corresponds to the centered and scaled moving average of $\widetilde{X}$, converges in distribution to a Gaussian random variable with mean 0 and variance σ^2.*

Proof. 这一卓越结果的证明超出了这些笔记的范围。它可以在任何高级概率论教材中找到。然而，我们仍然希望提供一些直观解释，说明为什么该定理成立。定理 3.5.2 表明，两个相互独立的随机变量之和的概率密度函数（pdf）等于它们各自 pdf 的卷积。对于离散随机变量也是如此：只要随机变量相互独立，其和的概率质量函数（pmf）就等于这些 pmf 的卷积。

如果 iid 序列中的每个元素的概率密度函数为 f，那么前 i 个元素之和的概率密度函数可以通过将 f 与其自身进行 i 次卷积得到

$$f_{\sum_{j=1}^{1}\widetilde{X}(j)}(x) = (f * f * \cdots * f)(x). \tag{6.26}$$

如果序列具有离散状态，并且每个条目的 pmf 为 p，那么前 i 个元素的和的 pmf 可以通过将 p 与自身卷积 i 次得到。

$$p_{\sum_{j=1}^{1}\widetilde{X}(j)}(x) = (p * p * \cdots * p)(x). \tag{6.27}$$

通过i标准化仅会导致卷积结果的缩放，因此移动平均$\widetilde{A}$的pmf或pdf是固定函数重复卷积的结果。这些卷积具有平滑效果，最终将pmf/pdf转化为高斯分布！我们在图6.3中通过数值方式展示了这一点，针对两种非常不同的分布：均匀分布和非常不规则的分布。两者在仅进行3或4次卷积后都收敛到类似高斯的形状。中心极限定理使这一点得到了精确的描述，确立了pmf或pdf的形状渐近地变为高斯分布。

□

在统计学中，中心极限定理常常被用来证明将平均值视为具有高斯分布是合理的。其思想是，对于足够大的 n $\sqrt{n}(\widetilde{A}-\mu)$，它大致呈高斯分布，均值为 0，方差为 σ^2，这意味着$\widetilde{A}$ 大致呈高斯分布，均值为 μ，方差为 σ^2/n。重要的是要记住，我们已 *not* 严格地证明了这一点。收敛的速率将取决于独立同分布序列的各项分布。

在实践中，收敛通常非常快。图6.4展示了一个指数分布和一个几何分布的独立同分布序列的移动平均的经验分布。在这两种情况下，即使只对100个样本取平均，由中心极限定理得到的近似也非常准确。该

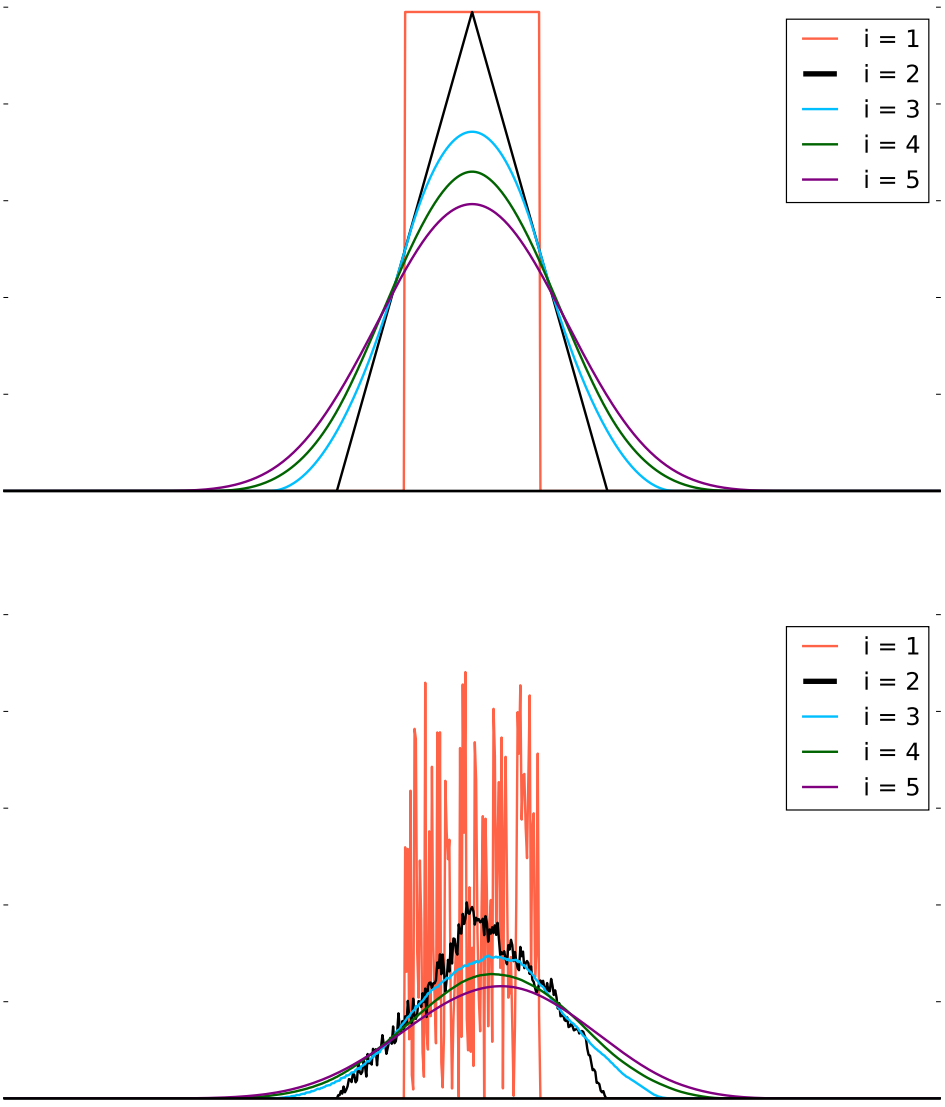


Figure 6.3: 将两种不同的分布分别与自身多次卷积的结果。其形状很快变得类似高斯分布。

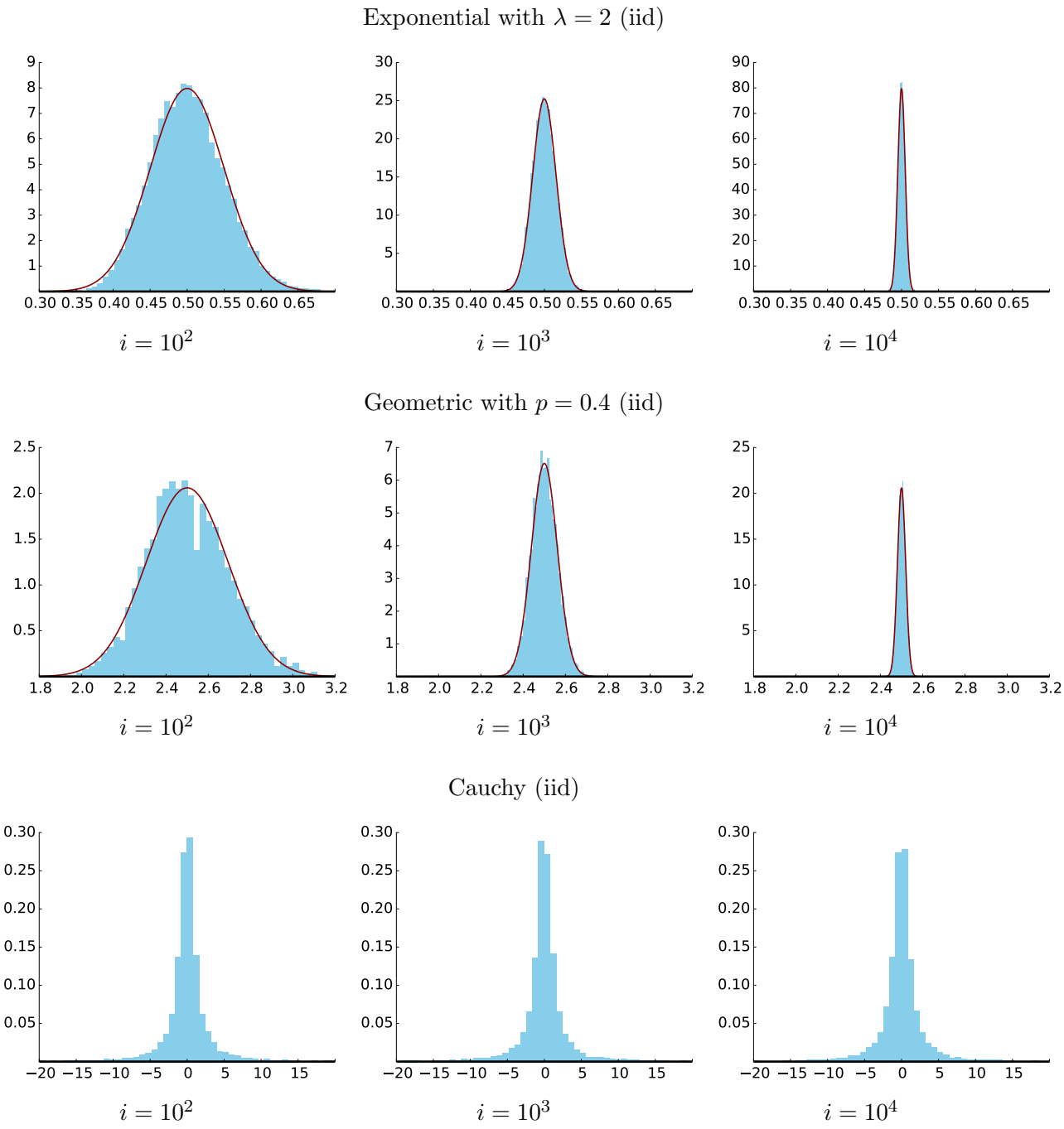


Figure 6.4: 独立同分布标准高斯序列的移动平均的经验分布（顶部），参数为 $p = 0.4$ 的独立同分布几何序列的经验分布（中间）和独立同分布柯西序列的经验分布（底部）。所有情况下，经验分布是通过 10^4 个样本计算得到的。对于前两行，中央极限定理提供的估计以红色绘制。

图中还显示，对于一个柯西 iid 序列，移动平均的分布并不会变成高斯分布；这并不与中心极限定理相矛盾，因为该分布并没有良好定义的均值。为结束本节，我们利用中心极限定理推导二项分布的一个有用近似。

Example 6.3.2 (二项分布的高斯近似)**.** 设 X 服从参数为 n 和 p 的二项分布，其中 n 很大。计算 X 落在某一区间内的概率需要对该区间内所有取值的 pmf 进行求和。或者，我们可以利用这样一个事实得到一个快速近似：当 n 很大时，二项随机变量的分布近似为高斯分布。事实上，我们可以将 X 写成 n 个参数为 p 的相互独立的伯努利随机变量之和，

$$X = \sum_{i=1}^{n} B_i. \tag{6.28}$$

B_i的均值是p，其方差是$p\left(1-p\right)$。根据中心极限定理，$\frac{1}{n}X$大致服从高斯分布，均值为p，方差为$p\left(1-p\right)/n$。等效地，根据引理2.5.1，X大致服从高斯分布，均值为np，方差为$np\left(1-p\right)$。

假设一名篮球运动员每次出手命中的概率为 $p = 0.4$。如果我们假设每次出手相互独立，那么她在 1000 次出手中命中超过 420 次的概率是多少？我们可以将命中次数建模为一个二项 X，其参数为 1000 和 0.4。精确答案是

$$\begin{aligned} \mathrm{P}\left(X \geq 420\right) &= \sum_{x=420}^{1000} p_X\left(x\right) && (6.29)\\ &= \sum_{x=420}^{1000} \binom{1000}{x} 0.4^x 0.6^{(n-x)} && (6.30)\\ &= 10.4\,10^{-2}. && (6.31) \end{aligned}$$

如果我们应用高斯近似，根据引理 2.5.1，X 大于 420 等同于标准高斯 U 大于 $\frac{420-\mu}{\sigma}$，其中 μ 和 σ 分别是 X 的均值和标准差，分别等于 $np = 400$ 和 $\sqrt{np(1-p)} = 15.5$。

$$\begin{aligned} \mathrm{P}\left(X \geq 420\right) &\approx \mathrm{P}\left(\sqrt{np\left(1-p\right)}U + np \geq 420\right) && (6.32)\\ &= \mathrm{P}\left(U \geq 1.29\right) && (6.33)\\ &= 1 - \Phi\left(1.29\right) && (6.34)\\ &= 9.85\,10^{-2}. && (6.35) \end{aligned}$$

△

6.4 Monte Carlo simulation

模拟是概率论和统计学中的一种强大工具。概率模型往往过于复杂，我们无法像在作业题中那样推导出感兴趣量的分布或期望的闭式解。

作为一个例子，假设你建立了一个概率模型来确定玩纸牌接龙游戏的胜率。如果牌被很好地洗牌，那么这个概率等于

$$\mathrm{P}\left(\text{Win}\right) = \frac{\text{Number of permutations that lead to a win}}{\text{Total number}}. \tag{6.36}$$

问题在于，要刻画哪些排列会导致获胜，在不实际把游戏完整地玩一遍并观察结果的情况下是非常困难的。对每一种可能的排列都这样做在计算上是不可行的，因为它们共有 $52! \approx 8\ 10^{67}$ 种。然而，有一种简单的方法可以近似我们感兴趣的概率：模拟大量的游戏，并记录其中有多少比例以获胜告终。正是纸牌接龙游戏启发了斯坦尼斯瓦夫·乌拉姆在20世纪40年代核武器研究的背景下提出基于模拟的方法，即所谓的蒙特卡罗方法（这是一个代号，灵感来自摩纳哥的蒙特卡罗赌场）：

The first thoughts and attempts I made to practice (the Monte Carlo Method) were suggested by a question which occurred to me in 1946 as I was convalescing from an illness and playing solitaires. The question was what are the chances that a Canfield solitaire laid out with 52 cards will come out successfully? After spending a lot of time trying to estimate them by pure combinatorial calculations, I wondered whether a more practical method than "abstract thinking" might not be to lay it out say one hundred times and simply observe and count the number of successful plays.

This was already possible to envisage with the beginning of the new era of fast computers, and I immediately thought of problems of neutron diffusion and other questions of mathematical physics, and more generally how to change processes described by certain differential equations into an equivalent form interpretable as a succession of random operations. Later, I described the idea to John von Neumann, and we began to plan actual calculations.[1]

蒙特卡罗方法通过模拟来估计那些难以精确计算的量。在本节中，我们考虑近似估计事件 $\mathcal{E}$ 的概率的问题，如 *game of solitaire* 示例所示。

Algorithm 6.4.1 (蒙特卡洛近似)**.** *To approximate the probability of an event $\mathcal{E}$, we:*

1. *Generate n independent samples from the indicator function $1_{\mathcal{E}}$ associated to the event: I_1, I_2, ..., I_n.*
2. *Compute the average of the n samples*

$$\widetilde{A}\left(n\right) := \frac{1}{n}\sum_{i=1}^{n} I_i \tag{6.37}$$

which is the estimate for the probability of $\mathcal{E}$

兴趣的概率可以解释为与事件相关的指示函数 $1_{\mathcal{E}}$ 的期望。

$$\mathrm{E}\left(1_{\mathcal{E}}\right) = \mathrm{P}\left(\mathcal{E}\right). \tag{6.38}$$

根据大数定律，随着 $n \to \infty$，估计 $\widetilde{A}$ 收敛到真实概率。下面的例子展示了这种简单技术的威力。

[1] http://en.wikipedia.org/wiki/Monte_Carlo_method#History

Game outcomes			Rank			Probability
1-2	1-3	2-3	R_1	R_2	R_3	
1	1	2	1	2	3	1/6
1	1	3	1	3	2	1/6
1	3	2	1	1	1	1/12
1	3	3	2	3	1	1/12
2	1	2	2	1	3	1/6
2	1	3	1	1	1	1/6
2	3	2	3	1	2	1/12
2	3	3	3	2	1	1/12

概率质量函数

	R_1	R_2	R_3
1	7/12	1/2	5/12
2	1/4	1/4	1/4
3	1/6	1/4	1/3

Table 6.1: 左侧的表格展示了由三支球队组成的联赛中的所有可能结果（$m = 3$），每支球队的最终名次以及对应的概率。右侧的表格展示了各支球队名次的概率质量函数（pmf）。

Example 6.4.2 (篮球联赛)**.** 在一场校内篮球联赛中，m 队伍每个赛季互相对抗一次。队伍根据过去的成绩进行排名：队伍 1 是最强的，队伍 m 是最弱的。我们对队伍 i 战胜队伍 j 的概率进行建模，对于 $1 \leq i < j \leq m$，表示为

$$\mathrm{P}\left(\text{team } j \text{ beats team } i\right) := \frac{1}{j-i+1}. \tag{6.39}$$

最强的队伍以 1/2 的概率击败第二名、以 2/3 的概率击败第三名；第二名以 1/2 的概率击败第三名、以 2/3 的概率击败第四名、以 3/4 的概率击败第五名，依此类推。我们假设不同比赛的结果相互独立。

在赛季结束时，当每支球队都与其他每支球队进行过比赛之后，球队将根据其获胜场次数进行排名。如果有多支球队的获胜场次数相同，那么它们共享相同的名次。例如，如果有两支球队的胜场数最多，那么它们都排名第1，下一支球队则排名第3。目标是计算联盟中每支球队最终名次的分布，我们将其建模为随机变量 R_1、R_2、. . .、R_m。通过应用全概率定律，我们拥有计算这些随机变量联合 pmf 所需的全部信息。如表6.1中针对 $m = 3$ 所示，我们所需要做的只是枚举所有可能的比赛结果，并将那些导致某一特定名次的结果的概率加总。

不幸的是，可能结果的数量随着 m 急剧增长。比赛场次数等于 $m\left(m-1\right)/2$，因此可能的结果共有 $2^{m(m-1)/2}$ 种。当只有 10 支球队时，这个数量就超过 10^{13}。因此，对于规模不小的联赛，计算最终排名的精确分布在计算上非常耗费资源。幸运的是，算法 6.4.1 提供了一种更为可行的替代方案：我们可以通过将每场比赛模拟为参数由公式 (6.39) 给定的伯努利随机变量，来抽样大量的赛季 n，并用每支球队最终落在各个位置的频率来近似 pmf。模拟一个完整赛季只需要抽样 $m\left(m-1\right)/2$ 场比赛，这可以非常快速地完成。

表6.2说明了$m = 3$的蒙特卡洛方法。如果我们仅使用$n = 10$个模拟季节，近似值会相当粗略，但当$n = 2,000$时，精确度非常高。图6.5

Game outcomes			Rank		
1-2	1-3	2-3	R_1	R_2	R_3
1	3	2	1	1	1
1	1	3	1	3	2
2	1	2	2	1	3
2	3	2	3	1	2
2	1	3	1	1	1
1	1	2	1	2	3
2	1	3	1	1	1
2	3	2	3	1	2
1	1	2	1	2	3
2	3	2	3	1	2

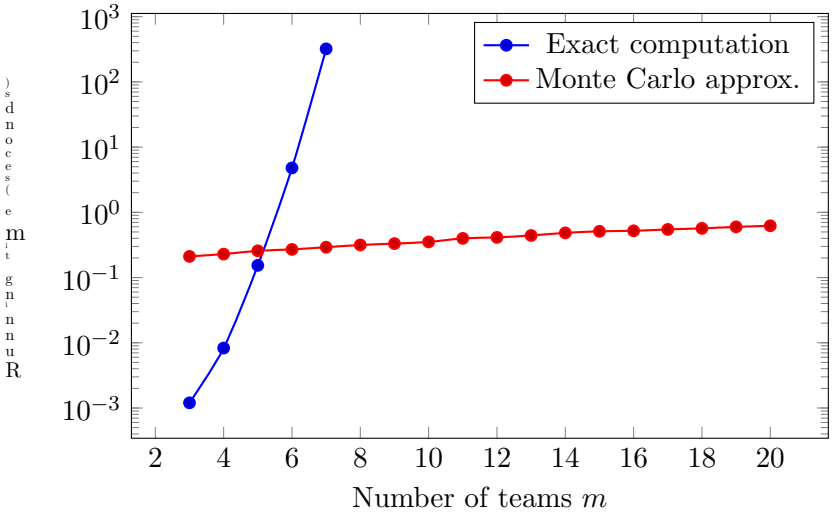
估计的 pmf ($n = 10$)

	R_1	R_2	R_3
1	0.6 (0.583)	0.7 (0.5)	0.3 (0.417)
2	0.1 (0.25)	0.2 (0.25)	0.4 (0.25)
3	0.3 (0.167)	0.1 (0.25)	0.3 (0.333)

估计的pmf ($n = 2,000$)

	R_1	R_2	R_3
1	0.582 (0.583)	0.496 (0.5)	0.417 (0.417)
2	0.248 (0.25)	0.261 (0.25)	0.244 (0.25)
3	0.171 (0.167)	0.245 (0.25)	0.339 (0.333)

Table 6.2: 左侧表格展示了一个由三支球队组成的联赛（$m = 3$）的 10 次模拟结果及其对应的排名。右侧表格展示了通过蒙特卡洛模拟得到的估计 pmf：上方基于左侧的模拟结果，下方基于 2,000 次模拟结果。括号中给出了精确值以供比较。



蒙特卡罗误差

m	Average error
3	$9.28\,10^{-3}$
4	$12.7\,10^{-3}$
5	$7.95\,10^{-3}$
6	$7.12\,10^{-3}$
7	$7.12\,10^{-3}$

Figure 6.5: 左侧的图展示了在例 6.4.2 中获得最终排名精确 pmf 所需的时间，以及使用 2,000 次模拟联赛结果进行蒙特卡洛近似所需的时间。右侧的表格展示了蒙特卡洛近似中每个条目的平均误差。

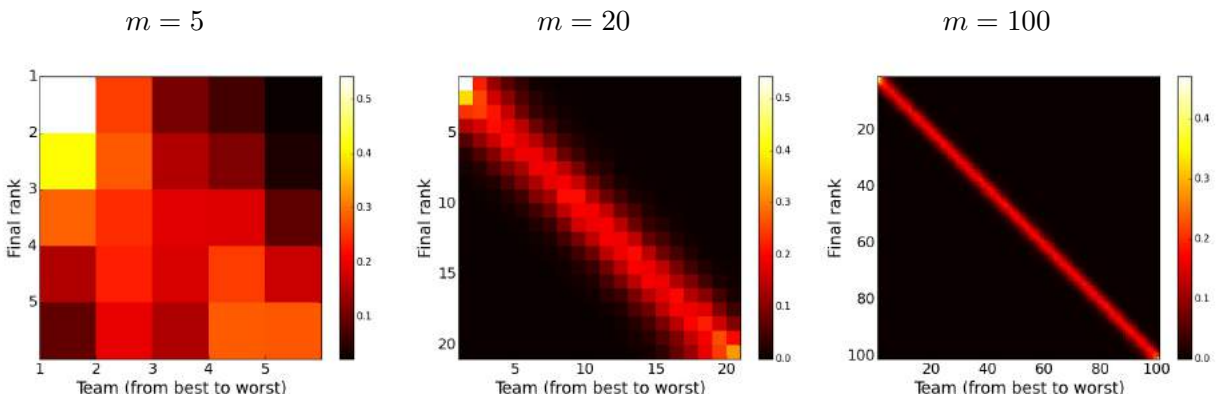


Figure 6.6: 使用2,000次模拟联赛结果得到的示例6.4.2中最终排名的近似pmf。

显示了在不同球队数量下，计算精确 pmf 以及使用蒙特卡洛方法进行近似所需的运行时间。当球队数量非常少时，精确计算非常快，但正如预期的那样，运行时间会随着 m 呈指数增长，因此对于 7 支球队，计算已经需要 5 分半钟。相比之下，蒙特卡洛近似要快得多。对于 $m = 20$，只需半秒钟。图 6.6 显示了 5、20 和 100 支球队时最终名次的近似 pmf。较高名次具有更高的概率，因为当两支球队打平时，会被授予较高的名次。

△

Chapter 7

Markov Chains

马尔可夫性质由任何随机过程满足，其中 *the future is conditionally independent from the past given the present*。

Definition 7.0.1 (马尔可夫性质)**.** *A random process satisfies the Markov property if* $\widetilde{X}(t_{i+1})$ *is conditionally independent of* $\widetilde{X}(t_1), \ldots, \widetilde{X}(t_{i-1})$ *given* $\widetilde{X}(t_i)$ *for any* $t_1 < t_2 < \ldots < t_i < t_{i+1}$*. If the state space of the random process is discrete, then for any* x_1*,* x_2*, . . . ,* x_{i+1}

$$p_{\widetilde{X}(t_{n+1})|\widetilde{X}(t_1),\widetilde{X}(t_2),\ldots,\widetilde{X}(t_i)}(x_{n+1}|x_1, x_2, \ldots, x_n) = p_{\widetilde{X}(t_{i+1})|\widetilde{X}(t_i)}(x_{i+1}|x_i). \tag{7.1}$$

If the state space of the random process is continuous (and the distribution has a joint pdf),

$$f_{\widetilde{X}(t_{i+1})|\widetilde{X}(t_1),\widetilde{X}(t_2),\ldots,\widetilde{X}(t_i)}(x_{i+1}|x_1, x_2, \ldots, x_i) = f_{\widetilde{X}(t_{i+1})|\widetilde{X}(t_i)}(x_{i+1}|x_i). \tag{7.2}$$

图7.1展示了与马尔可夫性质所隐含的依赖假设相对应的有向图模型。任何独立同分布序列都满足马尔可夫性质，因为所有条件概率质量函数或概率密度函数都等同于边缘分布（在这种情况下，图7.1中的有向无环图将不存在任何边）。随机游走同样满足该性质，因为一旦我们固定随机游走在某一时刻i所处的位置，在其接下来的步伐中，之前i所走过的路径不再产生任何影响。

Lemma 7.0.2. *The random walk satisfies the Markov property.*

Proof. 令$\widetilde{X}$表示在第 5.6 节中定义的随机游走。在给定$\widetilde{X}(j) = x_i$且对 $j \leq i$ 的条件下，$\widetilde{X}(i+1)$ 等于 $x_i + \widetilde{S}(i+1)$。这不依赖于 $x_1, \ldots, x_{i-1}$，从而推出 (7.1)。 □

7.1 Time-homogeneous discrete-time Markov chains

马尔可夫链是一种满足马尔可夫性质的随机过程。在这里，我们考虑具有**finite state space**的**discrete-time**马尔可夫链，这意味着该过程在任何给定时间点只能取有限数量的值。为了指定一个马尔可夫链，我们只需要定义随机过程在其起始点（我们假设在$i = 0$时）处的pmf及其转移概率。这源于马尔可夫性质，因为对于任何

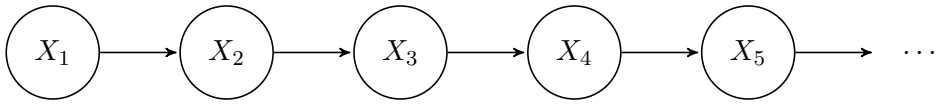


Figure 7.1: 描述马尔可夫性质所蕴含的依赖假设的有向图模型。

$n \geq 0$

$$p_{\widetilde{X}(0),\widetilde{X}(1),\ldots,\widetilde{X}(n)}\left(x_0, x_1, \ldots, x_n\right) := \prod_{i=0}^{n} p_{\widetilde{X}(i)|\widetilde{X}(0),\ldots,\widetilde{X}(i-1)}\left(x_i | x_0, \ldots, x_{i-1}\right) \tag{7.3}$$

$$= \prod_{i=0}^{n} p_{\widetilde{X}(i)|\widetilde{X}(i-1)}\left(x_i | x_{i-1}\right). \tag{7.4}$$

如果这些转移概率在每个时间步长上都是相同的（即它们是常数，并且不依赖于 i），则该马尔可夫链被称为 **time homogeneous**。在这种情况下，我们可以将每个可能转移的概率存储在一个 $s \times s$ 矩阵 $T_{\widetilde{X}}$ 中，其中 s 是状态的数量。

$$\left(T_{\widetilde{X}}\right)_{jk} := p_{\widetilde{X}(i+1)|\widetilde{X}(i)}\left(x_j | x_k\right). \tag{7.5}$$

在本章中，我们集中讨论时间齐次有限状态马尔可夫链。这些链的转移概率可以通过状态图来可视化，状态图显示了每个状态以及每个可能转移的概率。以下图7.2为例。状态图不应与表示模型依赖结构的有向无环图（DAG）混淆，后者如图7.1所示。在状态图中，每个节点对应一个状态，边表示状态之间的转移概率，而DAG仅表示随机过程随时间的依赖结构，通常对于所有马尔可夫链都是相同的。

为简化记号，我们定义一个称为 **state vector** 的 s 维向量 $\vec{p}_{\widetilde{X}(i)}$，其中包含马尔可夫链在每个时间 i 的边缘概率质量函数，

$$\vec{p}_{\widetilde{X}(i)} := \begin{bmatrix} p_{\widetilde{X}(i)}\left(x_1\right) \\ p_{\widetilde{X}(i)}\left(x_2\right) \\ \cdots \\ p_{\widetilde{X}(i)}\left(x_s\right) \end{bmatrix}. \tag{7.6}$$

每个状态向量中的条目包含马尔可夫链在时间 i 处于该特定状态的概率。它是 *not* 马尔可夫链的值，马尔可夫链是一个随机变量。

初始状态空间 $\vec{p}_{\widetilde{X}(0)}$ 和转移矩阵 $T_{\widetilde{X}}$ 足以完全刻画一个时间齐次的有限状态马尔可夫链。事实上，通过应用 (7.4) 并对我们不感兴趣的任何时间进行边缘化，我们可以计算该链在任意 n 个时间点 i_1、i_2、$\ldots$、i_n 的联合分布，对于任意 $n \geq 1$ 从 $\vec{p}_{\widetilde{X}(0)}$ 和 $T_{\widetilde{X}}$。我们在下面的例子中说明这一点。

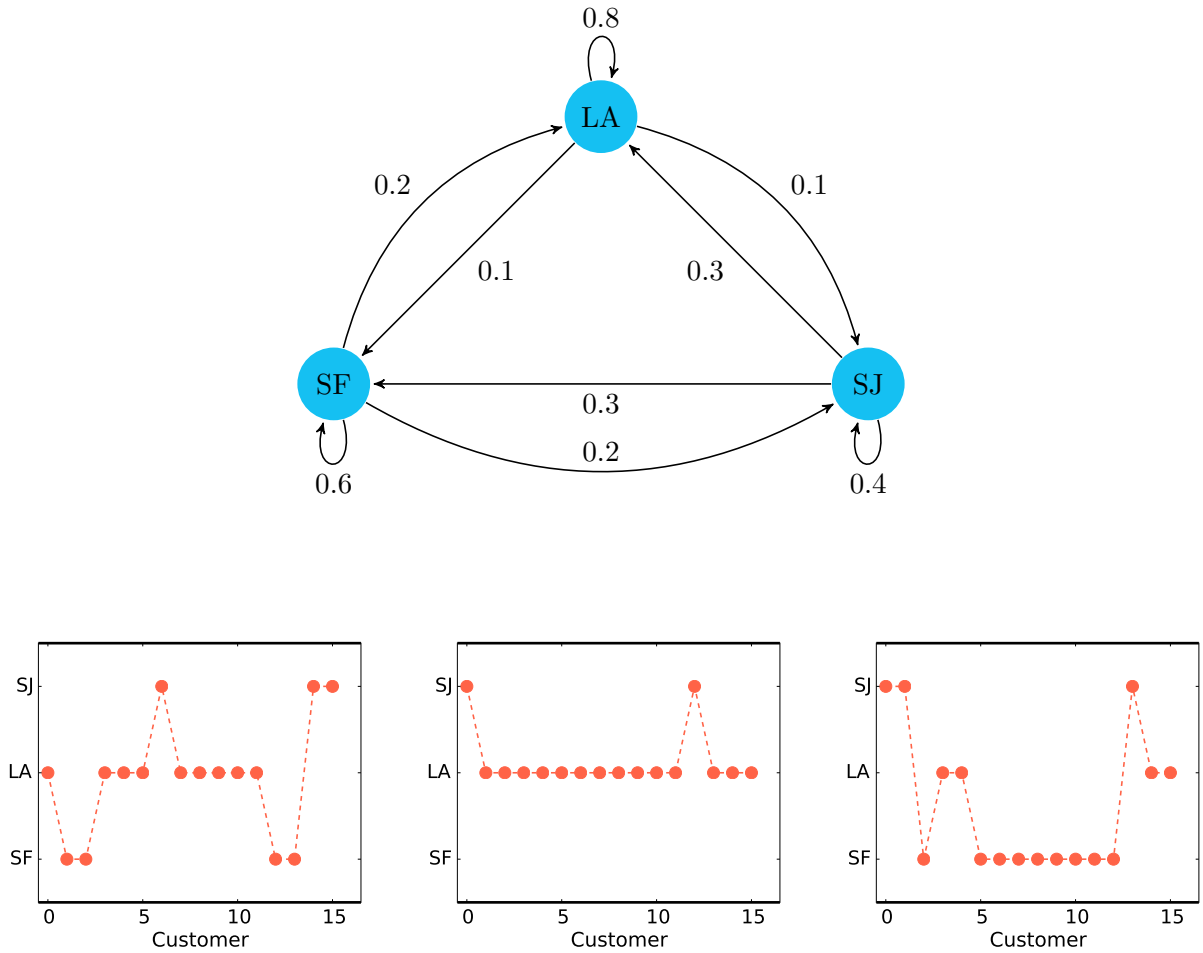


Figure 7.2: 例（7.1.1）中所描述的马尔可夫链的状态图（上）。每个箭头表示两个状态之间转移的概率。下方展示了该马尔可夫链的三条样本路径。

Example 7.1.1 (汽车租赁). 一家汽车租赁公司雇佣您来建模其汽车的位置。该公司在洛杉矶、旧金山和圣何塞运营。客户通常会在一个城市取车，并在另一个城市还车。对于公司来说，能够计算汽车最终到达特定城市的可能性将非常有用。您决定将汽车的位置建模为一个马尔可夫链，其中每个时间步对应一位新客户取车。公司将新车均匀分配到三个城市之间。根据过去的数据，过渡概率为

$$\begin{array}{c} \begin{array}{ccc} \text{San Francisco} & \text{Los Angeles} & \text{San Jose} \end{array} \\ \begin{pmatrix} 0.6 & 0.1 & 0.3 \\ 0.2 & 0.8 & 0.3 \\ 0.2 & 0.1 & 0.4 \end{pmatrix} \end{array} \begin{array}{l} \text{San Francisco} \\ \text{Los Angeles} \\ \text{San Jose} \end{array}$$

需要明确的是，客户将汽车从旧金山开到洛杉矶的概率是0.2，该

汽车停留在旧金山的概率为0.6，依此类推。
初始状态向量和马尔可夫链的转移矩阵为

$$\vec{p}_{\widetilde{X}(0)} := \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}, \qquad T_{\widetilde{X}} := \begin{bmatrix} 0.6 & 0.1 & 0.3 \\ 0.2 & 0.8 & 0.3 \\ 0.2 & 0.1 & 0.4 \end{bmatrix}. \tag{7.7}$$

状态 1 分配给 *San Francisco*，状态 2 分配给 *Los Angeles*，状态 3 分配给 *San Jose*。图 7.2 显示了马尔可夫链的状态图。图 7.2 显示了马尔可夫链的若干实现。

公司想要找出汽车在旧金山启动，但在第二个顾客之后立即到达圣何塞的概率。这可以表示为

$$\begin{aligned} p_{\widetilde{X}(0),\widetilde{X}(2)}(1,3) &= \sum_{i=1}^{3} p_{\widetilde{X}(0),\widetilde{X}(1),\widetilde{X}(2)}(1,i,3) && (7.8)\\ &= \sum_{i=1}^{3} p_{\widetilde{X}(0)}(1)\, p_{\widetilde{X}(1)|\widetilde{X}(0)}(i|1)\, p_{\widetilde{X}(2)|\widetilde{X}(1)}(3|i) && (7.9)\\ &= \left(\vec{p}_{\widetilde{X}(0)}\right)_1 \sum_{i=1}^{3} \left(T_{\widetilde{X}}\right)_{i1} \left(T_{\widetilde{X}}\right)_{3i} && (7.10)\\ &= \frac{0.6 \cdot 0.2 + 0.2 \cdot 0.1 + 0.2 \cdot 0.4}{3} \approx 7.33\,10^{-2}. && (7.11) \end{aligned}$$

概率是 7.33%。

△

以下引理给出了在时间i $\vec{p}_{\widetilde{X}(i)}$的状态向量关于$T_{\widetilde{X}}$和前一状态向量的一个简单表达式。

Lemma 7.1.2 (状态向量和转移矩阵)**.** *For a Markov chain* $\widetilde{X}$ *with transition matrix* $T_{\widetilde{X}}$

$$\vec{p}_{\widetilde{X}(i)} = T_{\widetilde{X}}\, \vec{p}_{\widetilde{X}(i-1)}. \tag{7.12}$$

If the Markov chain starts at time 0 *then*

$$\vec{p}_{\widetilde{X}(i)} = T_{\widetilde{X}}^{i}\, \vec{p}_{\widetilde{X}(0)}, \tag{7.13}$$

where $T_{\widetilde{X}}^{i}$ *denotes multiplying* i *times by matrix* $T_{\widetilde{X}}$*.*

Proof. 该证明直接由定义得出,

$$\vec{p}_{\widetilde{X}(i)} := \begin{bmatrix} p_{\widetilde{X}(i)}(x_1) \\ p_{\widetilde{X}(i)}(x_2) \\ \cdots \\ p_{\widetilde{X}(i)}(x_s) \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{s} p_{\widetilde{X}(i-1)}(x_j)\, p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_1|x_j) \\ \sum_{j=1}^{s} p_{\widetilde{X}(i-1)}(x_j)\, p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_2|x_j) \\ \cdots \\ \sum_{j=1}^{s} p_{\widetilde{X}(i-1)}(x_j)\, p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_s|x_j) \end{bmatrix} \tag{7.14}$$

$$= \begin{bmatrix} p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_1|x_1) & p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_1|x_2) & \cdots & p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_1|x_s) \\ p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_2|x_1) & p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_2|x_2) & \cdots & p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_2|x_s) \\ & & \cdots & \\ p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_s|x_1) & p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_s|x_2) & \cdots & p_{\widetilde{X}(i)|\widetilde{X}(i-1)}(x_s|x_s) \end{bmatrix} \begin{bmatrix} p_{\widetilde{X}(i-1)}(x_1) \\ p_{\widetilde{X}(i-1)}(x_2) \\ \cdots \\ p_{\widetilde{X}(i-1)}(x_s) \end{bmatrix}$$

$$= T_{\widetilde{X}}\, \vec{p}_{\widetilde{X}(i-1)} \tag{7.15}$$

式(7.13)是通过将(7.12)应用 i 次并考虑马尔可夫性质而得到的。

□

Example 7.1.3 (汽车租赁（续）)**.** 公司希望估计在第5位顾客使用完一辆车后，地点的分布。应用引理 7.1.2，我们得到

$$\vec{p}_{\widetilde{X}(5)} = T_{\widetilde{X}}^{5}\, \vec{p}_{\widetilde{X}(0)} \tag{7.16}$$

$$= \begin{bmatrix} 0.281 \\ 0.534 \\ 0.185 \end{bmatrix}. \tag{7.17}$$

模型估计 es 表示在 5 位客户之后，超过一半的汽车 在洛杉矶。

△

7.2 Recurrence

马尔可夫链的状态可以根据马尔可夫链是否必然总会返回这些状态，或者是否可能最终停止访问这些状态来进行分类。

Definition 7.2.1 (常返状态与暂态状态)**.** *Let $\widetilde{X}$ be a time-homogeneous finite-state Markov chain. We consider a particular state x. If*

$$\mathrm{P}\left(\widetilde{X}(j) = s \text{ for some } j > i \mid \widetilde{X}(i) = s\right) = 1 \tag{7.18}$$

*then the state is **recurrent**. In words, given that the Markov chain is at x, the probability that it returns to x is one. In contrast, if*

$$\mathrm{P}\left(\widetilde{X}(j) \neq s \text{ for all } j > i \mid \widetilde{X}(i) = s\right) > 0 \tag{7.19}$$

*the state is **transient**. Given that the Markov chain is at x, there is nonzero probability that it will never return.*

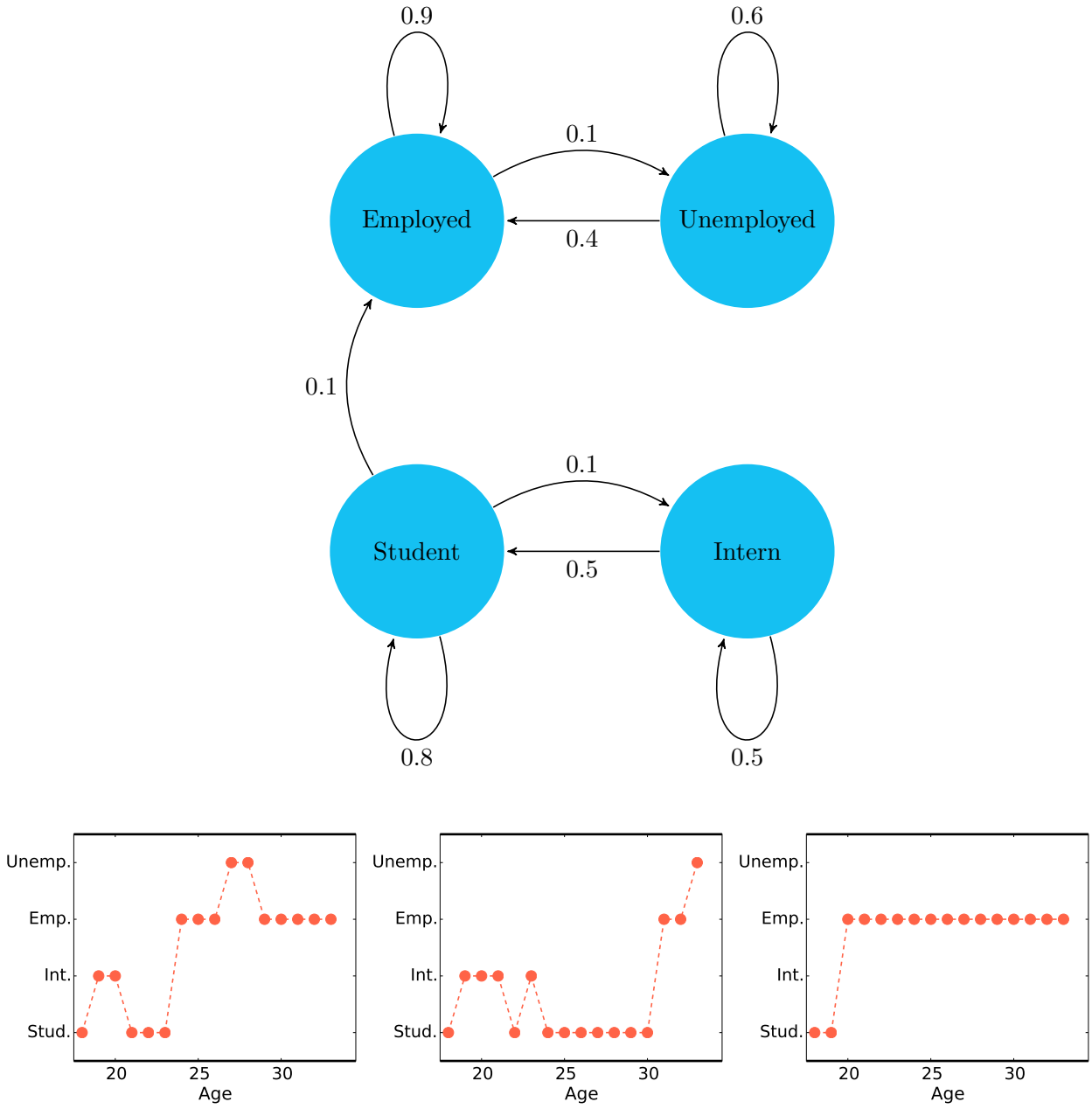


Figure 7.3: 示例（7.2.2）中描述的马尔可夫链的状态图（上）。下面我们展示该马尔可夫链的三种实现。

以下示例说明了常返状态与暂态状态之间的区别。

Example 7.2.2 (就业动态)**.** 一位研究人员有兴趣使用马尔可夫链来建模年轻人的就业动态。

她确定，在18岁时，一个人要么是学生，概率为0.9，要么是实习生，概率为0.1。随后，她估计了以下转移概率：

$$\begin{array}{cccc} \text{Student} & \text{Intern} & \text{Employed} & \text{Unemployed} & \\ \left(\begin{array}{cccc} 0.8 & 0.5 & 0 & 0 \\ 0.1 & 0.5 & 0 & 0 \\ 0.1 & 0 & 0.9 & 0.4 \\ 0 & 0 & 0.1 & 0.6 \end{array}\right) & \begin{array}{c} \text{Student} \\ \text{Intern} \\ \text{Employed} \\ \text{Unemployed} \end{array} \end{array}$$

