

| 第一部分 |

Performance Modeling and Design of Computer Systems: Queueing Theory in Action

排队论简介

第一部分是对分析建模的简介。

我们从第 1 章开始，介绍计算机系统设计中出现的一些自相矛盾的例子，展示分析建模在制定设计决策时的强大功能。

第 2 章向读者介绍本书其余部分使用的基本排队论术语和符号，包含开放和封闭排队网络以及标准性能指标，例如响应时间、吞吐量和系统中的作业数量。

第1章 |

Performance Modeling and Design of Computer Systems: Queueing Theory in Action

分析建模的功能及实例

1.1 什么是排队论

假设你有大量作业和稀缺资源，很快排起长队并产生延误——排队论研究的就是这一切背后的理论。这实际上是“队列理论”：是什么让队列出现以及如何使队列消失。

想象一下计算机系统，比如一个 Web 服务器，只有一个作业。作业到来时会使用某些资源(一些 CPU，一些 I/O)，然后作业离开。考虑到作业的资源需求，很容易准确预测作业何时离开。因为没有队列，所以没有延迟。如果每个作业确实要在自己的计算机上运行，就不需要排队论。不幸的是，这种情况很少发生。

排队论适用于队列出现的任何地方(见图 1-1)。我们都有在银行排队等候的经历，想知道为什么没有更多的柜员；或者在超市排队，想知道为什么快速通道支持不多于 8 件商品而不是不多于 15 件商品，或者是否实际上最好有两个快速通道，一个用于不多于 8 件商品，另一个用于不多于 15 件商品。队列也是任何计算机系统的核心。CPU 使用分时调度程序来为等待 CPU 时间的作业队列提供服务。计算机磁盘服务于等待读取或写入块的作业队列。网络中的路由器服务于等待路由的数据包队列。路由器队列是一个有限容量队列，当需求超过缓冲区空间时，数据包将被丢弃。内存库为请求内存块的线程队列提供服务。数据库有时包含有锁队列，事务在其中等待获取记录上的锁。服务器机群由许多服务器组成，每个服务器都有自己的作业队列。类似的例子层出不穷。



图 1-1 客户等待服务的队列和服务器的图示。图为一名顾客在服务器上接受服务，另外五名顾客在排队等候

排队论专家的目标是双重的。第一个是预测系统性能。通常，这意味着预测平均延迟、延迟可变性或延迟超过某些服务水平协议(SLA)的概率。然而，这还意味着可以预测将要排队的作业数量，或正在使用的服务器的平均数量(例如，总功率需求)，或任何其他此类指标。虽然预测很重要，但更重要的目标是找到一种卓越的系统设计来提高性能。通常采用容量规划的形式，其中确定要购买哪些额外资源以满足延迟目标(例如，购买更快的磁盘或更快的 CPU，或者添加第二个慢磁盘)。但是，很多时候可不购买任何额外的资源，只需部署更智能的调度策略或不同的路由策略来减少延迟，就可以提高性能。鉴于智能调度在计算机系统中的重要性，本书的第七部分致力于理解调度策略。

排队论基于更广泛的数学领域，称为随机建模和分析。随机建模将作业的服务需求和作业的间隔时间表示为随机变量。例如，UNIX 进程的 CPU 需求可使用帕累托分布^[84]进行建模，而繁忙的 Web 服务器上作业的到达过程可通过具有指数分布的到达间隔时间的泊松过程进行建模。随机模型也可以用来建立作业之间的依赖关系，以及表示为随机变量

的任何其他模型。

虽然通常可以提出一个随机模型来充分代表系统中的作业或客户及其服务动态，但这些随机模型在求解性能方面并不总是易于分析的。正如我们在第四部分中讨论的马尔可夫假设，例如假设指数分布的服务需求或泊松到达过程，将会大大简化分析。因此，许多现有的排队文献依赖于马尔可夫的这种假设。在许多情况下，这些是合理的近似值。例如，由于有许多独立用户，每个独立用户以低费率提交请求，所以亚马逊上图书订单的到达过程可以通过泊松过程进行合理的近似而估算出来（尽管这一切都在《哈利·波特》推出新版时失败了）。然而，在某些情况下，马尔可夫的假设与现实相差甚远，例如在作业的服务需求高度可变或相关的情况下。

虽然许多讨论排队的文献都淡化了马尔可夫的假设，但本书恰恰相反。我自己的大部分研究都致力于证明工作负载假设对正确预测系统性能的影响。我发现在很多情况下，对工作负载进行简化假设可能会导致非常不准确的性能结果和糟糕的系统设计。因此，在我自己的研究中，我非常重视将测量的工作负载分布集成到分析中。本书不是试图隐藏所做的假设，而是强调了有关工作负载的所有假设。我们将具体讨论工作负载模型是否准确以及我们的模型假设如何影响性能和设计，并且寻找更准确的工作负载模型。在我看来，计算机科学家采用排队论的速度如此之慢的一个主要原因是马尔可夫标准假设通常不合适。然而，通常有办法解决这些假设，其中许多假设都将在本书中展示，例如使用第 21 章介绍的相位型分布和矩阵分析方法。