马尔可夫假设显然并非完全精确，学习时间更长的人可能不太可能继续保持学生身份，但这类马尔可夫模型更容易拟合（我们只需要估计转移概率），并且往往能产生有用的洞见。

初始状态向量和马尔可夫链的转移矩阵为

$$\vec{p}_{\widetilde{X}(0)} := \begin{bmatrix} 0.9 \\ 0.1 \\ 0 \\ 0 \end{bmatrix}, \qquad T_{\widetilde{X}} := \begin{bmatrix} 0.8 & 0.5 & 0 & 0 \\ 0.1 & 0.5 & 0 & 0 \\ 0.1 & 0 & 0.9 & 0.4 \\ 0 & 0 & 0.1 & 0.6 \end{bmatrix}. \tag{7.20}$$

图7.3展示了马尔可夫链的状态图及其一些实现。

状态1（学生）和状态2（实习生）是暂态。注意，马尔可夫链在访问状态3（就业）之后返回到这些状态的概率为零，因此

$$\begin{aligned} \mathrm{P}\left(\widetilde{X}(j) \neq 1 \text{ for all } j > i \mid \widetilde{X}(i) = 1\right) &\geq \mathrm{P}\left(\widetilde{X}(i+1) = 3 \mid \widetilde{X}(i) = 1\right) && (7.21) \\ &= 0.1 > 0, && (7.22) \\ \mathrm{P}\left(\widetilde{X}(j) \neq 2 \text{ for all } j > i \mid \widetilde{X}(i) = 2\right) &\geq \mathrm{P}\left(\widetilde{X}(i+2) = 3 \mid \widetilde{X}(i) = 2\right) && (7.23) \\ &= 0.5 \cdot 0.1 > 0. && (7.24) \end{aligned}$$

相比之下，状态 3 和 4（失业）是重复出现的。我们证明了状态 3 的情况（状态 4 的论证完全相同）：

$$\begin{aligned} &\mathrm{P}\left(\widetilde{X}(j) \neq 3 \text{ for all } j > i \mid \widetilde{X}(i) = 3\right) && (7.25) \\ &= \mathrm{P}\left(\widetilde{X}(j) = 4 \text{ for all } j > i \mid \widetilde{X}(i) = 3\right) && (7.26) \\ &= \lim_{k\to\infty} \mathrm{P}\left(\widetilde{X}(i+1) = 4 \mid \widetilde{X}(i) = 3\right) \prod_{j=1}^{k} \mathrm{P}\left(\widetilde{X}(i+j+1) = 4 \mid \widetilde{X}(i+j) = 4\right) && (7.27) \\ &= \text{极限}_{k\to\infty}\ 1 \cdot 0.6^k = 0. && (7.28) \\ &= 0. && (7.29) \end{aligned}$$

△

在这个例子中，从状态 *employed* 或 *unemployed* 无法到达状态 *student* 和 *intern*。对于任何两个状态之间都存在可能的转移（即使不是直接的）的马尔可夫链，称为不可约的。

Definition 7.2.3 (不可约马尔可夫链)**.** *A time-homogeneous finite-state Markov chain is irreducible if for any state x, the probability of reaching every other state $y \neq x$ in a finite number of steps is nonzero, i.e. there exists $m \geq 0$ such that*

$$\mathrm{P}\left(\widetilde{X}\left(i+m\right)=y \mid \widetilde{X}\left(i\right)=x\right)>0. \tag{7.30}$$

可以很容易地检查出示例7.1.1中的马尔可夫链是不可约的，而示例7.2.2中的则不是。一个重要的结果是，在不可约的马尔可夫链中，所有状态都是重返的。

Theorem 7.2.4 (不可约马尔可夫链)**.** *All states in an irreducible Markov chain are recurrent.*

Proof. 在任何有限状态马尔可夫链中，至少必须有一个状态是常返的。如果所有状态都是暂态的，那么就存在一个非零概率使其永远离开所有状态，这是不可能的。不失一般性，假设状态x是常返的。现在我们给出一个证明思路，说明另一个任意状态y也必然是常返的。为简化记号，令

$$p_{x,x} := \mathrm{P}\left(\widetilde{X}\left(j\right)=x \text{ for some } j>i \mid \widetilde{X}\left(i\right)=x\right), \tag{7.31}$$

$$p_{x,y} := \mathrm{P}\left(\widetilde{X}\left(j\right)=y \text{ for some } j>i \mid \widetilde{X}\left(i\right)=x\right), \tag{7.32}$$

$$p_{y,x} := \mathrm{P}\left(\widetilde{X}\left(k\right)=x \text{ for some } j>i \mid \widetilde{X}\left(i\right)=y\right). \tag{7.33}$$

该链是不可约的，因此存在一个非零概率 $p_m > 0$，使得在至多 m 步内从 x 到达 y，对于某个 $m > 0$。因此，该链从 x 到 y 且再也不返回 x 的概率至少为 $p_m\left(1-p_{y,x}\right)$。然而，$x$ 是常返的，因此这个概率必须为零！由于 $p_m > 0$，这意味着 $p_{y,x} = 1$。

考虑以下事件：

1.$\widetilde{X}$ 从 y 到 x。2.$\widetilde{X}$ 在到达 x 之后的 m 步内不会返回到 y。3.$\widetilde{X}$ 最终在时间 $m' > m$ 再次到达 x。

该事件的概率等于 $p_{y,x}\left(1-p_m\right)p_{x,x} = 1-p_m$ (回忆 x 是常返的，因此 $p_{x,x} = 1$)。现在设想第 2 和第 3 步重复 k 次，即$\widetilde{X}$ 在 m 步内从 x 到 y 失败 k 次。该事件的概率是 $p_{y,x}\left(1-p_m\right)^k p_{x,x}^k = \left(1-p_m\right)^k$。取 $k \to \infty$，对于任意 m 这都等于零，因此$\widetilde{X}$ 最终不返回 x 的概率必定为零（这可以严格化，但细节超出了这些笔记的范围）。

□

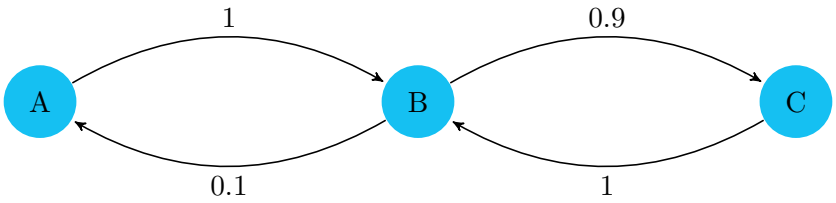


Figure 7.4: 马尔可夫链的状态图，其中状态具有周期二。

7.3 Periodicity

另一个重要的考虑因素是马尔可夫链是否始终以固定间隔访问给定的状态。如果是这样，那么该状态的周期大于一。

Definition 7.3.1 (状态的周期)**.** *Let $\widetilde{X}$ be a time-homogeneous finite-state Markov chain and x a state of the Markov chain. The period m of x is the largest integer such that it is only possible to return to x in a number of steps that is a multiple of m, i.e. we can only return in km steps with nonzero probability where k is a positive integer.*

图7.4展示了一个马尔可夫链，其中各个状态的周期等于2。非周期马尔可夫链不包含周期大于1的状态。

Definition 7.3.2 (非周期马尔可夫链)**.** *A time-homogeneous finite-state Markov chain $\widetilde{X}$ is aperiodic if all states have period equal to one.*

例 7.1.1 和 7.2.2 中的马尔可夫链都是非周期的。

7.4 Convergence

在本节中，我们研究一个有限状态、时间齐次的马尔可夫链$\widetilde{X}$ 在何种条件下在分布意义下收敛。如果一个马尔可夫链在分布意义下收敛，那么其状态向量 $\vec{p}_{\widetilde{X}(i)}$（其中包含$\widetilde{X}$ 的一阶概率质量函数）将收敛到一个固定向量 $\vec{p}_\infty$。

$$\vec{p}_\infty := \lim_{i\to\infty} \vec{p}_{\widetilde{X}(i)}. \tag{7.34}$$

在那种情况下，马尔可夫链处于每个状态的概率最终趋向于一个固定值（这并不意味着*not*马尔可夫链将停留在某个给定的状态！）。

由引理 7.1.2 可知，我们可以用马尔可夫链的初始状态向量和转移矩阵来表示式 (7.34)。

$$\vec{p}_\infty = \lim_{i\to\infty} T_{\widetilde{X}}^{\,i}\, \vec{p}_{\widetilde{X}(0)}. \tag{7.35}$$

对特定的 $T_{\widetilde{X}}$ 和 $\vec{p}_{\widetilde{X}(0)}$ 进行解析地计算该极限，乍看之下可能颇具挑战。然而，通常可以利用转移矩阵的特征分解（如果存在）来求得 $\vec{p}_\infty$。下面的例子对此进行了说明。

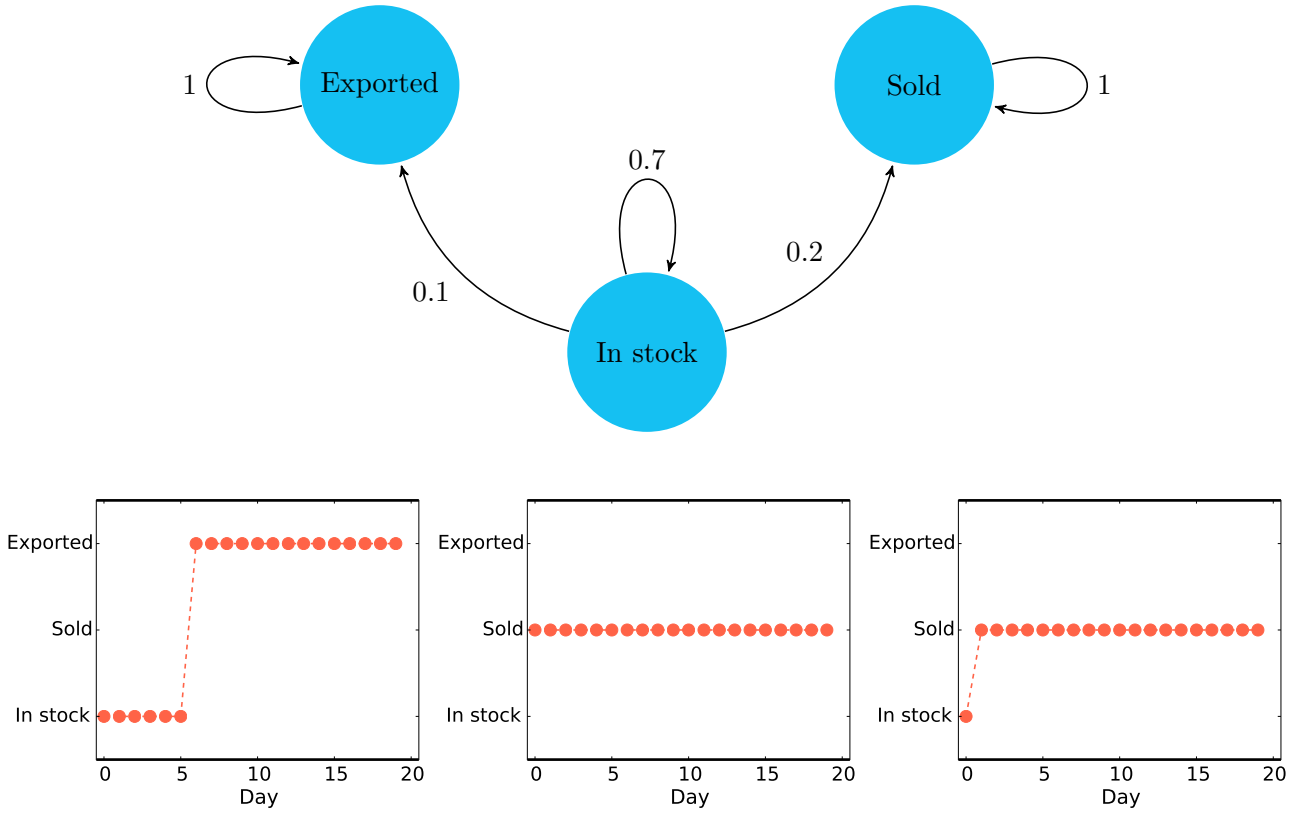


Figure 7.5: 例（7.4.1）中描述的马尔可夫链的状态图（上）。下方展示了该马尔可夫链的三个实现。

Example 7.4.1 (手机). 一家制造手机的公司希望对他们刚发布的新型号的销售情况进行建模。目前，90%的手机仍有库存，10%已经在本地售出，且没有任何手机被出口。根据过去的数据，该公司确定每天每部手机以0.2的概率被售出，并以0.1的概率被出口。初始状态向量和马尔可夫链的转移矩阵为

$$\vec{a} := \begin{bmatrix} 0.9 \\ 0.1 \\ 0 \end{bmatrix}, \qquad T_{\widetilde{X}} = \begin{bmatrix} 0.7 & 0 & 0 \\ 0.2 & 1 & 0 \\ 0.1 & 0 & 1 \end{bmatrix}. \tag{7.36}$$

我们使用 $\vec{a}$ 来表示 $\vec{p}_{\widetilde{X}(0)}$，因为稍后我们将考虑其他可能的初始状态向量。图 7.6 显示了状态图以及马尔可夫链的一些实现。

公司对新型号的命运感兴趣。特别是，它希望计算出最终将出口的手机所占的比例，以及将本地销售的手机所占的比例。

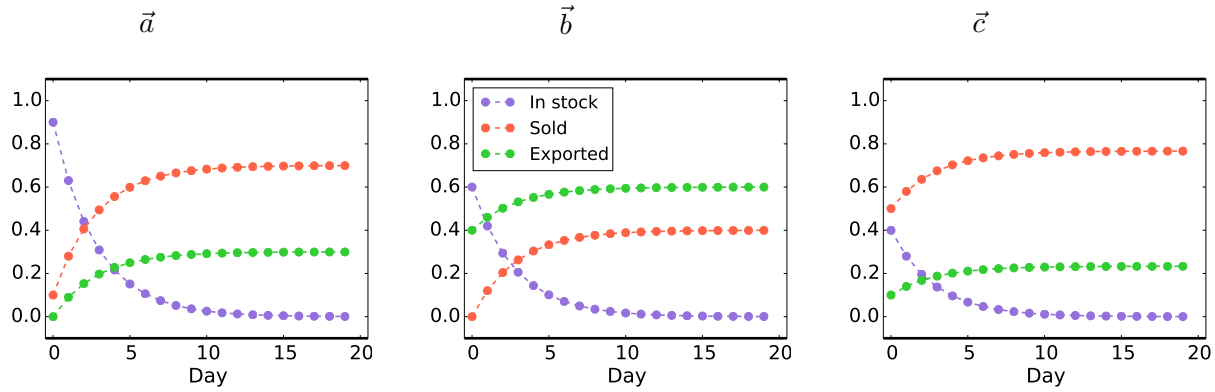


Figure 7.6: 马尔可夫链在例子 (7.4.1) 中的状态向量演变，对于不同的初始状态向量 $\vec{p}_{\widetilde{X}(0)}$ 值。

相当于计算

$$\lim_{i\to\infty} \vec{p}_{\widetilde{X}(i)} = \lim_{i\to\infty} T_{\widetilde{X}}^{i}\, \vec{p}_{\widetilde{X}(0)} \tag{7.37}$$

$$= \lim_{i\to\infty} T_{\widetilde{X}}^{i}\, \vec{a}. \tag{7.38}$$

转移矩阵 $T_{\widetilde{X}}$ 有三个特征向量

$$\vec{q}_1 := \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \qquad \vec{q}_2 := \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \qquad \vec{q}_3 := \begin{bmatrix} 0.80 \\ -0.53 \\ -0.27 \end{bmatrix}. \tag{7.39}$$

对应的特征值为 $\lambda_1 := 1$, $\lambda_2 := 1$ 和 $\lambda_3 := 0.7$。我们将特征向量和特征值聚集成两个矩阵。

$$Q := \begin{bmatrix} \vec{q}_1 & \vec{q}_2 & \vec{q}_3 \end{bmatrix}, \qquad \Lambda := \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}, \tag{7.40}$$

因此，$T_{\widetilde{X}}$ 的特征分解为

$$T_{\widetilde{X}} := Q\Lambda Q^{-1}. \tag{7.41}$$

将初始状态向量 $\vec{a}$ 用不同的特征向量表示是有用的。这可以通过计算实现

$$Q^{-1}\vec{p}_{\widetilde{X}(0)} = \begin{bmatrix} 0.3 \\ 0.7 \\ 1.122 \end{bmatrix}, \tag{7.42}$$

以便

$$\vec{a} = 0.3\,\vec{q}_1 + 0.7\,\vec{q}_2 + 1.122\,\vec{q}_3. \tag{7.43}$$

我们得出结论，

$$\lim_{i\to\infty} T_{\widetilde{X}}^{i}\,\vec{a} = \lim_{i\to\infty} T_{\widetilde{X}}^{i}\,(0.3\,\vec{q}_1 + 0.7\,\vec{q}_2 + 1.122\,\vec{q}_3) \tag{7.44}$$

$$= \lim_{i\to\infty} 0.3\,T_{\widetilde{X}}^{i}\,\vec{q}_1 + 0.7\,T_{\widetilde{X}}^{i}\,\vec{q}_2 + 1.122\,T_{\widetilde{X}}^{i}\,\vec{q}_3 \tag{7.45}$$

$$= \lim_{i\to\infty} 0.3\,\lambda_1^{\,i}\,\vec{q}_1 + 0.7\,\lambda_2^{\,i}\,\vec{q}_2 + 1.122\,\lambda_3^{\,i}\,\vec{q}_3 \tag{7.46}$$

$$= \lim_{i\to\infty} 0.3\,\vec{q}_1 + 0.7\,\vec{q}_2 + 1.122\,0.5^{\,i}\,\vec{q}_3 \tag{7.47}$$

$$= 0.3\,\vec{q}_1 + 0.7\,\vec{q}_2 \tag{7.48}$$

$$= \begin{bmatrix} 0 \\ 0.7 \\ 0.3 \end{bmatrix}. \tag{7.49}$$

这意味着最终每部手机在本地售出的概率为 0.7，而被出口的概率为 0.3。图 7.6 左侧的图展示了状态向量的演化过程。正如预测的那样，它最终收敛到方程（7.49）中的向量。

一般来说，由于本例中两个特征值等于 1 的特征向量具有特殊的结构，我们有

$$\lim_{i\to\infty} T_{\widetilde{X}}^{i}\,\vec{p}_{\widetilde{X}(0)} = \begin{bmatrix} 0 \\ \left(Q^{-1}\vec{p}_{\widetilde{X}(0)}\right)_2 \\ \left(Q^{-1}\vec{p}_{\widetilde{X}(0)}\right)_1 \end{bmatrix}. \tag{7.50}$$

这在图 7.6 中得到了说明，在那里你可以看到当状态向量被初始化为另外这两种分布时的演化过程：

$$\vec{b} := \begin{bmatrix} 0.6 \\ 0 \\ 0.4 \end{bmatrix}, \qquad Q^{-1}\vec{b} = \begin{bmatrix} 0.6 \\ 0.4 \\ 0.75 \end{bmatrix}, \tag{7.51}$$

$$\vec{c} := \begin{bmatrix} 0.4 \\ 0.5 \\ 0.1 \end{bmatrix}, \qquad Q^{-1}\vec{c} = \begin{bmatrix} 0.23 \\ 0.77 \\ 0.50 \end{bmatrix}. \tag{7.52}$$

△

例 7.4.1 中马尔可夫链的转移矩阵有两个特征值等于 1 的特征向量。如果我们将初始状态向量设为其中任意一个特征向量（注意必须将其归一化，以确保状态向量包含一个有效的 pmf），那么

$$T_{\widetilde{X}}\,\vec{p}_{\widetilde{X}(0)} = \vec{p}_{\widetilde{X}(0)}, \tag{7.53}$$

以便

$$\vec{p}_{\widetilde{X}(i)} = T_{\widetilde{X}}^{i}\,\vec{p}_{\widetilde{X}(0)} \tag{7.54}$$

$$= \vec{p}_{\widetilde{X}(0)} \tag{7.55}$$

对于所有 i。特别是,

$$\lim_{i\to\infty} \vec{p}_{\widetilde{X}(i)} = \vec{p}_{\widetilde{X}(0)}, \tag{7.56}$$

因此，$\vec{X}$ 收敛到具有 pmf $\vec{p}_{\widetilde{X}(0)}$ 的随机变量。满足 (7.56) 的分布称为马尔可夫链的 *stationary* 分布。

Definition 7.4.2 (平稳分布)**.** *Let $\widetilde{X}$ be a finite-state time-homogeneous Markov chain and let $\vec{p}_{stat}$ be a state vector containing a valid pmf over the possible states of $\widetilde{X}$. If $\vec{p}_{stat}$ is an eigenvector associated to an eigenvalue equal to one, so that*

$$T_{\widetilde{X}}\, \vec{p}_{stat} = \vec{p}_{stat}, \tag{7.57}$$

then the distribution corresponding to $\vec{p}_{stat}$ is a stationary or steady-state distribution of $\widetilde{X}$.

通过检查(7.57)是否成立来确定一个分布是否是平稳的，如果状态空间非常大，这在计算上可能是具有挑战性的。现在我们推导出一个暗示平稳性的替代条件。让我们首先定义马尔可夫链的可逆性。

Definition 7.4.3 (可逆性)**.** *Let $\widetilde{X}$ be a finite-state time-homogeneous Markov chain with s states and transition matrix $T_{\widetilde{X}}$. Assume that $\widetilde{X}(i)$ is distributed according to the state vector $\vec{p} \in \mathbb{R}^s$. If*

$$\mathrm{P}\left(\widetilde{X}(i) = x_j, \widetilde{X}(i+1) = x_k\right) = \mathrm{P}\left(\widetilde{X}(i) = x_k, \widetilde{X}(i+1) = x_j\right), \quad \text{for all } 1 \le j,k \le s, \tag{7.58}$$

then we say that $\widetilde{X}$ is reversible with respect to $\vec{p}$. This is equivalent to the detailed-balance condition

$$\left(T_{\widetilde{X}}\right)_{kj} \vec{p}_j = \left(T_{\widetilde{X}}\right)_{jk} \vec{p}_k, \quad \text{for all } 1 \le j,k \le s. \tag{7.59}$$

如下面的定理所证明的，可逆性蕴含平稳性，但反之不成立。马尔可夫链相对于某个平稳分布并不一定是可逆的（而且往往不是）。因此，详细平衡条件只提供了平稳性的一个充分条件。

Theorem 7.4.4 (可逆性蕴含平稳性)**.** *If a time-homogeneous Markov chain $\widetilde{X}$ is reversible with respect to a distribution p_X, then p_X is a stationary distribution of $\widetilde{X}$.*

Proof. 设 $\vec{p}$ 为包含 p_X 的状态向量。根据假设，$T_{\widetilde{X}}$ 和 $\vec{p}$ 满足 (7.59)，因此对于 $1 \le j \le s$

$$\left(T_{\widetilde{X}}\vec{p}\right)_j = \sum_{k=1}^{s} \left(T_{\widetilde{X}}\right)_{jk} \vec{p}_k \tag{7.60}$$

$$= \sum_{k=1}^{s} \left(T_{\widetilde{X}}\right)_{kj} \vec{p}_j \tag{7.61}$$

$$= \vec{p}_j \sum_{k=1}^{s} \left(T_{\widetilde{X}}\right)_{kj} \tag{7.62}$$

$$= \vec{p}_j. \tag{7.63}$$

$$\vec{p}_{\widetilde{X}(0)} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \qquad \vec{p}_{\widetilde{X}(0)} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \qquad \vec{p}_{\widetilde{X}(0)} = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.7 \end{bmatrix}$$

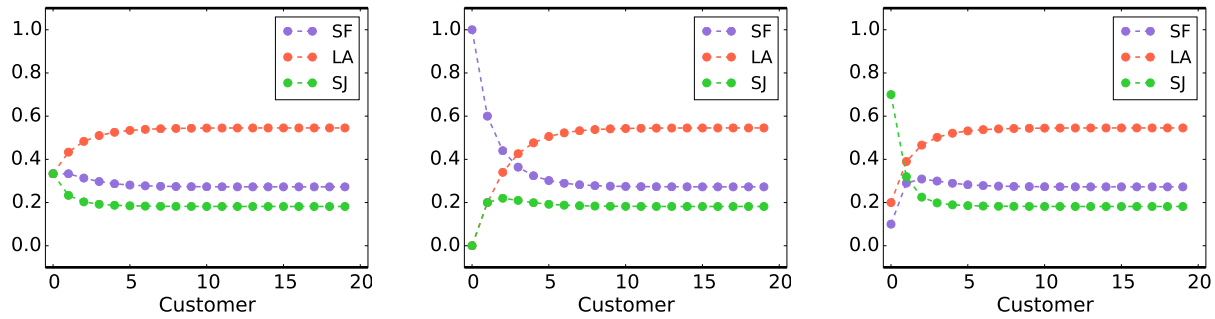


Figure 7.7: 示例（7.4.7）中马尔可夫链状态向量的演化。

最后一步来自于这样一个事实：有效转移矩阵的列之和必须为1（链总是必须有一个去向）。

□

在示例 7.4.1 中，马尔可夫链有两个平稳分布。事实证明，对于不可约马尔可夫链，这是不可能的。

Theorem 7.4.5. *Irreducible Markov chains have a single stationary distribution.*

Proof. 这源自佩龙-弗罗贝纽斯定理，该定理表明，某不可约马尔可夫链的转移矩阵具有一个特征值为1且所有条目非负的特征向量。

□

如果此外，马尔可夫链是非周期的，那么它保证会以其平稳分布 *for any initial state vector* 收敛到一个随机变量。这样的马尔可夫链称为 **ergodic**。

Theorem 7.4.6 (马尔可夫链的收敛性)**.** *If a discrete-time time-homogeneous Markov chain $\widetilde{X}$ is irreducible and aperiodic its state vector converges to the stationary distribution $\vec{p}_{stat}$ of $\widetilde{X}$ for any initial state vector $\vec{p}_{\widetilde{X}(0)}$. This implies that $\widetilde{X}$ converges in distribution to a random variable with pmf given by $\vec{p}_{stat}$.*

这个结果的证明超出了本文的范围。

Example 7.4.7 (汽车租赁（续））**.** 汽车租赁示例中的马尔可夫链是不可约的且是非周期的。现在我们将检查它是否确实在分布上收敛。其转移矩阵具有以下特征向量

$$\vec{q}_1 := \begin{bmatrix} 0.273 \\ 0.545 \\ 0.182 \end{bmatrix}, \qquad \vec{q}_2 := \begin{bmatrix} -0.577 \\ 0.789 \\ -0.211 \end{bmatrix}, \qquad \vec{q}_3 := \begin{bmatrix} -0.577 \\ -0.211 \\ 0.789 \end{bmatrix}. \tag{7.64}$$

相应的特征值为 $\lambda_1 := 1$, $\lambda_2 := 0.573$ 和 $\lambda_3 := 0.227$。如定理 7.4.5 所预测，马尔可夫链具有唯一的平稳分布。

对于任何初始状态向量，与 $\vec{q}_1$ 共线的分量将会在马尔可夫链的过渡过程中被保留，但其他两个分量会在一段时间后变得可以忽略不计。因此，链将收敛到一个具有概率质量函数 $\vec{q}_1$ (的随机变量。请注意，$\vec{q}_1$ 已被归一化为有效的概率质量函数)，正如定理 7.4.6 所预测的那样。这一点在图 7.7 中有所展示。无论公司如何分配新车，最终 27.3% 的车将进入旧金山，54.5% 进入洛杉矶，18.2% 进入圣何塞。△

7.5 Markov-chain Monte Carlo

马尔可夫链收敛到平稳分布对于模拟随机变量非常有用。马尔可夫链蒙特卡罗（MCMC）方法通过构建马尔可夫链，使得平稳分布等于目标分布，从而从目标分布中生成样本。这些技术在现代统计学中具有重要意义，尤其是在贝叶斯建模中。在本节中，我们将描述最受欢迎的MCMC方法之一，并通过一个简单的例子进行说明。

MCMC方法的关键挑战是设计一个不可约非周期性的马尔可夫链，使得目标分布是平稳的。Metropolis-Hastings算法使用一个辅助马尔可夫链来实现这一点。

Algorithm 7.5.1 (Metropolis-Hastings 算法). *We store the pmf p_X of the target distribution in a vector $\vec{p} \in \mathbb{R}^s$, such that*

$$\vec{p}_j := p_X(x_j), \qquad 1 \leq j \leq s. \tag{7.65}$$

Let T denote the transition matrix of an irreducible Markov chain with the same state space $\{x_1, \ldots, x_s\}$ as $\vec{p}$.
Initialize $\widetilde{X}(0)$ randomly or to a fixed state, then repeat the following steps for $i = 1, 2, 3, \ldots$.

1. *Generate a candidate random variable C from $\widetilde{X}(i-1)$ by using the transition matrix T, i.e.*

$$\mathrm{P}\left(C = k \,|\, \widetilde{X}(i-1) = j\right) = T_{kj}, \qquad 1 \leq j, k \leq s. \tag{7.66}$$

2. *Set*

$$\widetilde{X}(i) := \begin{cases} C & \text{with probability } p_{\mathrm{acc}}\left(\widetilde{X}(i-1), C\right), \\ \widetilde{X}(i-1) & \text{otherwise,} \end{cases} \tag{7.67}$$

where the 接受概率 *is defined as*

$$p_{\mathrm{acc}}(j,k) := \min\left\{\frac{T_{jk}\,\vec{p}_k}{T_{kj}\,\vec{p}_j}, 1\right\} \qquad 1 \leq j, k \leq s. \tag{7.68}$$

事实证明，该算法产生了一个马尔可夫链，该链相对于感兴趣的分布是可逆的，这确保了该分布是平稳的。

Theorem 7.5.2. *The pmf in $\vec{p}$ corresponds to a stationary distribution of the Markov chain $\widetilde{X}$ obtained by the Metropolis-Hastings algorithm.*

Proof. 我们证明了马尔可夫链$\widetilde{X}$关于$\vec{p}$是可逆的，即

$$\left(T_{\widetilde{X}}\right)_{kj}\vec{p}_j = \left(T_{\widetilde{X}}\right)_{jk}\vec{p}_k, \tag{7.69}$$

对所有$1 \leq j,k \leq s$都成立。这通过定理 7.4.4 确立了该结果。若$j=k$，则详细平衡条件显然成立。如果$j \neq k$，则有

$$\begin{aligned}
\left(T_{\widetilde{X}}\right)_{kj} &:= \mathrm{P}\left(\widetilde{X}(i)=k \,|\, \widetilde{X}(i-1)=j\right) && (7.70)\\
&= \mathrm{P}\left(\widetilde{X}(i)=C, C=k \,|\, \widetilde{X}(i-1)=j\right) && (7.71)\\
&= \mathrm{P}\left(\widetilde{X}(i)=C \,|\, C=k, \widetilde{X}(i-1)=j\right) P\left(C=k \,|\, \widetilde{X}(i-1)=j\right) && (7.72)\\
&= p_{\mathrm{acc}}(j,k)\, T_{kj} && (7.73)
\end{aligned}$$

并且通过完全相同的论证$\left(T_{\widetilde{X}}\right)_{jk} = p_{\mathrm{acc}}(k,j)\, T_{jk}$，我们得出结论：

$$\begin{aligned}
\left(T_{\widetilde{X}}\right)_{kj}\vec{p}_j &= p_{\mathrm{acc}}(j,k)\, T_{kj}\,\vec{p}_j && (7.74)\\
&= T_{kj}\,\vec{p}_j \min\left\{\frac{T_{jk}\,\vec{p}_k}{T_{kj}\,\vec{p}_j}, 1\right\} && (7.75)\\
&= \min\left\{T_{jk}\,\vec{p}_k, T_{kj}\,\vec{p}_j\right\} && (7.76)\\
&= T_{jk}\,\vec{p}_k \min\left\{1, \frac{T_{kj}\,\vec{p}_j}{T_{jk}\,\vec{p}_k}\right\} && (7.77)\\
&= p_{\mathrm{acc}}(k,j)\, T_{jk}\,\vec{p}_k && (7.78)\\
&= \left(T_{\widetilde{X}}\right)_{jk}\vec{p}_k. && (7.79)
\end{aligned}$$

□

以下示例摘自Hastings的开创性论文*Monte Carlo Sampling Methods Using Markov Chains and Their Applications*。

Example 7.5.3 (生成泊松随机变量)**.** 我们的目标是生成泊松随机变量。请注意，我们不需要知道泊松概率质量函数中的归一化常数e^{λ}，只要知道它与以下内容成正比即可。

$$p_X(x) \propto \frac{\lambda^x}{x!} \tag{7.80}$$

辅助马尔可夫链必须能够达到X的所有可能值，即所有正整数。我们将使用一种修改过的随机游走，它以1/2的概率向上和向下移动，但永远不会低于0。其转移矩阵等于

$$T_{kj} := \begin{cases} \frac{1}{2} & \text{if } j=0 \text{ and } k=0,\\ \frac{1}{2} & \text{if } k=j+1,\\ \frac{1}{2} & \text{if } j>0 \text{ and } k=j-1,\\ 0 & \text{otherwise.} \end{cases} \tag{7.81}$$

T 是对称的，因此接受概率等于pmf的比率:

$$p_{\text{acc}}(j,k) := \min\left\{\frac{T_{jk}\, p_X(k)}{T_{kj}\, p_X(j)}, 1\right\} \tag{7.82}$$

$$= \min\left\{\frac{p_X(k)}{p_X(j)}, 1\right\}. \tag{7.83}$$

为了计算接受概率，我们仅考虑在随机游走下可能发生的转移。如果 $j = 0$ 且 $k = 0$

$$p_{\text{acc}}(j,k) = 1. \tag{7.84}$$

如果 $k = j + 1$

$$p_{\text{acc}}(j, j+1) = \min\left\{\frac{\frac{\lambda^{j+1}}{(j+1)!}}{\frac{\lambda^j}{j!}}, 1\right\} \tag{7.85}$$

$$= \min\left\{\frac{\lambda}{j+1}, 1\right\}. \tag{7.86}$$

如果 $k = j - 1$

$$p_{\text{acc}}(j, j-1) = \min\left\{\frac{\frac{\lambda^{j-1}}{(j-1)!}}{\frac{\lambda^j}{j!}}, 1\right\} \tag{7.87}$$

$$= \min\left\{\frac{j}{\lambda}, 1\right\}. \tag{7.88}$$

我们现在详细说明 Metropolis-Hastings 方法的步骤。为了模拟辅助随机游走，我们使用一系列伯努利随机变量，用以指示随机游走是尝试向上还是向下（或停留在零）。我们将链初始化在 $x_0 = 0$。然后，对于 $i = 1, 2, \ldots$，我们

- 生成一个参数为 1/2 的伯努利分布样本 b，以及一个在 [0, 1] 区间内均匀分布的样本 u。

- 如果 $b = 0$: – 如果 $x_{i-1} = 0$, $x_i := 0$.
  – 如果 $x_{i-1} > 0$: $*$ 如果 $u < \frac{x_{i-1}}{\lambda}$, x_i $:= x_{i-1} - 1$. $*$ 否则 $x_i := x_{i-1}$.

- If $b = 1$:
  - If $u < \frac{\lambda}{x_{i-1}+1}$, $x_i := x_{i-1} + 1$.
  - Otherwise $x_i := x_{i-1}$.

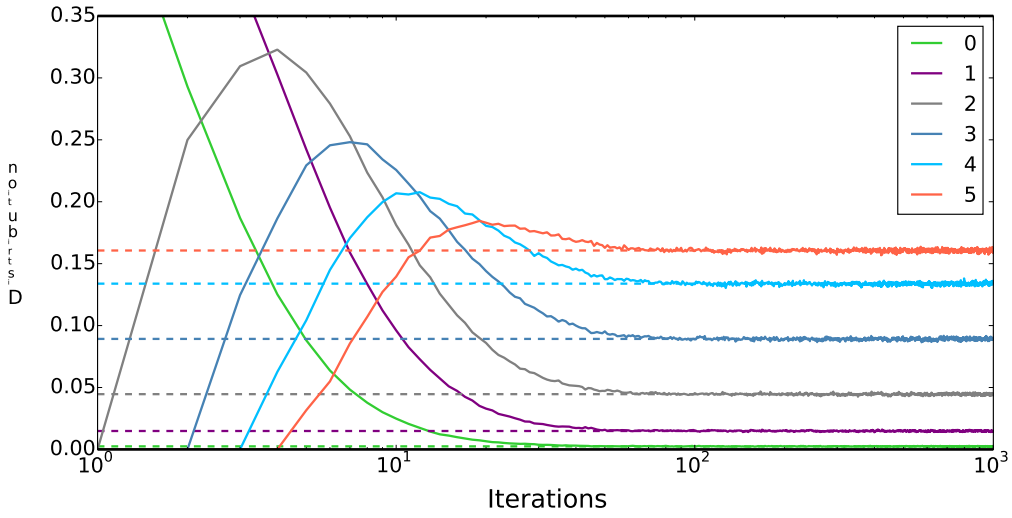


Figure 7.8: 马尔可夫链在示例7.8中为$\lambda:=6$构建的分布收敛性。为了避免杂乱，我们仅绘制了6个状态的经验分布，该分布通过运行马尔可夫链10^4次计算得出。

我们构建的马尔可夫链是不可约的：从任何非负整数到任何其他非负整数都有非零的概率（尽管可能需要一段时间！）。我们并没有真正证明该链应该收敛到期望的分布，因为我们没有讨论具有无限状态空间的马尔可夫链的收敛性，但图 7.8 显示该方法确实可以从一个泊松分布中采样，参数为 $\lambda:=6$。

△

对于图 7.8 中的示例，大约在 100 次迭代之后会出现分布上的近似收敛。这被称为马尔可夫链的**mixing time**。为此，MCMC 方法通常会在一个称为$burn\text{-}in$时间的初始阶段丢弃链中的样本。

仔细的读者可能会想，如果我们已经能够访问到所需的分布，使用MCMC方法的意义何在。似乎直接应用第2.6.1节中描述的方法会更简单。然而，Metropolis-Hastings方法可以应用于具有无限支持的离散分布，也可以应用于连续分布（证明这一点超出了本笔记的范围）。至关重要的是，与逆变换采样和拒绝采样不同，Metropolis-Hastings不需要访问目标分布的pmf p_X或pdf f_X，而是需要访问每个$x \neq y$的比率$p_X(x)/p_X(y)$或$f_X(x)/f_X(y)$。这在计算概率模型中的条件分布时非常有用。

假设我们可以访问连续随机变量 A 的边际分布，以及在给定 A 的条件下，另一个连续随机变量 B 的条件分布。计算

条件概率密度函数

$$f_{A|B}\left(a|b\right)=\frac{f_{A}\left(a\right)f_{B|A}\left(b|a\right)}{\int_{u=-\infty}^{\infty}f_{A}\left(u\right)f_{B|A}\left(b|u\right)\,\mathrm{d}u} \tag{7.89}$$

由于分母中的积分，实际上是不可行的。然而，如果我们应用Metropolis-Hastings从$f_{A|B}$中采样，我们就不需要计算归一化因子，因为对于任何$a_1 \neq a_2$

$$\frac{f_{A|B}\left(a_1|b\right)}{f_{A|B}\left(a_2|b\right)}=\frac{f_{A}\left(a_1\right)f_{B|A}\left(b|a_1\right)}{f_{A}\left(a_2\right)f_{B|A}\left(b|a_2\right)}. \tag{7.90}$$

Chapter 8

Descriptive statistics

在本章中，我们描述了几种数据可视化的技术，以及计算有效总结数据的量化方法。这些量化方法被称为描述性统计量。正如我们将在接下来的章节中看到的，这些统计量通常可以在概率框架内进行解释，但当没有合理的概率假设时，它们仍然是有用的。因此，我们从 **deterministic** 的角度来介绍它们。

8.1 Histogram

我们首先考虑包含一维数据的数据集。可视化一维数据最自然的方法之一是绘制它们的直方图。直方图通过将数据范围划分为若干区间并统计落入每个区间的实例数量来获得。区间的宽度是一个可以调整的参数，用以得到更高或更低的分辨率。如果我们将数据解释为来自某个随机变量的样本，那么直方图就是其概率质量函数（pmf）或概率密度函数（pdf）的分段常数近似。

图 8.1 显示了从牛津气象站收集的温度数据计算出的两个直方图，这些数据涵盖了 150 年的时间。[1] 每个数据点代表某一年 1 月或 8 月的最高气温。图 8.2 显示了根据联合国数据，2014 年全球所有国家的每 capita GDP 直方图。[2]

8.2 Sample mean and variance

对一维数据集中各元素取平均，可以得到对数据的一个单一数值总结，它是随机变量均值的确定性对应物（回想一下，本章中我们不作任何概率假设）。这一方法可以通过分别对每个维度取平均而扩展到多维数据。

Definition 8.2.1 (样本均值)**.** *Let* $\{x_1, x_2, \ldots, x_n\}$ *be a set of real-valued data. The sample*

[1] The data is available at `http://www.metoffice.gov.uk/pub/data/weather/uk/climate/stationdata/oxforddata.txt`.

[2] The data is available at `http://unstats.un.org/unsd/snaama/selbasicFast.asp`.

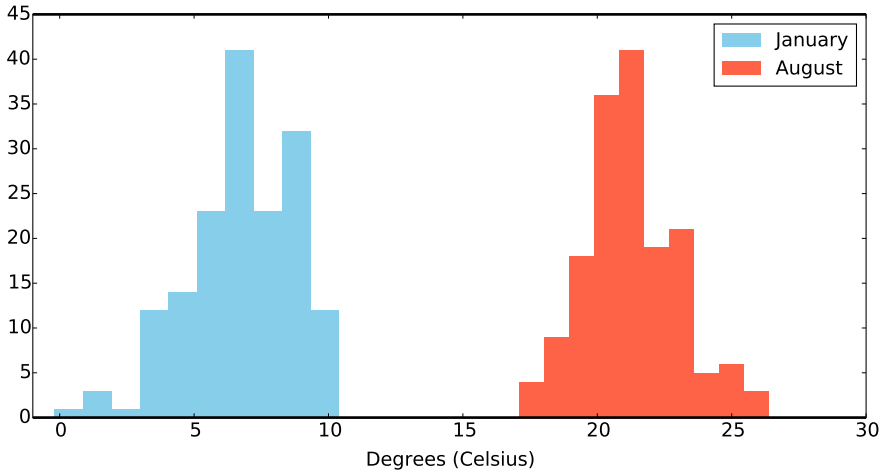


Figure 8.1: 奥克斯福德气象站150年来温度数据的直方图。每个数据点代表某年某月记录的最高温度。

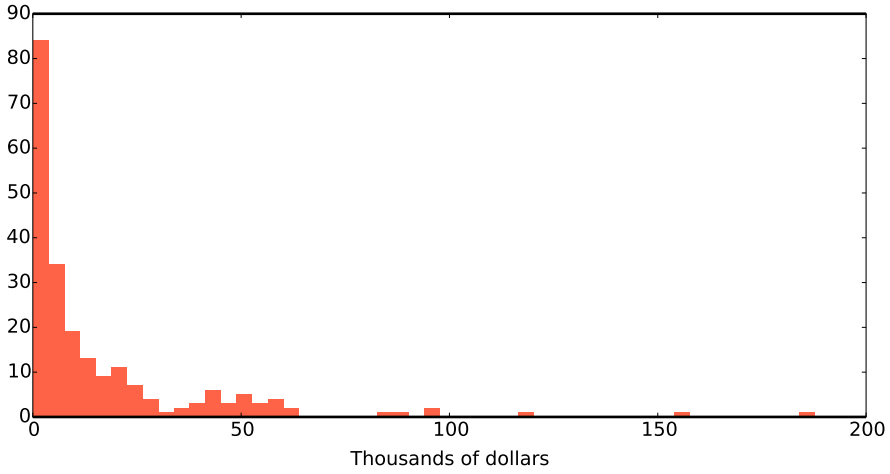


Figure 8.2: 2014年世界各国人均GDP的直方图。

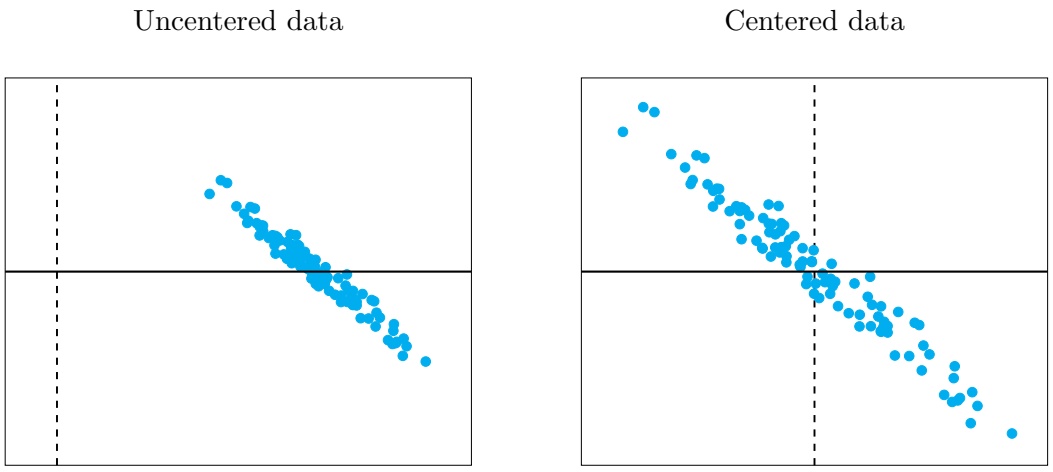


Figure 8.3: 二维数据集中心化的影响。轴线使用虚线表示。

mean of the data is defined as

$$\operatorname{av}(x_1, x_2, \ldots, x_n) := \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{8.1}$$

Let $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$ *be a set of d-dimensional real-valued data vectors. The sample mean is*

$$\operatorname{av}(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n) := \frac{1}{n}\sum_{i=1}^{n} \vec{x}_i. \tag{8.2}$$

图8.1中数据的样本均值为1月6.73 °C，8月21.3 °C。图8.2中人均GDP的样本均值为16,500美元。

从几何角度看，平均值（也称为样本均值）是数据的质心。数据分析中一个常见的预处理步骤是通过减去其样本均值来对一组数据进行**center**。图 8.3 展示了一个示例。

Algorithm 8.2.2 (居中)**.** *Let* $\vec{x}_1, \ldots, \vec{x}_n$ *be a set of d-dimensional data. To center the data set we:*

1. *Compute the sample mean following Definition 8.2.1.*
2. *Subtract the sample mean from each vector of data. For* $1 \leq i \leq n$

$$\vec{y}_i := \vec{x}_i - \operatorname{av}(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n). \tag{8.3}$$

The resulting data set $\vec{y}_1, \ldots, \vec{y}_n$ *has sample mean equal to zero; it is centered at the origin.*

样本方差是各观测值相对于样本均值的偏差平方的平均值。从几何上看，它量化了数据集围绕其中心的平均变动程度。它是随机变量方差的确定性对应物。

Definition 8.2.3 (样本方差和标准差). *Let $\{x_1, x_2, \ldots, x_n\}$ be a set of real-valued data. The sample variance is defined as*

$$\operatorname{var}(x_1, x_2, \ldots, x_n) := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \operatorname{av}(x_1, x_2, \ldots, x_n))^2 \tag{8.4}$$

The sample standard deviation is the square root of the sample variance

$$\operatorname{std}(x_1, x_2, \ldots, x_n) := \sqrt{\operatorname{var}(x_1, x_2, \ldots, x_n)}. \tag{8.5}$$

你可能会想知道为什么归一化常数是 $1/(n-1)$ 而不是 $1/n$。原因是这样可以确保当数据是独立同分布 (iid) 时，样本方差的期望值等于真实方差（见引理 9.2.5）。在实践中，这两种归一化方式之间没有太大区别。

图8.1中的温度数据的样本标准差为1.99 °C（1月）和1.73 °C（8月）。图8.2中的GDP数据的样本标准差为$25,300。

8.3 Order statistics

在某些情况下，一个数据集可以通过其均值和标准差得到很好的描述。

In January the temperature in Oxford is around 6.73 °C give or take 2 °C.