1.2 排队论实例

本章的其余部分致力于展示排队论的一些具体例子，但不必理解例子中的所有内容。本书稍后将详细介绍这些示例。你可能不熟悉的“泊松过程”等术语也将在本书的后面部分进行说明。这些例子仅用于突出本书中涵盖的课程类型。

如前所述，排队论的一种用途是作为预测工具，允许人们预测给定系统的性能。例如，可能正在分析具有特定带宽的网络，其中不同类别的数据包以特定速率到达并且同时遵循整个网络中的某些路由。然后，排队论可以用来计算数量，例如数据包在特定路由器 i 上等待的平均时间，路由器 i 上队列建立时的分布，或者从网络中的路由器 i 到路由器 j 的平均总时间。

我们现在转向将排队论作为选择最佳系统设计以最小化响应时间的设计工具，看看其有用性如何。以下例子说明系统设计通常是违反直觉的过程。

设计实例 1——到达率翻倍

考虑一个由单个 CPU 组成的系统，它以先来先服务（FCFS）顺序服务于一个作业队列，如图 1-2 所示。这些作业根据一些随机过程到达，具有平均到达率，比如说每秒 $\lambda = 3$ 个作业。每个作业都有一些 CPU 服务要求，独立于作业服务要求的某些分布（我们可以假设在此例子中对作业服务要求进行任何分配）。假设平均服务率为每秒 $\mu = 5$ 个作业（即每个作业平均需要 $1/5$ 秒的服务）。请注意，系统未处于过载状态 ($3 < 5$)。设 $E[T]$ 表示该系统的平均响应时间，其中响应时间是从作业到达完成服务的时间，即逗留时间。

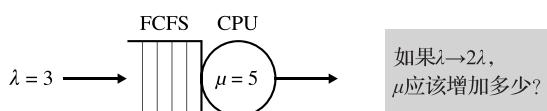


图 1-2 具有单个 CPU 的系统，按 FCFS 顺序服务于作业

问题：老板告诉你，从明天开始，到达率将加倍。你被告知须购买更快的 CPU 以确保作业经历相同的平均响应时间 $E[T]$ 。也就是说，客户不应该注意到增加到达率的影响。你应该将 CPU 的速度提高多少？(a) CPU 速度加倍；(b) CPU 速度的两倍以上；(c) CPU 速度不到两倍。

答案：(c)，不到两倍。

问题：为什么不是(a)？

答案：事实证明，CPU 速度加倍以及到达率加倍通常会导致平均响应时间缩短一半！我们在第 13 章证明了这一点。因此，CPU 速度不需要加倍。

问题：你能否立即看到这个结果的粗略论证，不涉及任何排队论公式？如果我们将服务率翻倍并且到达率加倍会怎样？

答案：想象一下有两种类型的时间：联邦时间和克林贡时间。克林贡秒比联邦秒快。事实上，每个克林贡秒在联邦时间中相当于半秒。现在，假设在联邦时间中，有一个 CPU 服务作业。作业以每秒 λ 个作业的速率到达，并以每秒 μ 个作业的速率服务。克林贡人窃取了系统规格并重新设计了克林贡世界的同一系统。在克林贡系统中，到达率是每克林贡秒 λ 个作业，服务率是每克林贡秒 μ 个作业。请注意，两个系统具有相同的平均响应时间 $E[T]$ ，除了克林贡系统响应时间以克林贡秒测量，而联邦系统响应时间以联邦秒测量。现在考虑一下柯克船长正在观察联邦系统和克林贡重新设计的系统。从他的角度来看，克林贡系统的到达率是联邦系统的两倍，服务率也是其两倍；然而，克林贡系统的平均响应时间减半（因为克林贡秒在联邦时间中是半秒）。

问题：假设 CPU 采用分时服务命令（称为处理器共享或 PS），而不是 FCFS。答案会改变吗？

答案：不。相同的基本论点仍然有效。

设计实例 2——有时“改进”什么都不做

考虑图 1-3 所示的批处理系统。该系统中总有 $N=6$ 个作业（这称为多道程序设计水平）。一旦作业完成服务，就会启动一个新作业（这称为“关闭”系统）。每个作业都必须通过“服务设施”。在服务设施中，有 $1/2$ 的概率作业进入服务器 1，有 $1/2$ 的概率作业进入服务器 2。服务器 1 以每 3 秒 1 个作业的平均速率为作业提供服务，服务器 2 也以每 3 秒 1 个作业的平均速率为作业提供服务。作业服务时间的分配与此问题无关。响应时间通常定义为从作业首次到达服务设施（在分支处）到完成服务的时间。

问题：将服务器 1 替换为速度提高一倍的服务器（新服务器服务作业时平均每 3 秒执行 2 次作业）。这种“改进”是否会影响系统的平均响应时间？它会影响吞吐量吗？（假设进入服务器 1 和服务器 2 的概率始终均为 $1/2$ 。）