这是对上一节温度数据的相当准确的描述。然而，假设有人将图8.2中的GDP数据集描述为：

Countries typically have a GDP per capita of about $16 500 give or take $25 300.

这个描述很糟糕。问题在于大多数国家的人均GDP非常小，而少数国家的人均GDP非常大，样本均值和标准差并不能很好地传达这些信息。顺序统计量提供了一种替代的描述，当数据中存在极端值时，这种描述通常更加有用。

Definition 8.3.1 (分位数和百分位数). *Let $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ denote the ordered elements of a set of data $\{x_1, x_2, \ldots, x_n\}$. The q quantile of the data for $0 < q < 1$ is $x_{([q(n+1)])}$, where $[q(n+1)]$ is the result of rounding $q(n+1)$ to the closest integer. The $100\,p$ quantile is known as the p percentile.*

*The 0.25 and 0.75 quantiles are known as the first and third **quartiles**, whereas the 0.5 quantile is known as the **sample median**. A quarter of the data are smaller than the 0.25 quantile, half are smaller (or larger) than the median and three quarters are smaller than the 0.75 quartile. If n is even, the sample median is usually set to*

$$\frac{x_{(n/2)} + x_{(n/2+1)}}{2}. \tag{8.6}$$

*The difference between the third and the first quartile is known as the **interquartile range** (IQR).*

事实证明，对于图8.1中的温度数据集，1月的样本中位数为6.80 °C，8月为21.2 °C，几乎与样本均值相同。四分位距（IQR）是

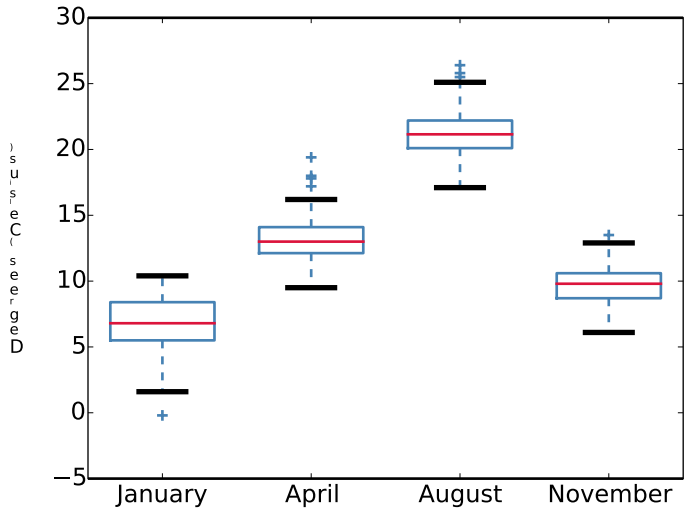


Figure 8.4: 箱型图展示了图 8.1 中使用的牛津温度数据集。每个箱型图对应过去 150 年中某一特定月份（1 月、4 月、8 月和 11 月）的最高温度。

1月为 2.9 °C，8月为 2.1 °C。这使得围绕中位数的分布范围与样本均值非常相似。在这个特定的例子中，使用顺序统计量似乎并没有优势。

对于GDP数据集，中位数为6,350美元。这意味着一半国家的GDP低于6,350美元。相比之下，71 %的国家人均GDP低于样本均值！这些数据的四分位距（IQR）为18,200美元。为了更完整地描述数据集，我们可以列出一个**five-number summary**的顺序统计量：最小值$x_{(1)}$、第一四分位数、样本中位数、第三四分位数和最大值$x_{(n)}$。对于GDP数据集，这些值分别是130美元、1,960美元、6,350美元、20,100美元和188,000美元。

我们可以使用**box plot**来可视化数据集的主要顺序统计量，它在一个盒子中显示数据的中位数。盒子的底部和顶部分别是第一四分位数和第三四分位数。这种可视化数据集的方法由数学家约翰·图基提出。图基的箱线图还包括*whiskers*。下须是一条从盒子底部延伸到位于第一四分位数1.5 IQR范围内的最小值的线。上须从盒子顶部延伸到位于第三四分位数1.5 IQR范围内的最大值。超出须线的值被视为**outliers**，并单独绘制。

图8.4 使用箱线图对图8.1中所用的温度数据集进行可视化。每个箱线图对应过去150年中某一特定月份（1月、4月、8月和11月）的最高气温。箱线图使我们能够快速比较不同月份气温的离散程度。图8.5 展示了图8.2 中 GDP 数据的箱线图。从该箱线图可以立刻看出，大多数国家的人均 GDP 非常低；随着人均 GDP 增大，国家之间的差异也随之增大；同时，只有少数国家拥有非常高的人均 GDP。

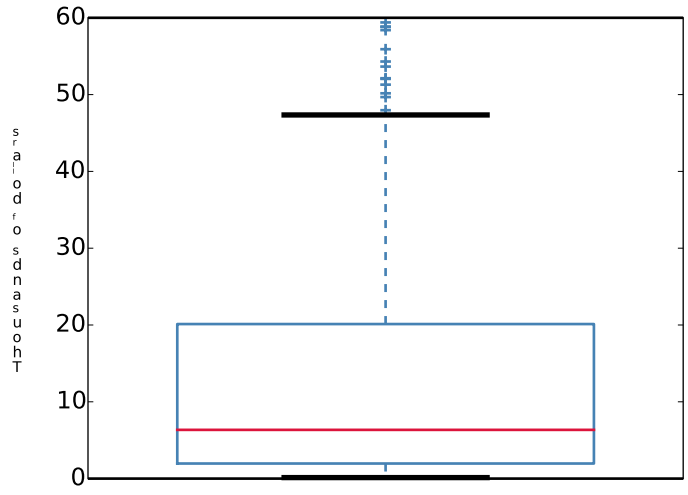


Figure 8.5: 2014年全球各国人均GDP的箱线图。并非所有离群值都显示。

8.4 Sample covariance

在前面的章节中，我们主要考虑由一维数据组成的数据集（除非在讨论多维数据集的样本均值时）。用机器学习的术语来说，每个数据点只有一个特征。现在我们研究一种多维情形，其中每个数据点都关联着多个特征。

如果数据集的维度等于二（即每个数据点有两个特征），我们可以使用**scatter plot**对数据进行可视化，其中每一条轴表示一个特征。图8.6展示了若干温度数据的散点图。这些数据与图8.1中的相同，但现在我们将它们整理成二维数据集。在左侧的图中，一个维度对应一月的温度，另一个维度对应八月的温度（每年一个数据点）。在右侧的图中，一个维度表示某一月份的最低温度，另一个维度表示同一月份的最高温度（每个月一个数据点）。样本协方差用于量化二维数据集中两个特征在平均意义上是否倾向于以相似的方式变化，正如协方差用于量化两个随机变量期望的联合变化一样。

Definition 8.4.1 (样本协方差)**.** *Let* $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ *be a data set where each example consists of a measurement of two different features. The sample covariance is defined as*

$$\operatorname{cov}((x_1, y_1), \ldots, (x_n, y_n)) := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \operatorname{av}(x_1, \ldots, x_n))(y_i - \operatorname{av}(y_1, \ldots, y_n)). \tag{8.7}$$

为了考虑到每个特征可能在不同的尺度上变化，一个常见的预处理步骤是对每个特征进行*normalize*处理，方法是将其除以样本标准差。

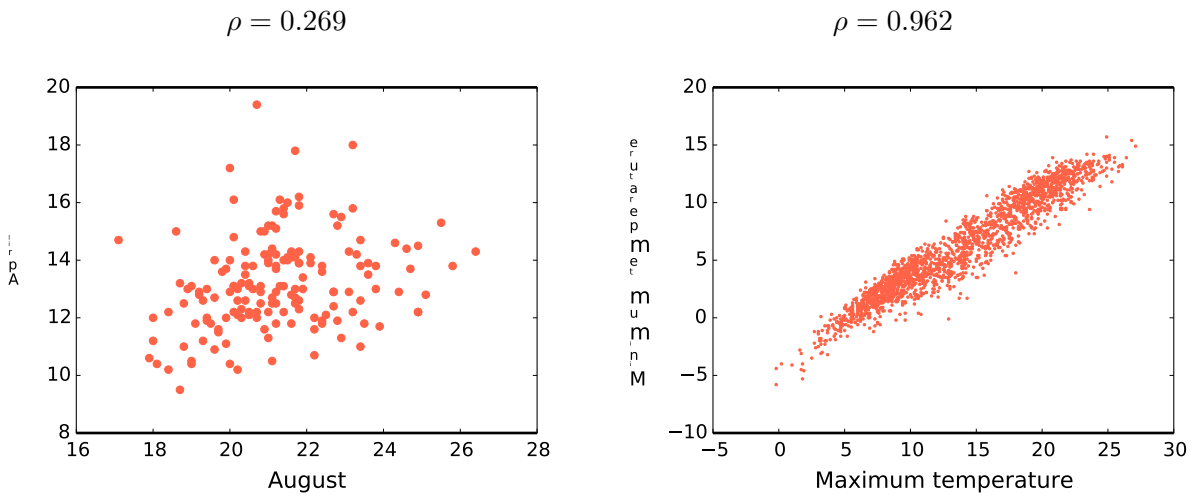


Figure 8.6: 散点图显示了牛津过去150年1月和8月的温度（左）以及每月最高和最低温度（右）。

如果我们在计算协方差之前进行归一化，我们得到的是两个特征的样本相关系数。相关系数的一个优点是，我们不需要担心特征的测量单位。相比之下，如果我们不相应地缩放另一个特征，表示距离的特征如果以英寸或英里为单位进行测量，可能会严重扭曲协方差。

Definition 8.4.2 (样本相关系数)**.** *Let* $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ *be a data set where each example consists of two features. The sample correlation coefficient is defined as*

$$\rho((x_1, y_1), \ldots, (x_n, y_n)) := \frac{\operatorname{cov}((x_1, y_1), \ldots, (x_n, y_n))}{\operatorname{std}(x_1, \ldots, x_n) \operatorname{std}(y_1, \ldots, y_n)}. \tag{8.8}$$

根据柯西—施瓦茨不等式（定理 B.2.4），该不等式表明对于任意向量 $\vec{a}$ 和 $\vec{b}$

$$-1 \leq \frac{\vec{a}^T \vec{b}}{||a||_2 \, ||b||_2} \leq 1, \tag{8.9}$$

样本相关系数的大小被限定在1以内。如果它等于1或-1，那么这两个中心化的数据集是共线的。Cauchy-Schwarz不等式与随机变量的Cauchy-Schwarz不等式（定理4.3.7）相关，但在这里它适用于确定性向量。

图 8.6 标注了与这两个图对应的样本相关系数。同一月份内的最高温度和最低温度高度相关，而同一年中一月和八月的最高温度仅呈现一定程度的相关性。

8.5 Sample covariance matrix

8.5.1 Definition

我们现在考虑多维数据集。具体而言，我们关注对数据中变异性的分析。一个数据集的样本协方差矩阵包含每一对特征之间的两两样本协方差。

Definition 8.5.1 (样本协方差矩阵). *Let $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$ be a set of d-dimensional real-valued data vectors.The sample covariance matrix of these data is the $d \times d$ matrix*

$$\Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right) := \frac{1}{n-1} \sum_{i=1}^{n} \left(\vec{x}_i - \operatorname{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)\left(\vec{x}_i - \operatorname{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)^T. \tag{8.10}$$

The (i, j) entry of the covariance matrix, where $1 \leq i, j \leq d$, is given by

$$\Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right)_{ij} = \begin{cases} \operatorname{var}\left(\left(\vec{x}_1\right)_i, \ldots, \left(\vec{x}_n\right)_i\right) & \text{if } i = j, \\ \operatorname{cov}\left(\left(\left(\vec{x}_1\right)_i, \left(\vec{x}_1\right)_j\right), \ldots, \left(\left(\vec{x}_n\right)_i, \left(\vec{x}_n\right)_j\right)\right) & \text{if } i \neq j. \end{cases} \tag{8.11}$$

为了刻画多维数据集围绕其中心的变化，我们考虑其在不同方向上的变化。数据在某一方向上的平均变化由数据在该方向上的投影的样本方差来量化。设 $\vec{v}$ 为与感兴趣方向对齐的单位范数向量，则数据集在 $\vec{v}$ 方向上的样本方差由下式给出

$$\operatorname{var}\left(\vec{v}^T\vec{x}_1, \ldots, \vec{v}^T\vec{x}_n\right) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\vec{v}^T\vec{x}_i - \operatorname{av}\left(\vec{v}^T\vec{x}_1, \ldots, \vec{v}^T\vec{x}_n\right)\right)^2 \tag{8.12}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \left(\vec{v}^T\left(\vec{x}_i - \operatorname{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)\right)^2 \tag{8.13}$$

$$= \vec{v}^T \left(\frac{1}{n-1} \sum_{i=1}^{n} \left(\vec{x}_i - \operatorname{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)\left(\vec{x}_i - \operatorname{av}\left(\vec{x}_1, \ldots, \vec{x}_n\right)\right)^T \right) \vec{v}$$

$$= \vec{v}^T \Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right) \vec{v}. \tag{8.14}$$

利用样本协方差矩阵，我们可以刻画各个方向上的变化！这是对这样一个事实的确定性类比：随机向量的协方差矩阵编码了其在各个方向上的方差。

8.5.2 Principal component analysis

考虑协方差矩阵的特征分解

$$\Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right) = \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n \end{bmatrix}^T. \tag{8.15}$$

根据定义，$\Sigma\left(\vec{x}_1, \ldots, \vec{x}_n\right)$ 是对称的，因此其特征向量 u_1、u_2、……、u_n 相互正交。根据方程 (8.14) 和定理 B.7.2，特征向量和特征值完全刻画了数据在各个方向上的变化。

$\lambda_1/n = 0.497$ $\lambda_1/n = 0.967$ $\lambda_1/n = 1.820$
$\lambda_2/n = 0.476$ $\lambda_2/n = 0.127$ $\lambda_2/n = 0.021$

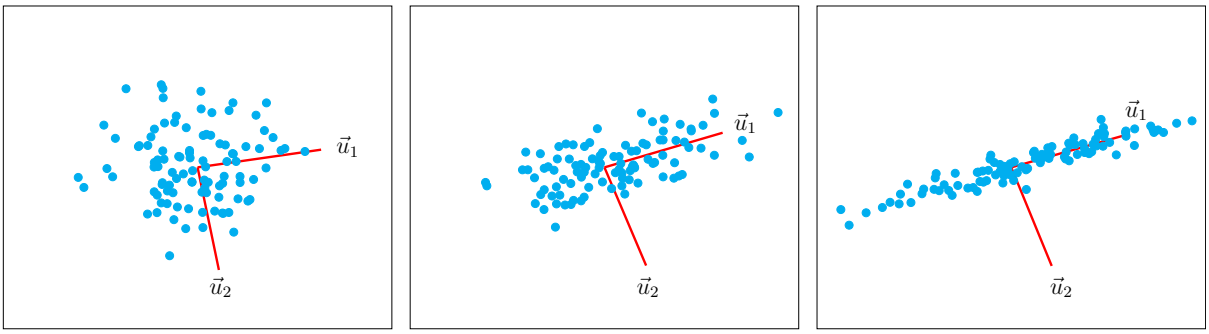


Figure 8.7: 一组包含$n = 100$个具有不同配置的二维数据点的主成分分析（PCA）。

Theorem 8.5.2. *Let the sample covariance of a set of vectors* $\Sigma(\vec{x}_1, \ldots, \vec{x}_n)$ *have an eigendecomposition given by* (8.15) *where the eigenvalues are ordered* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. *Then,*

$$\lambda_1 = \max_{||\vec{v}||_2=1} \text{var}\left(\vec{v}^T\vec{x}_1, \ldots, \vec{v}^T\vec{x}_n\right), \tag{8.16}$$

$$\vec{u}_1 = \arg\max_{||\vec{v}||_2=1} \text{var}\left(\vec{v}^T\vec{x}_1, \ldots, \vec{v}^T\vec{x}_n\right), \tag{8.17}$$

$$\lambda_k = \max_{||\vec{v}||_2=1, \vec{u}\perp\vec{u}_1,\ldots,\vec{u}_{k-1}} \text{var}\left(\vec{v}^T\vec{x}_1, \ldots, \vec{v}^T\vec{x}_n\right), \tag{8.18}$$

$$\vec{u}_k = \arg\max_{||\vec{v}||_2=1, \vec{u}\perp\vec{u}_1,\ldots,\vec{u}_{k-1}} \text{var}\left(\vec{v}^T\vec{x}_1, \ldots, \vec{v}^T\vec{x}_n\right). \tag{8.19}$$

这意味着 $\vec{u}_1$ 是最大变化的方向。对应于第二大特征值 λ_2 的特征向量 $\vec{u}_2$ 是与 $\vec{u}_1$ 正交的最大变化方向。一般而言，对应于第 k 大特征值 λ_k 的特征向量 $\vec{u}_k$ 揭示了与 $\vec{u}_1$、$\vec{u}_2$、$\cdots$、$\vec{u}_{k-1}$ 正交的最大变化方向。最后，$\vec{u_n}$ 是最小变化的方向。

在数据分析中，样本协方差矩阵的特征向量通常称为主方向。通过计算这些特征向量来量化数据集在不同方向上的变化，称为**principal component analysis** (PCA)。图 8.7 展示了若干二维示例中的主方向。

图 8.8 说明了在应用 PCA 之前进行中心化的重要性。即使数据未进行中心化，定理 8.5.2 仍然成立。然而，投影到某一方向上的范数不再反映数据的变异性。事实上，如果数据集中在一个远离原点的点附近，第一个主方向往往会与该点对齐。这是合理的，因为投影到该方向能够捕获更多能量。结果是，主方向并不能反映数据云中最大变异的方向 *within*。在应用 PCA 之前对数据集进行中心化可以解决这一问题。

下面的例子解释了如何将主成分分析应用于降维。其动机在于，在许多情况下，具有较大方差的方向对数据集的结构更具信息量。

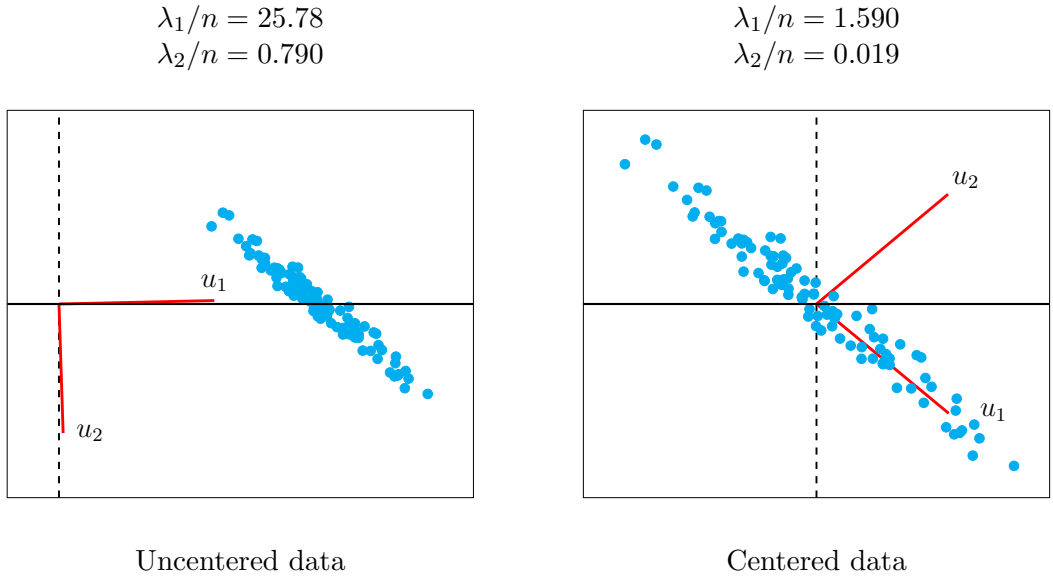


Figure 8.8: PCA 应用于 $n = 100$ 个二维数据点。左侧的数据未进行中心化。因此，主导的主方向 u_1 位于数据均值的方向上，PCA 并未反映实际结构。一旦进行中心化，u_1 便与最大变化方向对齐。

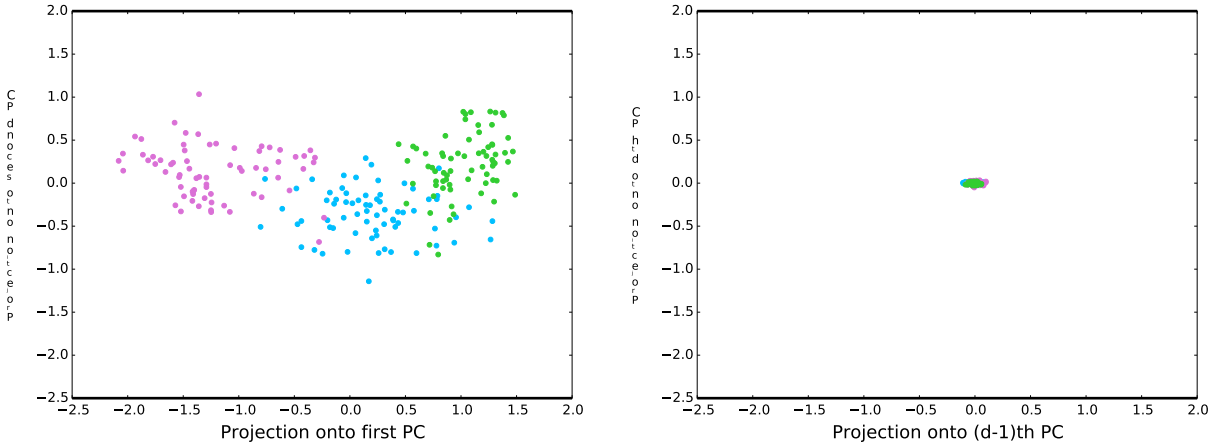


Figure 8.9: 将描述不同小麦籽粒的7维向量投影到数据集的前两个（左）和最后两个（右）主方向上。每种颜色代表一种小麦品种。

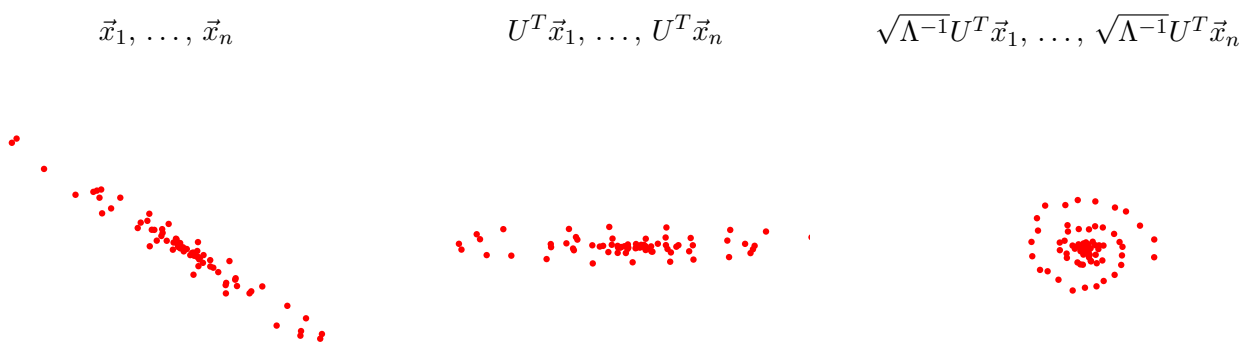


Figure 8.10: 对一组数据进行白化的效果。原始数据主要受线性偏斜主导（左）。应用 U^T 使坐标轴与样本协方差矩阵的特征向量对齐（中）。最后，$\sqrt{\Lambda^{-1}}$ 沿这些轴对数据重新加权，使其具有相同的平均变异，从而揭示被线性偏斜所掩盖的非线性结构（右）。

Example 8.5.3 (通过 PCA 进行降维)**.** 我们考虑一个数据集，其中每个数据点对应一粒种子，具有七个特征：面积、周长、紧致度、籽粒长度、籽粒宽度、非对称系数以及籽粒沟槽长度。种子属于三种不同的小麦品种：Kama、Rosa 和 Canadian。[3] 我们的目标是通过将数据投影到二维来实现可视化，并尽可能保留最大的变异性。这可以通过将每个点投影到数据集的前两个主成分维度上来实现。

图 8.9 展示了数据在前两个以及最后两个主方向上的投影。在后一种情况下，几乎没有可辨识的变化。前两个方向更好地保留了数据的结构，使得三种种子的差异可以被清晰地可视化。不过需要注意的是，投影到前几个主方向只保证尽可能保留方差，但并不一定保留对诸如分类等任务有用的特征。 △

8.5.3 Whitening

白化是一种用于对包含非线性模式的数据进行预处理的有用过程。其目标是通过沿不同方向对数据进行旋转和收缩来消除数据中的线性偏斜，从而揭示其潜在的非线性结构。这可以通过应用一种本质上对样本协方差矩阵求逆的线性变换来实现，使得结果彼此不相关。该过程称为 **whitening**，因为具有不相关分量的随机向量通常被称为白噪声。它与用于给随机向量着色的算法8.5.4密切相关。

Algorithm 8.5.4 (美白)**.** *Let $\vec{x}_1, \ldots, \vec{x}_n$ be a set of d-dimensional data, which we assume to be centered and to have a full-rank covariance matrix. To whiten the data set we:*

1. *Compute the eigendecomposition of the sample covariance matrix $\Sigma(\vec{x}_1, \ldots, \vec{x}_n) = U\Lambda U^T$.*

[3] The data can be found at `https://archive.ics.uci.edu/ml/datasets/seeds`.

2. Set $\vec{y}_i :_= \sqrt{\Lambda}^{-1} U^T \vec{x}_i$, *for* $i = 1, \ldots, n$, *where*

$$\sqrt{\Lambda} := \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \sqrt{\lambda_n} \end{bmatrix}, \tag{8.20}$$

so that $\Sigma(\vec{x}_1, \ldots, \vec{x}_n) = U\sqrt{\Lambda}\sqrt{\Lambda}U^T$.

whitened 数据集 $\vec{y}_1, \ldots, \vec{y}_n$ 的样本协方差矩阵等于单位矩阵,

$$\Sigma(\vec{y}_1, \ldots, \vec{y}_n) := \frac{1}{n-1} \sum_{i=1}^{n} \vec{y}_i \vec{y}_i^T \tag{8.21}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \sqrt{\Lambda}^{-1} U^T \vec{x}_i \left(\sqrt{\Lambda}^{-1} U^T \vec{x}_i \right)^T \tag{8.22}$$

$$= \sqrt{\Lambda}^{-1} U^T \left(\frac{1}{n-1} \sum_{i=1}^{n} \vec{x}_i \vec{x}_i^T \right) U \sqrt{\Lambda}^{-1} \tag{8.23}$$

$$= \sqrt{\Lambda}^{-1} U^T \Sigma(\vec{x}_1, \ldots, \vec{x}_n) U \sqrt{\Lambda}^{-1} \tag{8.24}$$

$$= \sqrt{\Lambda}^{-1} U^T U \sqrt{\Lambda} \sqrt{\Lambda} U^T U \sqrt{\Lambda}^{-1} \tag{8.25}$$

$$= I. \tag{8.26}$$

直观地说，白化首先对数据进行旋转，然后对其进行收缩或扩展，使得各个方向上的平均变化相同。因此，非线性模式变得更加明显，如图 8.10 所示。

Chapter 9

Frequentist Statistics

统计分析的目标是通过计算**statistics**从数据中*extract information*，而**statistics**是数据的确定性函数。在第8章中，我们从确定性和几何的角度描述了若干统计量，而不对数据生成过程作任何假设。这使得评估所获得信息的准确性变得非常具有挑战性。

在本章中，我们以概率的方式建模数据采集过程。这使得我们能够分析统计技术并推导出它们在性能上的理论保证。数据被解释为**realizations**的随机变量、向量或过程（取决于维度）。我们希望提取的信息可以通过这些量的联合分布来表达。我们认为这个分布是未知的，但**fixed**，采用**frequentist**的视角。贝叶斯统计的替代框架在第10章中进行了描述。

9.1 Independent identically-distributed sampling

在本章中，我们考虑一维实值数据，将其建模为独立同分布序列的实现。图 9.1 描绘了相应的图形模型。这是一个非常流行的假设，适用于受控实验，如随机试验药物测试，并且在其他设置中通常也是一个不错的近似。然而，在实践中，评估模型的独立性假设在多大程度上实际成立是至关重要的。

以下示例表明，通过从一个大群体中随机抽取一部分个体来测量某个量，可以得到满足iid假设的数据，只要我们进行有放回抽样（如果群体很大，无放回抽样的影响可以忽略不计）。

Example 9.1.1 (从总体).中抽样 假设我们正在研究一个由m个个体组成的总体。我们对与每个人相关的某个量感兴趣，例如他们的胆固醇水平、薪水或他们在选举中投票给谁。该量$\{z_1, z_2, \ldots, z_k\}$有$k$个可能的取值，其中$k$可以等于$m$或远小于$m$。我们用$m_j$表示数量等于$z_j$的人的数量，$1 \leq j \leq k$。在有两个候选人的选举中，$k$将等于二，$m_1$和$m_2$将代表投票给每个候选人的人。

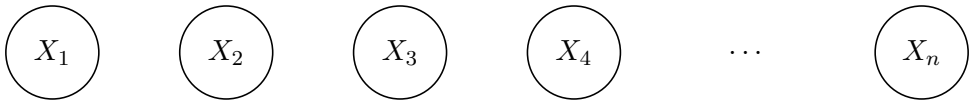


Figure 9.1: 对应于独立序列的有向图模型。如果该序列还是同分布的，那么 $X_1, X_2, \ldots, X_n$ 都具有相同的分布。

假设我们独立地随机选择 n 个个体，并且允许重复选择，这意味着一个个体可能被选择多次，并记录感兴趣量的值。在这些假设下，测量值可以被建模为独立变量$\widetilde{X}$ 的随机序列。由于每次选择时选择任何个体的概率是相同的，因此序列的一阶概率质量函数（pmf）是

$$p_{\widetilde{X}(i)}(z_j) = P(\text{The } i\text{th measurement equals } z_j) \tag{9.1}$$

$$= \frac{\text{People such that the quantity equals } z_j}{\text{Total number of people}} \tag{9.2}$$

$$= \frac{m_j}{m}, \qquad 1 \leq j \leq k, \tag{9.3}$$

根据全概率法则，对于$1 \leq i \leq n$，我们得出结论，数据可以建模为iid序列的一个实现。△

9.2 Mean square error

我们定义一个**estimator**为可用数据$x_1, x_2, \ldots, x_n$的一个确定性函数，提供与生成数据的分布相关的一个量的近似值。

$$y := h(x_1, x_2, \ldots, x_n). \tag{9.4}$$

例如，正如我们将要看到的，如果我们想要估计潜在分布的期望，一个合理的估计量是数据的平均值。由于我们采取频率学派的观点，感兴趣的量被建模为确定性的（与贝叶斯观点相对，后者会将其建模为随机变量）。对于一个固定的数据集，估计量是数据的一个确定性函数。然而，如果我们将数据建模为一系列随机变量的实现，那么估计量也是该随机变量的一个实现。

$$Y := h(X_1, X_2, \ldots, X_n). \tag{9.5}$$

这使得可以以概率方式评估估计量（通常在对底层分布作出某些假设的情况下）。例如，我们可以通过计算估计量与真实感兴趣量之间差值的均方来度量估计量所产生的误差。

Definition 9.2.1 (均方误差)**.** *The mean square error (MSE) of an estimator Y that approximates a deterministic quantity $\gamma \in \mathbb{R}$ is*

$$\text{MSE}(Y) := \text{E}\left((Y - \gamma)^2\right). \tag{9.6}$$

均方误差（MSE）可以分解为一个 **bias** 项和一个 **variance** 项。偏差项是感兴趣的量与估计量期望值之间的差异。方差项对应于估计量围绕其期望值的波动。

Lemma 9.2.2 (偏差-方差分解)**.** *The MSE of an estimator Y that approximates $\gamma \in \mathbb{R}$ satisfies*

$$MSE(Y) = \underbrace{\mathrm{E}\left((Y - \mathrm{E}(Y))^2\right)}_{\text{variance}} + \underbrace{(\mathrm{E}(Y) - \gamma)^2}_{\text{bias}}. \tag{9.7}$$

Proof. 引理是期望线性性的直接结果。 □

如果偏差为零，则该估计量在期望意义下等于所关注的量。

Definition 9.2.3 (无偏估计量)**.** *An estimator Y that approximates $\gamma \in \mathbb{R}$ is unbiased if its bias is equal to zero, i.e. if and only if*

$$\mathrm{E}(Y) = \gamma. \tag{9.8}$$

一个估计量可能是无偏的，但仍然由于其方差而产生较大的均方误差。

以下引理表明，样本均值和样本方差是独立同分布（iid）随机变量序列的真实均值和方差的无偏估计量。

Lemma 9.2.4 (样本均值是无偏的)**.** *The sample mean is an unbiased estimator of the mean of an iid sequence of random variables.*

Proof. 我们考虑独立同分布序列$\widetilde{X}$ 的样本均值，其中 均值 μ,

$$\widetilde{Y}(n) := \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}(i). \tag{9.9}$$

根据期望的线性性

$$\mathrm{E}\left(\widetilde{Y}(n)\right) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left(\widetilde{X}(i)\right) \tag{9.10}$$

$$= \mu. \tag{9.11}$$

□

Lemma 9.2.5 (样本方差是无偏的)**.** *The sample variance is an unbiased estimator of the variance of an iid sequence of random variables.*

该结果的证明见第 9.7.1 节。

9.3 Consistency

如果我们正在估计一个标量量，随着收集到更多数据，估计应该得到改进。理想情况下，当数据数量 $n \to \infty$ 的极限时，估计应当收敛到真实值。能够实现这一点的估计量被称为一致的。

Definition 9.3.1 (一致性)**.** *An estimator* $\widetilde{Y}(n) := h\left(\widetilde{X}(1), \widetilde{X}(2), \ldots, \widetilde{X}(n)\right)$ *that approximates* $\gamma \in \mathbb{R}$ *is consistent if it converges to* γ *as* $n \to \infty$ *in mean square, with probability one or in probability.*

下面的定理表明，均值是一致的。

Theorem 9.3.2 (样本均值是一致的)**.** *The sample mean is a consistent estimator of the mean of an iid sequence of random variables as long as the variance of the sequence is bounded.*

Proof. 我们考虑均值为 μ 的 iid 序列$\widetilde{X}$ 的样本均值,

$$\widetilde{Y}(n) := \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}(i). \tag{9.12}$$

估计量等于数据的移动平均。因此，根据大数法则（定理6.2.2），它在均方意义下（并且以概率1）收敛到μ，前提是iid序列中每个条目的方差σ^2是有界的。

□

Example 9.3.3 (估计平均身高)**.** 在这个例子中，我们说明了样本均值的一致性。假设我们想要估计一个人群的平均身高。为了具体说明，我们考虑一个包含$m := 25000$人的人群。图9.2展示了他们身高的直方图[1]。如示例9.1.1所解释，如果我们从这个人群中进行有放回的抽样n，那么他们的身高将形成一个独立同分布的序列$\widetilde{X}$。这个序列的均值是

$$\mathrm{E}\left(\widetilde{X}(i)\right) := \sum_{j=1}^{m} \mathrm{P}(\text{Person } j \text{ is chosen}) \cdot \text{height of person } j \tag{9.13}$$

$$= \frac{1}{m}\sum_{j=1}^{m} h_j \tag{9.14}$$

$$= \mathrm{av}(h_1, \ldots, h_m) \tag{9.15}$$

对于 $1 \leq i \leq n$，其中 h_1，...，h_m 是人们的身高。此外，由于身高是有限的，方差是有界的。根据定理 9.3.2，n 数据的样本均值应该收敛到 iid 序列的均值，从而收敛到整个群体的平均身高。图 9.3 数值上展示了这一点。

△

如果基础分布的均值没有明确定义，或者其方差是无界的，那么样本均值不一定是一个一致的估计量。这与以下事实有关

[1]The data are available here: `wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights`.

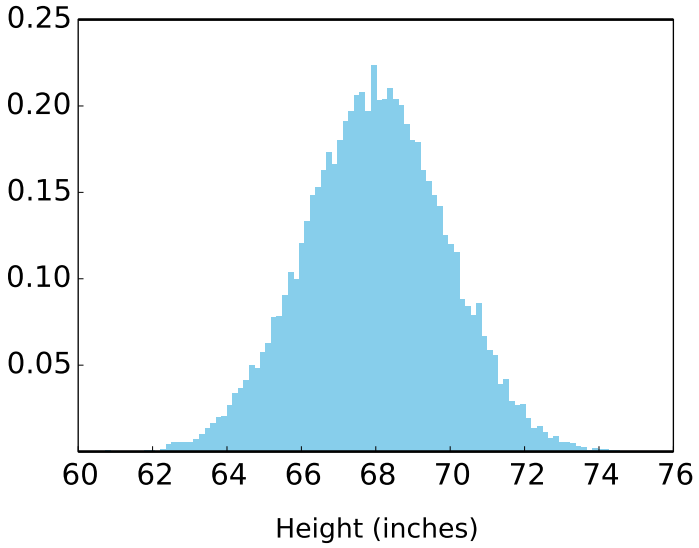


Figure 9.2: 一组25,000人身高的直方图。

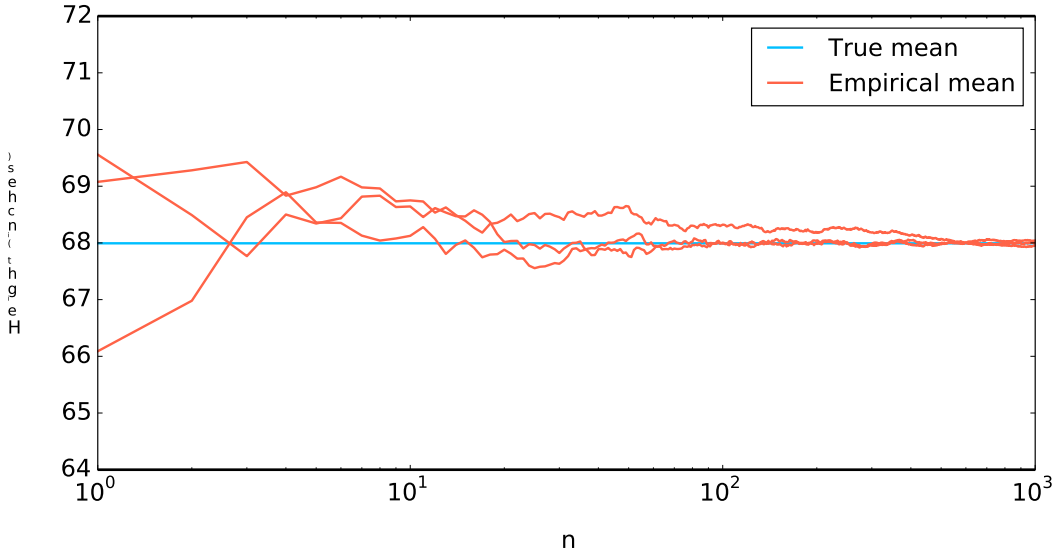


Figure 9.3: 不同的样本均值实现，当从图 9.2 中的总体中按有放回的方式抽取个体时。

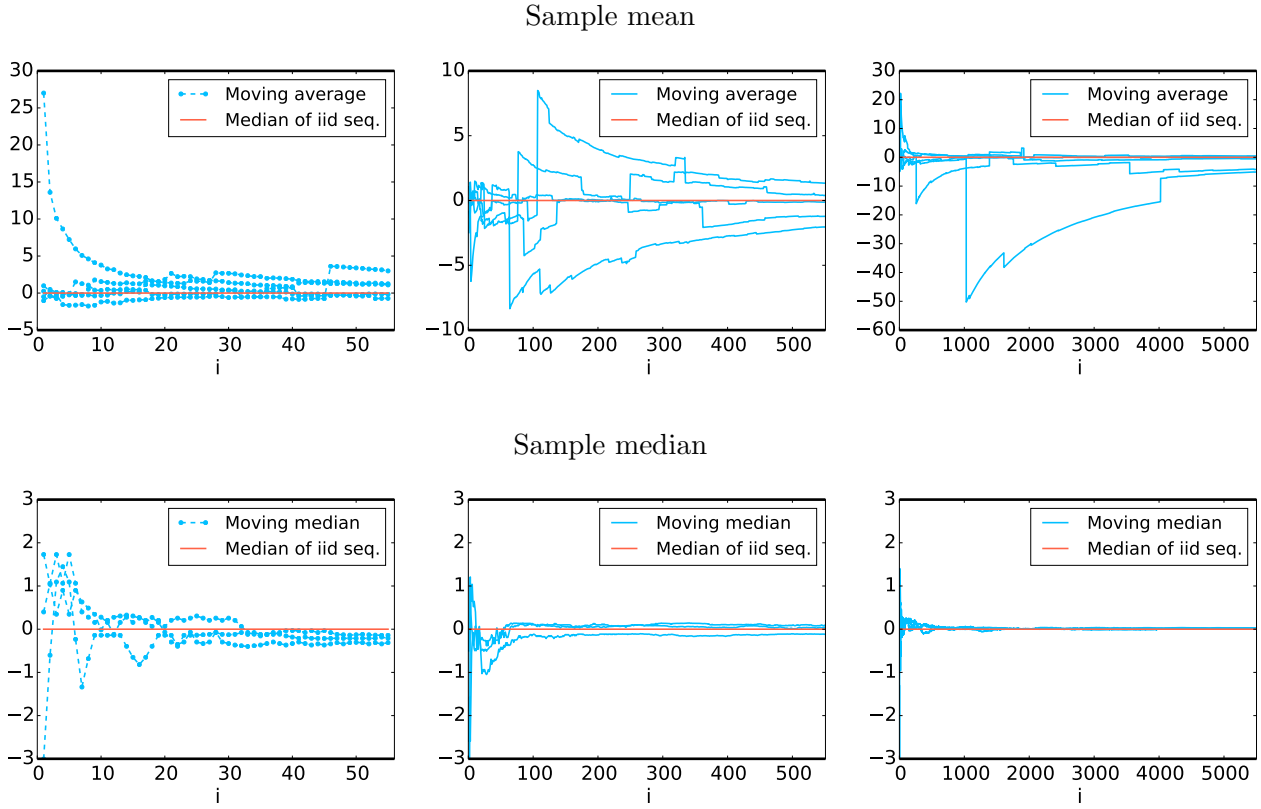


Figure 9.4: 独立同分布柯西序列的移动平均的一个实现（上），与移动中位数（下）进行比较。

正如我们在第 8.2 节中讨论的那样，样本均值可能会受到极端值存在的严重影响。相比之下，如第 8.3 节所述，样本中位数在这种情况下往往更加稳健。下面的定理表明，在独立同分布（iid）假设下，样本中位数是一致的，即使均值并未良好定义或方差是无界的。证明见第 9.7.2 节。

Theorem 9.3.4 (样本中位数作为中位数的估计量). *The sample median is a consistent estimator of the median of an iid sequence of random variables.*

图 9.4 比较了三种不同实现下，独立同分布的 Cauchy 随机变量序列的滑动平均和滑动中位数。滑动平均是不稳定的，无论有多少数据可用，都不会收敛，这并不令人惊讶，因为均值定义不明确。相比之下，滑动中位数最终会收敛到真实中位数，正如定理 9.3.4 所预测的那样。

在对底层分布的高阶矩作出某些假设的条件下，样本方差和样本协方差分别是方差和协方差的一致估计量。这为主成分分析提供了一种直观的解释（见第 8.5.2 节）：在假设数据是随机向量的一个独立同分布（iid）序列的实现的情况下，主成分近似于真实协方差矩阵的特征向量（见第 4.3.3 节），因此对应于多维分布中方差最大的方向。图 9.5

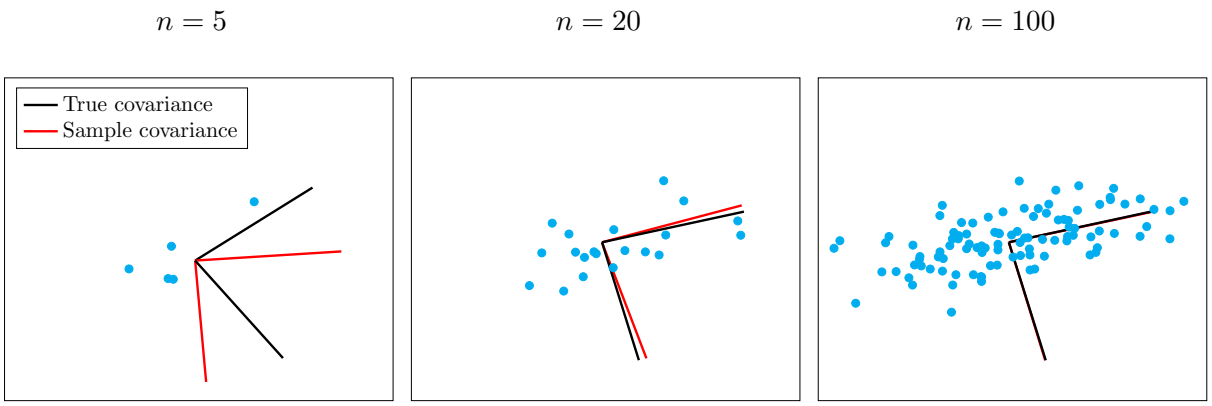


Figure 9.5: n 样本的主方向来自双变量高斯分布（红色），与分布协方差矩阵的特征向量（黑色）进行比较。

通过一个数值例子说明这一点，其中主成分确实随着数据量的增加而收敛到特征向量。

9.4 Confidence intervals

一致性意味着，如果我们获得无限的数据，估计量将是完美的，但这在实践中当然是不可能的。因此，量化固定数据量下估计量的准确性是很重要的。从频率学派的角度来看，置信区间可以用来做到这一点。置信区间可以被解释为对感兴趣的确定性量的*soft estimate*，它保证真实值以一定的概率属于该区间。

Definition 9.4.1 (置信区间)**.** *A* $1-\alpha$ *confidence interval* $\mathcal{I}$ *for* $\gamma \in \mathbb{R}$ *satisfies*

$$\mathrm{P}\left(\gamma \in \mathcal{I}\right) \geq 1-\alpha, \tag{9.16}$$

where $0 < \alpha < 1$*.*

置信区间通常呈现为 $[Y-c, Y+c]$ 形式，其中 Y 是感兴趣量的估计量，c 是一个常数，依赖于数据的数量。以下定理推导了一个关于独立同分布序列均值的置信区间。该置信区间以样本均值为中心。

Theorem 9.4.2 (独立同分布序列的均值置信区间)**.** *Let* $\widetilde{X}$ *be an iid sequence with mean* μ *and variance* $\sigma^2 \leq b^2$ *for some* $b > 0$*. For any* $0 < \alpha < 1$

$$\mathcal{I}_n := \left[Y_n - \frac{b}{\sqrt{\alpha\, n}}, Y_n + \frac{b}{\sqrt{\alpha\, n}} \right], \qquad Y_n := \mathrm{av}\left(\widetilde{X}\left(1\right), \widetilde{X}\left(2\right), \ldots, \widetilde{X}\left(n\right) \right), \tag{9.17}$$

is a $1-\alpha$ *confidence interval for* μ*.*

Proof. 请记住，Y_n 的方差等于 $\text{Var}\left(\bar{X}_n\right) = \sigma^2/n$ (，见定理 6.2.2 证明中的公式 (6.21))。我们有

$$P\left(\mu \in \left[Y_n - \frac{b}{\sqrt{\alpha\, n}}, Y_n + \frac{\sigma}{\sqrt{\alpha\, n}}\right]\right) = 1 - P\left(|Y_n - \mu| > \frac{b}{\sqrt{\alpha\, n}}\right) \tag{9.18}$$

$$\geq 1 - \frac{\alpha\, n \text{Var}\left(Y_n\right)}{b^2} \quad \text{by Chebyshev's inequality} \tag{9.19}$$

$$= 1 - \frac{\alpha\, \sigma^2}{b^2} \tag{9.20}$$

$$\geq 1 - \alpha. \tag{9.21}$$

□

定理中给定的区间宽度随着n的增大而减小，保持α不变，这很有道理，因为引入更多数据减少了估计量的方差，从而减少了我们对其的不确定性。

Example 9.4.3 (优胜美地的熊)**.** 一位科学家正试图估计优胜美地国家公园中黑熊的平均体重。她成功捕获了300只熊。我们假设这些熊是在有放回的情况下均匀随机抽样的（同一只熊可能被称重多次）。在这些假设下，在例9.1.1中我们展示了这些数据可以建模为 iid 样本，而在例9.3.3中我们表明样本均值是总体均值的一致估计量。

被捕获的300只熊的平均体重是 $Y := 200$ 磅。要从这些信息推导出一个置信区间，我们需要一个关于方差的界限。历史上记录到的黑熊最大体重为 880 磅。设 μ 和 σ^2 分别为整个总体体重的（未知）均值和方差。若 X 表示从整个总体中均匀随机选取的一只熊的体重，则 X 的均值为 μ，方差为 σ^2，因此

$$\sigma^2 = \text{E}\left(X^2\right) - \text{E}^2\left(X\right) \tag{9.22}$$

$$\leq \text{E}\left(X^2\right) \tag{9.23}$$

$$\leq 880^2 \quad \text{because } X \leq 880. \tag{9.24}$$

因此，880 是标准差的上界。应用定理 9.4.2,

$$\left[Y - \frac{b}{\sqrt{\alpha\, n}}, Y + \frac{b}{\sqrt{\alpha\, n}}\right] = [-27.2, 427.2] \tag{9.25}$$

是总体平均体重的95%置信区间。该区间不是非常精确，因为n不够大。△

如本例所示，基于切比雪夫不等式推导出的置信区间往往非常保守。一种替代方法是利用中心极限定理（CLT）。CLT 从渐近意义上刻画了样本均值的分布，因此从中推导出的置信区间不一定能保证精确。然而，CLT 通常能为有限 n 提供非常准确的样本均值分布近似，正如我们在第六章的一些数值例子中所展示的那样。为了根据定理 6.3.1 从 CLT 获得 iid 序列均值的置信区间，我们需要知道真实的方差。

该序列在实践中不切实际。然而，以下结果表明我们可以用样本方差替代真实方差。证明超出了这些笔记的范围。

Theorem 9.4.4 (中心极限定理与样本标准差)**.** *Let $\widetilde{X}$ be an iid discrete random process with mean $\mu_{\widetilde{X}} := \mu$ such that its variance and fourth moment $\mathrm{E}(\widetilde{X}(i)^4)$ are bounded. The sequence*

$$\frac{\sqrt{n}\left(\mathrm{av}\left(\widetilde{X}(1),\ldots,\widetilde{X}(n)\right)-\mu\right)}{\mathrm{std}\left(\widetilde{X}(1),\ldots,\widetilde{X}(n)\right)} \tag{9.26}$$

converges in distribution to a standard Gaussian random variable.

请记住，标准高斯分布的累积分布函数没有封闭形式的表达式。为了简化符号，我们将置信区间表示为Q函数。

Definition 9.4.5 (Q函数)**.** *$Q(x)$ is the probability that a standard Gaussian random variable is greater than x for positive x,*

$$Q(x) := \int_{u=x}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du, \quad x > 0. \tag{9.27}$$

By symmetry, if U is a standard Gaussian random variable and $y < 0$

$$\mathrm{P}(U < y) = Q(-y). \tag{9.28}$$

Corollary 9.4.6 (均值)**.** *Let $\widetilde{X}$ be an iid sequence that satisfies the conditions of Theorem 9.4.4. For any $0 < \alpha < 1$*的近似置信区间

$$\mathcal{I}_n := \left[Y_n - \frac{S_n}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right), Y_n + \frac{S_n}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right)\right], \tag{9.29}$$

$$Y_n := \mathrm{av}\left(\widetilde{X}(1),\widetilde{X}(2),\ldots,\widetilde{X}(n)\right), \tag{9.30}$$

$$S_n := \mathrm{std}\left(\widetilde{X}(1),\widetilde{X}(2),\ldots,\widetilde{X}(n)\right), \tag{9.31}$$

is an approximate $1-\alpha$ confidence interval for μ, i.e.

$$P(\mu \in \mathcal{I}_n) \approx 1 - \alpha. \tag{9.32}$$

Proof. 根据中心极限定理，当 $n \to \infty$ $\bar{X}_n$ 服从均值为 μ、方差为 σ^2 的高斯随机变量分布时。因此

$$P(\mu \in \mathcal{I}_n) = 1 - P\left(Y_n > \mu + \frac{S_n}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right)\right) - P\left(Y_n < \mu - \frac{S_n}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right)\right) \tag{9.33}$$

$$= 1 - P\left(\frac{\sqrt{n}(Y_n-\mu)}{S_n} > Q^{-1}\left(\frac{\alpha}{2}\right)\right) - P\left(\frac{\sqrt{n}(Y_n-\mu)}{S_n} < -Q^{-1}\left(\frac{\alpha}{2}\right)\right) \tag{9.34}$$

$$\approx 1 - 2Q\left(Q^{-1}\left(\frac{\alpha}{2}\right)\right) \quad \text{by Theorem 9.4.4} \tag{9.35}$$

$$= 1 - \alpha. \tag{9.36}$$

□

需要强调的是，只有当 n 足够大，使样本方差收敛到真实方差并且中心极限定理（CLT）开始生效时，该结果才会提供准确的置信区间。

Example 9.4.7 (优胜美地的熊（续））**.** 科学家捕获的熊的样本标准差为100磅。我们应用推论9.4.6推导出一个比应用切比雪夫不等式得到的置信区间更紧的近似值。鉴于$Q\left(1.95\right) \approx 0.025$,

$$\left[Y - \frac{\sigma}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right), Y + \frac{\sigma}{\sqrt{n}} Q^{-1}\left(\frac{\alpha}{2}\right)\right] \approx [188.8, 211.3] \tag{9.37}$$

是熊群体平均体重的约95%置信区间。

△

解释置信区间有些棘手。在计算了例子 9.4.7 中的置信区间后，人们很容易产生以下结论:

The probability that the average weight is between 188.8 and 211.3 lbs is 0.95.

然而，我们将平均体重建模为一个确定性量，因此在这个表述中不存在任何随机量！正确的解释是：如果我们反复进行从总体中抽样并计算置信区间的过程，那么真实值将在 95% 的次数中落在该区间内。下面的示例和图 9.6 对此作了说明。

Example 9.4.8 (估计平均身高（续））**.** 图9.6显示了示例9.3.3中身高总体平均值的几个95%置信区间。为了计算每个区间，我们选择n个个体，然后应用推论9.4.6。随着n增大，区间的宽度变小，但因为它们都是95%置信区间，所以它们都以0.95的概率包含真实的平均值。实际上，这对120个区间中有113个（94%）是成立的。

△

9.5 Nonparametric model estimation

在本节中，我们考虑从多个独立同分布（iid）样本估计分布的问题。这需要近似分布的累积分布函数（cdf）、概率质量函数（pmf）或概率密度函数（pdf）。如果我们假设分布属于一个预定义的家族，那么问题就转化为估计描述该家族的参数，正如我们在9.6节中详细解释的那样。在这里，我们不做这样的假设。直接估计一个分布是非常具有挑战性的；显然，许多（无限多！）不同的分布都可能生成这些数据。然而，只要样本足够，通常可以获得能够产生准确近似的模型，前提是iid假设成立。

9.5.1 Empirical cdf

在假设一个数据集对应于来自某一分布的独立同分布（iid）样本的前提下，在给定点 x 处对该分布的累积分布函数（cdf）的一个合理估计，是小于 x 的样本所占的比例。这由此得到一个分段常数的估计量，称为经验累积分布函数。

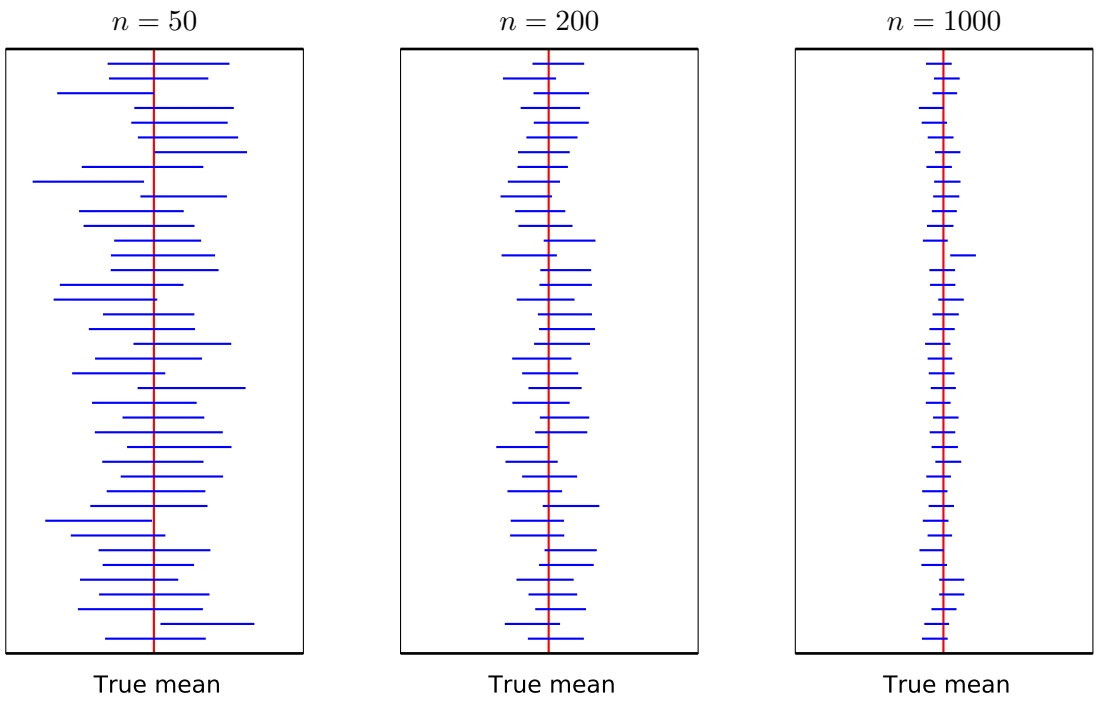


Figure 9.6: 95%置信区间用于示例9.3.3中身高人群的平均值。

Definition 9.5.1 (经验累积分布函数). *The empirical cdf corresponding to data* $x_1, \ldots, x_n$ *is*

$$\widehat{F}_n(x) := \frac{1}{n}\sum_{i=1}^{n} 1_{x_i \le x}, \tag{9.38}$$

where $x \in \mathbb{R}$.

经验累积分布函数（cdf）是真实累积分布函数的无偏和一致估计量。这个结果在下面的定理 9.5.2 中得到了严格证明，并在图 9.7 中通过实证例子进行了说明。来自25,000人身高数据的累积分布函数与从不同数量的独立同分布（iid）样本计算得到的三种经验累积分布函数的实现进行了比较。随着可用样本数量的增加，近似变得非常准确。

Theorem 9.5.2. *Let* $\widetilde{X}$ *be an iid sequence with marginal cdf* F_X. *For any fixed* $x \in \mathbb{R}$ $\widehat{F}_n(x)$ *is an unbiased and consistent estimator of* $F_X(x)$. *In fact,* $\widehat{F}_n(x)$ *converges in mean square to* $F_X(x)$.

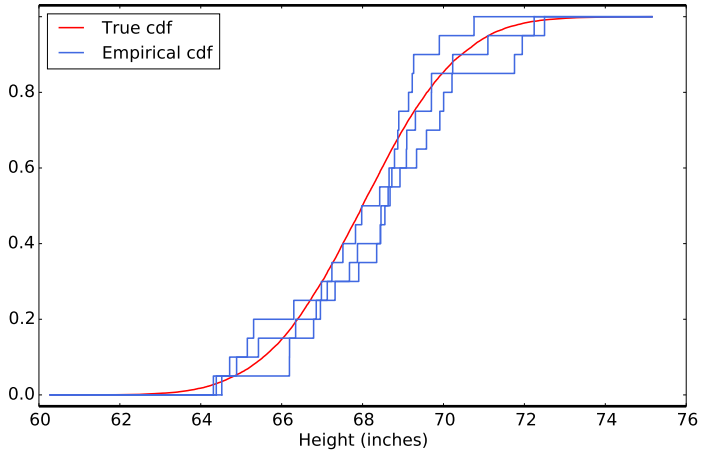
Proof. 首先，我们验证

$$\mathrm{E}\left(\widehat{F}_n(x)\right) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} 1_{\widetilde{X}(i)\le x}\right) \tag{9.39}$$

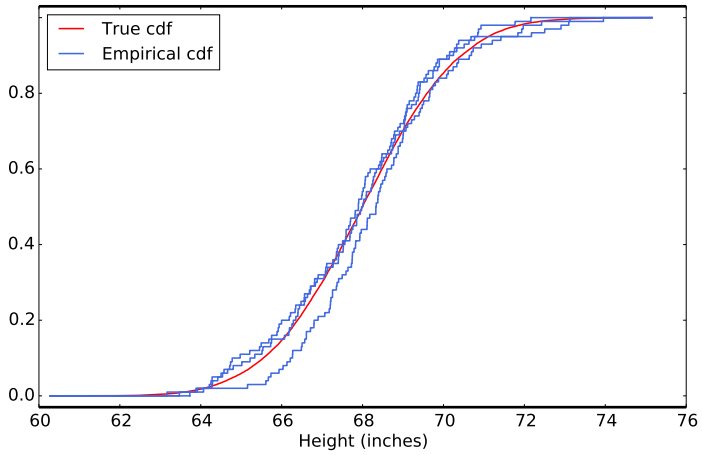
$$= \frac{1}{n}\sum_{i=1}^{n} \mathrm{P}\left(\widetilde{X}(i) \le x\right) \quad \text{by linearity of expectation} \tag{9.40}$$

$$= F_X(x), \tag{9.41}$$

$n = 10$



$n = 100$



$n = 1000$

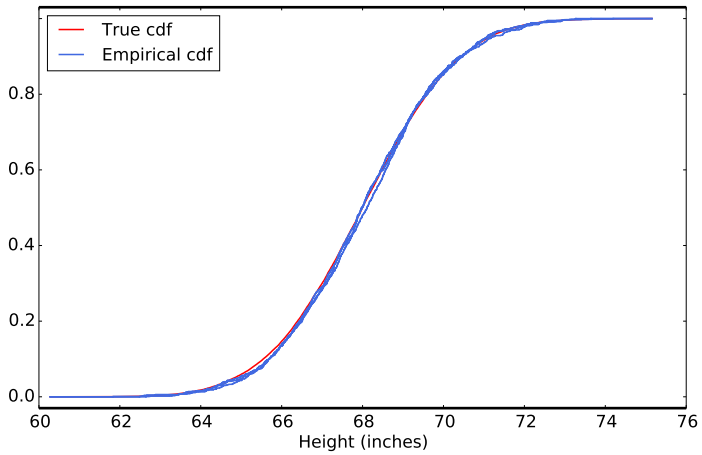


Figure 9.7: 图2.13中高度数据的累积分布函数（Cdf），以及通过n独立同分布（iid）样本计算的经验累积分布函数的三次实现，样本量为$n = 10, 100, 1000$。

所以估计量是无偏的。我们现在估计它的均方

$$\mathrm{E}\left(\widehat{F}_n^2(x)\right) = \mathrm{E}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} 1_{\widetilde{X}(i)\leq x} 1_{\widetilde{X}(j)\leq x}\right) \tag{9.42}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{P}\left(\widetilde{X}(i)\leq x\right) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}\mathrm{P}\left(\widetilde{X}(i)\leq x, \widetilde{X}(j)\leq x\right) \tag{9.43}$$

$$= \frac{F_X(x)}{n} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} F_{\widetilde{X}(i)}(x) F_{\widetilde{X}(j)}(x) \quad \text{by independence} \tag{9.44}$$

$$= \frac{F_X(x)}{n} + \frac{n-1}{n} F_X(x)^2 \tag{9.45}$$

因此，方差等于

$$\mathrm{Var}\left(\widehat{F}_n(x)\right) = \mathrm{E}\left(\widehat{F}_n(x)^2\right) - \mathrm{E}^2\left(\widehat{F}_n(x)\right) \tag{9.46}$$

$$= \frac{F_X(x)(1-F_X(x))}{n}. \tag{9.47}$$

我们得出结论，

$$\lim_{n\to\infty}\mathrm{E}\left(\left(F_X(x)-\widehat{F}_n(x)\right)^2\right) = \lim_{n\to\infty}\mathrm{Var}\left(\widehat{F}_n(x)\right) = 0. \tag{9.48}$$

□

9.5.2 Density estimation

估计连续变量的 pdf 比估计 cdf 要困难得多。如果我们有足够的数据，小于某个 x 的样本比例可以很好地估计该点的 cdf。然而，无论我们有多少数据，几乎不可能恰好在 x 处看到任何样本：逐点的经验密度估计在几乎所有地方都会等于零（除了已有样本处）。

我们唯一能够得到准确估计量的希望是，如果我们要估计的pdf是平滑的。在这种情况下，我们可以通过位于邻近位置的观察样本来估计在某一点 x 处的值。如果有许多样本接近 x，那么这表明在 x 处的估计应该较大，而如果所有样本都很远，那么估计应该较小。**Kernel density estimation** 通过对样本进行平均来实现这一点。

Definition 9.5.3 (核密度估计器)**.** *The kernel density estimate with bandwidth h of the distribution of $x_1, \ldots, x_n$ at $x \in \mathbb{R}$ is*

$$\widehat{f}_{h,n}(x) := \frac{1}{n\,h}\sum_{i=1}^{n} k\left(\frac{x-x_i}{h}\right), \tag{9.49}$$

where k is a kernel function centered at the origin that satisfies

$$k(x) \geq 0 \quad \textit{for all } x \in \mathbb{R}, \tag{9.50}$$

$$\int_{\mathbb{R}} k(x)\, dx = 1. \tag{9.51}$$

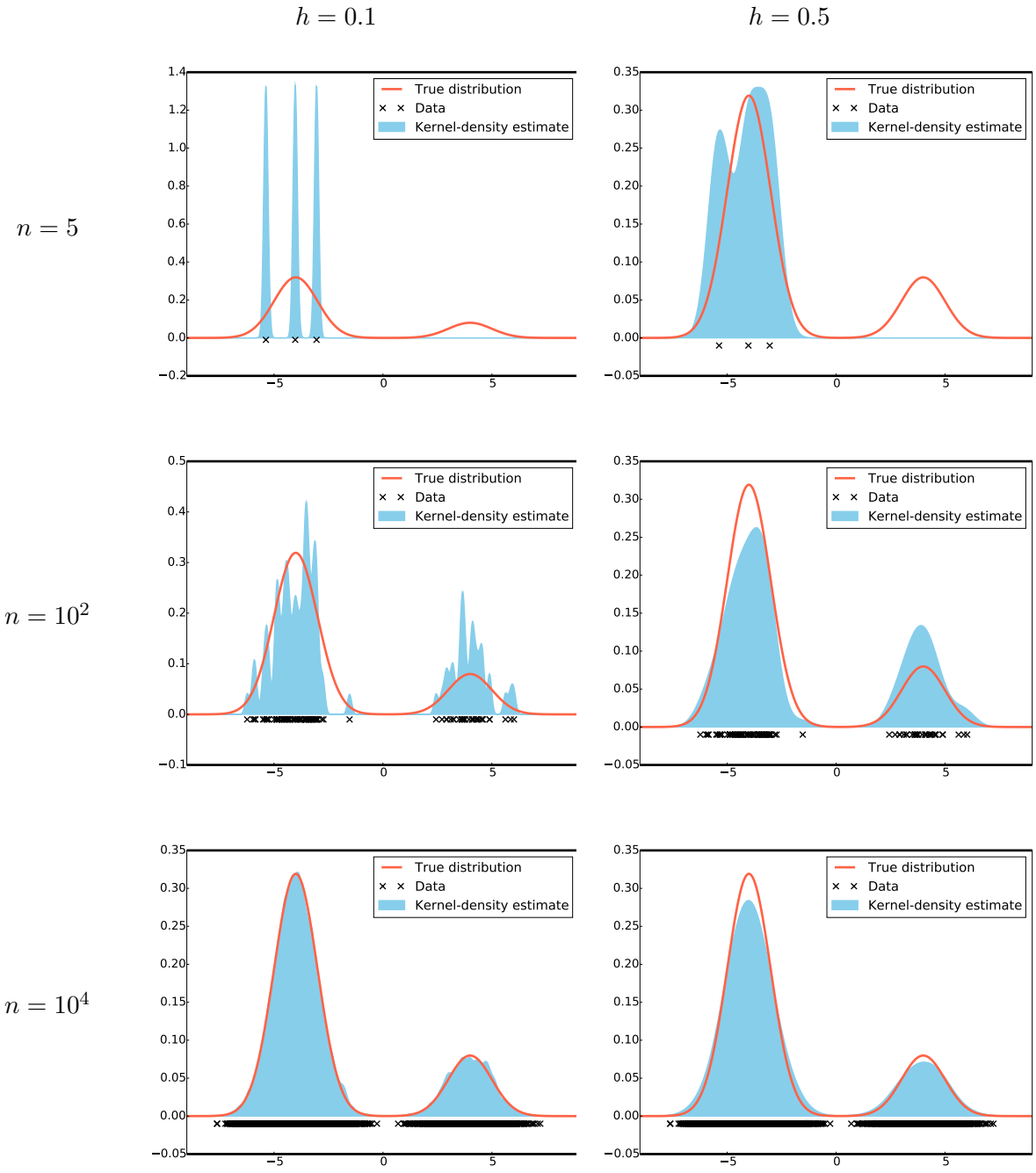


Figure 9.8: 高斯混合的核密度估计，如示例9.6.5所述，适用于不同数量的独立同分布样本以及不同的核带宽值h。

核函数的作用是根据每个样本与我们估计概率密度函数（pdf）x的点之间的距离对样本进行加权。选择矩形核会产生一个经验密度估计，该估计是分段常数的，粗略看起来像一个直方图（对应的权重是常数或等于零）。一个流行的替代方案是高斯核$k(x) = \exp\left(-x^2\right)/\sqrt{\pi}$，它生成一个平滑的密度估计。核函数应该衰减，使得当样本x_i接近x时$k\left(\left(x - x_i\right)/h\right)$较大，而当其远离时$k\left(\left(x - x_i\right)/h\right)$较小。这种衰减由带宽$h$控制，带宽是在事先根据我们对pdf平滑度的期望以及可用数据量来选择的。如果带宽非常小，个别样本对密度估计有较大影响。这允许更容易地重现不规则形状，但也会导致与真实曲线不符的虚假波动，特别是在样本量较少时。增加带宽可以平滑这些波动，并在数据量较少时提供更稳定的估计。然而，它也可能会过度平滑估计。作为经验法则，随着数据量的增加，我们应该减小核函数的带宽。

图9.8和图9.9展示了在不同采样率下改变带宽h的影响。图9.8中，采用高斯核密度估计来估计示例9.6.5中描述的高斯混合模型。图9.9给出了将同一技术用于真实数据的一个例子：目标是估计海螺种群体重的密度。[2]整个种群由4,177个个体组成。核密度估计基于200个独立同分布（iid）样本，并在不同的核带宽取值下计算得到。

9.6 Parametric model estimation

在前一节中，我们描述了如何通过直接估计生成数据的累积分布函数（cdf）或概率密度函数（pdf）来估计分布。在本节中，我们讨论了另一种基于已知生成数据的分布类型的假设的途径。如果是这种情况，那么问题就简化为将描述分布的参数拟合到数据上。回想一下，从频率学派的观点来看，真实分布是固定的，因此相应的参数被建模为确定性量（与此相对，在贝叶斯框架中，它们被建模为随机变量）。

9.6.1 The method of moments

矩量法调整分布的参数，使得分布的矩与数据的样本矩一致（即其均值、均方或方差等）。如果分布仅依赖于一个参数，则我们使用样本均值作为真实均值的代理，并计算相应的参数值。对于参数为λ且均值为μ的指数分布，我们有

$$\mu = \frac{1}{\lambda}. \tag{9.52}$$

假设我们有来自指数分布的 n 个独立同分布样本 $x_1, \ldots, x_n$，则 λ 的矩估计为

$$\lambda_{\mathrm{MM}} := \frac{1}{\mathrm{av}\left(x_1, \ldots, x_n\right)}. \tag{9.53}$$

[2]The data are available at `archive.ics.uci.edu/ml/datasets/Abalone`

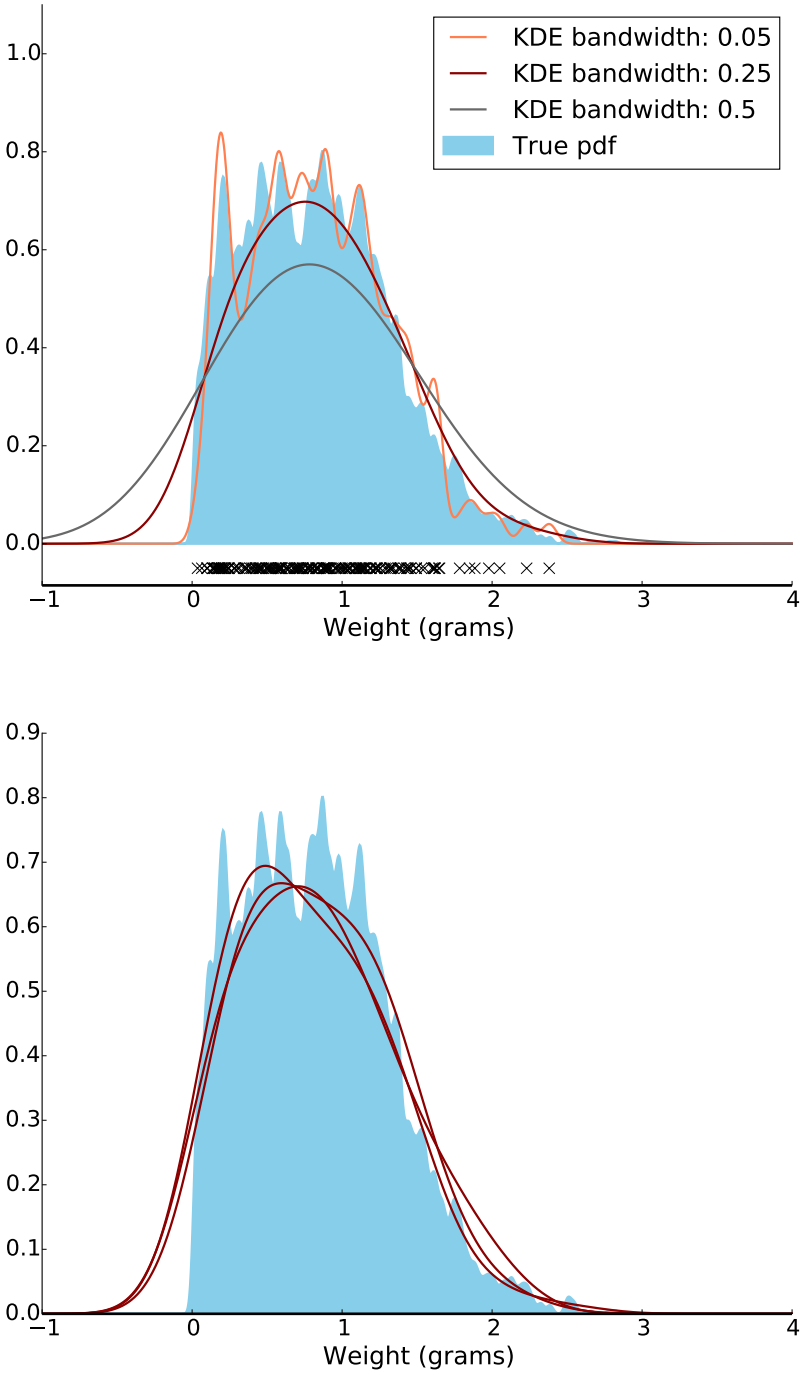


Figure 9.9: 鲍鱼（一种海螺）种群重量的核密度估计。在上图中，密度是基于 200 个独立同分布样本，使用具有三种不同带宽的高斯核进行估计的。下方显示的黑色叉号表示各个单独的样本。在下图中，我们可以看到在带宽固定为 0.25 的情况下，将该过程重复三次的结果。

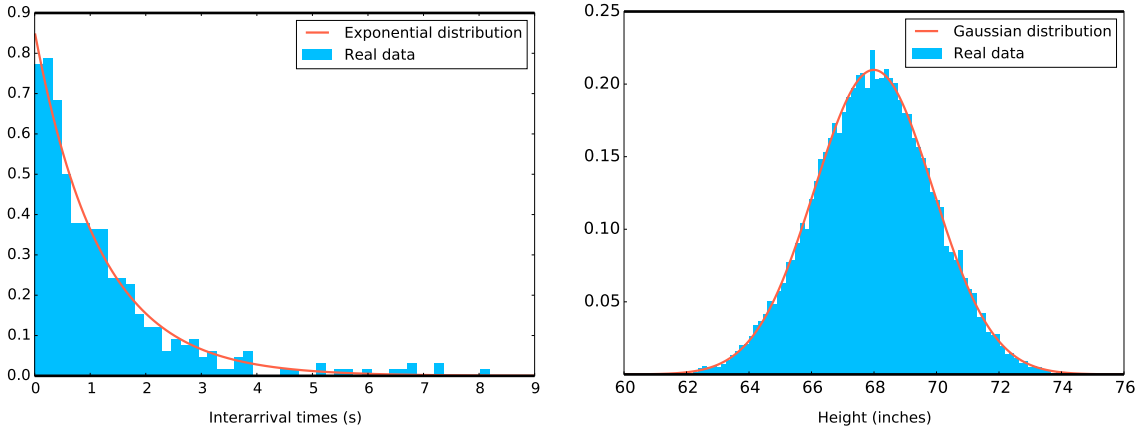


Figure 9.10: 指数分布拟合以色列一家呼叫中心来电到达间隔时间的数据（左）。高斯分布拟合身高数据（右）。

图9.10右侧的图展示了将指数分布拟合到图2.11中的呼叫中心数据所得到的结果。类似地，使用矩方法拟合高斯分布时，我们将均值设为其样本均值，将方差设为样本方差；如图9.10右侧所示，该图使用了图2.13中的数据进行说明。

9.6.2 Maximum likelihood

学习参数化模型最常用的方法是最大似然拟合。**likelihood** 函数是数据的联合 pmf 或 pdf，被解释为一个 *function of the unknown parameters*。更具体地说，我们用 $x_1, \ldots, x_n$ 表示数据，并假设它们是一组离散随机变量 $X_1, \ldots, X_n$ 的实现，这些随机变量具有一个依赖于参数向量 $\vec{\theta}$ 的联合 pmf。为了强调联合 pmf 依赖于 $\vec{\theta}$，我们将其记为 $p_{\vec{\theta}} := p_{X_1,\ldots,X_n}$。在观测数据处求值的该 pmf

$$p_{\vec{\theta}}(x_1, \ldots, x_n) \tag{9.54}$$

当我们将其解释为 $\vec{\theta}$ 的函数时，它就是似然函数。对于连续随机变量，我们则使用数据的联合概率密度函数（pdf）。

Definition 9.6.1 (似然函数)**.** *Given a realization $x_1, \ldots, x_n$ of a set of discrete random variables $X_1, \ldots, X_n$ with joint pmf $p_{\vec{\theta}}$, where $\vec{\theta} \in \mathbb{R}^m$ is a vector of parameters, the likelihood function is*

$$\mathcal{L}_{x_1,\ldots,x_n}\left(\vec{\theta}\right) := p_{\vec{\theta}}(x_1, \ldots, x_n)\,. \tag{9.55}$$

If the random variables are continuous with pdf $f_{\vec{\theta}}$, where $\vec{\theta} \in \mathbb{R}^m$, the likelihood function is

$$\mathcal{L}_{x_1,\ldots,x_n}\left(\vec{\theta}\right) := f_{\vec{\theta}}(x_1, \ldots, x_n)\,. \tag{9.56}$$

*The **log-likelihood function** is equal to the logarithm of the likelihood function* $\log \mathcal{L}_{x_1,\ldots,x_n}\left(\vec{\theta}\right)$.

当数据被建模为独立同分布（iid）样本时，似然函数可以分解为边际pmf或pdf的乘积，因此对数似然可以分解为一个和。

在离散分布的情况下，对于固定的 $\vec{\theta}$，似然是 $X_1, \ldots, X_n$ 等于观测数据的概率。如果我们不知道 $\vec{\theta}$，那么选择一个使该概率尽可能大的 $\vec{\theta}$ 是有意义的，即最大化似然。对于连续分布，我们将同样的原理应用于数据的联合概率密度函数。

Definition 9.6.2 (最大似然估计量)**.** *The* ***maximum likelihood*** *(ML) estimator for the vector of parameters $\vec{\theta} \in \mathbb{R}^m$ is*

$$\vec{\theta}_{\mathrm{ML}}(x_1, \ldots, x_n) := \arg\max_{\vec{\theta}} \mathcal{L}_{x_1,\ldots,x_n}\left(\vec{\theta}\right) \tag{9.57}$$

$$= \arg\max_{\vec{\theta}} \log \mathcal{L}_{x_1,\ldots,x_n}\left(\vec{\theta}\right). \tag{9.58}$$

The maximum of the likelihood function and that of the log-likelihood function are at the same location because the logarithm is monotone.