答案：不是。平均响应时间和吞吐量几乎都不受影响。这将在第 7 章中解释。

问题：假设系统具有更高的多道程序设计水平 N ，答案是否会改变？

答案：否。随着 N 的增加，对响应时间和吞吐量的影响已经可以忽略不计。

问题：假设系统的 N 值较低。答案会改变吗？

答案：是的。如果 N 足够低，则“改进”有帮助。例如，考虑 $N=1$ 的情况。

问题：假设系统变为开放系统，而不是封闭系统，如图 1-4 所示，其中到达时间与服

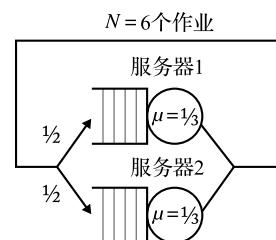


图 1-3 一个封闭的批处理系统

务完成无关。现在，“改进”是否会减少平均响应时间？

答案：绝对是！

设计实例 3——一台机器还是多台机器？

可以在速度为 s 的一个快速 CPU 或速度为 s/n 的 n 个慢速 CPU 之间进行选择（参见图 1-5）。目标是尽量缩短平均响应时间。首先，假设作业是不可抢占的（即每个作业必须运行完成）。

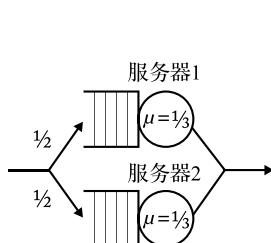


图 1-4 开放系统

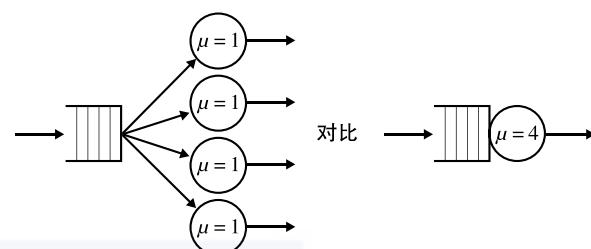


图 1-5 哪种方法可以最大限度地缩短平均响应时间：
许多慢速服务器还是一台快速服务器

问题：哪个是更好的选择：一台快速机器还是许多台慢速机器？

提示：假设我告诉你答案是“这取决于工作负载”。你认为答案取决于工作负载的哪些方面？

答案：事实证明，答案取决于作业规模分布的可变性以及系统负载。

问题：当作业规模变化很大时，你更喜欢哪种系统？

答案：当作业规模变化很大时，我们更喜欢许多慢速服务器，因为我们不希望短作业被困在长作业之后。

问题：当负载较低时，你更喜欢哪种系统？

答案：当负载较低时，并非所有服务器都会被利用，所以最好配合一台快速服务器。在整本书中我们将多次重新审视这些观察结果。

问题：现在假设我们问同样的问题，但作业是可以抢占的，也就是说，它们可以在停止的地方重新启动。与单台快速机器相比，我们何时更喜欢多台慢速机器？

答案：如果作业是可抢占的，你可以始终使用一台快速机器来模拟 n 台慢速机器的效果。因此，单台快速机器至少同样好。

许多慢速服务器与一些快速服务器的问题适用于很多领域，因为任何东西都可以被视为一种资源，包括 CPU、功率和带宽。

有关数据中心电源管理的例子，请考虑来自[69]的问题，其中假设具有固定功率预算 P 和由 n 个服务器组成的服务器机群。你必须决定为每个服务器分配多少功率，以便最小化到达服务器机群的作业的总体平均响应时间。有一个函数指定分配给服务器的功率与其运行的速度（频率）之间的关系——通常，你分配给服务器的功率越大，运行的速度越快（其频率越高），这受制于打开服务器所需的最大可能频率和一些最低功率水平。要回答如何分配电源的问题，你需要考虑是否需要许多慢速服务器（仅为每台服务器分配一点功率）或一些快速服务器（在少量服务器之间分配所有功率）。在[69]中，排队论用于在各种参数设置下最优地回答这个问题。

作为另一个例子，如果带宽是资源，我们可以询问何时将带宽划分为更小的块以及何时最好不要这样做。当性能与价格相结合时，问题也很有趣。例如，购买许多慢速服务器

而不是几台快速服务器通常更便宜。然而，在某些情况下，许多慢速服务器可以消耗比一些快速服务器更多的总功率。所有这些因素都可以进一步影响架构的选择。

设计实例 4——服务器机群中的任务分配

考虑具有中央调度程序和多个主机的服务器机群。每个到达的作业立即被分配到其中一个主机进行处理。图 1-6 说明了这样一个系统。

像这样的服务器机群随处可见。Web 服务器机群通常部署前端调度器，如 Cisco 的本地控制器或 IBM 的网络调度器。超级计算站点可能使用 LoadLeveler 或其他调度器来平衡负载并将作业分配给主机。

目前，我们假设所有主机都是相同的（同类），并且所有作业只使用单个资源。我们还假设一旦将作业分配给主机，它们就会以 FCFS 顺序处理，并且不可抢占。

有许多可能的任务分配策略可用于将作业分配给主机。以下是一些示例：

随机：每个作业翻转一枚公平的硬币，以确定它的路由。

轮转：第 i 个作业转到主机 $i \bmod n$ ，其中 n 是主机数，主机编号为 $0, 1, \dots, n-1$ 。

最短队列：每个作业都进入拥有最少作业的主机。

规模-间隔-任务分配(SITA)：对于某些“短”“中”“长”的定义，“短”作业转到第一个主机，“中”作业转到第二个主机，“长”作业转到第三个主机，等等。

最少剩余工作(LWL)：每个作业进入剩余工作总量最少的主机，其中主机的“工作”是那里作业规模的总和。

中央队列：不是在每个主机上都有一个队列，而是将作业集中在一个中央队列中。当主机完成某个作业时，它会抓取中央队列中的第一个作业进行处理。

问题：哪些任务分配策略产生最低的平均响应时间？

答案：尽管服务器机群无处不在，人们对这个问题的了解却少得惊人。如果作业规模的可变性很低，那么 LWL 策略是最好的。如果作业规模的变化很大，那么保持短作业不会落后于长作业是很重要的，因此类似 SITA 的策略可以更好地与长作业隔离。事实上，长期以来人们认为，当作业规模变化很大时，SITA 总是优于 LWL。然而，最近发现（见 [90]）即使作业规模变化趋于无穷大，SITA 也可能远远低于 LWL。事实证明，工作负载的其他属性（包括负载和作业规模分布的分数矩）也很重要。

问题：对于上一个问题，了解作业规模有多重要？例如，需要知道作业规模的 LWL 与不需该信息的中央队列相比如何？

答案：实际上，大多数任务分配策略都不需要知道作业的规模。例如，可以通过归纳证明 LWL 等同于中央队列。即使像 SITA 这样的策略——其定义是基于对作业规模的了解，也可以通过其他不需要了解作业规模的策略来很好地近似，见 [82]。

问题：现在考虑一种不同的模式，其中作业是可抢占的。具体来说，假设服务器是处理器共享(PS)服务器，它在服务器上的所有作业之间共享时间，而不是以 FCFS 顺序提供服务。现在哪个任务分配策略更合适？答案与 FCFS 服务器的答案相同吗？

答案：对于 PS 服务器，最适合 FCFS 服务器的任务分配策略通常是灾难性的。最短队列策略接近最优（见 [79]），而如果作业规模可变性很高，则该策略对于 FCFS 服务器来

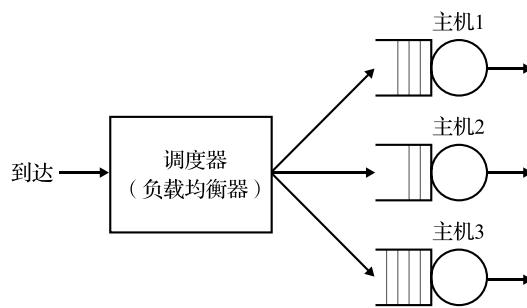


图 1-6 带有中央调度器的分布式服务器系统

说则非常糟糕。

关于任务分配策略有许多未解决的问题。例如，具有 PS 服务器的服务器机群的情况几乎没有受到关注，甚至对 FCFS 服务器的情况仍然只是部分了解。还有许多其他未提及的任务分配策略。例如，周期窃取(利用空闲主机来处理某些其他队列中的作业)可以与许多现有任务分配策略组合以创建改进的策略。还需要考虑其他指标，例如最小化响应时间的方差，而不是平均响应时间或最大化公平性。最后，当工作负载随时间变化时，任务分配可能变得更加复杂，而且更重要。

在我们有机会研究经验工作负载之后，第 24 章将详细分析任务分配。

设计实例 5——调度策略

假设有一台服务器，作业根据泊松过程到达。可以假设有关作业规模分布的任何内容。以下是服务作业的一些可能的服务顺序(调度顺序)。

先来先服务(FCFS)：当服务器完成一项作业时，它开始处理最早到达的作业。

非抢占式后到先服务(LCFS)：当服务器完成一项作业时，它开始处理最后到达的作业。

随机：当服务器完成一项作业时，它开始处理随机作业。

问题：这些非抢占式服务顺序中的哪一个将导致最低的平均响应时间？

答案：信不信由你，它们都有相同的平均响应时间。

问题：假设我们将非抢占式 LCFS 策略更改为抢占式 LCFS 策略(PLCFS)，其工作方式如下：每当新到达作业进入系统时，立即抢占服务中的作业。该策略的平均响应时间与其他策略相比如何？

答案：这取决于作业规模分布的可变性。如果作业规模分布至少是适度变化的，那么 PLCFS 将是一个巨大的进步。如果作业规模分布几乎是不可变的(基本上是不变的)，那么 PLCFS 策略将比原来差 2 倍。