在某些条件下，可以证明最大似然估计量是一致的：随着数据量的增加，它会以概率收敛到真实参数。还可以证明它的分布会收敛到高斯随机变量（或向量）的分布，就像样本均值的分布一样。这些结果超出了本课程的范围。然而，请记住，这些结果仅在数据确实是由我们考虑的分布类型生成时成立。

我们现在展示如何推导伯努利分布和高斯分布的最大似然估计。所得的参数估计量与矩方法估计量相同（除了高斯方差参数估计上有一些细微差异）。

Example 9.6.3 (伯努利分布)**.**参数的ML估计器 我们将一组数据$x_1, \ldots, x_n$建模为来自伯努利分布的独立同分布样本，参数为θ (，在这种情况下只有一个参数)。似然函数等于

$$\mathcal{L}_{x_1,\ldots,x_n}(\theta) = p_\theta(x_1, \ldots, x_n) \tag{9.59}$$

$$= \prod_{i=1} \left(1_{x_i=1}\theta + 1_{x_i=0}(1-\theta)\right) \tag{9.60}$$

$$= \theta^{n_1}(1-\theta)^{n_0} \tag{9.61}$$

并且对数似然函数为

$$\log \mathcal{L}_{x_1,\ldots,x_n}(\theta) = n_1 \log\theta + n_0 \log(1-\theta), \tag{9.62}$$

其中 n_1 是等于一的样本数量，n_0 是等于零的样本数量。参数 θ 的最大似然估计量是

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} \log \mathcal{L}_{x_1,\ldots,x_n}(\theta) \tag{9.63}$$

$$= \arg\max_{\theta} n_1 \log\theta + n_0 \log(1-\theta). \tag{9.64}$$

我们计算对数似然函数的一阶导数和二阶导数,

$$\frac{\mathrm{d}\log\mathcal{L}_{x_1,\ldots,x_n}(\theta)}{\mathrm{d}\theta}=\frac{n_1}{\theta}-\frac{n_0}{1-\theta}, \tag{9.65}$$

$$\frac{\mathrm{d}^2\log\mathcal{L}_{x_1,\ldots,x_n}(\theta)}{\mathrm{d}\theta^2}=-\frac{n_1}{\theta^2}-\frac{n_0}{(1-\theta)^2}<0. \tag{9.66}$$

该函数是凹的，因为二阶导数为负。因此，最大值出现在一阶导数等于零的点，即

$$\theta_{\mathrm{ML}}=\frac{n_1}{n_0+n_1}. \tag{9.67}$$

该估计等于取值为一的样本所占的比例。

△

Example 9.6.4 (高斯分布)**.**参数的最大似然估计器 设$x_1, x_2, \ldots$为我们希望建模的独立同分布样本数据，来自均值为μ、标准差为σ的高斯分布。似然函数等于

$$\mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma)=f_{\mu,\sigma}(x_1,\ldots,x_n) \tag{9.68}$$

$$=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{9.69}$$

并且对数似然函数为

$$\log\mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma)=-\frac{n\log(2\pi)}{2}-n\log\sigma-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}. \tag{9.70}$$

参数μ和σ的ML估计量为

$$\{\mu_{\mathrm{ML}},\sigma_{\mathrm{ML}}\}=\arg\max_{\{\mu,\sigma\}}\log\mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma) \tag{9.71}$$

$$=\arg\max_{\{\mu,\sigma\}}-n\log\sigma-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}. \tag{9.72}$$

我们计算对数似然函数的偏导数,

$$\frac{\partial\log\mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma)}{\partial\mu}=-\sum_{i=1}^{n}\frac{x_i-\mu}{\sigma^2}, \tag{9.73}$$

$$\frac{\partial\log\mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma)}{\partial\sigma}=-\frac{n}{\sigma}+\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^3}. \tag{9.74}$$

我们试图最大化的函数在 $\{\mu,\sigma\}$ 上是严格凹的。为了证明这一点，我们需要证明该函数的 Hessian 矩阵是正定的。我们省略了表明情况如此的计算。将偏导数设为零，我们得到

$$\mu_{\mathrm{ML}}=\frac{1}{n}\sum_{i=1}^{n}x_i, \tag{9.75}$$

$$\sigma^2_{\mathrm{ML}}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_{\mathrm{ML}})^2. \tag{9.76}$$

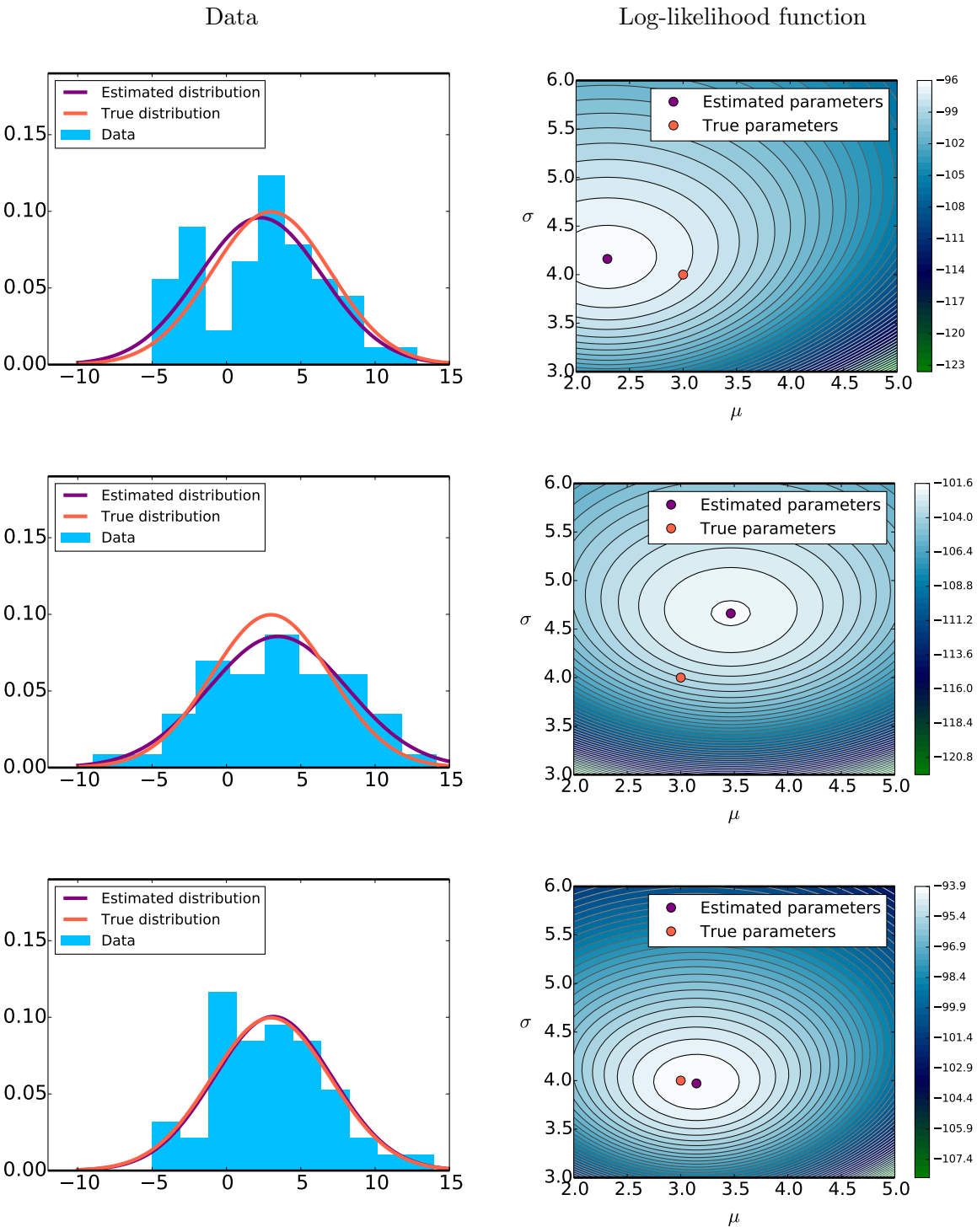


Figure 9.11: 左列显示了来自高斯分布的 50 个独立同分布（iid）样本的直方图，以及原始分布的概率密度函数（pdf）和最大似然估计。右列显示了与数据对应的对数似然函数，以及其最大值的位置和与真实参数对应的点。

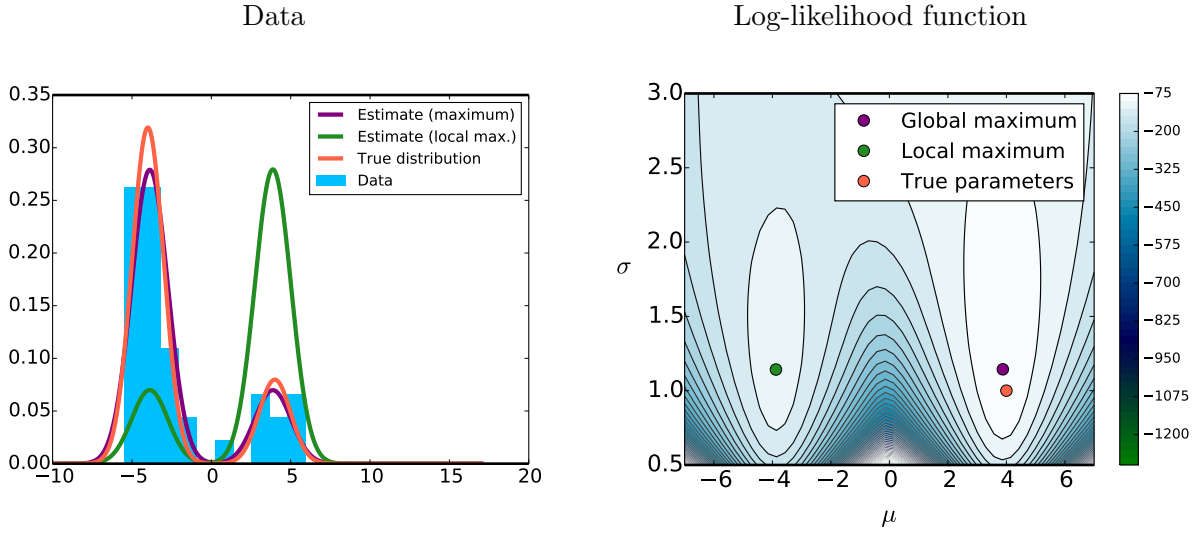


Figure 9.12: 左图显示了来自示例 9.6.5 中定义的高斯混合分布的 40 个独立同分布样本的直方图，以及原始分布的概率密度函数（pdf）。右图显示了与数据对应的对数似然函数，除了全局最大值外，它还有一个局部最大值。对应于这两个最大值的密度估计显示在左侧。

均值的估计量就是样本均值。方差的估计量是经过重新缩放的样本方差。

△

图9.11展示了对应于来自一个高斯分布的50个iid样本的对数似然函数，其中$\mu := 3$，$\sigma := 4$。它还展示了通过最大似然获得的对真实pdf的近似。在示例9.6.3和9.6.4中，对数似然函数是严格凹的。这意味着该函数具有唯一的最大值，可以通过将梯度设为零来定位。当这会产生无法直接求解的非线性方程时，我们可以利用诸如梯度上升之类的优化方法，它们将收敛到最大值。然而，对数似然函数并不总是凹的。如下例所示，在这种情况下它可能具有多个局部极大值，这可能使得计算最大似然估计量变得难以处理。

Example 9.6.5 (高斯混合).的对数似然函数 设X为一个定义为的高斯混合

$$X := \begin{cases} G_1 & \text{with probability } \frac{1}{5}, \\ G_2 & \text{with probability } \frac{4}{5}, \end{cases} \tag{9.77}$$

其中 G_1 是均值为 $-\mu$ 方差为 σ^2 的高斯随机变量，而 G_2 也是高斯分布，均值为 μ 方差为 σ^2。我们仅用两个参数对混合模型进行了参数化，以便我们能够在二维中可视化对数似然。设 $x_1, x_2, \ldots$ 为数据。

被建模为来自 X 的独立同分布（i.i.d.）样本。似然函数等于

$$\mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma) = f_{\mu,\sigma}(x_1,\ldots,x_n) \tag{9.78}$$

$$= \prod_{i=1}^{n} \frac{1}{5\sqrt{2\pi}\sigma} e^{-\frac{(x_i+\mu)^2}{2\sigma^2}} + \frac{4}{5\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{9.79}$$

并且对数似然函数为

$$\log \mathcal{L}_{x_1,\ldots,x_n}(\mu,\sigma) = \sum_{i=1}^{n} \log\left(\frac{1}{5\sqrt{2\pi}\sigma} e^{-\frac{(x_i+\mu)^2}{2\sigma^2}} + \frac{4}{5\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right). \tag{9.80}$$

图 9.12 展示了当 $\mu := 4$ 且 $\sigma := 1$ 时，该分布的 40 个 iid 样本的对数似然函数。该函数在全局最大值之外还存在一个局部最大值。这意味着如果我们使用局部上升方法来寻找 ML 估计量，可能无法找到全局最大值，而是会停留在局部最大值处。对应于局部最大值的估计（左图所示）与全局最大值具有相同的方差，但 μ 接近 -4 而不是 4。尽管该估计对数据的拟合并不理想，但它在局部是最优的，对 μ 和 σ 进行微小的偏移都会导致更差的拟合（就似然而言）。

△

作为本节的结尾，我们描述一种基于参数化拟合并采用最大似然（ML）估计的监督学习机器学习算法。

Example 9.6.6 (二次判别分析)**.** 二次判别分析是一种监督学习算法。该算法的输入是两组训练数据，分别由属于两个不同类别的d维向量$\vec{a}_1,\ldots,\vec{a}_n$和$\vec{b}_1,\ldots,\vec{b}_n$组成（该方法可以很容易地扩展到处理更多类别）。目标是根据数据的结构对新的实例进行分类。

为了进行二次判别分析，我们首先使用均值和协方差矩阵的最大似然（ML）估计器，为每个类别的数据拟合一个 d 维高斯分布；该估计器对应于训练数据的样本均值和样本协方差矩阵（样本协方差仅存在轻微的重缩放）。更具体地，$\vec{a}_1,\ldots,\vec{a}_n$ 用于估计均值 $\vec{\mu}_a$ 和协方差矩阵 Σ_a，而 $\vec{b}_1,\ldots,\vec{b}_n$ 用于估计 $\vec{\mu}_b$ 和 Σ_b,

$$\{\vec{\mu}_a,\Sigma_a\} := \arg\max_{\vec{\mu},\Sigma} \mathcal{L}_{\vec{a}_1,\ldots,\vec{a}_n}(\vec{\mu},\Sigma), \tag{9.81}$$

$$\{\vec{\mu}_b,\Sigma_b\} := \arg\max_{\vec{\mu},\Sigma} \mathcal{L}_{\vec{b}_1,\ldots,\vec{b}_n}(\vec{\mu},\Sigma). \tag{9.82}$$

然后，对于每个新的示例 $\vec{x}$，评估该示例在两个类别下的密度函数值。如果

$$f_{\vec{\mu}_a,\Sigma_a}(\vec{x}) > f_{\vec{\mu}_b,\Sigma_b}(\vec{x}) \tag{9.83}$$

则判定 $\vec{x}$ 属于第一类，否则判定其属于第二类。图9.13 显示了将该方法应用于由两个高斯分布模拟的数据所得的结果。

△

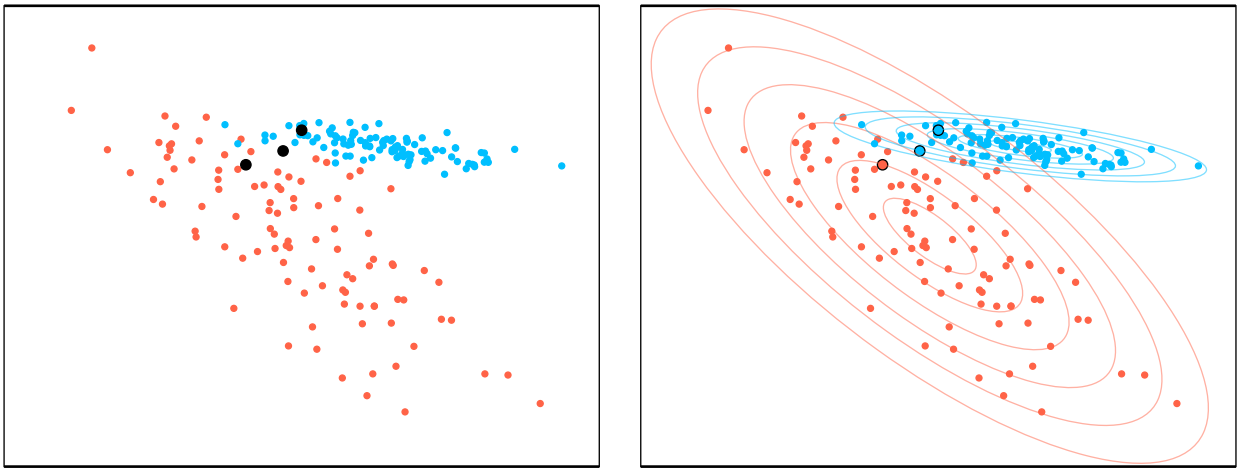


Figure 9.13: 将二次判别分析应用于来自两个不同类别的数据（左）。对应于这两个不同类别的数据分别用橙色和蓝色表示。三个新的样本用黑色表示。对数据拟合了两个二维高斯分布。它们的等高线在右侧以各自类别的颜色显示。利用这些分布对新样本进行分类，并根据其估计的类别着色。

9.7 Proofs

9.7.1 Proof of Lemma 9.2.5

我们考虑一个独立同分布序列$\widetilde{X}$的样本方差，均值为μ，方差为σ^2，

$$\widetilde{Y}(n) := \frac{1}{n-1}\sum_{i=1}^{n}\left(\widetilde{X}(i) - \frac{1}{n}\sum_{j=1}^{1}\widetilde{X}(j)\right) \tag{9.84}$$

$$= \frac{1}{n-1}\left(\widetilde{X}(i) - \frac{1}{n}\sum_{j=1}^{n}\widetilde{X}(j)\right)^2 \tag{9.85}$$

$$= \frac{1}{n-1}\left(\widetilde{X}(i)^2 + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\widetilde{X}(j)\widetilde{X}(k) - \frac{2}{n}\sum_{j=1}^{n}\widetilde{X}(i)\widetilde{X}(j)\right) \tag{9.86}$$

为简化记号，我们用 ξ 表示均方 $\mathrm{E}\left(\widetilde{X}(i)^2\right) = \mu^2 + \sigma^2$。我们有

$$\mathrm{E}\left(\widetilde{Y}(n)\right) = \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{E}\left(\widetilde{X}(i)^2\right) + \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{E}\left(\widetilde{X}(j)^2\right) + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}\mathrm{E}\left(\widetilde{X}(j)\widetilde{X}(k)\right) \tag{9.87}$$

$$-\frac{2}{n}\mathrm{E}\left(\widetilde{X}(i)^2\right) - \frac{2}{n}\sum_{\substack{j=1\\j\neq i}}^{n}\mathrm{E}\left(\widetilde{X}(i)\widetilde{X}(j)\right) \tag{9.88}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\xi + \frac{n\xi}{n^2} + \frac{n(n-1)\mu^2}{n^2} - \frac{2\xi}{n} - \frac{2(n-1)\mu^2}{n} \tag{9.89}$$

$$= \frac{1}{n}\frac{n-1}{\left(\xi - \sum_{i=1}^{n}\mu^2\right)} \tag{9.90}$$

$$= \sigma^2. \tag{9.91}$$

9.7.2 Proof of Theorem 9.3.4

我们表示该 sam 通过$\widetilde{Y}(n)$计算中位数。我们的目标是展示对于一个 y $\epsilon > 0$

$$\lim_{n\to\infty}\mathrm{P}\left(\left|\widetilde{Y}(n) - \gamma\right| \geq \epsilon\right) = 0. \tag{9.92}$$

我们将证明

$$\lim_{n\to\infty}\mathrm{P}\left(\widetilde{Y}(n) \geq \gamma + \epsilon\right) = 0. \tag{9.93}$$

相同的论证可以确立

$$\lim_{n\to\infty}\mathrm{P}\left(\widetilde{Y}(n) \leq \gamma - \epsilon\right) = 0. \tag{9.94}$$

如果我们对集合 $\left\{\widetilde{X}(1), \ldots, \widetilde{X}(n)\right\}$ 进行排序，那么当 n 为奇数时，$\widetilde{Y}(n)$ 等于第 $(n+1)/2$th 个元素；当 n 为偶数时，$\widetilde{Y}(n)$ 等于第 $n/2$th 个元素与第 $(n/2+1)$th 个元素的平均值。事件 $\widetilde{Y}(n) \geq \gamma + \epsilon$ 因此意味着至少有 $(n+1)/2$ 个元素大于 $\gamma + \epsilon$。

对于每个个体$\widetilde{X}(i)$，$\widetilde{X}(i) > \gamma + \epsilon$ 的概率是

$$p := 1 - F_{\widetilde{X}(i)}(\gamma + \epsilon) = 1/2 - \epsilon' \tag{9.95}$$

我们假设 $\epsilon' >$ 为 0。如果不是这种情况，那么独立同分布序列的累积分布函数在 γ 处为 *flat*，并且中位数没有明确的定义。集合 $\left\{\widetilde{X}(1), \ldots, \widetilde{X}(n)\right\}$ 中大于 $\gamma + \epsilon$ 的随机变量数量服从参数为 n 的二项分布随机变量 B_n。

以及 p。因此，我们有

$$\begin{aligned}
\mathrm{P}\left(\widetilde{Y}(n) \geq \gamma + \epsilon\right) &\leq \mathrm{P}\left(\frac{n+1}{2} \text{ or more samples are greater or equal to } \gamma + \epsilon\right) && (9.96)\\
&= \mathrm{P}\left(B_n \geq \frac{n+1}{2}\right) && (9.97)\\
&= \mathrm{P}\left(B_n - np \geq \frac{n+1}{2} - np\right) && (9.98)\\
&\leq \mathrm{P}\left(|B_n - np| \geq n\epsilon' + \frac{1}{2}\right) && (9.99)\\
&\leq \frac{\mathrm{Var}(B_n)}{\left(n\epsilon' + \frac{1}{2}\right)^2} \quad \text{by Chebyshev's inequality} && (9.100)\\
&= \frac{np(1-p)}{n^2\left(\epsilon' + \frac{1}{2n}\right)^2} && (9.101)\\
&= \frac{p(1-p)}{n\left(\epsilon' + \frac{1}{2n}\right)^2}, && (9.102)
\end{aligned}$$

其在 $n \to \infty$ 时收敛到零。这就证明了（9.93）。

Chapter 10

Bayesian Statistics

在频率主义范式中，我们将数据建模为来自一个固定分布的实现。特别地，如果模型是参数化的，参数是 *deterministic* 数量。相比之下，在贝叶斯参数化建模中，参数被建模为 **random variables**。目标是具有灵活性，在事先量化我们对基础分布的不确定性，例如为了整合关于数据的可用先验信息。

10.1 Bayesian parametric models

在本节中，我们描述如何在贝叶斯框架下将参数化模型拟合到一个数据集。与第9.6节一样，我们假设数据是通过从具有未知参数的已知分布中抽样生成的。关键的区别在于，我们将参数建模为随机的，而非确定性的。这需要在拟合数据之前选择它们的先验分布，从而能够事先量化我们对参数取值的不确定性。贝叶斯参数模型由以下内容指定：

1. **prior** 分布是 $\vec{\Theta}$ 的分布，它在看到数据之前对模型的不确定性进行编码。 2. **likelihood** 是在给定 $\vec{\Theta}$ 条件下 $\vec{X}$ 的条件分布，它规定了数据如何依赖于参数。与频率学派框架不同，似然是将 *not* 解释为参数的确定性函数。

我们的目标在于学习一个贝叶斯模型，即计算在给定$\vec{X}$的情况下，参数Θ的 **posterior distribution**。在实现$\vec{x}$时评估这个后验分布，能够利用数据更新我们关于Θ的不确定性。

以下示例将贝叶斯模型拟合到来自伯努利随机变量的独立同分布样本。

Example 10.1.1 (伯努利分布)**.** 设 $\vec{x}$ 为一个数据向量，我们希望将其建模为来自伯努利分布的 iid（独立同分布）样本。由于我们采用贝叶斯方法，因此为伯努利分布的参数选择一个先验分布。我们将考虑两种不同的贝叶斯估计器 Θ_1 和 Θ_2：

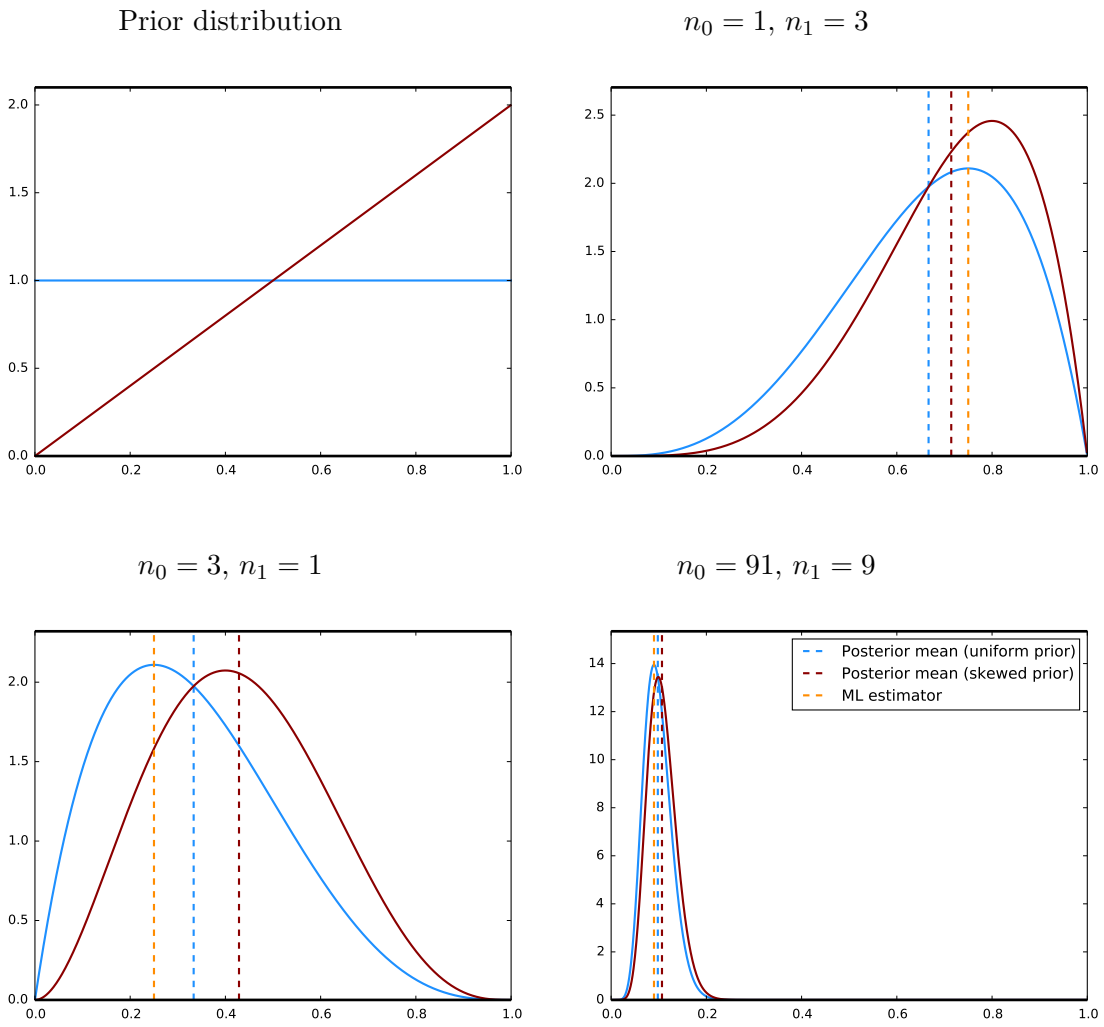


Figure 10.1: Θ_1 (蓝色)和Θ_2 (深红色)的先验分布在例 10.1.1 中的左上图中显示。其余的图展示了不同数据集的对应后验分布。

1. Θ_1 在先验信息方面表示一种保守估计器。我们为该参数分配一个均匀的概率密度函数（pdf）。单位区间内的任何取值都具有相同的概率密度：

$$f_{\Theta_1}(\theta) = \begin{cases} 1 & \text{for } 0 \leq \theta \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{10.1}$$

2. Θ_2 是一种估计量，它假设该参数更接近 1 而不是 0。例如，我们可以用它来刻画对一枚硬币偏向正面的怀疑。我们选择一个偏斜的 pdf，使其从 0 到 1 线性递增，

$$f_{\Theta_2}(\theta) = \begin{cases} 2\,\theta & \text{for } 0 \leq \theta \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{10.2}$$

根据独立同分布假设，似然函数，即在给定伯努利分布参数下的数据的条件概率质量函数，等于

$$p_{\vec{X}|\Theta}\left(\vec{x}|\theta\right)=\theta^{n_1}\left(1-\theta\right)^{n_0}, \tag{10.3}$$

其中 n_1 是数据中 1 的数量，n_0 是 0 的数量（见例 9.6.3）。因此，两个估计量的后验概率密度函数分别为

$$\begin{aligned}
f_{\Theta_1|\vec{X}}\left(\theta|\vec{x}\right) &= \frac{f_{\Theta_1}\left(\theta\right)p_{\vec{X}|\Theta_1}\left(\vec{x}|\theta\right)}{p_{\vec{X}}\left(\vec{x}\right)} && (10.4)\\
&= \frac{f_{\Theta_1}\left(\theta\right)p_{\vec{X}|\Theta_1}\left(\vec{x}|\theta\right)}{\int_u f_{\Theta_1}\left(u\right)p_{\vec{X}|\Theta_1}\left(\vec{x}|u\right)\,\mathrm{d}u} && (10.5)\\
&= \frac{\theta^{n_1}\left(1-\theta\right)^{n_0}}{\int_u u^{n_1}\left(1-u\right)^{n_0}\,\mathrm{d}u} && (10.6)\\
&= \frac{\theta^{n_1}\left(1-\theta\right)^{n_0}}{\beta\left(n_1+1,n_0+1\right)}, && (10.7)\\
f_{\Theta_2|\vec{X}}\left(\theta|\vec{x}\right) &= \frac{f_{\Theta_2}\left(\theta\right)p_{\vec{X}|\Theta_2}\left(\vec{x}|\theta\right)}{p_{\vec{X}}\left(\vec{x}\right)} && (10.8)\\
&= \frac{\theta^{n_1+1}\left(1-\theta\right)^{n_0}}{\int_u u^{n}1{+}1\left(1-u\right)^{n_0}\mathrm{d}u} && (10.9)\\
&= \frac{\theta^{n_1+1}\left(1-\theta\right)^{n_0}}{\beta\left(n_1+2,n_0+1\right)}, && (10.10)\\
& && (10.11)
\end{aligned}$$

在哪里

$$\beta\left(a,b\right):=\int_u u^{a-1}\left(1-u\right)^{b-1}\,\mathrm{d}u \tag{10.12}$$

是一个特殊函数，称为贝塔函数或欧拉第一类积分，已被列入表格。

图 10.1 显示了不同 n_1 和 n_0 值的后验分布图。它还展示了参数的最大似然估计值，即 $n_1/\left(n_0+n_1\right)$ (，参见示例 9.6.3)。对于较少的翻转次数，Θ_2 的后验概率密度函数相对于 Θ_1 的后验概率密度函数偏向右侧，反映了先验信念，即该参数更接近 1。然而，对于较多的翻转次数，两个后验密度非常接近。

△

10.2 Conjugate prior

示例 10.1.1 中的两个后验分布都是 Beta 分布（参见定义 2.3.12），先验分布也是如此。Θ_1 的均匀先验是具有参数 $a=1$ 和 $b=1$ 的 Beta 分布，而 Θ_2 的偏斜先验是具有参数 $a=2$ 和 $b=1$ 的 Beta 分布。由于先验和后验属于同一家族，计算后验分布等同于仅更新参数。当对于特定的似然函数，先验和后验分布保证属于同一家族时，这些分布被称为共轭先验。

Definition 10.2.1 (共轭先验)**.** *A conjugate family of distributions for a certain likelihood satisfies the following property: if the prior belongs to the family, then the posterior also belongs to the family.*

当似然为二项分布时，贝塔分布是共轭先验。

Theorem 10.2.2 (贝塔分布是二项式似然)**.** *If the prior distribution of Θ is a beta distributions with parameters a and b and the likelihood of the data X given Θ is binomial with parameters n and x, then the posterior distribution of Θ given X is a beta distribution with parameters $x+a$ and $n-x+b$.*的共轭分布

Proof.

$$
\begin{aligned}
f_{\Theta \mid X}(\theta \mid x) &= \frac{f_{\Theta}(\theta)\, p_{X \mid \Theta}(x \mid \theta)}{p_X(x)} && (10.13)\\
&= \frac{f_{\Theta}(\theta)\, p_{X \mid \Theta}(x \mid \theta)}{\int_u f_{\Theta}(u)\, p_{X \mid \Theta}(x \mid u)\, \mathrm{d}u} && (10.14)\\
&= \frac{\theta^{a-1}(1-\theta)^{b-1}\binom{n}{x}\theta^x(1-\theta)^{n-x}}{\int_u u^{a-1}(1-u)^{b-1}\binom{n}{x}u^x(1-u)^{n-x}\,\mathrm{d}u} && (10.15)\\
&= \frac{\theta^{x+a-1}(1-\theta)^{n-x+b-1}}{\int_u u^{x+a-1}(1-u)^{n-x+b-1}\,\mathrm{d}u} && (10.16)\\
&= f_{\beta}(\theta; x+a, n-x+b). && (10.17)
\end{aligned}
$$

□

注意，例10.1.1中得到的后验分布可直接由该定理推出。

Example 10.2.3 (新墨西哥州的民意调查)**.** 在2016年美国大选的一项新墨西哥州民意调查中，共有429名参与者，其中227人打算投票给克林顿，202人打算投票给特朗普（数据来自一次真实的民调[1]，但为简化起见，我们忽略了其他候选人以及尚未做出决定的人）。我们的目标是使用贝叶斯框架，基于这些数据来预测新墨西哥州的选举结果。

我们将支持特朗普的选民比例建模为一个随机变量 Θ。我们假设在民调中的 n 人是从总体中按均匀随机抽取并且有放回的，因此，给定 $\Theta=\theta$，特朗普选民的数量是一个具有参数 n 和 θ 的二项分布。我们没有关于 Θ 可能值的额外信息，因此我们假设它是均匀分布，或者等效地是具有参数 $a:=1$ 和 $b:=1$ 的贝塔分布。

根据定理10.2.2，在我们观察到的数据条件下，Θ 的后验分布是一个参数为 $a:=203$ 和 $b:=228$ 的贝塔分布，如图10.2所示。对应的 $\Theta \geq 0.5$ 的概率为11.4%，这是我们对特朗普在新墨西哥州获胜概率的估计。

△

[1]The poll results are taken from
https://www.abqjournal.com/883092/clinton-still-ahead-in-new-mexico.html

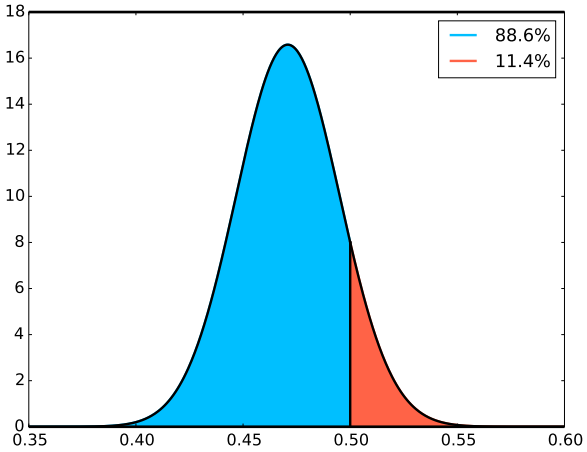


Figure 10.2: 新墨西哥州特朗普选民比例的后验分布，基于示例 10.2.3 中的民调数据。

10.3 Bayesian estimators

贝叶斯方法学习概率模型会得出感兴趣参数的完整后验分布。在本节中，我们描述了两种从后验分布中推导参数单一估计值的替代方法。

10.3.1 Minimum mean-square-error estimation

后验分布的均值是在给定数据条件下参数的条件期望。将后验均值作为参数 $\vec{\Theta}$ 的估计量具有坚实的理论依据：它保证在 *all possible estimators* 中达到最小的均方误差（MSE）。当然，这只有在所有假设都成立时才成立，即参数按照先验生成，数据随后按照似然生成，而真实数据未必如此。

Theorem 10.3.1 (后验均值最小化了 MSE)**.** *The posterior mean is the minimum mean-square-error (MMSE) estimate of the parameter $\vec{\Theta}$ given the data $\vec{X}$. To be more precise, let us define*

$$\theta_{\text{MMSE}}\left(\vec{x}\right) := \mathrm{E}\big(\vec{\Theta}|\vec{X} = \vec{x}\big). \tag{10.18}$$

For any arbitrary estimator $\theta_{\text{其他}}\left(\vec{x}\right)$,

$$\mathrm{E}\left(\left(\theta_{\text{other}}(\vec{X}) - \vec{\Theta}\right)^2\right) \geq \mathrm{E}\left(\left(\theta_{\text{MMSE}}(\vec{X}) - \vec{\Theta}\right)^2\right). \tag{10.19}$$

Proof. 我们首先计算在$\vec{X} = \vec{x}$ 条件下任意估计器的 MSE，在

以在给定$\vec{X}$条件下Θ的条件期望来表示,

$$\mathrm{E}\left(\left(\theta_{\text{other}}(\vec{X})-\vec{\Theta}\right)^2\middle|\vec{X}=\vec{x}\right) \tag{10.20}$$

$$=\mathrm{E}\left(\left(\theta_{\text{other}}(\vec{X})-\theta_{\text{MMSE}}(\vec{X})+\theta_{\text{MMSE}}(\vec{X})-\vec{\Theta}\right)^2\middle|\vec{X}=\vec{x}\right) \tag{10.21}$$

$$=\left(\theta_{\text{other}}(\vec{x})-\theta_{\text{MMSE}}(\vec{x})\right)^2+\mathrm{E}\Big(\left(\theta_{\text{MMSE}}(\vec{X})-\vec{\Theta}\right)^2\middle|\vec{X}=\vec{x}\Big) \tag{10.22}$$

$$+2\left(\theta_{\text{other}}(\vec{x})-\theta_{\text{MMSE}}(\vec{x})\right)\mathrm{E}\left(\theta_{\text{MMSE}}(\vec{x})-\mathrm{E}\left(\vec{\Theta}\middle|\vec{X}=\vec{x}\right)\right)$$

$$=\left(\theta_{\text{other}}(\vec{x})-\theta_{\text{MMSE}}(\vec{x})\right)^2+\mathrm{E}\Big(\left(\theta_{\text{MMSE}}(\vec{X})-\vec{\Theta}\right)^2\middle|\vec{X}=\vec{x}\Big). \tag{10.23}$$

由迭代期望定理,

$$\mathrm{E}\left(\left(\theta_{\text{other}}(\vec{X})-\Theta\right)^2\right)=\mathrm{E}\left(\mathrm{E}\left(\left(\theta_{\text{other}}(\vec{X})-\vec{\Theta}\right)^2\middle|\vec{X}\right)\right) \tag{10.24}$$

$$=\mathrm{E}\left(\left(\theta_{\text{other}}(\vec{X})-\theta_{\text{MMSE}}(\vec{X})\right)^2\right)+\mathrm{E}\left(\mathrm{E}\Big(\left(\theta_{\text{MMSE}}(\vec{X})-\vec{\Theta}\right)^2\middle|\vec{X}\Big)\right)$$

$$=\mathrm{E}\left(\left(\theta_{\text{other}}(\vec{X})-\theta_{\text{MMSE}}(\vec{X})\right)^2\right)+\mathrm{E}\left(\left(\theta_{\text{MMSE}}(\vec{X})-\vec{\Theta}\right)^2\right) \tag{10.25}$$

$$\geq\mathrm{E}\left(\left(\theta_{\text{MMSE}}(\vec{X})-\vec{\Theta}\right)^2\right), \tag{10.26}$$

因为非负量的期望是非负的。 □

Example 10.3.2 (伯努利分布（续）)**.** 为了在例 10.1.1 中获得该参数的点估计，我们计算后验均值:

$$\mathrm{E}\left(\Theta_1|\vec{X}=\vec{x}\right)=\int_0^1\theta f_{\Theta_1|\vec{X}}\left(\theta|\vec{x}\right)\,\mathrm{d}\theta \tag{10.27}$$

$$=\frac{\int 10\;\theta^{n_1+1}\left(1-\theta\right)^{n_0}\mathrm{d}\theta}{\beta\left(n_1+1,n_0+1\right)} \tag{10.28}$$

$$=\frac{\beta\left(n_1+2,n_0+1\right)}{\beta\left(n_1+1,n_0+1\right)}, \tag{10.29}$$

$$\mathrm{E}\left(\Theta_2|\vec{X}=\vec{x}\right)=\int_0^1\theta f_{\Theta_2|\vec{X}}\left(\theta|\vec{x}\right)\,\mathrm{d}\theta \tag{10.30}$$

$$=\frac{\beta\left(n_1+3,n_0+1\right)}{\beta\left(n_1+2,n_0+1\right)}. \tag{10.31}$$

图 10.1 显示了不同 n_0 和 n_1 值的后验均值。

△

10.3.2 Maximum-a-posteriori estimation

后验均值的替代方法是后验众数，它是后验分布的概率密度函数（pdf）或概率质量函数（pmf）的最大值。

Definition 10.3.3 (最大后验估计量)**.** *The maximum-a-posteriori (MAP) estimator of a parameter $\vec{\Theta}$ given data $\vec{x}$ modeled as a realization of a random vector $\vec{X}$ is*

$$\theta_{\mathrm{MAP}}\left(\vec{x}\right) := \arg\max_{\vec{\theta}} p_{\vec{\Theta}\,|\,\vec{X}}\left(\vec{\theta}\,|\,\vec{x}\right) \tag{10.32}$$

if $\vec{\Theta}$ is modeled as a discrete random variable and

$$\theta_{\mathrm{MAP}}\left(\vec{x}\right) := \arg\max_{\vec{\theta}} f_{\vec{\Theta}\,|\,\vec{X}}\left(\vec{\theta}\,|\,\vec{x}\right) \tag{10.33}$$

if it is modeled as a continuous random variable.

在图10.1中，Θ的ML估计量是后验分布的众数（最大值），当先验为均匀分布时。这不是巧合，在均匀先验下，MAP和ML估计是相同的。

Lemma 10.3.4. *The maximum-likelihood estimator of a parameter Θ is the mode (maximum value) of the pdf of the posterior distribution given the data $\vec{X}$ if its prior distribution is uniform.*

Proof. 我们在数据模型和参数均为连续的情况下证明该结果；如果其中任一或二者是离散的，证明是完全相同的（在这种情况下，ML 估计量是后验分布 pmf 的众数）。如果参数的先验分布是均匀的，那么对于任意 $\vec{\theta}$，$f_{\vec{\Theta}}\big(\vec{\theta}\big)$ 都是常数，这意味着

$$\begin{aligned}
\arg\max_{\vec{\theta}} f_{\vec{\Theta}\,|\,\vec{X}}\left(\vec{\theta}|\vec{x}\right) &= \arg\max_{\vec{\theta}} \frac{f_{\vec{\Theta}}\big(\vec{\theta}\big) f_{\vec{X}|\vec{\Theta}}\left(\vec{x}|\vec{\theta}\right)}{\int_u f_{\vec{\Theta}}\left(u\right) f_{\vec{X}|\vec{\Theta}}\left(\vec{x}|u\right)\,\mathrm{d}u} && (10.34)\\
&= \arg\max_{\vec{\theta}} f_{\vec{X}|\vec{\Theta}}\left(\vec{x}|\vec{\theta}\right) \qquad \text{(the rest of the terms do not depend on } \vec{\theta}) \\
&= \arg\max_{\vec{\theta}} \mathcal{L}_{\vec{x}}\big(\vec{\theta}\big). && (10.35)
\end{aligned}$$

□

请注意，均匀先验仅在参数被限制在一个有界集合中的情况下才是良好定义的。

我们现在描述一种 MAP 估计器最优的情形。如果参数 Θ 只能取离散的一组取值，那么 MAP 估计器会最小化做出错误选择的概率。

Theorem 10.3.5 (MAP估计量最小化错误概率)**.** *Let $\vec{\Theta}$ be a discrete random vector and let $\vec{X}$ be a random vector modeling the data. We define*

$$\theta_{\mathrm{MAP}}\left(\vec{x}\right) := \arg\max_{\vec{\theta}} p_{\vec{\Theta}\,|\,\vec{X}}\big(\vec{\theta}\,|\,\vec{X}=\vec{x}\big). \tag{10.36}$$

For any arbitrary estimator $\theta_{\text{其他}}\left(\vec{x}\right)$,

$$\mathrm{P}\left(\theta_{\mathrm{other}}(\vec{X}) \neq \vec{\Theta}\right) \geq \mathrm{P}\left(\theta_{\mathrm{MAP}}(\vec{X}) \neq \vec{\Theta}\right). \tag{10.37}$$

In words, the MAP estimator minimizes the probability of error.