在本书的最后，我们在第 28 章到第 33 章研究了许多违反直觉的调度理论结果。

更多设计实例

计算机系统设计中还有许多问题可以用于排队论解决方案。

一个例子是设置成本的概念。事实证明，打开关闭的服务器可能需要大量的时间和电力。在设计有效的电源管理策略时，我们经常希望关闭服务器(以节省电量)，但是我们必须支付设置成本，以便在作业到达时重新启动它们。考虑到响应时间和功耗的性能目标，一个重要的问题是关闭服务器是否有用。如果是这样，那么可以确切地询问应该留下多少台服务器。这些问题将在第 15 章和第 27 章中讨论。

当作业具有优先级时，还存在涉及优化调度的问题(例如，某些用户为其作业支付更多以使其优先于其他用户的作业，或者某些作业本质上比其他作业更重要)。同样，排队论在设计正确的优先级方案以最大化已完成工作的价值方面非常有用。

然而，排队论(以及更普通的分析建模)目前还不是万能的！有许多非常简单的问题，我们最多只能近似地分析。例如，考虑图 1-7 所示的简单双服务器网络，其中作业规模来自一般分布。没有人知道如何获得该网络的平均响应时间。存在近似值，但它们非常差，特别是当作业规模变化很大时^[76]。我们在本书中提到了许多这样的开放性问题，并鼓励读者尝试解决这些问题！

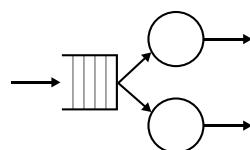


图 1-7 难题示例：M/G/2
队列由单个队列和
两个服务器组成。
当服务器完成作业
时，它将开始处理
队列头部的作业。
作业规模遵循一般
分布 G

第 2 章 |

Performance Modeling and Design of Computer Systems: Queueing Theory in Action

排队论术语

2.1 我们将去向何方

排队论是对网络和系统中排队行为的研究。图 2-1 显示了解决方案流程。

在第 1 章中，我们研究了排队论作为设计工具所具备的能力。在本章中，我们从头开始并定义排队论中使用的术语。

2.2 单服务器网络

排队网络由服务器组成。

排队网络的最简单例子是单服务器网络，如图 2-2 所示。本节中的讨论仅限于具有先来先服务(FCFS)服务顺序的单服务器网络。可以将服务器视为 CPU。

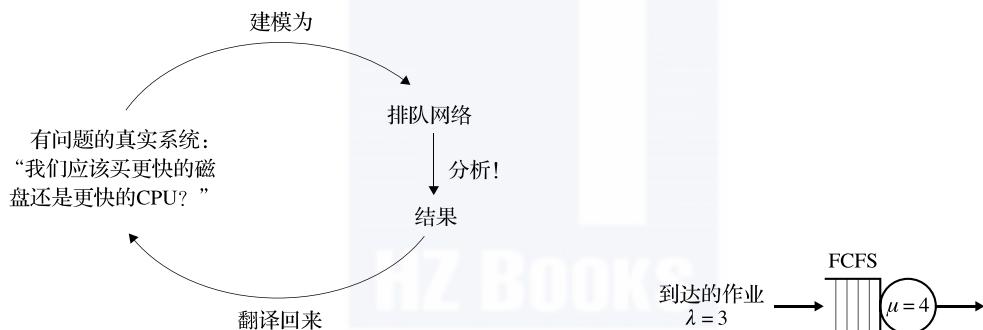


图 2-1 解决方案流程

图 2-2 单服务器网络

有一些与单服务器网络相关的参数：

服务顺序。这是服务器将服务作业的顺序。除非另有说明，否则假设为先来先服务。

平均到达率。这是作业到达服务器的平均速率 λ (例如， $\lambda=3$ 个作业/秒)。

平均到达时间。这是连续作业到达之间的平均时间(例如， $1/\lambda=1/3$ 秒)。

服务要求、规模。作业的规模通常用随机变量 S 表示。如果周围没有其他作业(没有排队)，则是在此服务器上运行作业所需的时间。在排队模型中，规模(也称为服务要求)通常与服务器相关联(例如，此作业将在服务器上花费 5 秒)。