Proof. 我们假设$\vec{X}$是一个连续随机向量，但如果它是离散的，同样的论证也适用。我们有

$$\mathrm{P}\left(\Theta = \theta_{\text{other}}(\vec{X})\right) = \int_{\vec{x}} f_{\vec{X}}(\vec{x})\,\mathrm{P}\left(\Theta = \theta_{\text{other}}(\vec{x}) \mid \vec{X} = \vec{x}\right)\,\mathrm{d}\vec{x} \tag{10.38}$$

$$= \int_{\vec{x}} f_{\vec{X}}(\vec{x})\, p_{\vec{\Theta}\mid\vec{X}}(\theta_{\text{other}}(\vec{x}) \mid \vec{x})\,\mathrm{d}\vec{x} \tag{10.39}$$

$$\leq \int_{\vec{x}} f_{\vec{X}}(\vec{x})\, p_{\vec{\Theta}\mid\vec{X}}(\theta_{\text{MAP}}(\vec{x}) \mid \vec{x})\,\mathrm{d}\vec{x} \tag{10.40}$$

$$= \mathrm{P}\left(\Theta = \theta_{\text{MAP}}(\vec{X})\right), \tag{10.41}$$

其中 (10.40) 来自 fr　　从将 MAP 估计器定义为众数这一点出发　　后验的。 □

Example 10.3.6 (发送比特)**.** 我们考虑一个非常简单的通信信道模型，在该模型中，我们旨在发送一个由单个比特组成的信号Θ。我们的先验知识表明，信号等于1的概率为1/4。

$$p_\Theta(1) = \frac{1}{4}, \qquad p_\Theta(0) = \frac{3}{4}. \tag{10.42}$$

由于信道中存在噪声，我们将信号发送 n 次。在接收端我们观察到

$$\vec{X}_i = \Theta + \vec{Z}_i, \quad 1 \leq i \leq n, \tag{10.43}$$

其中$\vec{Z}$包含 n 个独立同分布的标准高斯随机变量。将扰动建模为高斯在通信领域是一种常见的选择。在噪声由许多近似独立的小效应叠加而成的假设下，这一做法由中心极限定理所证明。

我们现在将计算并比较给定观测值的 Θ 的 ML 和 MAP 估计量。
似然等于

$$\mathcal{L}_{\vec{x}}(\theta) = \prod_{i=1}^{n} f_{\vec{X}_i|\Theta}(\vec{x}_i \mid \theta) \tag{10.44}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\vec{x}_i-\theta)^2}{2}}. \tag{10.45}$$

处理对数似然函数更容易，

$$\log \mathcal{L}_{\vec{x}}(\theta) = -\sum_{i=1}^{n} \frac{(\vec{x}_i - \theta)^2}{2} - \frac{n}{2}\log 2\pi. \tag{10.46}$$

由于Θ只取两个值，我们可以直接比较。我们将选择$\theta_{\text{ML}}(\vec{x}) =$为1，如果

$$\log \mathcal{L}_{\vec{x}}(1) = -\sum_{i=1}^{n} \frac{\vec{x}_i^2 - 2\vec{x}_i + 1}{2} - \frac{n}{2}\log 2\pi \tag{10.47}$$

$$\geq -\sum_{i=1}^{n} \frac{\vec{x}_i^2}{2} - \frac{n}{2}\log 2\pi \tag{10.48}$$

$$= \log \mathcal{L}_{\vec{x}}(0). \tag{10.49}$$

等价地,

$$\theta_{\mathrm{ML}}\left(\vec{x}\right)=\begin{cases}1 & \text{if } \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i>\frac{1}{2},\\ 0 & \text{otherwise.}\end{cases} \tag{10.50}$$

这个规则非常合理：如果数据的样本均值更接近 1 而不是 0，那么我们的估计就等于 1。根据全概率公式，该估计量的误差概率等于

$$\begin{aligned}\mathrm{P}\left(\Theta\neq\theta_{\mathrm{ML}}(\vec{X})\right)&=\mathrm{P}\left(\Theta\neq\theta_{\mathrm{ML}}(\vec{X})\middle|\Theta=0\right)\mathrm{P}\left(\Theta=0\right)+\mathrm{P}\left(\Theta\neq\theta_{\mathrm{ML}}(\vec{X})\middle|\Theta=1\right)\mathrm{P}\left(\Theta=1\right)\\&=\mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n}\vec{X}_i>\frac{1}{2}\middle|\Theta=0\right)\mathrm{P}\left(\Theta=0\right)+\mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n}\vec{X}_i<\frac{1}{2}\middle|\Theta=1\right)\mathrm{P}\left(\Theta=1\right)\\&=Q\left(\sqrt{n}/2\right),\end{aligned} \tag{10.51}$$

其中最后一个等式源于这样一个事实：如果我们在 $\Theta=\theta$ 条件下，则经验均值服从高斯分布，其方差为 σ^2/n，均值为 θ (；参见定理 6.2.2) 的证明。

为了计算MAP估计，我们必须在给定观测数据的情况下，找到Θ的后验概率密度函数的最大值。等价地，我们可以找到其对数的最大值（这是等价的，因为对数是单调函数），

$$\log p_{\Theta|\vec{X}}\left(\theta|\vec{x}\right)=\log\frac{\prod_{i=1}^{n}f_{\vec{X}_i|\Theta}\left(\vec{x}_i|\theta\right)p_{\Theta}\left(\theta\right)}{f_{\vec{X}}\left(\vec{x}\right)} \tag{10.52}$$

$$=\sum_{i=1}^{n}\log f_{\vec{X}_i|\Theta}\left(\vec{x}_i|\theta\right)p_{\Theta}\left(\theta\right)-\log f_{\vec{X}}\left(\vec{x}\right) \tag{10.53}$$

$$=-\sum_{i=1}^{n}\frac{\vec{x}_i^2-2\vec{x}_i\theta+\theta^2}{2}-\frac{n}{2}\log 2\pi+\log p_{\Theta}\left(\theta\right)-\log f_{\vec{X}}\left(\vec{x}\right). \tag{10.54}$$

我们比较该函数在 Θ 的两种可能取值（0 和 1）下的值。如果，我们选择 $\theta_{\mathrm{MAP}}\left(\vec{x}\right)=$ 为 1

$$\log p_{\Theta|\vec{X}}\left(1|\vec{x}\right)+\log f_{\vec{X}}\left(\vec{x}\right)=-\sum_{i=1}^{n}\frac{\vec{x}_i^2-2\vec{x}_i+1}{2}-\frac{n}{2}\log 2\pi-\log 4 \tag{10.55}$$

$$\geq-\sum_{i=1}^{n}\frac{\vec{x}_i^2}{2}-\frac{n}{2}\log 2\pi-\log 4+\log 3 \tag{10.56}$$

$$=\log p_{\Theta|\vec{X}}\left(0|\vec{x}\right)+\log f_{\vec{X}}\left(\vec{x}\right). \tag{10.57}$$

等价地,

$$\theta_{\mathrm{MAP}}\left(\vec{x}\right)=\begin{cases}1 & \text{if } \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i>\frac{1}{2}+\frac{\log 3}{n},\\ 0 & \text{otherwise.}\end{cases} \tag{10.58}$$

MAP估计相对于ML估计会调整阈值，以考虑到Θ更容易等于零。然而，随着我们收集到更多证据，修正项趋于零，因此如果有大量数据，两种估计器将非常相似。

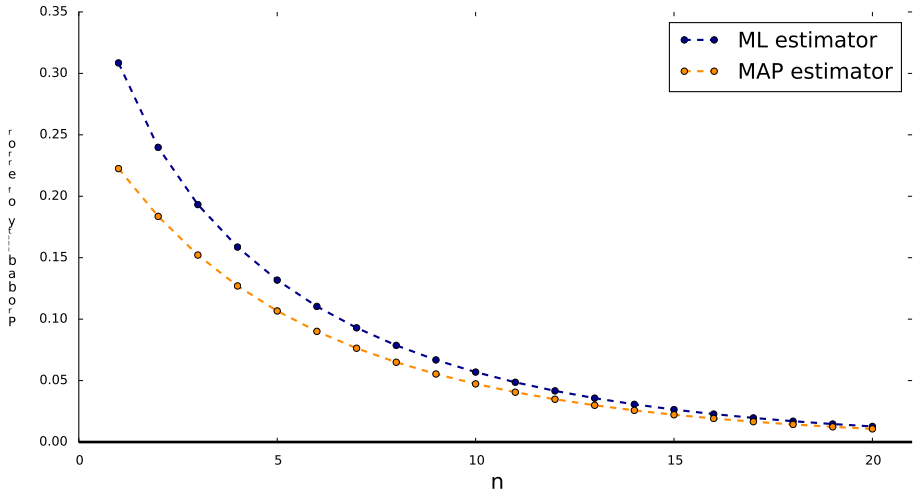


Figure 10.3: 示例 10.3.6 中 ML 和 MAP 估计器在不同 n 取值下的误差概率。

MAP 估计器的错误概率等于

$$\begin{aligned}
\mathrm{P}\left(\Theta \neq \theta_{\mathrm{MAP}}(\vec{X})\right) &= \mathrm{P}\left(\Theta \neq \theta_{\mathrm{MAP}}(\vec{X})|\Theta=0\right)\mathrm{P}\left(\Theta=0\right) + \mathrm{P}\left(\Theta \neq \theta_{\mathrm{MAP}}(\vec{X})|\Theta=1\right)\mathrm{P}\left(\Theta=1\right) \\
&= \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n}\vec{X}_i > \frac{1}{2}+\frac{\log 3}{n}\middle|\Theta=0\right)\mathrm{P}\left(\Theta=0\right) \qquad (10.59) \\
&\quad + \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n}\vec{X}_i < \frac{1}{2}+\frac{\log 3}{n}\middle|\Theta=1\right)\mathrm{P}\left(\Theta=1\right) \\
&= \frac{3}{4}Q\left(\sqrt{n}/2+\frac{\log 3}{\sqrt{n}}\right)+\frac{1}{4}Q\left(\sqrt{n}/2-\frac{\log 3}{\sqrt{n}}\right). \qquad (10.60)
\end{aligned}$$

我们在图10.3中比较了 ML 和 MAP 估计器的误差概率。MAP 估计具有更好的性能，但随着 n 的增大，这种差异变得很小。

△

Chapter 11

Hypothesis testing

在一项医学研究中，我们观察到10%的女性和12.5%的男性患有心脏病。如果研究中有20人，我们可能会犹豫是否声明女性比男性更不容易患心脏病；结果很可能是偶然发生的。然而，如果研究中有20,000人，那么我们更有可能认为我们观察到的现象是真实的。假设检验使这一直觉变得更加精确；它是一个框架，允许我们判断在数据中观察到的模式是否可能是随机波动的结果。

11.1 The hypothesis-testing framework

假设检验的目标是评估一个预先定义的猜想。在上面的例子中，这个猜想可以是心脏病在男性中的患病率高于女性。我们的猜想为假的假设称为**null hypothesis**，记作H_0。在我们的例子中，零假设将是心脏病在男性中的患病率至少与女性一样高。如果零假设成立，那么我们在数据中检测到的、似乎支持我们猜想的任何模式都只是偶然现象。研究中恰好有很多患心脏病的男性（或没有心脏病的女性）。相比之下，使我们的猜想为真的假设称为**alternative hypothesis**，记作H_1。在本章中我们采取频率学派的视角：*the hypotheses either hold or not*，它们以*not*的方式进行概率建模。

一个 **test** 是一种程序，用于根据数据确定我们是否应该 *reject* 原假设。拒绝原假设意味着我们认为它发生的可能性很小，这是支持备择假设的证据。如果我们未能拒绝原假设，这并不意味着我们认为它很可能成立，只是我们没有足够的信息来排除它。大多数检验通过对 **test statistic** 进行阈值化来做出决定，**test statistic** 是一个将数据（即 $\mathbb{R}^n$ 中的一个向量）映射到一个单一数字的函数。如果检验统计量属于 **rejection region** $\mathcal{R}$，则检验拒绝原假设。例如，我们可能会有

$$\mathcal{R} := \{t \mid t \geq \eta\}, \tag{11.1}$$

其中 t 是从数据中计算出的检验统计量，η 是预定的阈值。在这种情况下，只有当 t 大于 η 时，我们才会拒绝原假设。

如表11.1所示，我们可能会犯两种错误。**Type I error** 是一种 *false positive*：我们的猜想是错误的，但我们拒绝了零假设。**Type II error** 是

拒绝 H_0?

	No	Yes
H_0 is true	☺	Type I error
H_1 is true	Type II error	☺

Table 11.1: I型和II型错误。

一个 *false negative*：我们的假设成立，但我们并不拒绝零假设。在假设检验中，我们的优先任务是控制 I 型错误。当你在一项研究中看到一个结果在 0.05 水平下是 **statistically significant** 时，这意味着犯 I 型错误的概率被限制在 5% 以内。

Definition 11.1.1 (显著性水平和规模)**.** *The size of a test is the probability of making a Type I error. The significance level of a test is an upper bound on the size.*

拒绝原假设并不能给出数据与原假设不相容程度的定量刻画。**p value** 是一个由数据决定、起到这一作用的函数。

Definition 11.1.2 (p 值)**.** *The p value is the smallest significance level at which we would reject the null hypothesis for the data we observe.*

对于一个固定的显著性水平，理想的做法是选择一个能够最小化第二类错误概率的检验。换句话说，我们希望最大化在原假设不成立时拒绝原假设的概率。这个概率称为该检验的 **power**。

Definition 11.1.3 (功率)**.** *The power of a test is the probability of rejecting the null hypothesis if it does not hold.*

请注意，为了表征检验的功效，我们需要知道在备择假设下数据的分布，而这通常是不现实的（回想一下，备择假设仅仅是零假设的补集，因此包含了许多不同的可能性）。

在应用科学中应用假设检验的标准流程如下:

1. 选择一个猜想。 2. 确定相应的零假设。 3. 选择一个检验。

4. 收集数据。

5. 从数据中计算检验统计量。

6. 计算 p 值，并在其低于预先定义的阈值（通常为 1% 或 5%）时拒绝原假设。

Example 11.1.4 (关键时刻)**.** 我们想要验证一个猜想，即NBA中某个球员在关键时刻的表现是$clutch$，也就是说，他在比赛接近结束时得分比比赛其余时间要多。原假设是他的表现没有差异。我们选择的检验统计量t是他在最后一节比在比赛其余时间每分钟得分更多还是更少。

$$t\left(\vec{x}\right)=\sum_{i=1}^{n}1_{\vec{x}_i>0}, \tag{11.2}$$

其中 $\vec{x}_i$ 是他在比赛第4节和其他节次中每分钟得分的差值，i 为 $1\leq i\leq n$。

该检验的拒绝域为如下形式

$$\mathcal{R}:=\left\{t\left(\vec{x}\right)\mid t\left(\vec{x}\right)\geq\eta\right\}, \tag{11.3}$$

对于一个固定的阈值η。在原假设下，在第四节每分钟得到更多分数的概率为1/2（为简化起见，我们忽略他得到相同分数的可能性），因此我们可以将原假设下的检验统计量建模为参数为n和1/2的二项随机变量。如果η是介于0和n之间的一个整数，那么在原假设成立时，检验统计量落入拒绝域的概率为

$$P\left(T_0\geq\eta\right)=\frac{1}{2^n}\sum_{k=\eta}^{n}\binom{n}{k}. \tag{11.4}$$

因此检验的尺寸为$\frac{1}{2^n}\sum_{k=\eta}^{n}\binom{n}{k}$。表 11.2 显示了 η 所有可能取值下的该值。如果我们希望显著性水平为 1% 或 5%，则需要将阈值分别设为 $\eta=16$ 或 $\eta=15$。

我们从20场比赛$\vec{x}$中收集数据，并计算检验统计量$t\left(\vec{x}\right)$ (，注意我们使用小写字母，因为它是一个特定的实现)，其结果为14（在20场比赛中，有14场他在第四节每分钟得分更高）。这不足以在我们预先设定的1%或5%的水平下拒绝零假设。因此，该结果不具有统计显著性。

无论如何，我们计算 p 值，即结果达到显著性的最小水平。由表可知，它等于 0.058。请注意，在频率主义框架下，我们 $cannot$ 将其解释为原假设成立的概率（即该球员在第四节并没有表现得更好），因为假设本身不是随机的，它要么成立，要么不成立。我们的结果几乎显著，尽管我们没有足够的证据来支持我们的猜想，但该球员在第四节表现更好的说法似乎是合理的。

△

11.2 Parametric testing

在本节中，我们讨论假设检验，假设我们的数据是从具有$unknown$参数的已知分布中抽样的。我们再次采用频率学派的视角，

η	1	2	3	4	5	6	7	8	9	10
$P(T_0 \geq \eta)$	1.000	1.000	1.000	0.999	0.994	0.979	0.942	0.868	0.748	0.588
η	11	12	13	14	15	16	17	18	19	20
$P(T_0 \geq \eta)$	0.412	0.252	0.132	0.058	0.021	0.006	0.001	0.000	0.000	0.000

Table 11.2: 在示例 11.1.4 中，犯第一类错误的概率取决于阈值的值。值已四舍五入到小数点后三位。

通常在应用科学的多数研究中都是这样做的。因此，参数是确定性的，假设也是如此：零假设要么为真，要么为假，*the probability that the null hypothesis holds* 并不存在。

为了简化叙述，我们假设概率分布仅依赖于一个参数，我们用θ表示。P_θ是我们概率空间的概率测度，当θ是该参数的值时。$\vec{X}$是一个根据P_θ分布的随机向量。我们观察到的实际数据，我们用$\vec{x}$表示，假设是从这个随机向量中得到的一个实现。

假设原假设为 $\theta = \theta_0$。在这种情况下，具有检验统计量 T 和拒绝区域 $\mathcal{R}$ 的检验的大小等于

$$\alpha = \mathrm{P}_{\theta_0}\left(T(\vec{X}) \in \mathcal{R}\right). \tag{11.5}$$

对于形如 (11.1) 的拒绝域，我们有

$$\alpha := \mathrm{P}_{\theta_0}\left(T(\vec{X}) \geq \eta\right). \tag{11.6}$$

如果检验统计量的取值为 $T(x_1, \ldots, x_n)$，那么我们将拒绝 H_0 的显著性水平将是

$$p = \mathrm{P}_{\theta_0}\left(T(\vec{X}) \geq T(\vec{x})\right), \tag{11.7}$$

如果我们观察到 $\vec{x}$，其 p 值是多少。 因此，p 值可以被解释为：在 *if the null hypothesis holds* 的情况下，观察到一个比我们在数据中观察到的结果更 *more extreme* 的结果的概率。

形式为$\theta = \theta_0$的假设称为**simple**假设。如果某个集合$\mathcal{S}$下的假设具有$\theta \in \mathcal{S}$的形式，则该假设是**composite**。对于复合原假设$\theta \in \mathcal{H}_0$，我们以下述方式重新定义显著性水平和 p 值，

$$\alpha = \sup_{\theta \in \mathcal{H}_0} \mathrm{P}_\theta\left(T(\vec{X}) \geq \eta\right), \tag{11.8}$$

$$p = \sup_{\theta \in \mathcal{H}_0} \mathrm{P}_\theta\left(T(\vec{X}) \geq T(\vec{x})\right). \tag{11.9}$$

为了表征在某一显著性水平下的检验效能，我们计算效能函数。

Definition 11.2.1 (幂函数)**.** *Let P_θ be the probability measure parametrized by θ and let $\mathcal{R}$ the rejection region for a test based on the test statistic $T(\vec{x})$. The power function of the test is defined as*

$$\beta(\theta) := P_\theta\left(T(\vec{X}) \in \mathcal{R}\right) \tag{11.10}$$

理想情况下，我们希望 $\beta(\theta) \approx 0$ 对于 $\theta \in \mathcal{H}_0$ 和 $\beta(\theta) \approx 1$ 对于 $\theta \in \mathcal{H}_1$。

Example 11.2.2 (抛硬币)**.** 我们感兴趣的是检查硬币是否偏向正面。零假设是，对于每次抛硬币，获得正面的概率是$\theta \leq 1/2$。因此，备择假设是$\theta > 1/2$。我们考虑一个检验统计量，该统计量等于在一系列n 独立同分布（iid）抛掷中观察到的正面次数,

$$T(\vec{x}) = \sum_{i=1}^{n} 1_{\vec{x}_i=1}, \tag{11.11}$$

其中 $\vec{x}_i$ 为 1，如果第 i 次抛硬币为正面，反之为 0。一个自然的拒绝区域是

$$T(\vec{x}) \geq \eta. \tag{11.12}$$

特别地，我们考虑两个可能的阈值

1. $\eta = n$，即只有在all硬币抛掷为正面时，我们才会拒绝零假设,
2. $\eta = 3n/5$，即当至少五分之三的抛硬币结果为正面时，我们拒绝原假设。

我们应该使用什么检验方法，如果抛硬币的次数是5次、50次或100次？这些检验的显著性水平是5%吗？对于这些n的值，检验的效能是多少?

为回答这些问题，我们计算两种方案下检验的功效函数。如果 $\eta = n$,

$$\beta_1(\theta) = P_\theta\left(T(\vec{X}) \in \mathcal{R}\right) \tag{11.13}$$

$$= \theta^n. \tag{11.14}$$

如果 $\eta = 3n/5$,

$$\beta_2(\theta) = \sum_{k=3n/5}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k}. \tag{11.15}$$

图11.1显示了两个功效函数。如果$\eta = n$，那么对于n的三个值，检验的显著性水平为5%。然而，功效非常低，尤其是对于较大的n。这很有道理：即使硬币偏向很大，得到n个正面朝上的概率仍然极低。如果$\eta = 3n/5$，那么对于$n = 5$，检验的显著性水平远高于5%，因为即使硬币不偏，观察到5次抛掷中有3次正面朝上的概率也相当高。然而，对于较大的n，检验的功效比第一个选项高得多。如果硬币的偏向大于0.7，我们以很高的概率拒绝原假设。

△

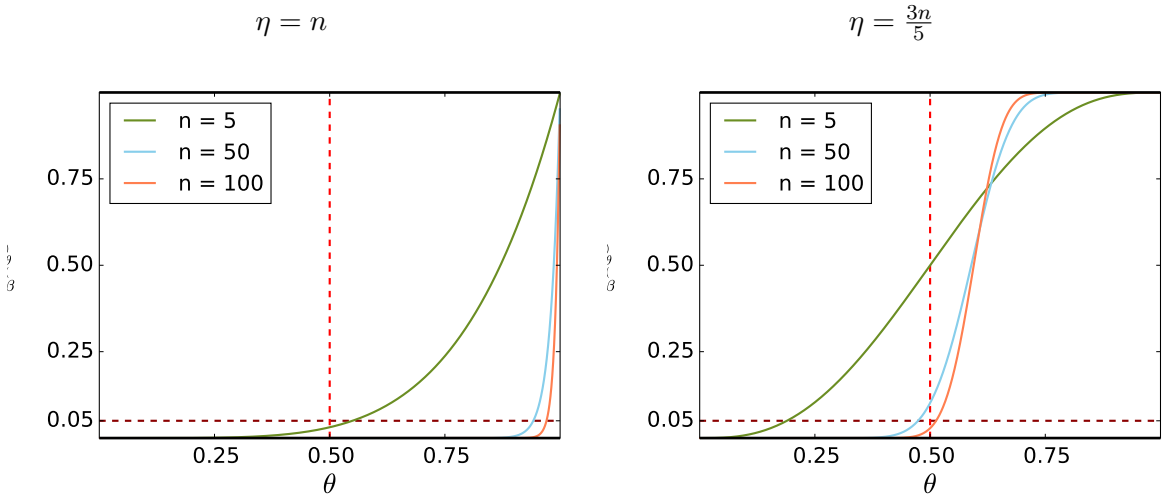


Figure 11.1: 示例11.2.2中所描述的检验的功效函数。

在参数化假设下构建检验的一种系统方法，是对数据在零假设下的似然与数据在备择假设下的似然之比进行阈值判断。如果该比值较高，则数据与零假设相容，因此不应拒绝零假设。

Definition 11.2.3 (似然比检验). *Let $\mathcal{L}_{\vec{x}}(\theta)$ denote the likelihood function corresponding to a data vector $\vec{x}$. $\mathcal{H}_0$ and $\mathcal{H}_1$ are the sets corresponding to the null and alternative hypotheses respectively. The likelihood ratio is*

$$\Lambda(\vec{x}) := \frac{\sup_{\theta \in \mathcal{H}_0} \mathcal{L}_{\vec{x}}(\theta)}{\sup_{\theta \in \mathcal{H}_1} \mathcal{L}_{\vec{x}}(\theta)}. \tag{11.16}$$

A likelihood-ratio test has a rejection region of the form $\{\Lambda(\vec{x}) \leq \eta\}$, for a constant threshold η.

Example 11.2.4 (已知方差的高斯分布). 假设你有一些数据，这些数据可以很好地建模为具有已知方差σ的独立同分布高斯分布。均值是未知的，我们感兴趣的是确定它是否等于某个特定值μ_0。相应的似然比检验是什么，应该如何设置阈值，以便达到显著性水平α?

首先，由示例 9.6.4 可知，样本均值达到了高斯分布似然函数的最大值。

$$\text{av}(\vec{x}) := \arg\max_{\mu} \mathcal{L}_{\vec{x}}(\mu, \sigma) \tag{11.17}$$

对于σ的任何值。利用这个结果，我们得到

$$\Lambda(\vec{x}) = \frac{\sup_{\mu \in \mathcal{H}_0} \mathcal{L}_{\vec{x}}(\mu)}{\sup_{\mu \in \mathcal{H}_1} \mathcal{L}_{\vec{x}}(\mu)} \tag{11.18}$$

$$= \frac{\mathcal{L}_{\vec{x}}(\mu_0)}{\mathcal{L}_{\vec{x}}(\text{av}(\vec{x}))}. \tag{11.19}$$

将似然的表达式代入，我们得到，

$$\Lambda(\vec{x}) = \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left((\vec{x}_i - \operatorname{av}(\vec{x}))^2 - (\vec{x}_i - \mu_0)^2\right)\right\} \tag{11.20}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}\left(-2\operatorname{av}(\vec{x})\sum_{i=1}^{n}\vec{x}_i + n\operatorname{av}(\vec{x})^2 - 2\mu_0\sum_{i=1}^{n}\vec{x}_i + n\mu_0^2\right)\right\} \tag{11.21}$$

$$= \exp\left\{-\frac{n(\operatorname{av}(\vec{x}) - \mu_0)^2}{2\sigma^2}\right\}. \tag{11.22}$$

取对数后，检验的形式为

$$\frac{|\operatorname{av}(\vec{x}) - \mu_0|}{\sigma} \geq \sqrt{\frac{-2\log\eta}{n}}. \tag{11.23}$$

n 个独立的高斯随机变量，其均值为 μ_0，方差为 σ^2，它们的样本均值服从均值为 μ_0，方差为 σ^2/n 的高斯分布，这意味着

$$\alpha = \mathrm{P}_{\mu_0}\left(\left|\frac{\operatorname{av}\left(\vec{X}\right) - \mu_0}{\sigma/\sqrt{n}}\right| \geq \sqrt{-2\log\eta}\right) \tag{11.24}$$

$$= 2\,Q\left(\sqrt{-2\log\eta}\right). \tag{11.25}$$

如果我们固定一个期望的大小 α，那么测试变为

$$\frac{|\operatorname{av}(\vec{x}) - \mu_0|}{\sigma} \geq \frac{Q^{-1}(\alpha/2)}{\sqrt{n}}. \tag{11.26}$$

△

一个激励采用似然比检验的论点是，如果原假设和备择假设是简单的，那么在检验力方面它是最优的。

Lemma 11.2.5 (奈曼-皮尔森引理)**.** *If both the null hypothesis and the alternative hypothesis are simple, i.e. the parameter θ can only have two values θ_0 and θ_1, then the likelihood-ratio test has the highest power among all tests with a fixed size.*

Proof. 回顾一下，检验功效是当零假设不成立时拒绝零假设的概率。如果我们用 $\mathcal{R}_{LR}$ 表示似然比检验的拒绝域，那么其功效为

$$P_{\theta_1}\left(\vec{X} \in \mathcal{R}_{LR}\right). \tag{11.27}$$

假设我们 ha 再进行一个拒绝域为 $\mathcal{R}$ 的检验。其功效为 等于

$$P_{\theta_1}\left(\vec{X} \in \mathcal{R}\right). \tag{11.28}$$

为了证明似然比检验的功效更大，我们只需要证明

$$P_{\theta_1}\left(\vec{X} \in \mathcal{R}^c \cap \mathcal{R}_{LR}\right) \geq P_{\theta_1}\left(\vec{X} \in \mathcal{R}_{LR}^c \cap \mathcal{R}\right). \tag{11.29}$$

假设数据是连续随机变量（对于离散随机变量的论证几乎相同），并且在原假设和备择假设成立时的概率密度函数分别为 f_{θ_0} 和 f_{θ_1}。根据似然比检验拒绝域的定义，若 $\Lambda(\vec{x}) \in \mathcal{R}_{LR}$

$$f_{\theta_1}(\vec{x}) \geq \frac{f_{\theta_0}(\vec{x})}{\eta}, \tag{11.30}$$

而如果 $\Lambda(\vec{x}) \in \mathcal{R}_{LR}^c$

$$f_{\theta_1}(\vec{x}) \leq \frac{f_{\theta_0}(\vec{x})}{\eta}. \tag{11.31}$$

如果两个测试的大小为 α，则

$$P_{\theta_0}\left(\vec{X} \in \mathcal{R}\right) = \alpha = P_{\theta_0}\left(\vec{X} \in \mathcal{R}_{LR}\right). \tag{11.32}$$

因此

$$\begin{aligned}
P_{\theta_0}\left(\vec{X} \in \mathcal{R}^c \cap \mathcal{R}_{LR}\right) &= P_{\theta_0}\left(\vec{X} \in \mathcal{R}_{LR}\right) - P_{\theta_0}\left(\vec{X} \in \mathcal{R} \cap \mathcal{R}_{LR}\right) && (11.33)\\
&= P_{\theta_0}\left(\vec{X} \in \mathcal{R}\right) - P_{\theta_0}\left(\vec{X} \in \mathcal{R} \cap \mathcal{R}_{LR}\right) && (11.34)\\
&= P_{\theta_0}\left(\vec{X} \in \mathcal{R} \cap \mathcal{R}_{LR}^c\right). && (11.35)
\end{aligned}$$

现在让我们证明 (11.29) 成立,

$$\begin{aligned}
P_{\theta_1}\left(\vec{X} \in \mathcal{R}^c \cap \mathcal{R}_{LR}\right) &= \int_{\vec{x} \in \mathcal{R}^c \cap \mathcal{R}_{LR}} f_{\theta_1}(\vec{x})\, \mathrm{d}\vec{x} && (11.36)\\
&\geq \frac{1}{\eta} \int_{\vec{x} \in \mathcal{R}^c \cap \mathcal{R}_{LR}} f_{\theta_0}(\vec{x})\, \mathrm{d}\vec{x} \qquad \text{by (11.30)} && (11.37)\\
&= \frac{1}{\eta} P_{\theta_0}\left(\vec{X} \in \mathcal{R}^c \cap \mathcal{R}_{LR}\right) && (11.38)\\
&= \frac{1}{\eta} P_{\theta_0}\left(\vec{X} \in \mathcal{R} \cap \mathcal{R}_{LR}^c\right) \qquad \text{by (11.35)} && (11.39)\\
&= \frac{1}{\eta} \int_{\vec{x} \in \mathcal{R} \cap \mathcal{R}_{LR}^c} f_{\theta_0}(\vec{x})\, \mathrm{d}\vec{x} && (11.40)\\
&\geq \int_{\vec{x} \in \mathcal{R} \cap \mathcal{R}_{LR}^c} f_{\theta_1}(\vec{x})\, \mathrm{d}\vec{x} \qquad \text{by (11.31)} && (11.41)\\
&= P_{\theta_1}\left(\vec{X} \in \mathcal{R} \cap \mathcal{R}_{LR}^c\right). && (11.42)
\end{aligned}$$

□

11.3 Nonparametric testing: The permutation test

在实际情况下，我们可能无法设计一个对我们的数据足够合适的参数化模型。非参数检验是假设检验，它们不假定数据遵循

任何具有预定义形式的分布。在本节中我们描述置换检验，这是一种非参数检验，可用于比较两个数据集 $\vec{x}_A$ 和 $\vec{x}_B$，以评估形如
$\vec{x}_A$ *is sampled from a distribution that has a higher mean than* $\vec{x}_B$ 或 $\vec{x}_B$ *is sampled from a distribution that has a higher variance than* $\vec{x}_A$ 的猜想。零假设是这两个数据集实际上是从同一分布中抽样得到的。

置换检验中的检验统计量是对两个数据集上评估的感兴趣检验统计量 t 的值之间的差异。

$$t_{\text{diff}}\left(\vec{x}\right) := t\left(\vec{x}_A\right) - t\left(\vec{x}_B\right), \tag{11.43}$$

其中 $\vec{x}$ 是所有数据合并在一起。我们的目标是检验 $t\left(\vec{x}_A\right)$ 是否大于 $t\left(\vec{x}_B\right)$ 在某个显著性水平下。相应的拒绝区域的形式为 $\mathcal{R} := \{t \mid t \geq \eta\}$。问题是如何设定阈值，以使得检验具有所需的显著性水平。

设想我们在合并的数据集 $\vec{x}$ 中随机置换标签 A 和 B。结果是，一些原本标记为 A 的数据将被标记为 B，反之亦然。如果我们重新计算 $t_{\text{diff}}\left(\vec{x}\right)$，显然会得到一个不同的值。然而，在假设数据是从同一分布中抽样的前提下，随机变量 $t_{\text{diff}}(\vec{X})$ 的 *distribution* 已经 *not* 改变。的确，零假设意味着，任何仅依赖于分配给每个变量的类别的$\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_n$ 的函数，其分布是 *invariant to permu- tations*。更正式地说，对于此类函数，该随机序列是 **exchangeable**。

考虑在所有可能的标签排列 $t_{\text{diff},1}, t_{\text{diff},2}, \ldots t_{\text{diff},n!}$ 下 t_{diff} 的取值。如果原假设成立，那么发现 $t_{\text{diff}}\left(\vec{x}\right)$ 大于大多数 $t_{\text{diff},i}$ 将会令人惊讶。事实上，在原假设下，随机变量 $t_{\text{diff}}(\vec{X})$ 在集合 $\{t_{\text{diff},1}, t_{\text{diff},2}, \ldots t_{\text{diff},n!}\}$ 上服从均匀分布，因此

$$\mathrm{P}\left(t_{\text{diff}}(\vec{X}) \geq \eta\right) = \frac{1}{n!}\sum_{i=1}^{n!} 1_{t_{\text{diff},i} \geq \eta}. \tag{11.44}$$

这正是测试的大小。因此，我们可以计算观察到的统计量 $t_{\text{diff}}\left(\vec{x}\right)$ 的 p 值，如下所示

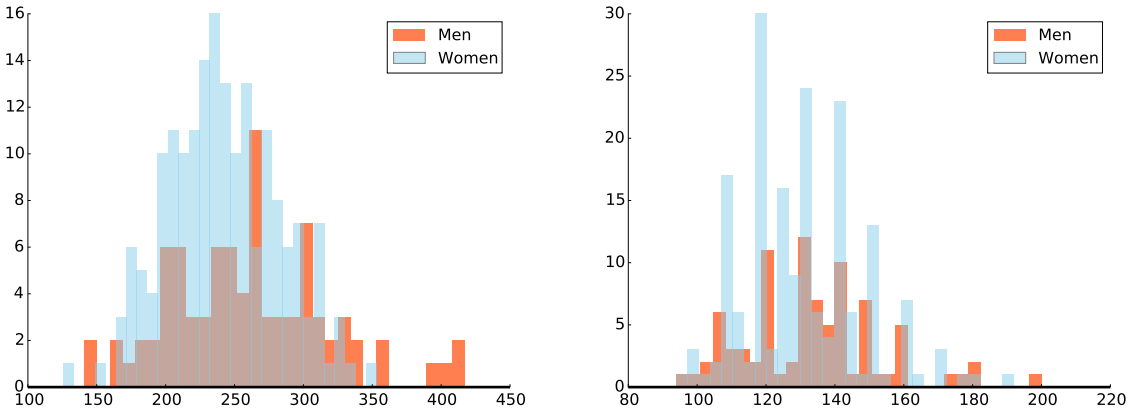
$$p = \mathrm{P}\left(t_{\text{diff}}(\vec{X}) \geq t_{\text{diff}}\left(\vec{x}\right)\right) \tag{11.45}$$

$$= \frac{1}{n!}\sum_{i=1}^{n!} 1_{t_{\text{diff},i} \geq t_{\text{diff}}(\vec{x})}. \tag{11.46}$$

用文字来说，p 值是产生比我们观测到的检验统计量更极端结果的置换所占的比例。不幸的是，精确计算 (11.46) 往往具有挑战性。即使对于中等规模的数据集，可能的置换数量通常也过于庞大（例如，$40! > 8\,10^{47}$），从而在计算上不可行。在这种情况下，可以通过抽样大量置换，并用其平均值对 (11.46) 进行蒙特卡罗近似来近似计算 p 值。

在查看示例之前，让我们回顾在应用置换检验时需要遵循的步骤。

1. 选择一个关于 $\vec{x}_A$ 和 $\vec{x}_B$ 有何不同的猜想。



Cholesterol Blood pressure

Figure 11.2: 示例 11.3.1 中男女的胆固醇和血压的直方图。

2. 选择一个检验统计量 t_{diff}。 3. 计算 $t_{\text{diff}}(\vec{x})$。 4. 对标签进行 m 次置换，并计算相应的 t_{diff} 值：$t_{\text{diff},1}$、$t_{\text{diff},2}$、$\cdots\cdots t_{\text{diff},m}$。 5. 计算近似 p 值。

$$p = \mathrm{P}\left(t_{\text{diff}}(\vec{X}) \geq t_{\text{diff}}(\vec{x})\right) \tag{11.47}$$

$$= \frac{1}{m}\sum_{i=1}^{m} 1_{t_{\text{diff},i} \geq t_{\text{diff}}(\vec{x})} \tag{11.48}$$

如果它低于预先设定的阈值（通常为1%或5%），则拒绝零假设。

Example 11.3.1 (胆固醇和血压)**.** 一位科学家想要确定男性是否有更高的胆固醇和血压。她从86名男性和182名女性中收集数据。图11.2显示了男性和女性的胆固醇和血压直方图。从直方图来看，似乎男性的胆固醇和血压水平较高。男性的胆固醇样本均值为261.3 mg/dl，女性为242.0 mg/dl。男性的血压样本均值为133.2 mmHg，女性为130.6 mmHg。

为了量化这些差异是否显著，我们计算了样本均值差异的样本置换分布，使用了10^6次置换。为了确保结果的稳定性，我们重复了该过程三次。结果如图11.3所示。对于胆固醇，p值约为0.1%，因此我们有非常强的证据反对零假设。相比之下，血压的p值为13%，因此结果并不非常明确，我们不能排除差异仅仅是由于随机波动所致。

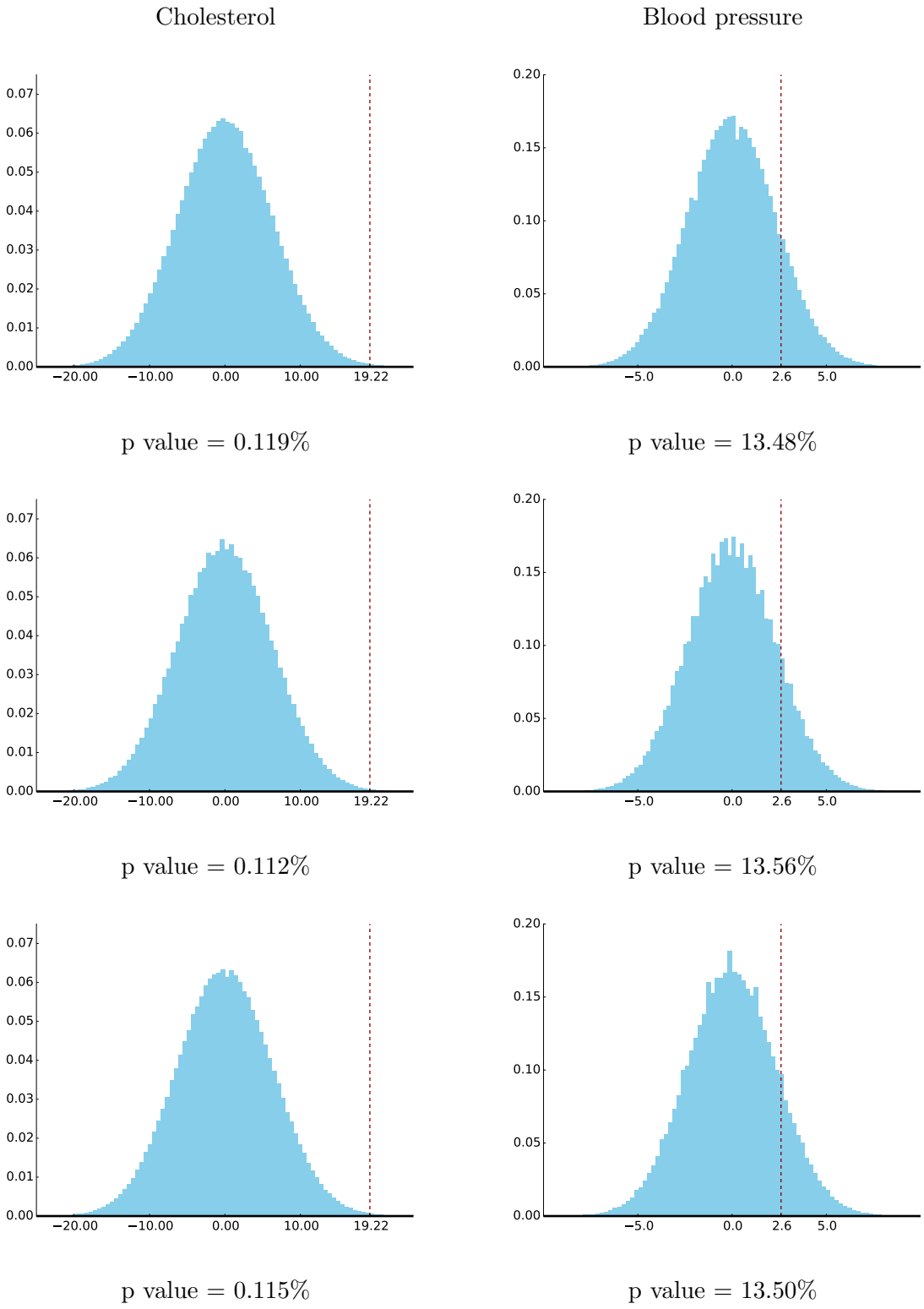


Figure 11.3: 在零假设下，男性和女性的胆固醇与血压样本均值差异的近似分布。检验统计量的观察值由虚线标出。

△

11.4 Multiple testing

在一些应用中，进行许多同时的假设检验是常见的。例如，在计算基因组学中，研究人员可能会有兴趣测试一组几千个基因中是否有任何一个与某种疾病相关。如果我们在这种情况下应用一个大小为 α 的假设检验，那么获得某个特定基因的假阳性概率为 α。现在，假设我们检验 n 个基因，并且事件 *gene i is a false positive*、$1 \leq i \leq n$ 都是相互独立的。获得至少一个假阳性的概率是

$$\begin{aligned} \mathrm{P}\left(\text{at least one false positive}\right) &= 1 - \mathrm{P}\left(\text{no false positives}\right) && (11.49) \\ &= 1 - (1-\alpha)^n . && (11.50) \end{aligned}$$

对于 $\alpha = 0.01$ 和 $n = 500$，这个概率等于 0.99！如果我们想要控制犯第一类错误的概率，就必须考虑到我们同时在进行多重检验。实现这一点的一种常用方法是 **Bonferroni's method**。

Definition 11.4.1 (Bonferroni 方法)**.** *Given n hypothesis tests, compute the corresponding p values $p_1, \ldots, p_n$. For a fixed significance level α reject the ith null hypothesis if*

$$p_i \leq \frac{\alpha}{n}. \tag{11.51}$$

以下引理表明，该方法保证所有检验的期望显著性水平同时成立。

Lemma 11.4.2. *If we apply Bonferroni's method, the probability of making a Type I error is bounded by α.*

Proof. 该结果直接由并集界得到，并集界通过各个事件概率之和来控制事件并集的概率。

Theorem 11.4.3 (并集界)**.** *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $S_1, S_2, \ldots$ a collection of events in $\mathcal{F}$. Then*

$$\mathrm{P}\left(\cup_i S_i\right) \leq \sum_i \mathrm{P}\left(S_i\right). \tag{11.52}$$

Proof. 让我们定义这些集合：

$$\tilde{S}_i = S_i \cap \cap_{j=1}^{i-1} S_j^c. \tag{11.53}$$

通过归纳法可以直接证明，对于任意 n，都有 $\cup_{j=1}^n S_j = \cup_{j=1}^n \tilde{S}_j$，因此 $\cup_i S_i = \cup_i \tilde{S}_i$。集合 $\tilde{S}_1$、$\tilde{S}_2$、…… 在构造上是互不相交的，因此

$$\begin{aligned} \mathrm{P}\left(\cup_i S_i\right) = \mathrm{P}\left(\cup_i \tilde{S}_i\right) &= \sum_i \mathrm{P}\left(\tilde{S}_i\right) \quad \text{by Axiom 2 in Definition 1.1.4} && (11.54) \\ &\leq \sum_i \mathrm{P}\left(S_i\right) \quad \text{because } \tilde{S}_i \subseteq S_i. && (11.55) \end{aligned}$$

□

应用该界限,

$$\begin{aligned} \mathrm{P}\left(\text{Type I error}\right) &= \mathrm{P}\left(\cup_{i=1}^{n}\text{Type I error for test } i\right) && (11.56)\\ &\leq \sum_{i=1}^{n}\mathrm{P}\left(\text{Type I error for test } i\right) \qquad \text{由并集界} && (11.57)\\ &= n\cdot\frac{\alpha}{n}=\alpha. && (11.58) \end{aligned}$$

□

Example 11.4.4 (关键时刻（续)）**.** 如果我们将例 11.1.4 中的检验应用于 10 名球员，那么仅由于偶然因素而显得“关键”的概率将显著增加。为此，根据 Bonferroni 方法，我们必须将各个检验的 p 值除以 10。结果是，为了保持 0.05 的显著性水平，我们将要求每名球员在 20 场比赛中有 17 场（而不是 15 场，见表 11.2）在最后一节每分钟得到更多分数，才能拒绝原假设。

△

Chapter 12

Linear Regression

在统计学中，回归是刻画某一感兴趣的数量 y（称为 **response** 或 **dependent variable**）与若干观测变量 x_1、x_2、⋯、x_p 之间关系的问题，这些观测变量被称为 **covariates**、**features** 或 **independent variables**。例如，响应变量可以是房屋价格，而协变量可以对应于房屋面积、房间数量、建造年份等。回归模型将描述房屋价格如何受到所有这些因素的影响。

更正式地说，回归模型中的主要假设是，预测变量是通过将函数 h 应用于特征而生成的，随后受到某种未知噪声 z 的扰动，而该噪声通常是加性的，

$$y = h\left(\vec{x}\right) + z. \tag{12.1}$$

目标是从n个响应示例及其相应的特征中学习h

$$\left(y^{(1)}, \vec{x}^{\,(1)}\right), \left(y^{(2)}, \vec{x}^{\,(2)}\right), \ldots, \left(y^{(n)}, \vec{x}^{\,(n)}\right). \tag{12.2}$$

在此　在本章中，我们重点讨论h是线性函数的情况。 It seems like there might be a typo or an incomplete text i

12.1 Linear models

如果形式为12.1的模型中的回归函数h是线性的，那么响应被建模为预测变量的线性组合：

$$y^{(i)} = \vec{x}^{\,(i)\,T}\vec{\beta}^{\,*} + z^{(i)}, \quad 1 \le i \le n, \tag{12.3}$$

其中 $z^{(i)}$ 是未知噪声向量的一个分量。该函数由权重向量 $\vec{\beta}^{\,*} \in \mathbb{R}^p$ 参数化。将线性模型拟合到数据所需的一切就是估计这些权重。

将线性系统 (12.3) 表示为矩阵形式得到线性回归模型的以下表示

$$\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \cdots \\ y^{(n)} \end{bmatrix} = \begin{bmatrix} \vec{x}_1^{\,(1)} & \vec{x}_2^{\,(1)} & \cdots & \vec{x}_p^{\,(1)} \\ \vec{x}_1^{\,(2)} & \vec{x}_2^{\,(2)} & \cdots & \vec{x}_p^{\,(2)} \\ \cdots & \cdots & \cdots & \cdots \\ \vec{x}_1^{\,(n)} & \vec{x}_2^{\,(n)} & \cdots & \vec{x}_p^{\,(n)} \end{bmatrix} \begin{bmatrix} \vec{\beta}_1^{\,*} \\ \vec{\beta}_2^{\,*} \\ \cdots \\ \vec{\beta}_p^{\,*} \end{bmatrix} + \begin{bmatrix} z^{(1)} \\ z^{(2)} \\ \cdots \\ z^{(n)} \end{bmatrix}. \tag{12.4}$$

等价地,

$$\vec{y} = \mathcal{X}\vec{\beta}^* + \vec{z}, \tag{12.5}$$

其中 $\mathcal{X}$ 是一个包含特征的 $n \times p$ 矩阵，$\vec{y}$ 包含响应，$\vec{z} \in \mathbb{R}^n$ 表示噪声。

Example 12.1.1 (GDP).的线性模型 我们考虑构建一个线性模型的问题，用其人口规模和失业率来预测美国某一州的国内生产总值（GDP）。我们拥有如下数据:

	GDP (USD millions)	Population	Unemployment rate (%)
North Dakota	52 089	757 952	2.4
Alabama	204 861	4 863 300	3.8
Mississippi	107 680	2 988 726	5.2
Arkansas	120 689	2 988 248	3.5
Kansas	153 258	2 907 289	3.8
Georgia	525 360	10 310 371	4.5
Iowa	178 766	3 134 693	3.2
West Virginia	73 374	1 831 102	5.1
Kentucky	197 043	4 436 974	5.2
Tennessee	???	6 651 194	3.0

在这个例子中，GDP 是响应变量，人口和失业率是特征。我们的目标是对数据拟合一个线性模型，以便使用线性模型来预测田纳西州的 GDP。我们首先对数据进行中心化和归一化。响应变量和各个特征的平均值为

$$\operatorname{av}(\vec{y}) = 179\ 236, \qquad \operatorname{av}(X) = \begin{bmatrix} 3\ 802\ 073 & 4.1 \end{bmatrix}. \tag{12.6}$$

经验标准差为

$$\operatorname{std}(\vec{y}) = 396\ 701, \qquad \operatorname{std}(X) = \begin{bmatrix} 7\ 720\ 656 & 2.80 \end{bmatrix}. \tag{12.7}$$

我们减去平均值并除以标准差，以便响应和

特征被中心化并且具有相同的尺度,

$$\vec{y} = \begin{bmatrix} -0.321 \\ 0.065 \\ -0.180 \\ -0.148 \\ -0.065 \\ 0.872 \\ -0.001 \\ -0.267 \\ 0.045 \end{bmatrix}, \qquad X = \begin{bmatrix} -0.394 & -0.600 \\ 0.137 & -0.099 \\ -0.105 & 0.401 \\ -0.105 & -0.207 \\ -0.116 & -0.099 \\ 0.843 & 0.151 \\ -0.086 & -0.314 \\ -0.255 & 0.366 \\ 0.082 & 0.401 \end{bmatrix}. \tag{12.8}$$

为了获得田纳西州 GDP 的估计值，我们拟合该模型

$$\vec{y} \approx X\vec{\beta}, \tag{12.9}$$

根据标准差（12.7）进行重新缩放，并使用平均值（12.6）进行重新居中。最终估计为

$$\vec{y}^{\text{Ten}} = \text{av}\left(\vec{y}\right) + \text{std}\left(\vec{y}\right)\left\langle \vec{x}^{\text{Ten}}_{\text{norm}}, \vec{\beta} \right\rangle \tag{12.10}$$

其中 $\vec{x}^{\text{Ten}}_{\text{norm}}$ 使用 av (X) 进行居中，并使用 std (X) 进行标准化。 △

12.2 Least-squares estimation

为了校准线性回归模型，我们需要估计权重向量，以使其能够很好地拟合数据。我们可以通过误差平方和来评估特定选择的$\vec{\beta} \in \mathbb{R}^p$的拟合程度,

$$\sum_{i=1}^{n} \left(y^{(i)} - \vec{x}^{(i)\,T}\vec{\beta}\right)^2 = \left|\left|\vec{y} - \mathcal{X}\vec{\beta}\right|\right|_2^2. \tag{12.11}$$

最小二乘估计 $\vec{\beta}_{\text{LS}}$ 是使该代价函数最小化的权重向量,

$$\vec{\beta}_{\text{LS}} := \arg\min_{\vec{\beta}} \quad \left|\left|\vec{y} - \mathcal{X}\vec{\beta}\right|\right|_2. \tag{12.12}$$

最小二乘代价函数从计算角度来看非常方便，因为它是凸的，可以高效地最小化（实际上，正如我们稍后将看到的，它有一个封闭解）。此外，它还具有直观的几何和概率解释。图12.1显示了在一个简单示例中，使用最小二乘法学习的线性模型，其中只有一个特征（$p = 1$）和40个样本（$n = 40$）。

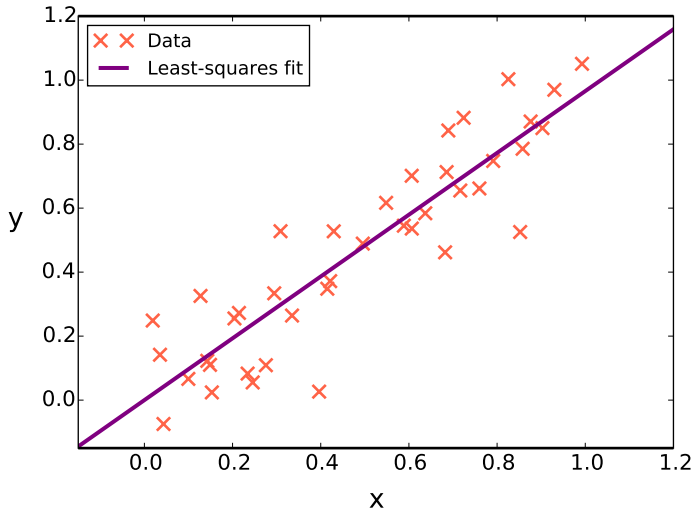


Figure 12.1: 通过最小二乘拟合学习的线性模型，用于一个简单的例子，其中只有一个特征 ($p = 1$) 和 40 个样本 ($n = 40$)。

Example 12.2.1 (线性模型用于GDP（续））. 线性GDP模型中回归系数的最小二乘估计等于

$$\vec{\beta}_{\text{LS}} = \begin{bmatrix} 1.019 \\ -0.111 \end{bmatrix}. \tag{12.13}$$

GDP似乎与人口成正比，与失业率成反比。我们现在将线性模型提供的拟合与原始数据进行比较，并预测田纳西州的GDP:

	GDP	Estimate
North Dakota	52 089	46 241
Alabama	204 861	239 165
Mississippi	107 680	119 005
Arkansas	120 689	145 712
Kansas	153 258	136 756
Georgia	525 360	513 343
Iowa	178 766	158 097
West Virginia	73 374	59 969
Kentucky	197 043	194 829
Tennessee	328 770	345 352

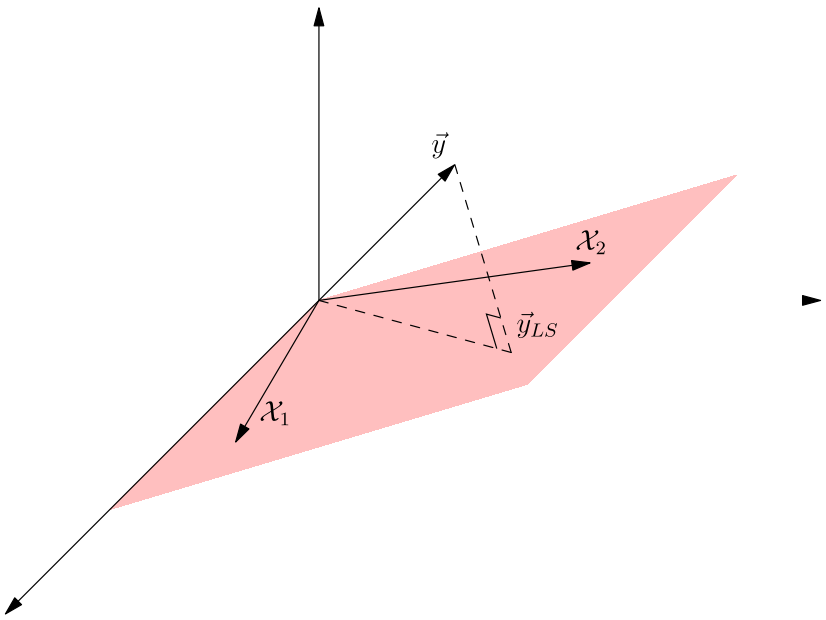


Figure 12.2: 推论12.2.3的示意图。最小二乘解是将数据投影到由 $\mathcal{X}$ 的列所张成的子空间上，该子空间记为 $\mathcal{X}_1$ 和 $\mathcal{X}_2$。

△

12.2.1 Geometric interpretation

下面的定理（在第12.2.2节中证明）表明，最小二乘问题具有闭式解。

Theorem 12.2.2 (最小二乘解). *For $p \geq n$, if $\mathcal{X}$ is full rank then the solution to the least-squares problem* (12.12) *is*

$$\vec{\beta}_{\text{LS}} := \left(\mathcal{X}^T \mathcal{X}\right)^{-1} \mathcal{X}^T \vec{y}. \tag{12.14}$$

该结果的一个推论为 $\vec{y}$ 的最小二乘估计提供了一个几何解释：它是通过将响应向量投影到由预测变量构成的矩阵的列空间上得到的。

Corollary 12.2.3. *For $p \geq n$, if $\mathcal{X}$ is full rank then $\mathcal{X}\vec{\beta}_{\text{LS}}$ is the projection of $\vec{y}$ onto the column space of $\mathcal{X}$.*

我们在附录的第12.5.2节中提供了正式的证明，但结果是非常直观的。任何形如 $\mathcal{X}\vec{\beta}$ 的向量都在 $\mathcal{X}$ 的列空间中。根据定义，最小二乘估计是与 $\vec{y}$ 最接近的向量，并且可以以这种方式表示，因此它是 $\vec{y}$ 在 $\mathcal{X}$ 列空间上的投影。这个过程在图12.2中进行了说明。

12.2.2 Probabilistic interpretation

如果我们将(12.5)中的噪声建模为来自一个随机向量 $\vec{Z}$ 的实现，该向量的元素是均值为零且具有一定方差 σ^2 的独立高斯随机变量，那么我们可以

将最小二乘估计解释为最大似然估计。在这一假设下，数据是随机向量 {v*} 的一个实现。

$$\vec{Y} := \mathcal{X}\vec{\beta} + \vec{Z}, \tag{12.15}$$

这是一个均值为 $\mathcal{X}\vec{\beta}$ 和协方差矩阵为 $\sigma^2 I$ 的独立同分布高斯随机向量。$\vec{Y}$ 的联合概率密度函数等于

$$f_{\vec{Y}}(\vec{a}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(\vec{a}_i - \left(\mathcal{X}\vec{\beta}\right)_i\right)^2\right) \tag{12.16}$$

$$= \frac{1}{\sqrt{(2\pi)^n}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\left|\left|\vec{a} - \mathcal{X}\vec{\beta}\right|\right|_2^2\right). \tag{12.17}$$

似然函数是$\vec{Y}$在观测数据$\vec{y}$处的概率密度函数，并被解释为权重向量$\vec{\beta}$的函数。

$$\mathcal{L}_{\vec{y}}\left(\vec{\beta}\right) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}\left|\left|\vec{y} - \mathcal{X}\vec{\beta}\right|\right|_2^2\right). \tag{12.18}$$

为了找到最大似然估计，我们最大化对数似然函数。我们得出结论，它由最小二乘问题的解给出，因为

$$\vec{\beta}_{\mathrm{ML}} = \arg\max_{\vec{\beta}} \mathcal{L}_{\vec{y}}\left(\vec{\beta}\right) \tag{12.19}$$

$$= \arg\max_{\vec{\beta}} \log \mathcal{L}_{\vec{y}}\left(\vec{\beta}\right) \tag{12.20}$$

$$= \arg\min_{\vec{\beta}} \left|\left|\vec{y} - \mathcal{X}\vec{\beta}\right|\right|_2^2 \tag{12.21}$$

$$= \vec{\beta}_{\mathrm{LS}}. \tag{12.22}$$

12.3 Overfitting

想象一下，一个朋友对你说：

I found a cool way to predict the temperature in New York: It's just a linear combination of the temperature in every other state. I fit the model on data from the last month and a half and it's perfect!

你的朋友并没有撒谎，但问题在于她使用的数据点数量与参数数量大致相同来拟合线性模型。如果 $n \leq p$，我们可以找到一个 $\vec{\beta}$，使得 $\vec{y} = \mathcal{X}\vec{\beta}$ 精确成立，即使 $\vec{y}$ 和 $\mathcal{X}$ 彼此毫无关系！这被称为过拟合，通常是由于在可用数据数量相对于模型而言过少时，使用了过于灵活的模型所导致的。

为了评估一个模型是否存在过拟合，我们将数据分为训练集和测试集。训练集用于拟合模型，测试集用于评估误差。一个过拟合训练集的模型在评估训练样本时会有非常低的误差，但在测试样本上无法很好地泛化。

图 12.3 显示了评估具有 $p =$ 个参数的线性模型的训练误差和测试误差的结果，这些参数是从 n 个训练样本中拟合得到的。训练数据和测试数据是通过固定一个权重向量 $\vec{\beta}^*$ 并计算得到的。

$$\vec{y}_{\text{train}} = \mathcal{X}_{\text{train}}\,\vec{\beta}^* + \vec{z}_{\text{train}}, \tag{12.23}$$

$$\vec{y}_{\text{test}} = \mathcal{X}_{\text{test}}\,\vec{\beta}^*, \tag{12.24}$$

其中，$\mathcal{X}_{\text{train}}$、$\mathcal{X}_{\text{test}}$、$\vec{z}_{\text{train}}$ 和 $\vec{\beta}^*$ 的条目是从均值为零、方差为一的高斯分布中独立随机抽样得到的。训练误差和测试误差定义为

$$\text{error}_{\text{train}} = \frac{\left|\left|\mathcal{X}_{\text{train}}\,\vec{\beta}_{\text{LS}} - \vec{y}_{\text{train}}\right|\right|_2}{||\vec{y}_{\text{train}}||_2}, \tag{12.25}$$

$$\text{error}_{\text{test}} = \frac{\left|\left|\mathcal{X}_{\text{test}}\,\vec{\beta}_{\text{LS}} - \vec{y}_{\text{test}}\right|\right|_2}{||\vec{y}_{\text{test}}||_2}. \tag{12.26}$$

请注意，即使真实的 $\vec{\beta}^*$ 由于噪声的存在也无法达到零训练误差，但如果我们能够精确估计 $\vec{\beta}^*$，测试误差实际上为零。

线性模型的训练误差随着 n 的增加而增长。这是合理的，因为模型必须用相同数量的参数去拟合更多的数据。当 n 接近 $p{:=}\,50$ 时，拟合得到的模型在复现训练数据方面比真实模型好得多（真实模型的误差以绿色显示）。这表明出现了过拟合：模型在适应噪声，而没有学习到真实的线性结构。确实，在这种情况下，测试误差极其高。在更大的 n 下，训练误差上升到真实线性模型所达到的水平，而测试误差下降，这表明我们正在学习潜在的模型。

12.4 Global warming

在本节中，我们描述了线性回归在气候数据中的应用。特别地，我们分析了在牛津的一座气象站收集的温度数据，覆盖了150年[1]。我们的目标不是进行预测，而是确定在过去150年中，牛津的温度是上升还是下降。

为了将温度分解为体现季节影响的不同组成部分，我们使用了一个包含三个预测变量和一个截距项的简单线性模型。

$$\vec{y}_t \approx \vec{\beta}_0 + \vec{\beta}_1 \cos\left(\frac{2\pi t}{12}\right) + \vec{\beta}_2 \sin\left(\frac{2\pi t}{12}\right) + \vec{\beta}_3\, t \tag{12.27}$$

其中 $1 \le t \le n$ 表示以月为单位的时间（n 等于 12 乘以 150）。相应的矩阵

[1]The data is available at `http://www.metoffice.gov.uk/pub/data/weather/uk/climate/stationdata/oxforddata.txt`.

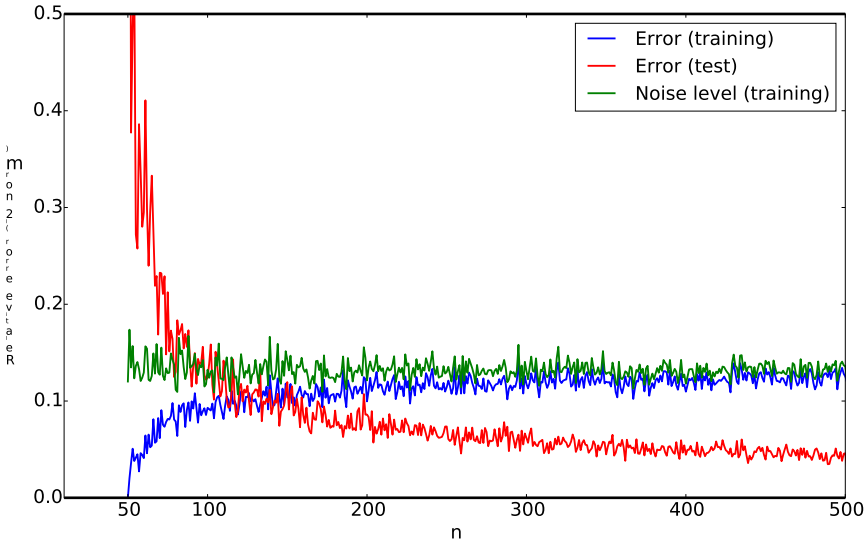


Figure 12.3: 相对 ℓ_2 范数误差，用于估计通过最小二乘回归得到的响应，针对不同的 n (训练数据数量)。训练误差用蓝色表示，而测试误差用红色表示。绿色线条表示用于生成数据的真实模型的训练误差。

预测因子的

$$\mathcal{X} := \begin{bmatrix} 1 & \cos\left(\frac{2\pi t_1}{12}\right) & \sin\left(\frac{2\pi t_1}{12}\right) & t_1 \\ 1 & \cos\left(\frac{2\pi t_2}{12}\right) & \sin\left(\frac{2\pi t_2}{12}\right) & t_2 \\ \dots & \dots & \dots & \dots \\ 1 & \cos\left(\frac{2\pi t_n}{12}\right) & \sin\left(\frac{2\pi t_n}{12}\right) & t_n \end{bmatrix}. \tag{12.28}$$

截距 $\vec{\beta}_0$ 表示平均温度，$\vec{\beta}_1$ 和 $\vec{\beta}_2$ 反映年度周期性波动，$\vec{\beta}_3$ 是总体趋势。如果 $\vec{\beta}_3$ 为正，则模型表明温度在上升；如果为负，则表明温度在下降。

拟合线性模型的结果如图12.4和12.5所示。拟合模型表明，最大和最小温度均呈现出大约0.8摄氏度（约1.4华氏度）的上升趋势。

12.5 Proofs

12.5.1 Proof of Proposition 12.2.2

让 $\mathcal{X} = U\Sigma V_T$ 为 $\mathcal{X}$ 的奇异值分解（SVD）。在定理的条件下，$\left(\mathcal{X}^T\mathcal{X}\right)^{-1}\mathcal{X}^T y = V\Sigma U^T$。我们首先将 $\vec{y}$ 分解为两个分量。

$$y = UU^T y + \left(I - UU^T\right) y \tag{12.29}$$

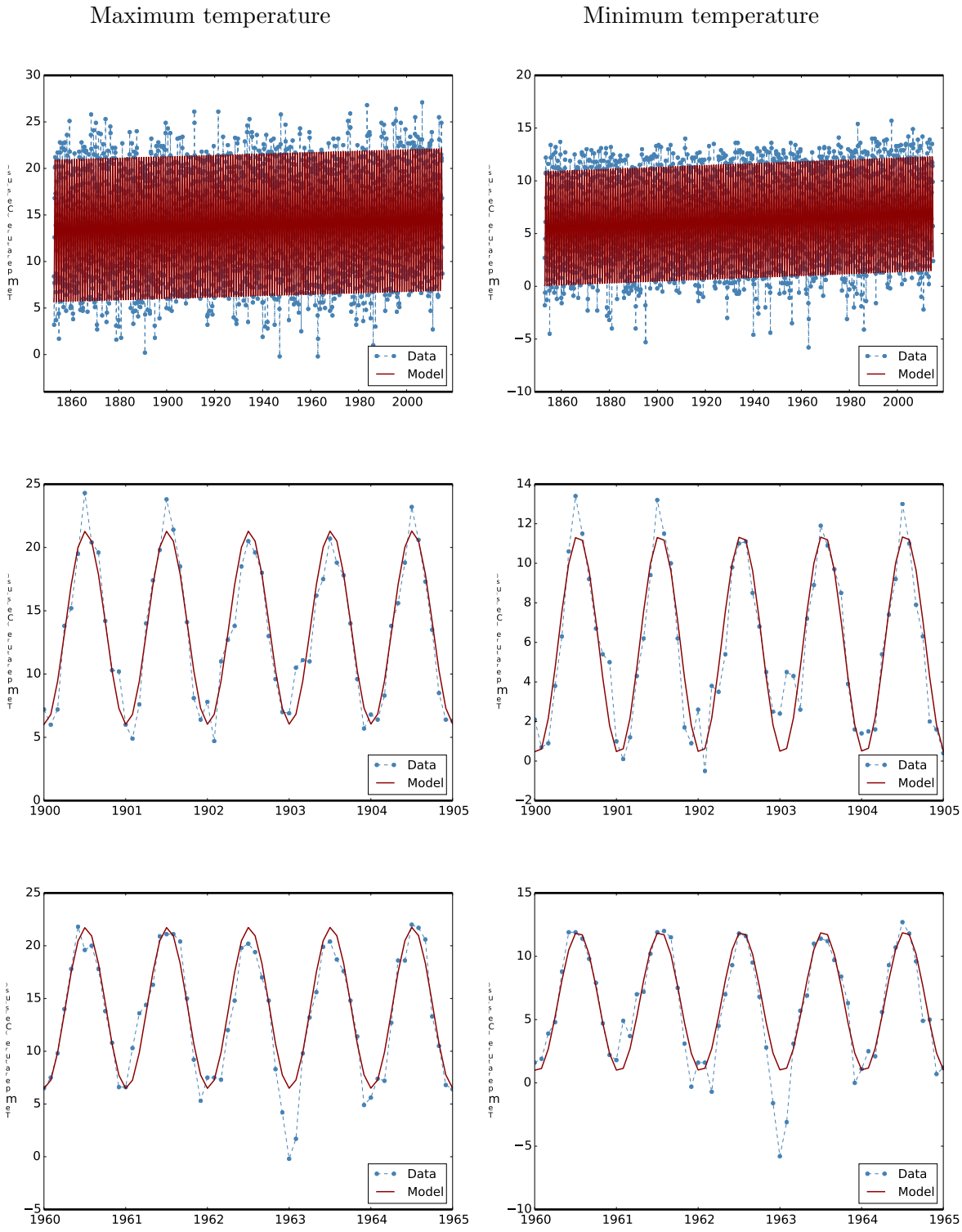


Figure 12.4: 温度数据与由(12.27)描述的线性模型一起，用于最大和最小温度。

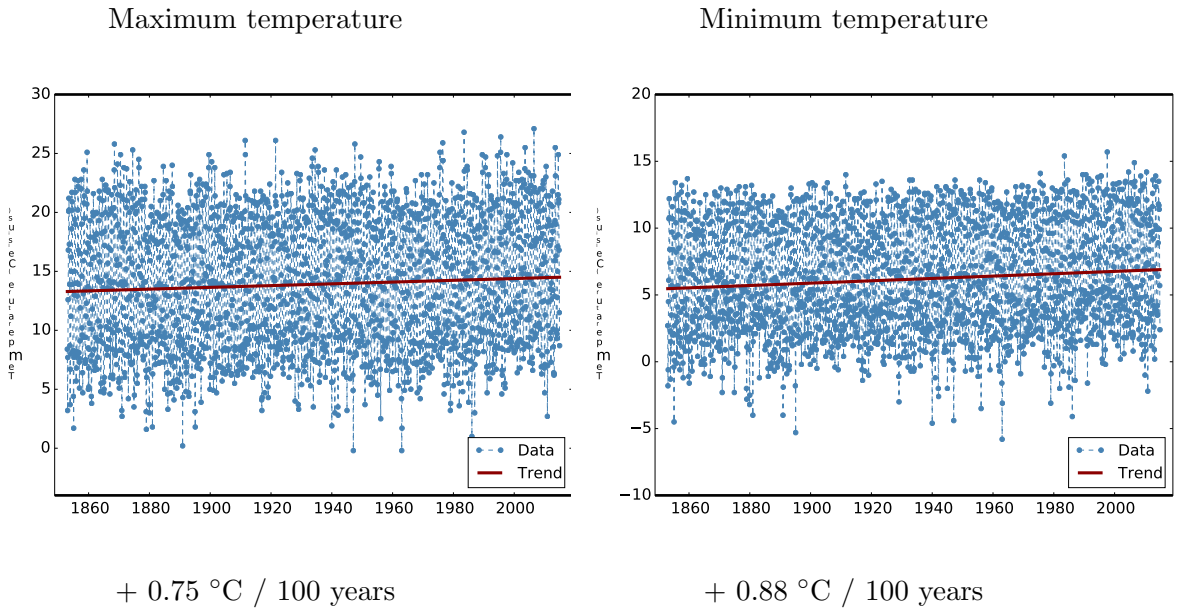


Figure 12.5: 通过拟合由（12.27）式描述的模型，获得的最大和最小温度的温度趋势。

其中 $UU^T y$ 是 $\vec{y}$ 在 $\mathcal{X}$ 的列空间上的投影。请注意，$\left(I - UU^T\right) y$ 与 $\mathcal{X}$ 的列空间正交，因此对于任何 $\vec{\beta}$，$\left(I - UU^T\right) y$ 也正交于 $UU^T y$ 和 $\mathcal{X}\vec{\beta}$。根据毕达哥拉斯定理

$$\left|\left|\vec{y} - \mathcal{X}\vec{\beta}\right|\right|_2^2 = \left|\left|\left(I - UU^T\right) y\right|\right|_2^2 + \left|\left|UU^T y - \mathcal{X}\vec{\beta}\right|\right|_2^2. \tag{12.30}$$

通过对 $\tilde{beta}$ 进行优化可以达到的该代价函数的最小值是 $||\vec{y}_{\mathcal{X}^\perp}||_2^2$。这可以通过求解该方程组来实现

$$UU^T y = \mathcal{X}\vec{\beta} = U\Sigma V_T \vec{\beta}. \tag{12.31}$$

由于 $U^T U = I$ 因为 $p \geq n$，将等式两边相乘得到等效系统

$$U^T y = \Sigma V_T \vec{\beta}. \tag{12.32}$$

由于 $\mathcal{X}$ 是满秩的，Σ 和 V 是方阵且可逆的（根据奇异值分解的定义 $V^{-1} = V^T$），所以

$$\vec{\beta}_{\text{LS}} = V\Sigma U^T y \tag{12.33}$$

是系统的唯一解，因此也是最小二乘问题的解。

12.5.2 Proof of Corollary 12.2.3

令 $\mathcal{X} = U\Sigma V^T$ 为 $\mathcal{X}$ 的奇异值分解。由于 $\mathcal{X}$ 是满秩的且 $p \geq n$，我们有 $U^TU = I$、$V^TV = I$，并且 Σ 是一个可逆的方阵，这意味着

$$\mathcal{X}\vec{\beta}_{\text{LS}} = \mathcal{X}\left(\mathcal{X}^T\mathcal{X}\right)^{-1}\mathcal{X}^T y \tag{12.34}$$

$$= U\Sigma V^T\left(V\Sigma U^T U\Sigma V^T\right)V\Sigma U^T y \tag{12.35}$$

$$= UU^T y. \tag{12.36}$$

Appendix A

Set theory

本章对集合论中的基本概念进行了回顾。

A.1 Basic definitions

一个 **set** 是一组对象。包含我们在某种情境中考虑的所有可能对象的集合称为 **universe**，通常用 Ω 表示。如果 Ω 中的对象 x 属于集合 S，我们称 x 是 S 的 **element**，并写作 $x \in S$。如果 x 不是 S 的元素，我们写作 $x \notin S$。**empty set**，通常用 $\emptyset$ 表示，是一个集合，对于所有 $x \in \in \Omega$ 都有 $x \notin$（即它没有元素）。如果集合 B 中的所有元素也属于集合 A，那么 B 是 A 的 **subset**，我们写作 $B \subseteq A$。如果另外还有至少一个 A 的元素不属于 B，则 B 是 $B \subset A$ 的真子集。

集合的元素可以是任意对象，尤其可以是集合本身。集合的幂集就是这种情况，其定义将在下一节给出。

一种定义集合的有用方法是通过关于其元素的陈述来实现。设 S 为满足某一特定陈述 $s(x)$ 的元素集合，为了定义 S，我们写作

$$S := \{x \mid s(x)\}\,. \tag{A.1}$$

例如，$A := \{x \mid 1 < x < 3\}$ 表示所有大于 1 且小于 3 的元素的集合。让我们使用这种记号来定义一些重要的集合和集合运算。

A.2 Basic operations

Definition A.2.1 (集合运算)**.**

- *The* ***complement*** S^c *of a set* S *contains all elements that are not in* S*.*

$$S^c := \{x \mid x \notin S\}\,. \tag{A.2}$$

- *The* ***union*** *of two sets* A *and* B *contains the objects that belong to* A *or* B*.*

$$A \cup B := \{x \mid x \in A \text{ or } x \in B\}\,. \tag{A.3}$$

This can be generalized to a sequence of sets A_1, A_2, ...

$$\bigcup_n A_n := \{x \mid x \in A_n \text{ for some } n\}, \tag{A.4}$$

where the sequence may be infinite.

- *The **intersection** of two sets A and B contains the objects that belong to A and B.*

$$A \cap B := \{x \mid x \in A \text{ and } x \in B\}. \tag{A.5}$$

Again, this can be generalized to a sequence,

$$\bigcap_n A_n := \{x \mid x \in A_n \text{ for all } n\}. \tag{A.6}$$

- *The **difference** of two sets A and B contains the elements in A that are not in B.*

$$A/B := \{x \mid x \in A \text{ and } x \notin B\}. \tag{A.7}$$

- *The **power set** 2^S of a set S is the set of all possible subsets of S, including $\emptyset$ and S.*

$$2^S := \left\{S' \mid S' \subseteq S\right\}. \tag{A.8}$$

- *The **cartesian product** of two sets S_1 and S_2 is the set of all ordered pairs of elements in the sets*

$$S_1 \times S_2 := \{(x_1, x_2) \mid x_1 \in S_1, x_2 \in S_2\}. \tag{A.9}$$

An example is $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, the set of all possible pairs of real numbers.

如果两个集合具有相同的元素，则它们相等，即 $A = B$ 当且仅当 $A \subseteq B$ 且 $B \subseteq A$。例如，很容易验证 $(A^c)^c = A$，$S \cup \Omega = \Omega$，$S \cap \Omega = S$，或者以下被称为 *De Morgan's laws* 的恒等式。

Theorem A.2.2 (德摩根定律)**.** *For any two sets A and B*

$$(A \cup B)^c = A^c \cap B^c, \tag{A.10}$$
$$(A \cap B)^c = A^c \cup B^c. \tag{A.11}$$

Proof. 让我们证明第一个恒等式；第二个的证明几乎是相同的。

首先我们证明 $(A \cup B)^c \subseteq A^c \cap B^c$。证明一个集合包含于另一个集合的一个标准方法是说明：如果一个元素属于第一个集合，那么它也必然属于第二个集合。任取 $(A \cup B)^c$ (中的任一元素 x ；如果该集合为空，则包含关系显然成立，因为对于任意集合 S)，$\emptyset \subseteq S$ 都属于 A^c；否则它将属于 A，从而也属于 $A \cup B$。同样地，x 也属于 B^c。因此我们得出 x 属于 $A^c \cap B^c$，这就证明了该包含关系。

为完成证明，我们建立 $A^c \cap B^c \subseteq (A \cup B)^c$。如果 $x \in A^c \cap B^c$，则 $x \notin A$ 且 $x \notin B$，因此 $x \notin A \cup B$，从而 $x \in (A \cup B)^c$。 □

Appendix B

Linear Algebra

本章回顾了线性代数中的基本概念。

B.1 Vector spaces

你无疑熟悉**vectors**在$\mathbb{R}^2$或$\mathbb{R}^3$中的应用，即

$$\vec{x} = \begin{bmatrix} 2.2 \\ 3 \end{bmatrix}, \quad \vec{y} = \begin{bmatrix} -1 \\ 0 \\ 5 \end{bmatrix}. \tag{B.1}$$

从代数的角度来看，向量是更为一般的对象。它们是叫做 **vector spaces** 的集合的元素，满足以下定义。

Definition B.1.1 (向量空间)**.** *A vector space consists of a set $\mathcal{V}$ and two operations $+$ and $\cdot$ satisfying the following conditions.*

1. *For any pair of elements $\vec{x}, \vec{y} \in \mathcal{V}$ the **vector sum** $\vec{x} + \vec{y}$ belongs to $\mathcal{V}$.*
2. *For any $\vec{x} \in \mathcal{V}$ and any scalar $\alpha \in \mathbb{R}$ the **scalar multiple** $\alpha \cdot \vec{x} \in \mathcal{V}$.*
3. *There exists a **zero vector** or **origin** $\vec{0}$ such that $\vec{x} + \vec{0} = \vec{x}$ for any $\vec{x} \in \mathcal{V}$.*
4. *For any $\vec{x} \in \mathcal{V}$ there exists an additive inverse $\vec{y}$ such that $\vec{x} + \vec{y} = \vec{0}$, usually denoted as $-\vec{x}$.*
5. *The vector sum is commutative and associative, i.e. for all $\vec{x}, \vec{y} \in \mathcal{V}$*
$$\vec{x} + \vec{y} = \vec{y} + \vec{x}, \quad (\vec{x} + \vec{y}) + z = \vec{x} + (\vec{y} + z). \tag{B.2}$$
6. *Scalar multiplication is associative, for any $\alpha, \beta \in \mathbb{R}$ and $\vec{x} \in \mathcal{V}$*
$$\alpha\,(\beta \cdot \vec{x}) = (\alpha\,\beta) \cdot \vec{x}. \tag{B.3}$$
7. *Scalar and vector sums are both distributive, i.e. for all $\alpha, \beta \in \mathbb{R}$ and $\vec{x}, \vec{y} \in \mathcal{V}$*
$$(\alpha + \beta) \cdot \vec{x} = \alpha \cdot \vec{x} + \beta \cdot \vec{x}, \quad \alpha \cdot (\vec{x} + \vec{y}) = \alpha \cdot \vec{x} + \alpha \cdot \vec{y}. \tag{B.4}$$

*A **subspace** of a vector space $\mathcal{V}$ is a subset of $\mathcal{V}$ that is also itself a vector space.*

从现在起，为了书写方便，我们将忽略标量积的符号 $\cdot$，并将 $\alpha \cdot \vec{x}$ 记作 $\alpha\,\vec{x}$。

Remark B.1.2 (更一般的定义)**.** *We can define vector spaces over an arbitrary field, instead of $\mathbb{R}$, such as the complex numbers $\mathbb{C}$. We refer to any linear algebra text for more details.*

我们可以轻松检查 $\mathbb{R}^n$ 是一个有效的向量空间，配合通常的向量加法和向量-标量乘积。在这种情况下，零向量是全零向量 $\begin{bmatrix}0\,0\,0\,\ldots\end{bmatrix}^T$。在思考向量空间时，最好把 $\mathbb{R}^2$ 或 $\mathbb{R}^3$ 放在脑海中以获得直觉，但同样重要的是要记住，我们可以在许多其他对象上定义向量集，例如无限序列、多项式、函数，甚至是随机变量，如以下示例所示。

向量空间的定义保证，在向量空间 $\mathcal{V}$ 中，任何通过先将向量乘以标量系数再相加而得到的向量的 **linear combination** 都属于 $\mathcal{V}$。给定一组向量，一个自然的问题是它们是否可以表示为彼此的线性组合，即它们是 **linearly dependent** 还是 **independent**。

Definition B.1.3 (线性相关/无关)**.** *A set of m vectors $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_m$ is linearly dependent if there exist m scalar coefficients $\alpha_1, \alpha_2, \ldots, \alpha_m$ which are **not** all equal to zero and such that*

$$\sum_{i=1}^{m} \alpha_i\,\vec{x}_i = \vec{0}. \tag{B.5}$$

*Otherwise, the vectors are **linearly independent**.*

Equivalently, at least one vector in a linearly dependent set can be expressed as the linear combination of the rest, whereas this is not the case for linearly independent sets.

让我们检查等价性。方程（B.5）在某些j下成立，且$\alpha_j \neq$为0，当且仅当

$$\vec{x}_j = \frac{1}{\alpha_j} \sum_{i \in \{1,\ldots,m\}/\{j\}} \alpha_i\,\vec{x}_i. \tag{B.6}$$

我们将一组向量 $\{\vec{x}_1, \ldots, \vec{x}_m\}$ 的 **span** 定义为所有可能的向量线性组合的集合：

$$\operatorname{span}(\vec{x}_1, \ldots, \vec{x}_m) := \left\{ \vec{y} \mid \vec{y} = \sum_{i=1}^{m} \alpha_i\,\vec{x}_i \quad \text{for some } \alpha_1, \alpha_2, \ldots, \alpha_m \in \mathbb{R} \right\}. \tag{B.7}$$

这证明它是一个向量空间。

Lemma B.1.4. *The span of any set of vectors $\vec{x}_1, \ldots, \vec{x}_m$ belonging to a vector space $\mathcal{V}$ is a subspace of $\mathcal{V}$.*

Proof. 跨度是 $\mathcal{V}$ 的一个子集，因为在定义 B.1.1 中的条件 1 和 2。我们现在证明它是一个向量空间。由于 $\mathcal{V}$ 是一个向量空间，定义 B.1.1 中的条件 5、6 和 7 成立。我们通过证明对于跨度中的任意两个元素，条件 1、2、3 和 4 成立来验证这些条件。

$$\vec{y}_1 = \sum_{i=1}^{m} \alpha_i \, \vec{x}_i, \quad \vec{y}_2 = \sum_{i=1}^{m} \beta_i \, \vec{x}_i, \quad \alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_m \in \mathbb{R}, \tag{B.8}$$

$\gamma_1 \, \vec{y}1 + \gamma_2 \, \vec{y}2$ 也属于该跨度。这是因为

$$\gamma_1 \, \vec{y}_1 + \gamma_2 \, \vec{y}_2 = \sum_{i=1}^{m} (\gamma_1 \, \alpha_i + \gamma_2 \, \beta_i) \, \vec{x}_i, \tag{B.9}$$

因此，$\gamma_1 \, \vec{y}_1 + \gamma_2 \, \vec{y}_2$ 位于区间 $(\vec{x}_1, \ldots, \vec{x}_m)$ 中。现在为了证明条件 1，我们设定 $\gamma_1 = \gamma_2 = 1$，条件 2 为 $\gamma_2 = 0$，条件 3 为 $\gamma_1 = \gamma_2 = 0$，条件 4 为 $\gamma_1 = -1, \gamma_2 = 0$。 □

在处理向量空间时，考虑基数最小且能够张成该空间的向量集合是很有用的。这称为该向量空间的**basis**。

Definition B.1.5 (基础)**.** *A basis of a vector space $\mathcal{V}$ is a set of independent vectors $\{\vec{x}_1, \ldots, \vec{x}_m\}$ such that*

$$\mathcal{V} = span\,(\vec{x}_1, \ldots, \vec{x}_m)\,. \tag{B.10}$$

所有向量空间中基的一个重要性质是它们具有相同的基数。

Theorem B.1.6. *If a vector space $\mathcal{V}$ has a basis with finite cardinality then every basis of $\mathcal{V}$ contains the same number of vectors.*

该定理（其证明见 B.8.1 节）使我们能够定义向量空间的 **dimension**。

Definition B.1.7 (维度)**.** *The dimension $dim\,(\mathcal{V})$ of a vector space $\mathcal{V}$ is the cardinality of any of its bases, or equivalently the smallest number of linearly independent vectors that span $\mathcal{V}$.*

该定义与$\mathbb{R}^2$和$\mathbb{R}^3$中的通常几何维数概念一致：一条直线的维数为1，而一个平面的维数为2（只要它们包含原点）。注意，存在无限维向量空间，例如定义在 [0, 1] 上的连续实值函数。

我们用来建模某个问题的向量空间通常称为 **ambient space**，它的维度是 **ambient dimension**。在 $\mathbb{R}^n$ 的情况下，环境维度是 n。

Lemma B.1.8 ($\mathbb{R}^n$)**.** *The dimension of* $\mathbb{R}$ 的维度 ^{n} *is n.*

Proof. 考虑由以下定义的向量集 $\vec{e}_1, \ldots, \vec{e}_n \subseteq \mathbb{R}^n$

$$\vec{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \cdots \\ 0 \end{bmatrix}, \quad \vec{e}_2 = \begin{bmatrix} 0 \\ 1 \\ \cdots \\ 0 \end{bmatrix}, \quad \ldots, \quad \vec{e}_n = \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 1 \end{bmatrix}. \tag{B.11}$$

可以很容易地验证这个集合是一个基。它实际上是 **standard basis** 的 $\mathbb{R}^n$。 □

B.2 Inner product and norm

到目前为止，我们只考虑了加法和标量乘法这两种运算。在本节中，我们介绍第三种运算：两个向量之间的**inner product**。

Definition B.2.1 (内积)**.** *An inner product on a vector space $\mathcal{V}$ is an operation $\langle \cdot, \cdot \rangle$ that maps pairs of vectors to $\mathbb{R}$ and satisfies the following conditions.*

- *It is symmetric, for any $\vec{x}, \vec{y} \in \mathcal{V}$*

$$\langle \vec{x}, \vec{y} \rangle = \langle \vec{y}, \vec{x} \rangle . \tag{B.12}$$

- *It is linear, i.e. for any $\alpha \in \mathbb{R}$ and any $\vec{x}, \vec{y}, \vec{z} \in \mathcal{V}$*

$$\langle \alpha\, \vec{x}, \vec{y} \rangle = \alpha \langle \vec{y}, \vec{x} \rangle , \tag{B.13}$$
$$\langle \vec{x} + \vec{y}, \vec{z} \rangle = \langle \vec{x}, \vec{z} \rangle + \langle \vec{y}, \vec{z} \rangle . \tag{B.14}$$

- *It is positive semidefinite: $\langle \vec{x}, \vec{x} \rangle$ is nonnegative for all $\vec{x} \in \mathcal{V}$ and if $\langle \vec{x}, \vec{x} \rangle = 0$ then $\vec{x} = 0$.*

一个赋予内积的向量空间称为一个**inner-product space**。内积的一个重要例子是两个向量 $\vec{x}, \vec{y} \in \mathbb{R}^n$之间的**dot product**，如下所示

$$\vec{x} \cdot \vec{y} := \sum_i \vec{x}\,[i]\ \vec{y}\,[i] , \tag{B.15}$$

其中 $\vec{x}[i]$ 是 $\vec{x}$ 的第 i 个分量。在本节中我们用 $\vec{x}_i$ 表示一个向量，但在笔记的其他部分它也可能表示向量 $\vec{x}$ 的一个分量；这一点将由上下文清楚。很容易验证点积是一个合法的内积。赋予点积的 $\mathbb{R}^n$ 通常称为维数为 n 的欧几里得空间。

norm 的向量是 *length* 概念的一个推广。

Definition B.2.2 (Norm)**.** *Let $\mathcal{V}$ be a vector space, a norm is a function $||\cdot||$ from $\mathcal{V}$ to $\mathbb{R}$ that satisfies the following conditions.*

- *It is homogeneous. For all $\alpha \in \mathbb{R}$ and $\vec{x} \in \mathcal{V}$*

$$||\alpha\, \vec{x}|| = |\alpha|\, ||\vec{x}|| . \tag{B.16}$$

- *It satisfies the **triangle inequality***

$$||\vec{x} + \vec{y}|| \leq ||\vec{x}|| + ||\vec{y}|| . \tag{B.17}$$

  In particular, it is nonnegative (set $\vec{y} = -\vec{x}$).