平均服务时间。这是 S 的预期值，即在此 CPU 上服务作业所需的平均时间，其中“服务”不包括排队时间。在图 2-2 中， $E[S]=1/4$ 秒。

平均服务率。这是服务作业的平均速率 μ (例如， $\mu=4$ 个作业/秒= $1/(E[S])$)。

请注意，这种说话方式与我们通常在谈话中谈论服务器的方式不同。例如，我们没有提到 CPU 的绝对速度；相反，我们只根据它正在处理的作业集来定义 CPU 的速度。

在正常的对话中，我们可能会说：

- 作业的平均到达率是每秒 3 个作业。
- 作业具有不同的服务要求，但作业所需的平均周期数为每个作业 5000 个周期。

- CPU 速度为每秒 20 000 个周期。

也就是说，每秒平均有 15 000 个周期的作业到达 CPU，CPU 每秒可以处理 20 000 个周期的作业。

按照排队论的说法，我们永远不会提到“周期”这个词。相反，我们会简单地说：

- 作业的平均到达率是每秒 3 个作业。
- CPU 可以为作业提供服务的平均速率为每秒 4 个作业。

第二种说法略去了一些细节，从而使问题更容易思考。在两者之间来回切换应该感到舒服一些。

我们在单服务器系统的上下文中考虑这些常见的性能指标：

响应时间、周转时间、系统时间或逗留时间(T)。我们通过 $T = t_{\text{离开}} - t_{\text{到达}}$ 定义作业的响应时间，其中 $t_{\text{离开}}$ 是作业离开系统的时间， $t_{\text{到达}}$ 是作业到达系统的时间。我们感兴趣的是平均响应时间 $E[T]$ 、响应时间的方差 $\text{Var}(T)$ 以及 T 的尾部行为 $P\{T > t\}$ 。

等待时间或延迟(T_Q)。这是作业在队列中花费的时间，而不是服务的时间。这也被称为“队列中的时间”或“浪费的时间”。注意 $E[T] = E[T_Q] + E[S]$ 。在 FCFS 服务顺序下，等待时间可以定义为从作业到达系统到第一次接收服务的时间。

系统中的作业数量(N)。这包括队列中的那些作业以及正在服务的作业(如果有)。

队列中的作业数(N_Q)。这仅表示等待(队列中)的作业数。

我们可以对单服务器网络进行一些即时观察。首先，观察到当平均到达率 λ 增加时，前面提到的所有性能指标都会增加(变得更糟)。此外，当平均服务率 μ 增加时，前面提到的所有性能指标都会降低(改善)。

我们需要令 $\lambda \leq \mu$ (我们总是假设 $\lambda < \mu$)。

问题：如果 $\lambda > \mu$ 会发生什么？

答案：如果 $\lambda > \mu$ ，队列长度随时间变为无穷大。

问题：你对此的直觉如何？

答案：考虑很长一段时间 t 。那么，如果 $N(t)$ 是时间 t 时系统中的作业数，并且 $A(t)$ (分别为 $D(t)$) 表示时间 t 之前到达(离开)的数量，那么我们有：

$$E[N(t)] = E[A(t)] - E[D(t)] \geq \lambda t - \mu t = t(\lambda - \mu)$$

(不相等来自这样一个事实，即时间 t 之前的预期离开次数实际上小于 μt ，因为服务器并不总是很忙。)现在观察到如果 $\lambda > \mu$ ，那么有当 $t \rightarrow \infty$ 时， $t(\lambda - \mu) \rightarrow \infty$ 。

在整本书中，我们假设 $\lambda < \mu$ ，这是稳定性所需的(避免队列大小不受限制地增长)。在第 9 章中会涉及 $\lambda \geq \mu$ 的情况。

问题：给定先前的稳定性条件($\lambda < \mu$)，假设到达间隔分布和服务时间分布是确定性的(即两者都是常数)。 T_Q 和 T 的值是什么？

答案： $T_Q = 0$ 以及 $T = S$ 。

因此，排队(等待)是由服务时间或到达间隔时间分布的可变性引起的。以下是可变性如何导致队列的例子。让我们将时间离散化。假设在每个时间步，到达的概率是 $p = 1/6$ 。假设在每个时间步，离开的概率是 $q = 1/3$ 。那么，如果在没有离开的情况下发生多次到达，则队列将(暂时)以非零概率建立。

2.3 排队网络的分类

排队网络可以分为两类：开放网络和封闭网络。随机过程书籍(例如，[149-150])通

常将讨论局限于开放网络。相比之下，系统性能分析书籍(例如，[117, 125])几乎只讨论封闭网络。开放网络在 2.4 节中介绍。封闭网络在 2.6 节中介绍。

2.4 开放网络

开放排队网络具有外部到达和离开。本节说明了开放网络的四个示例。

示例：单服务器系统

如图 2-2 所示。

示例：具有概率路由的队列网络

如图 2-3 所示。这里服务器 i 接收具有速率为 r_i 的外部到达。服务器 i 还从其他一些服务器接收内部到达。接下来，在服务器 i 处完成服务的包以概率 p_{ij} 被路由到服务器 j 。我们甚至可以允许概率依赖于数据包的“类”，因此并非所有数据包都必须遵循相同的路由方案。

应用：例如，在模拟因特网中的包流时，可以使包的类(及其路由)依赖于其源和目标 IP 地址。在建模延迟中，每条线路可能会被用于模拟线路延迟的服务器所取代。目标可能是在给定其他数据包存在的情况下，预测特定路由上的数据包的平均往返时间。我们将在第 18 章解决这个问题。

示例：具有非概率路由的队列网络

如图 2-4 所示。这里所有作业都遵循预定的路由：CPU 到磁盘 1、到磁盘 2、到磁盘 1、到磁盘 2、到磁盘 1、到出去。

示例：有限缓冲区

图 2-5 显示了具有有限缓冲区的单服务器网络示例。任何没有空间的到达都会被放弃。

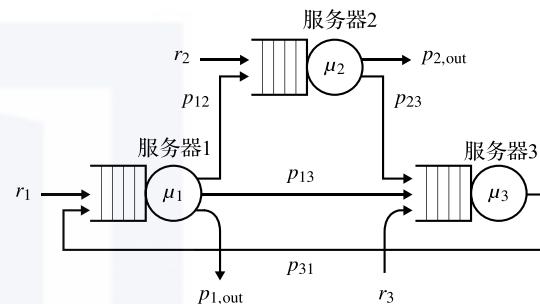


图 2-3 具有概率路由的队列网络

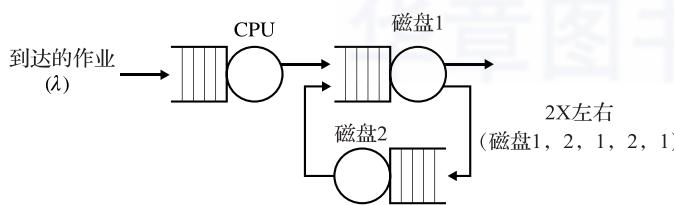


图 2-4 具有非概率路由的队列网络

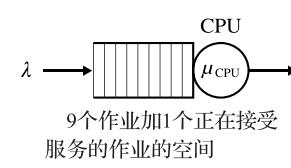


图 2-5 具有有限缓冲容量的单服务器网络

2.5 更多指标：吞吐量和利用率

我们已经看到了四个性能指标： $E[N]$ 、 $E[T]$ 、 $E[N_Q]$ 和 $E[T_Q]$ 。虽然这些应用于单服务器系统，但它们也可用于描述多服务器、多队列系统中的性能。例如， $E[T]$ 表示作业在整个系统中花费的平均时间，包括在各种队列中花费的所有时间以及在各种服务器上接收服务所花费的时间，而 $E[T_Q]$ 指的是在不同队列中等待作业所“浪费”的平均时间。如果想在这样的系统中仅引用第 i 个队列，通常用 $E[N_i]$ 来表示在服务器 i 排队和服务的预期作业数，并且用 $E[T_i]$ 表示一个作业在服务器 i 上排队和服务的预期时间。

现在我们引入两个新的性能指标：吞吐量和利用率。吞吐量可以说是最常用的性能指标，每个人都想要更高的吞吐量。让我们看看这是为什么。

问题：最大化吞吐量与最小化响应时间有何关系？例如，在图 2-6 中，哪个系统具有更高的吞吐量？

答案：我们很快就会看到。

让我们从定义利用率开始。

设备利用率(ρ_i)是设备 i 处于忙碌状态的时间占比。请注意，当前的利用率定义仅适用于单个设备(服务器)。当隐含设备时，我们只需写成 ρ (省略下标)。

假设我们长时间观看设备 i 。设 τ 表示观察期的长度。设 B 表示在观察期间设备非空闲(忙)的总时间。那么

$$\rho_i = \frac{B}{\tau}$$

设备吞吐量(X_i)是设备 i 的完成率(例如，作业/秒)。系统的吞吐量(X)是系统中的作业完成率。

设 C 表示在时间 τ 时在设备 i 处完成的作业总数。那么

$$X_i = \frac{C}{\tau}$$

那么 X_i 如何与 ρ_i 相关？

$$\frac{C}{\tau} = \left(\frac{C}{B}\right) \cdot \frac{B}{\tau}$$

问题：那么 C/B 是多少？

答案： $B/C = E[S]$ ，所以 $C/B = 1/E[S] = \mu_i$ 。所以有

$$X_i = \mu_i \cdot \rho_i$$

以下是另一种通过条件作用得到此表达式的方法：

X_i = 服务器 i 的平均完成速率

$$\begin{aligned} &= E[\text{服务器 } i \text{ 的完成速率} | \text{服务器 } i \text{ 处于忙碌状态}] \cdot P\{\text{服务器 } i \text{ 处于忙碌状态}\} \\ &\quad + E[\text{服务器 } i \text{ 的完成速率} | \text{服务器 } i \text{ 是处于空闲状态}] \cdot P\{\text{服务器 } i \text{ 处于空闲状态}\} \\ &= \mu_i \cdot P\{\text{服务器 } i \text{ 处于忙碌状态}\} + 0 = \mu_i \cdot \rho_i \end{aligned}$$

或者，等效地

$$\rho_i = X_i \cdot E[S]$$

后一种表述有一个名称：利用率定律。

示例：单服务器网络——吞吐量是多少？

在图 2-7 中，我们有一个单服务器系统。

问题： X 是多少？

答案： $X = \rho \cdot \mu$ 。但是 ρ 是多少？在第 6 章中，我们将证明 $\rho = \lambda / \mu$ 。现在，这里有一个潦草但直观的理解方式，但并不是证明！

$$\rho = \text{服务器处于忙碌状态的时间占比} = \frac{\text{一个作业所需的平均服务时间}}{\text{到达之间的平均时间}} = \frac{1/\mu}{1/\lambda} = \frac{\lambda}{\mu}$$