- $||\vec{x}|| = 0$ *implies that $\vec{x}$ is the zero vector $\vec{0}$.*

配备了范数的向量空间称为赋范空间。在赋范空间中，可以通过向量之差的范数来度量距离。

Definition B.2.3 (距离)**.** *The distance between two vectors $\vec{x}$ and $\vec{y}$ in a normed space with norm $||\cdot||$ is*

$$d(\vec{x},\vec{y}) := ||\vec{x}-\vec{y}||. \tag{B.18}$$

内积空间是赋范空间，因为我们可以使用内积定义一个有效的范数。由内积得到的范数 **induced** 是通过取向量与其自身的内积的平方根来获得的，

$$||\vec{x}||_{\langle\cdot,\cdot\rangle} := \sqrt{\langle\vec{x},\vec{x}\rangle}. \tag{B.19}$$

由内积诱导的范数显然是齐次的，因为内积具有线性和对称性。$||\vec{x}||_{\langle\cdot,\cdot\rangle}=0$ 说明 $\vec{x}=0$，因为内积是正半定的。我们只需要证明三角不等式成立，以确保内积是有效的范数。这可以从线性代数中的经典不等式得到，该不等式在第 B.8.2 节中得到了证明。

Theorem B.2.4 (柯西-施瓦茨不等式)**.** *For any two vectors $\vec{x}$ and $\vec{y}$ in an inner-product space*

$$|\langle\vec{x},\vec{y}\rangle| \le ||\vec{x}||_{\langle\cdot,\cdot\rangle}\,||\vec{y}||_{\langle\cdot,\cdot\rangle}. \tag{B.20}$$

Assume $||\vec{x}||_{\langle\cdot,\cdot\rangle}\neq 0$,

$$\langle\vec{x},\vec{y}\rangle = -\,||\vec{x}||_{\langle\cdot,\cdot\rangle}\,||\vec{y}||_{\langle\cdot,\cdot\rangle} \iff \vec{y} = -\frac{||\vec{y}||_{\langle\cdot,\cdot\rangle}}{||\vec{x}||_{\langle\cdot,\cdot\rangle}}\vec{x}, \tag{B.21}$$

$$\langle\vec{x},\vec{y}\rangle = ||\vec{x}||_{\langle\cdot,\cdot\rangle}\,||\vec{y}||_{\langle\cdot,\cdot\rangle} \iff \vec{y} = \frac{||\vec{y}||_{\langle\cdot,\cdot\rangle}}{||\vec{x}||_{\langle\cdot,\cdot\rangle}}\vec{x}. \tag{B.22}$$

Corollary B.2.5. *The norm induced by an inner product satisfies the triangle inequality.*

Proof.

$$\begin{aligned} ||\vec{x}+\vec{y}||^2_{\langle\cdot,\cdot\rangle} &= ||\vec{x}||^2_{\langle\cdot,\cdot\rangle} + ||\vec{y}||^2_{\langle\cdot,\cdot\rangle} + 2\langle\vec{x},\vec{y}\rangle && \text{(B.23)}\\ &\le ||\vec{x}||^2_{\langle\cdot,\cdot\rangle} + ||\vec{y}||^2_{\langle\cdot,\cdot\rangle} + 2\,||\vec{x}||_{\langle\cdot,\cdot\rangle}\,||\vec{y}||_{\langle\cdot,\cdot\rangle} \quad \text{by the Cauchy-Schwarz inequality}\\ &= \Big(||\vec{x}||_{\langle\cdot,\cdot\rangle} + ||\vec{y}||_{\langle\cdot,\cdot\rangle}\Big)^2. && \text{(B.24)} \end{aligned}$$

□

欧几里得或ℓ_2范数是由$\mathbb{R}^n$中的点积诱导的范数，

$$||\vec{x}||_2 := \sqrt{\vec{x}\cdot\vec{x}} = \sqrt{\sum_{i=1}^{n}\vec{x}[i]^2}. \tag{B.25}$$

在 $\mathbb{R}^2$ 或 $\mathbb{R}^3$ 的情况下，它就是我们通常所认为的向量的长度。

B.3 Orthogonality

线性代数中的一个重要概念是正交性。

Definition B.3.1 (正交性)**.** *Two vectors $\vec{x}$ and $\vec{y}$ are orthogonal if*

$$\langle \vec{x}, \vec{y} \rangle = 0. \tag{B.26}$$

A vector $\vec{x}$ is orthogonal to a set $\mathcal{S}$, if

$$\langle \vec{x}, \vec{s} \rangle = 0, \quad \text{for all } \vec{s} \in \mathcal{S}. \tag{B.27}$$

Two sets of $\mathcal{S}_1, \mathcal{S}_2$ are orthogonal if for any $\vec{x} \in \mathcal{S}_1, \vec{y} \in \mathcal{S}_2$

$$\langle \vec{x}, \vec{y} \rangle = 0. \tag{B.28}$$

The orthogonal complement of a subspace $\mathcal{S}$ is

$$\mathcal{S}^{\perp} := \{ \vec{x} \mid \langle \vec{x}, \vec{y} \rangle = 0 \quad \text{for all } \vec{y} \in \mathcal{S} \}. \tag{B.29}$$

用由内积诱导的范数来度量正交向量之间的距离很容易计算。

Theorem B.3.2 (勾股定理)**.** *If $\vec{x}$ and $\vec{y}$ are orthogonal vectors*

$$||\vec{x} + \vec{y}||^2_{\langle \cdot,\cdot \rangle} = ||\vec{x}||^2_{\langle \cdot,\cdot \rangle} + ||\vec{y}||^2_{\langle \cdot,\cdot \rangle}. \tag{B.30}$$

Proof. 通过内积的线性性

$$\begin{aligned} ||\vec{x} + \vec{y}||^2_{\langle \cdot,\cdot \rangle} &= ||\vec{x}||^2_{\langle \cdot,\cdot \rangle} + ||\vec{y}||^2_{\langle \cdot,\cdot \rangle} + 2 \langle \vec{x}, \vec{y} \rangle & \text{(B.31)} \\ &= ||\vec{x}||^2_{\langle \cdot,\cdot \rangle} + ||\vec{y}||^2_{\langle \cdot,\cdot \rangle}. & \text{(B.32)} \end{aligned}$$

□

如果我们想要证明一个向量与某个子空间正交，证明它与子空间基底中的每个向量正交就足够了。

Lemma B.3.3. *Let $\vec{x}$ be a vector and $\mathcal{S}$ a subspace of dimension n. If for any basis $\vec{b}_1, \vec{b}_2, \ldots, \vec{b}_n$ of $\mathcal{S}$,*

$$\left\langle \vec{x}, \vec{b}_i \right\rangle = 0, \quad 1 \le i \le n, \tag{B.33}$$

then $\vec{x}$ is orthogonal to $\mathcal{S}$.

Proof. 任何向量 $v \in \mathcal{S}$ 都可以表示为 $v = \sum_i \alpha^n_{i=1} \vec{b}_i$，用于 $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$，见 (B.33)

$$\langle \vec{x}, v \rangle = \left\langle \vec{x}, \sum_i \alpha^n_{i=1} \vec{b}_i \right\rangle = \sum_i \alpha^n_{i=1} \left\langle \vec{x}, \vec{b}_i \right\rangle = 0. \tag{B.34}$$

□

我们现在引入正交规范基。

Definition B.3.4 (正交规范基)**.** *A basis of mutually orthogonal vectors with norm equal to one is called an* ***orthonormal*** *basis.*

在正交规范基中，找到一个向量的系数是非常容易的：我们只需要与基向量计算点积。

Lemma B.3.5 (在正交标准基)**.** *If* $\{\vec{u}_1, \ldots, \vec{u}_n\}$ *is an orthonormal basis of a vector space* $\mathcal{V}$*, for any vector* $\vec{x} \in \mathcal{V}$中的系数

$$\vec{x} = \sum_{i=1}^{n} \langle \vec{u}_i, \vec{x} \rangle \, \vec{u}_i. \tag{B.35}$$

Proof. 由于 $\{\vec{u}_1, \ldots, \vec{u}_n\}$ 是一个基础,

$$\vec{x} = \sum_{i=1}^{m} \alpha_i \, \vec{u}_i \quad \text{for some } \alpha_1, \alpha_2, \ldots, \alpha_m \in \mathbb{R}. \tag{B.36}$$

立即,

$$\langle \vec{u}_i, \vec{x} \rangle = \left\langle \vec{u}_i, \sum_{i=1}^{m} \alpha_i \, \vec{u}_i \right\rangle = \sum_{i=1}^{m} \alpha_i \, \langle \vec{u}_i, \vec{u}_i \rangle = \alpha_i \tag{B.37}$$

因为 $\langle \vec{u}_i, \vec{u}_i \rangle = 1$ 和 $\langle \vec{u}_i, \vec{u}_j \rangle = 0$ 对于 $i \neq j$。 □

对于 $\mathbb{R}^n$ 的一个 *any* 维子空间，我们可以通过对张成该子空间的一组线性无关向量应用格拉姆–施密特方法来获得一个标准正交基。

Algorithm B.3.6 (Gram-Schmidt)**.** *Consider a set of linearly independent vectors* $\vec{x}_1, \ldots, \vec{x}_m$ *in* $\mathbb{R}^n$*. To obtain an orthonormal basis of the span of these vectors we:*

1. *Set* $\vec{u}_1 := \vec{x}_1 / \|\vec{x}_1\|_2$.
2. *For* $i = 1, \ldots, m$*, compute*

$$\vec{v}_i := \vec{x}_i - \sum_{j=1}^{i-1} \langle \vec{u}_j, \vec{x}_i \rangle \, \vec{u}_j. \tag{B.38}$$

and set $\vec{u}_i := \vec{v}_i / \|\vec{v}_i\|_2$.

不难证明，得到的向量集 $\vec{u}_1$，...，$\vec{u}_m$ 是 $\vec{x}_1$，...，$\vec{x}_m$ 的张成空间的正交规范基。这特别意味着我们总是可以假设一个子空间具有正交规范基。

Theorem B.3.7. *Every finite-dimensional vector space has an orthonormal basis.*

Proof. 要看到Gram-Schmidt方法为输入向量的跨度生成一个正交归一基，我们可以检查跨度($\vec{x}_1, \ldots, \vec{x}_i$) = 跨度($\vec{u}_1, \ldots, \vec{u}_i$)，并且$\vec{u}_1, \ldots, \vec{u}_i$是正交归一向量集。

□

B.4 Projections

向量 $\vec{x}$ 在子空间 $\mathcal{S}$ 上的投影是 $\mathcal{S}$ 中与 $\vec{x}$ 距离最近的向量。为了严格地定义这一点，我们首先引入直和的概念。如果两个子空间是不相交的，即它们唯一的公共点是原点，那么能够写成来自每个子空间的向量之和的向量，称为属于它们的直和。

Definition B.4.1 (直和)**.** *Let $\mathcal{V}$ be a vector space. For any subspaces $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{V}$ such that*

$$\mathcal{S}_1 \cap \mathcal{S}_2 = \{0\} \tag{B.39}$$

the direct sum is defined as

$$\mathcal{S}_1 \oplus \mathcal{S}_2 := \{\vec{x} \mid \vec{x} = \vec{s}_1 + \vec{s}_2 \quad \vec{s}_1 \in \mathcal{S}_1, \vec{s}_2 \in \mathcal{S}_2\}. \tag{B.40}$$

向量在两个子空间的直和中的表示是唯一的。

Lemma B.4.2. *Any vector $\vec{x} \in \mathcal{S}_1 \oplus \mathcal{S}_2$ has a **unique** representation*

$$\vec{x} = \vec{s}_1 + \vec{s}_2 \quad \vec{s}_1 \in \mathcal{S}_1, \vec{s}_2 \in \mathcal{S}_2. \tag{B.41}$$

Proof. 如果 $\vec{x} \in \mathcal{S}_1 \oplus \mathcal{S}_2$，则根据定义，存在 $\vec{s}_1 \in \mathcal{S}_1, \vec{s}_2 \in \mathcal{S}_2$ 使得 $\vec{x} = \vec{s}_1 + \vec{s}_2$。假设 $\vec{x} = \vec{v}_1 + \vec{v}_2$，$\vec{v}_1 \in \mathcal{S}_1, \vec{v}_2 \in \mathcal{S}_2$，则 $\vec{s}_1 - \vec{v}_1 = \vec{s}_2 - \vec{v}_2$。这意味着 $\vec{s}_1 - \vec{v}_1$ 和 $\vec{s}_2 - \vec{v}_2$ 属于 $\mathcal{S}_1$，并且也属于 $\mathcal{S}_2$。然而，$\mathcal{S}_1 \cap \mathcal{S}_2 = \{0\}$，因此我们得出 $\vec{s}_1 = \vec{v}_1$ 和 $\vec{s}_2 = \vec{v}_2$。 □

我们现在可以通过将向量 $\vec{x}$ 分解为属于 $\mathcal{S}$ 的分量和属于其正交补的另一个分量，来定义向量 $\vec{x}$ 在子空间 $\mathcal{S}$ 上的投影。

Definition B.4.3 (正交投影)**.** *Let $\mathcal{V}$ be a vector space. The orthogonal projection of a vector $\vec{x} \in \mathcal{V}$ onto a subspace $\mathcal{S} \subseteq \mathcal{V}$ is a vector denoted by $\mathcal{P}_{\mathcal{S}}\, \vec{x}$ such that $\vec{x} - \mathcal{P}_{\mathcal{S}}\, \vec{x} \in \mathcal{S}^{\perp}$.*

Theorem B.4.4 (正交投影的性质)**.** *Let $\mathcal{V}$ be a vector space. Every vector $\vec{x} \in \mathcal{V}$ has a **unique** orthogonal projection $\mathcal{P}_{\mathcal{S}}\, \vec{x}$ onto any subspace $\mathcal{S} \subseteq \mathcal{V}$ of finite dimension. In particular $\vec{x}$ can be expressed as*

$$\vec{x} = \mathcal{P}_{\mathcal{S}}\, \vec{x} + \mathcal{P}_{\mathcal{S}^{\perp}}\, \vec{x}. \tag{B.42}$$

For any vector $\vec{s} \in \mathcal{S}$

$$\langle \vec{x}, \vec{s} \rangle = \langle \mathcal{P}_{\mathcal{S}}\, \vec{x}, \vec{s} \rangle. \tag{B.43}$$

For any orthonormal basis $\vec{b}_1, \ldots, \vec{b}_m$ of $\mathcal{S}$,

$$\mathcal{P}_{\mathcal{S}}\, \vec{x} = \sum_{i=1}^{m} \left\langle \vec{x}, \vec{b}_i \right\rangle \vec{b}_i. \tag{B.44}$$

Proof. 我们将 $\mathcal{S}$ 的维度表示为 m。由于 m 是有限的，存在一个正交归一基 $\mathcal{S}$: $\vec{b}'_1, \ldots, \vec{b}'_m$。考虑向量

$$\vec{p} := \sum_{i=1}^{m} \left\langle \vec{x}, \vec{b}'_i \right\rangle \vec{b}'_i. \tag{B.45}$$

结果表明，$\vec{x} - \vec{p}$ 与基中的每个向量都正交。对于 $1 \leq j \leq m$,

$$\left\langle \vec{x} - \vec{p}, \vec{b}'_j \right\rangle = \left\langle \vec{x} - \sum_{i=1}^{m} \left\langle \vec{x}, \vec{b}'_i \right\rangle \vec{b}'_i, \vec{b}'_j \right\rangle \tag{B.46}$$

$$= \left\langle \vec{x}, \vec{b}'_j \right\rangle - \sum_{i=1}^{m} \left\langle \vec{x}, \vec{b}'_i \right\rangle \left\langle \vec{b}'_i, \vec{b}'_j \right\rangle \tag{B.47}$$

$$= \left\langle \vec{x}, \vec{b}'_j \right\rangle - \left\langle \vec{x}, \vec{b}'_j \right\rangle = 0, \tag{B.48}$$

因此，$\vec{x} - \vec{p} \in \mathcal{S}^\perp$ 和 $\vec{p}$ 是一个正交投影。由于 $\mathcal{S} \cap \mathcal{S}^\perp = \{0\}$ [1]，不可能存在另外两个向量 $\vec{x}_1 \in \mathcal{S}, \vec{x}_1 \in \mathcal{S}^\perp$，使得 $\vec{x} = \vec{x}_1 + \vec{x}_2$，因此正交投影是唯一的。

注意到 $\vec{o} := \vec{x} - \vec{p}$ 是 $\mathcal{S}^\perp$ 中的一个向量，使得 $\vec{x} - \vec{o} = \vec{p}$ 位于 S 中，因此也位于 $\left(\mathcal{S}^\perp\right)^\perp$ 中。这意味着 $\vec{o}$ 是 $\vec{x}$ 在 $\mathcal{S}^\perp$ 上的正交投影，并确立了 (B.42)。
公式 (B.43) 直接源于任意向量 $\vec{s} \in \mathcal{S}$ 与 $\mathcal{P}_\mathcal{S}\, \vec{x}$ 的正交性。公式 (B.44) 由 (B.43) 推出。

□

如果我们能够获得一个正交归一基，那么计算向量在子空间上的投影的范数就很容易（只要该范数是由内积诱导的）。

Lemma B.4.5 (投影)**.** *The norm of the projection of an arbitrary vector $\vec{x} \in \mathcal{V}$ onto a subspace $\mathcal{S} \subseteq \mathcal{V}$ of dimension d can be written as*的范数

$$||\mathcal{P}_\mathcal{S}\, \vec{x}||_{\langle \cdot, \cdot \rangle} = \sqrt{\sum_{i}^{d} \left\langle \vec{b}_i, \vec{x} \right\rangle^2} \tag{B.49}$$

for any orthonormal basis $\vec{b}_1, \ldots, \vec{b}_d$ of $\mathcal{S}$.

Proof. 通过 (B.44)

$$||\mathcal{P}_\mathcal{S}\, \vec{x}||^2_{\langle \cdot, \cdot \rangle} = \langle \mathcal{P}_\mathcal{S}\, \vec{x}, \mathcal{P}_\mathcal{S}\, \vec{x} \rangle \tag{B.50}$$

$$= \left\langle \sum_{i}^{d} \left\langle \vec{b}_i, \vec{x} \right\rangle \vec{b}_i, \sum_{j}^{d} \left\langle \vec{b}_j, \vec{x} \right\rangle \vec{b}_j \right\rangle \tag{B.51}$$

$$= \sum_{i}^{d} \sum_{j}^{d} \left\langle \vec{b}_i, \vec{x} \right\rangle \left\langle \vec{b}_j, \vec{x} \right\rangle \left\langle \vec{b}_i, \vec{b}_j \right\rangle \tag{B.52}$$

$$= \sum_{i}^{d} \left\langle \vec{b}_i, \vec{x} \right\rangle^2. \tag{B.53}$$

□

[1] For any vector $\vec{v}$ that belongs to both $\mathcal{S}$ and $\mathcal{S}^\perp$ $\langle \vec{v}, \vec{v} \rangle = ||\vec{v}||_2^2 = 0$, which implies $\vec{v} = 0$.

Example B.4.6 (投影到一维子空间)**.** 为了计算向量$\vec{x}$在由向量$\vec{v}$张成的一维子空间上的投影，我们利用$\left\{\vec{v}/\left|\left|\vec{v}\right|\right|_{\langle\cdot,\cdot\rangle}\right\}$是张成$(\vec{v})$的基（它是包含一个单位向量的集合，该单位向量张成该子空间）的事实，并应用（B.44）来得到

$$\mathcal{P}_{\text{span}(\vec{v})}\,\vec{x} = \frac{\langle\vec{v},\vec{x}\rangle}{\left|\left|\vec{v}\right|\right|^2_{\langle\cdot,\cdot\rangle}}\vec{v}. \tag{B.54}$$

△

最终，我们证明了向量 $\vec{x}$ 在子空间 $\mathcal{S}$ 上的投影确实是 $\mathcal{S}$ 中与 $\vec{x}$ 在内积范数诱导的距离下最接近的向量。

Theorem B.4.7 (正交投影是最接近的)**.** *The orthogonal projection of a vector $\vec{x}$ onto a subspace $\mathcal{S}$ belonging to the same inner-product space is the closest vector to $\vec{x}$ that belongs to $\mathcal{S}$ in terms of the norm induced by the inner product. More formally, $\mathcal{P}_{\mathcal{S}}\,\vec{x}$ is the solution to the optimization problem*

$$\underset{\vec{u}}{\text{minimize}} \qquad \left|\left|\vec{x}-\vec{u}\right|\right|_{\langle\cdot,\cdot\rangle} \tag{B.55}$$

$$\text{subject to} \qquad \vec{u}\in\mathcal{S}. \tag{B.56}$$

Proof. 取任意点 $\vec{s}\in\mathcal{S}$，使得 $\vec{s}\neq\mathcal{P}_{\mathcal{S}}\,\vec{x}$

$$\left|\left|\vec{x}-\vec{s}\right|\right|^2_{\langle\cdot,\cdot\rangle} = \left|\left|x-\mathcal{P}_{\mathcal{S}}\,\vec{x}+\mathcal{P}_{\mathcal{S}}\,\vec{x}-\vec{s}\right|\right|^2_{\langle\cdot,\cdot\rangle} \tag{B.57}$$

$$= \left|\left|x-\mathcal{P}_{\mathcal{S}}\,\vec{x}\right|\right|^2_{\langle\cdot,\cdot\rangle} + \left|\left|\mathcal{P}_{\mathcal{S}}\,\vec{x}-\vec{s}\right|\right|^2_{\langle\cdot,\cdot\rangle} \tag{B.58}$$

$$> \left|\left|x-\mathcal{P}_{\mathcal{S}}\,\vec{x}\right|\right|^2_{\langle\cdot,\cdot\rangle} \quad \text{because } \vec{s}\neq\mathcal{P}_{\mathcal{S}}\,\vec{x}, \tag{B.59}$$

其中（B.58）来源于毕达哥拉斯定理，因为$\mathcal{P}_{\mathcal{S}^\perp}\,\vec{x} := x-\mathcal{P}_{\mathcal{S}}\,\vec{x}$ 属于 $S^\perp$，而 $\mathcal{P}_{\mathcal{S}}\,\vec{x}-\vec{s}$ 属于 S。

□

B.5 Matrices

matrix 是一个由数字组成的矩形数组。我们用 $\mathbb{R}^{m\times n}$ 表示 $m\times n$ 矩阵的向量空间。我们将矩阵 A 的第 i 行记为 $A_{i:}$，第 j 列记为 $A_{:j}$，并将 (i,j) 处的元素记为 A_{ij}。矩阵的转置是通过交换其行和列得到的。

Definition B.5.1 (转置)**.** *The **transpose** A^T of a matrix $A\in\mathbb{R}^{m\times n}$ is a matrix in $A\in\mathbb{R}^{m\times n}$*

$$\left(A^T\right)_{ij} = A_{ji}. \tag{B.60}$$

一个**symmetric**矩阵是一个与其转置相等的矩阵。

矩阵通过一种称为矩阵-向量乘积的线性运算将向量映射到其他向量。

Definition B.5.2 (矩阵-向量积)**.** *The product of a matrix $A \in \mathbb{R}^{m\times n}$ and a vector $\vec{x} \in \mathbb{R}^n$ is a vector $A\vec{x} \in \mathbb{R}^n$, such that*

$$(A\vec{x})_i = \sum_{j=1}^{n} A_{ij}\vec{x}\,[j] \tag{B.61}$$

$$= \langle A_{i:}, \vec{x}\rangle\,, \tag{B.62}$$

i.e. the ith entry of $A\vec{x}$ is the dot product between the ith row of A and $\vec{x}$.
Equivalently,

$$A\vec{x} = \sum_{j=1}^{n} A_{:j}\vec{x}\,[j]\,, \tag{B.63}$$

i.e. $A\vec{x}$ is a linear combination of the columns of A weighted by the entries in $\vec{x}$.

可以轻松检查，两个矩阵 A 和 B 的乘积的转置等于转置后的矩阵按相反顺序相乘,

$$(AB)^T = B^T A^T. \tag{B.64}$$

我们可以将两个向量 $\vec{x}$ 和 $\vec{y}$ 之间的点积表示为

$$\langle \vec{x}, \vec{y}\rangle = \vec{x}^T\,\vec{y} = \vec{y}^T\,\vec{x}. \tag{B.65}$$

单位矩阵是一个将任何向量映射到自身的矩阵。

Definition B.5.3 (单位矩阵)**.** *The identity matrix in $\mathbb{R}^{n\times n}$ is*

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{B.66}$$

Clearly, for any $\vec{x} \in \mathbb{R}^n$ $I\vec{x} = \vec{x}$.

Definition B.5.4 (矩阵乘法)**.** *The product of two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$ is a matrix $AB \in \mathbb{R}^{m\times p}$, such that*

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik}B_{k_j} = \langle A_{i:}, B_{:,j}\rangle\,, \tag{B.67}$$

i.e. the (i, j) entry of AB is the dot product between the ith row of A and the jth column of B.
Equivalently, the jth column of AB is the result of multiplying A and the jth column of B

$$AB = \sum_{k=1}^{n} A_{ik}B_{k_j} = \langle A_{i:}, B_{:,j}\rangle\,, \tag{B.68}$$

and ith row of AB is the result of multiplying the ith row of A and B.

方阵可能有逆矩阵。如果有的话，逆矩阵是一个能够逆转任何向量的矩阵效果的矩阵。

Definition B.5.5 (矩阵逆). *The inverse of a square matrix $A \in \mathbb{R}^{n\times n}$ is a matrix $A^{-1} \in \mathbb{R}^{n\times n}$ such that*

$$AA^{-1} = A^{-1}A = \mathrm{I}. \tag{B.69}$$

Lemma B.5.6. *The inverse of a matrix is unique.*

Proof. 假设存在另一个矩阵 M，使得 $AM = \mathrm{I}$，那么

$$\begin{aligned} M &= A^{-1}AM \quad \text{by (B.69)} && \text{(B.70)}\\ &= A^{-1}. && \text{(B.71)} \end{aligned}$$

□

一个重要的矩阵类别是 **orthogonal matrices**。

Definition B.5.7 (正交矩阵). *An orthogonal matrix is a square matrix such that its inverse is equal to its transpose,*

$$U^TU = UU^T = \mathrm{I} \tag{B.72}$$

按照定义，任何正交矩阵的列 $U_{:1}, U_{:2}, \ldots, U_{:n}$ 都具有单位范数并且彼此正交，因此它们构成一组正交归一基（有点令人困惑的是，正交矩阵并不被称为正交归一矩阵）。我们可以将把 U^T 作用到向量 $\vec{x}$ 解释为：计算其在由 U 的列所形成的基中的表示系数。将 U 作用于 $U^T\vec{x}$ 通过用相应的系数缩放每个基向量来恢复 $\vec{x}$:

$$\vec{x} = UU^T\vec{x} = \sum_{i=1}^{n} \langle U_{:i}, \vec{x}\rangle\, U_{:i}. \tag{B.73}$$

将正交矩阵应用于一个向量不会影响其范数，它只是旋转该向量。

Lemma B.5.8 (正交矩阵保持范数). *For any orthogonal matrix $U \in \mathbb{R}^{n\times n}$ and any vector $\vec{x} \in \mathbb{R}^n$,*

$$||U\vec{x}||_2 = ||\vec{x}||_2 . \tag{B.74}$$

Proof. 根据正交矩阵的定义

$$\begin{aligned} ||U\vec{x}||_2^2 &= \vec{x}^TU^TU\vec{x} && \text{(B.75)}\\ &= \vec{x}^T\vec{x} && \text{(B.76)}\\ &= ||\vec{x}||_2^2 . && \text{(B.77)} \end{aligned}$$

□

B.6 Eigendecomposition

一个矩阵 A 的 **eigenvector** $\vec{v}$ 满足

$$A\vec{v} = \lambda\vec{v} \tag{B.78}$$

对于标量 λ，它是相应的 **eigenvalue**。即使 A 是实数，它的特征向量和特征值也可以是复数。

Lemma B.6.1 (特征分解)**.** *If a square matrix $A \in \mathbb{R}^{n\times n}$ has n linearly independent eigenvectors $\vec{v}_1, \ldots, \vec{v}_n$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ it can be expressed in terms of a matrix Q, whose columns are the eigenvectors, and a diagonal matrix containing the eigenvalues,*

$$A = \begin{bmatrix}\vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n\end{bmatrix} \begin{bmatrix}\lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_n\end{bmatrix} \begin{bmatrix}\vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n\end{bmatrix}^{-1} \tag{B.79}$$

$$= Q\Lambda Q^{-1} \tag{B.80}$$

Proof.

$$AQ = \begin{bmatrix}A\vec{v}_1 & A\vec{v}_2 & \cdots & A\vec{v}_n\end{bmatrix} \tag{B.81}$$

$$= \begin{bmatrix}\lambda_1\vec{v}_1 & \lambda_2\vec{v}_2 & \cdots & \lambda_2\vec{v}_n\end{bmatrix} \tag{B.82}$$

$$= Q\Lambda. \tag{B.83}$$

如果一个方阵的列向量全部线性无关，那么该矩阵存在逆矩阵，因此在等式两边乘以 Q^{-1} 即可完成证明。 □

Lemma B.6.2. *Not all matrices have an eigendecomposition*

Proof. 考虑例如矩阵

$$\begin{bmatrix}0 & 1 \\ 0 & 0\end{bmatrix}. \tag{B.84}$$

假设 λ 具有非零特征值，对应的特征向量的条目为 $\vec{v}[1]$ 和 $\vec{v}[2]$，则

$$\begin{bmatrix}\vec{v}[2] \\ 0\end{bmatrix} = \begin{bmatrix}0 & 1 \\ 0 & 0\end{bmatrix} \begin{bmatrix}\vec{v}[1] \\ \vec{v}[2]\end{bmatrix} = \begin{bmatrix}\lambda\vec{v}[2] \\ \lambda\vec{v}[2]\end{bmatrix}, \tag{B.85}$$

这意味着 $\vec{v}[2] = 0$，因此 $\vec{v}[1] = 0$，因为我们假设 $\lambda \neq 0$。这意味着该矩阵没有与非零特征向量相关联的非零特征值。 □

一种有趣的特征分解应用是非常快速地计算连续的矩阵乘积。假设我们想计算

$$AA\cdots A\vec{x} = A^k\vec{x}, \tag{B.86}$$

即，我们希望将 A 应用 $\vec{x}$ k 次。A^k *cannot* 可通过对其各个元素取幂来计算（尝试一个简单的例子来验证这一点）。然而，如果 A 具有特征分解，

$$\begin{aligned} A^k &= Q\Lambda Q^{-1}Q\Lambda Q^{-1}\cdots Q\Lambda Q^{-1} && \text{(B.87)}\\ &= Q\Lambda^k Q^{-1} && \text{(B.88)}\\ &= Q\begin{bmatrix} \lambda_1^k & 0 & \cdots & 0\\ 0 & \lambda_2^k & \cdots & 0\\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_n^k \end{bmatrix} Q^{-1}, && \text{(B.89)} \end{aligned}$$

利用对角矩阵的性质，反复应用矩阵相当于对角线元素的幂运算。这使得可以通过仅使用 3 次矩阵乘法和对 n 个数取幂来计算 k 矩阵乘积。

从高中或本科代数中，你可能记得如何使用行列式计算特征向量。实际上，由于稳定性问题，这通常不是一个可行的选择。一种计算特征向量的常用技术基于以下的洞察。设 $A \in \mathbb{R}^{n\times n}$ 是具有特征分解 $Q\Lambda Q^{-1}$ 的矩阵，且 $\vec{x}$ 是 $\mathbb{R}^n$ 中的一个任意向量。由于 Q 的列是线性无关的，它们构成了 $\mathbb{R}^n$ 的一组基，因此我们可以将 $\vec{x}$ 表示为

$$\vec{x} = \sum_{i=1}^{n} \alpha_i Q_{:i}, \quad \alpha_i \in \mathbb{R},\ 1 \le i \le n. \tag{B.90}$$

现在让我们将 A 应用到 $\vec{x}$ k 次，

$$\begin{aligned} A^k\vec{x} &= \sum_{i=1}^{n} \alpha_i A^k Q_{:i} && \text{(B.91)}\\ &= \sum_{i=1}^{n} \alpha_i \lambda_i^k Q_{:i}. && \text{(B.92)} \end{aligned}$$

如果我们假设特征向量按照其模长排序，并且其中有一个的模长大于其余的，$|\lambda_1| > |\lambda_2| \ge \ldots$，并且 $\alpha_1 \neq 0$（如果我们随机抽取一个 $\vec{x}$，这种情况以很高的概率发生），那么随着 k 变得越来越大，项 $\alpha_1\lambda_1^k Q_{:1}$ 将占据主导。除非在每次应用 A 之前都进行归一化，否则该项将发散或趋于零。将归一化步骤加入到这一过程中，就得到了 **power method**，或称幂迭代，这是数值线性代数中一个极其重要的算法。

Algorithm B.6.3 (幂法)**.**
输入*: A matrix A.*
输出*: An estimate of the eigenvector of A corresponding to the largest eigenvalue.*

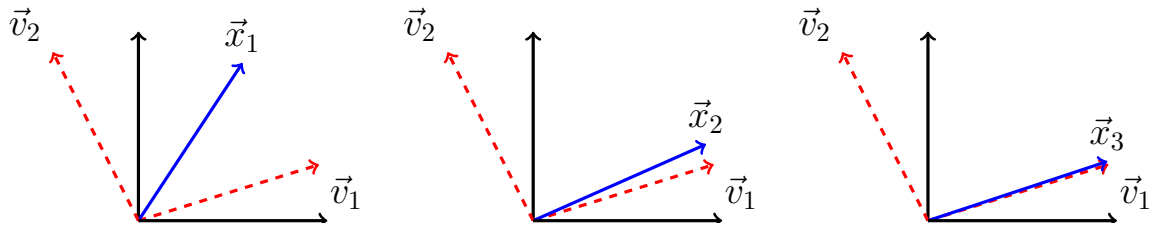


Figure B.1: 矩阵具有特征向量$\vec{v}_1$和$\vec{v}_2$的幂法前三次迭代的示意图，其对应的特征值为$\lambda_1 = 1.05$和$\lambda_2 = 0.1661$。

初始化: *Set* $\vec{x}_1 := \vec{x}/\left\|\vec{x}\right\|_2$, *where the entries of* $\vec{x}$ *are drawn at random. For* $i = 1, \ldots, k$, *compute*

$$\vec{x}_i := \frac{A\vec{x}_{i-1}}{\left\|A\vec{x}_{i-1}\right\|_2}. \tag{B.93}$$

图 B.1 展示了 th e 幂法在一个简单的例子上，其中 ma trix 等于

$$A = \begin{bmatrix} 0.930 & 0.388 \\ 0.237 & 0.286 \end{bmatrix}. \tag{B.94}$$

收敛到与模最大的特征值对应的特征向量的速度非常快。

B.7 Eigendecomposition of symmetric matrices

实对称矩阵总是具有特征分解。此外，它们的特征值都是实数，特征向量两两正交。

Theorem B.7.1 (实对称矩阵的谱定理)**.** *If* $A \in \mathbb{R}^{n\times n}$ *is symmetric, then it has an eigendecomposition of the form*

$$A = \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n \end{bmatrix}^T, \tag{B.95}$$

where the eigenvalues $\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_n$ *are real and the eigenvectors* $\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_n$ *are real and orthogonal.*

Proof. 每个实对称矩阵都有 n 个特征向量的证明超出了这些笔记的范围。在假设这一点成立的前提下，我们首先证明特征值是实数。考虑一个任意的特征值 λ_i 以及对应的归一化特征向量 $\vec{v}_i$ ，我们有

$$\vec{v}_i^* A\vec{v}_i = \lambda \vec{v}_i^* \vec{v}_i = \lambda, \tag{B.96}$$

$$\vec{v}_i^* A\vec{v}_i = \left(A\vec{v}_i\right)^* \vec{v}_i = \left(\lambda\vec{v}_i\right)^* = \bar{\lambda}\vec{v}_i^*\vec{v}_i = \bar{\lambda}. \tag{B.97}$$

这意味着 λ 是实数，因为 $\lambda = \bar{\lambda}$，所以我们可以将特征向量限制为实数（由于特征值是实数，特征向量的实部和虚部都是特征向量，并且至少有一个非零）。如果多个线性无关的特征向量具有相同的特征值，那么它们所张成的正交归一基也将由矩阵的特征向量组成。剩下需要证明的是，对应于不同特征值的特征向量是正交的。假设 $\vec{v}_i$ 和 $\vec{v}_j$ 是对应于不同特征值 $\lambda_i \neq \lambda_j$ 的特征向量，那么

$$\vec{u}_i^T \vec{u}_j = \frac{1}{\lambda_i} \left(A\vec{u}_i\right)^T \vec{u}_j \tag{B.98}$$

$$= \frac{1}{\lambda_i} \vec{u}_i^T A^T \vec{u}_j \tag{B.99}$$

$$= \frac{1}{\lambda_i} \vec{u}_i^T A \vec{u}_j \tag{B.100}$$

$$= \frac{\lambda_j}{\lambda_i} \vec{u}_i^T \vec{u}_j. \tag{B.101}$$

只有在$\vec{u}_i^T \vec{u}_j = 0$的情况下才可能。 □

对称矩阵的特征值决定了**quadratic form**的值：

$$q\left(\vec{x}\right) := \vec{x}^T A \vec{x} = \sum_{i=1}^{n} \lambda_i \left(\vec{x}^T \vec{u}_i\right)^2 \tag{B.102}$$

如果我们对特征值 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ 进行排序，那么第一个特征值是在其输入具有单位 ℓ_2 范数时该二次型所能达到的最大值；第二个特征值是在将其自变量限制为已归一化且与第一个特征向量正交时该二次型所能达到的最大值，依此类推。

Theorem B.7.2. *For any symmetric matrix $A \in \mathbb{R}^n$ with normalized eigenvectors $\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_n$ with corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$*

$$\lambda_1 = \max_{||\vec{u}||_2 = 1} \vec{u}^T A \vec{u}, \tag{B.103}$$

$$\vec{u}_1 = \arg\max_{||\vec{u}||_2 = 1} \vec{u}^T A \vec{u}, \tag{B.104}$$

$$\lambda_k = \max_{||\vec{u}||_2 = 1, \vec{u} \perp \vec{u}_1, \ldots, \vec{u}_{k-1}} \vec{u}^T A \vec{u}, \tag{B.105}$$

$$\vec{u}_k = \arg\max_{||\vec{u}||_2 = 1, \vec{u} \perp \vec{u}_1, \ldots, \vec{u}_{k-1}} \vec{u}^T A \vec{u}. \tag{B.106}$$

Proof. 特征向量构成一个正交归一基（它们彼此正交，并且我们假设它们已经被归一化），因此我们可以将任何与 $\vec{u}_1, \ldots, \vec{u}_{k-1}$ 正交的单位范数向量 $\vec{h}_k$ 表示为

$$\vec{h}_k = \sum_{i=k}^{m} \alpha_i \vec{u}_i \tag{B.107}$$

在哪里

$$\left|\left|\vec{h}_k\right|\right|_2^2 = \sum_{i=k}^{m} \alpha_i^2 = 1, \tag{B.108}$$

由引理 B.4.5 可得。注意，$\vec{h}_1$ 只是一个任意的单位范数向量。
现在我们将证明，当归一化的输入被限制为与 $\vec{u}_1, \ldots, \vec{u}_{k-1}$ 正交时，二次型的取值不可能大于 λ_k，

$$\vec{h}_k^T A \vec{h}_k = \sum_{i=1}^{n} \lambda_i \left(\sum_{j=k}^{m} \alpha_j \vec{u}_i^T \vec{u}_j \right)^2 \quad \text{by (B.102) and (B.107)} \tag{B.109}$$

$$= \sum_{i=1}^{n} \lambda_i \alpha_i^2 \quad \text{because } \vec{u}_1, \ldots, \vec{u}_m \text{ is an orthonormal basis} \tag{B.110}$$

$$\leq \lambda_k \sum_{i=k}^{m} \alpha_i^2 \quad \text{because } \lambda_k \geq \lambda_{k+1} \geq \ldots \geq \lambda_m \tag{B.111}$$

$$= \lambda_k, \quad \text{by (B.108).} \tag{B.112}$$

这就建立了（B.103）和（B.105）。为了证明（B.104）和（B.106），我们只需证明 $\vec{u}_k$ 达到最大值。

$$\vec{u}_k^T A \vec{u}_k = \sum_{i=1}^{n} \lambda_i \left(\vec{u}_i^T \vec{u}_k \right)^2 \tag{B.113}$$

$$= \lambda_k. \tag{B.114}$$

□

B.8 Proofs

B.8.1 Proof of Theorem B.1.6

我们通过反证法证明该断言。假设我们有两个基 $\{\vec{x}_1, \ldots, \vec{x}_m\}$ 和 $\{\vec{y}_1, \ldots, \vec{y}_n\}$，使得 $m < n$（或第二个集合具有无限基数)。该证明通过将下面的引理应用 m 次（设定 $r = 0, 1, \ldots, m-1$），以证明 $\{\vec{y}_1, \ldots, \vec{y}_m\}$ 张成 $\mathcal{V}$，因此 $\{\vec{y}_1, \ldots, \vec{y}_n\}$ 必然线性相关。

Lemma B.8.1. *Under the assumptions of the theorem, if* $\{\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_r, \vec{x}_{r+1}, \ldots, \vec{x}_m\}$ *spans* $\mathcal{V}$ *then* $\{\vec{y}_1, \ldots, \vec{y}_{r+1}, \vec{x}_{r+2}, \ldots, \vec{x}_m\}$ *also spans* $\mathcal{V}$ *(possibly after rearranging the indices* $r+1, \ldots, m$*) for* $r = 0, 1, \ldots, m-1$.

Proof. 由于 $\{\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_r, \vec{x}_{r+1}, \ldots, \vec{x}_m\}$ 跨越 $\mathcal{V}$

$$\vec{y}_{r+1} = \sum_{i=1}^{r} \beta_i \vec{y}_i + \sum_{i=r+1}^{m} \gamma_i \vec{x}_i, \quad \beta_1, \ldots, \beta_r, \gamma_{r+1}, \ldots, \gamma_m \in \mathbb{R}, \tag{B.115}$$

当至少有一个 γ_j 不为零时，因为 $\{\vec{y}_1, \ldots, \vec{y}_n\}$ 按假设是线性无关的。没有失去一般性（这里可能需要重新排列下标），我们假设 $\gamma_{r+1} \neq$ 为 0，因此

$$\vec{x}_{r+1} = \frac{1}{\gamma_{r+1}} \left(\sum_{i=1}^{r} \beta_i \, \vec{y}_i - \sum_{i=r+2}^{m} \gamma_i \vec{x}_i \right). \tag{B.116}$$

这意味着 $\{\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_r, \vec{x}_{r+1}, \ldots, \vec{x}_m\}$ 的跨度中的任何向量，即 $\mathcal{V}$ 中的向量，都可以表示为 $\{\vec{y}_1, \ldots, \vec{y}_{r+1}, \vec{x}_{r+2}, \ldots, \vec{x}_m\}$ 中向量的线性组合，从而完成证明。

□

B.8.2 Proof of Theorem B.2.4

如果 $||\vec{x}||_{\langle \cdot,\cdot \rangle} = 0$，则 $\vec{x} = \vec{0}$，因为内积是半正定的，这意味着 $\langle \vec{x}, \vec{y} \rangle = 0$，从而 (B.20) 以等式成立。如果 $||\vec{y}||_{\langle \cdot,\cdot \rangle} = 0$，情况也是如此。现在假设 $||\vec{x}||_{\langle \cdot,\cdot \rangle} \neq 0$ 且 $||\vec{y}||_{\langle \cdot,\cdot \rangle} \neq 0$。根据内积的半正定性，

$$0 \leq \left|\left| ||\vec{y}||_{\langle \cdot,\cdot \rangle} \, \vec{x} + ||\vec{x}||_{\langle \cdot,\cdot \rangle} \, \vec{y} \right|\right|^2 = 2 \, ||\vec{x}||^2_{\langle \cdot,\cdot \rangle} \, ||\vec{y}||^2_{\langle \cdot,\cdot \rangle} + 2 \, ||\vec{x}||_{\langle \cdot,\cdot \rangle} \, ||\vec{y}||_{\langle \cdot,\cdot \rangle} \, \langle \vec{x}, \vec{y} \rangle \,, \tag{B.117}$$

$$0 \leq \left|\left| ||\vec{y}||_{\langle \cdot,\cdot \rangle} \, \vec{x} - ||\vec{x}||_{\langle \cdot,\cdot \rangle} \, \vec{y} \right|\right|^2 = 2 \, ||\vec{x}||^2_{\langle \cdot,\cdot \rangle} \, ||\vec{y}||^2_{\langle \cdot,\cdot \rangle} - 2 \, ||\vec{x}||_{\langle \cdot,\cdot \rangle} \, ||\vec{y}||_{\langle \cdot,\cdot \rangle} \, \langle \vec{x}, \vec{y} \rangle \,. \tag{B.118}$$

这些不等式证明了（B.20）。

让我们通过证明两个推论来证明 (B.21)。

(⇒) 假设 $\langle \vec{x}, \vec{y} \rangle = -\,||\vec{x}||_{\langle \cdot,\cdot \rangle} \, ||\vec{y}||_{\langle \cdot,\cdot \rangle}$。然后 (B.117) 等于零，因此 $||\vec{y}||_{\langle \cdot,\cdot \rangle} \, \vec{x} = -\,||\vec{x}||_{\langle \cdot,\cdot \rangle} \, \vec{y}$，因为内积是半正定的。

（⇐）假设 $||\vec{y}||_{\langle \cdot,\cdot \rangle} \, \vec{x} = -\,||\vec{x}||_{\langle \cdot,\cdot \rangle} \, \vec{y}$。那么可以容易地验证 (B.117) 等于零，这意味着 $\langle \vec{x}, \vec{y} \rangle = -\,||\vec{x}||_{\langle \cdot,\cdot \rangle} \, ||\vec{y}||_{\langle \cdot,\cdot \rangle}$。

(B.22) 的证明是相同的（使用 (B.118) 替代 (B.117)）。