所以，我们得到

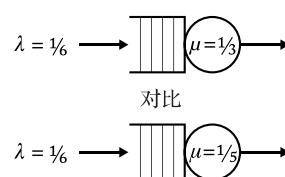


图 2-6 比较两个系统的吞吐量

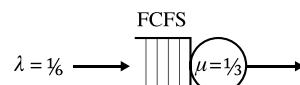


图 2-7 单服务器模型

$$X = \rho \cdot \mu = \frac{\lambda}{\mu} \cdot \mu = \lambda$$

因此吞吐量不依赖于服务率！

特别是，图 2-6 所示的例子在图 2-8 中再次重复给出，两个系统的吞吐量相同，为 $1/6$ 个作业/秒。在处理器速度较快的情况下，响应时间会下降，队列长度会下降，但 X 不会发生变化。因此，较低的响应时间与较高的吞吐量无关。

问题：解释为什么 X 不会改变。

答案：无论我们将 μ 提高多少，完成率仍然受到到达率的限制：到达率 = 离开率。改变 μ 影响最大可能的 X ，但不影响实际的 X 。注意，因为我们假设是稳定的系统，那么，对于较大的 t ， t 期间的到达次数大约是 t 期间的完成次数。

示例：队列的概率网络——吞吐量是多少？

对于图 2-3， r_i 表示进入服务器 i 的平均外部到达率， μ_i 表示服务器 i 处的平均服务率。

问题：图 2-3 中的系统吞吐量 X 是多少？

答案： $X = \sum_i r_i$ 。

问题：服务器 i 的吞吐量 X_i 是多少？

答案：设 λ_i 表示进入服务器 i 的总到达率。那么 $X_i = \lambda_i$ 。但要获得 λ_i ，我们需要求解这些联立方程：

$$\lambda_i = r_i + \sum_j \lambda_j P_{ji} \quad (2.1)$$

问题： r_i 如何在这些方程中受到约束？

答案：为了让网络达到“均衡”（服务器的流入 = 服务器的流出），必须有 $\lambda_i < \mu_i$ ， $\forall i$ ，这就约束了 r_i （参见习题 2.1）。

示例：具有非概率路由的队列网络——吞吐量是多少？

问题：图 2-4 中的 X 是多少？

答案： $X = \lambda$ 。

问题： $X_{\text{磁盘1}}$ 和 $X_{\text{磁盘2}}$ 是多少？

答案： $X_{\text{磁盘1}} = 3\lambda$ ， $X_{\text{磁盘2}} = 2\lambda$ 。

示例：有限缓冲区——吞吐量是多少？

对于图 2-5，外部到达率为 λ ，服务率为 μ 。

问题： X 是多少？

答案： $X = \rho\mu$ 。但我们需要随机分析来确定 ρ ，因为它不再仅仅是 λ/μ 。观察到 $X < \lambda$ ，因为一些到达的作业被丢弃了。

2.6 封闭网络

封闭排队网络没有外部到达或离开。它们可以分为两类，如图 2-9 所示。

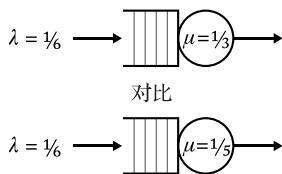


图 2-8 相同模型，但 μ 值不同。吞吐量 X 在两者中都是相同的

封闭网络	
交互式 (终端驱动)	批处理系统

图 2-9 封闭网络的类别

2.6.1 交互式(终端驱动)系统

交互式(终端驱动)系统的一个例子如图 2-10 所示。终端代表每个将作业发送到“中央子系统”然后等待响应的用户。中央子系统是队列网络。用户在上一份作业返回之前无法提交下一份作业。因此，系统中的作业数量是固定的(等于终端数量)。这个数字有时被称为负载或 MPL(多道程序设计水平)，不要与设备利用率混淆。

存在一个思考时间 Z ，它是一个随机变量，表示在接收一个作业的结果和发出下一个作业之间每个终端的时间。请注意，中央子系统中的作业数最多为终端数，因为某些用户可能处于“思考”状态。

交互式系统的一个例子如图 2-10 所示，是一个数据输入应用程序。 N 个用户在终端上填写屏幕上的条目。这里必须填写屏幕上的几个字段，然后将所有字段提交给中央子系统以进行适当的处理和数据库更新。在执行上一次更新之前，无法填写新字段。“思考时间” Z 是将数据键入屏幕的时间。

单个用户(终端)在思考状态和提交状态之间摆动，如图 2-11 所示。

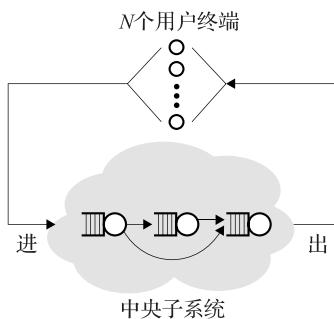


图 2-10 交互式系统

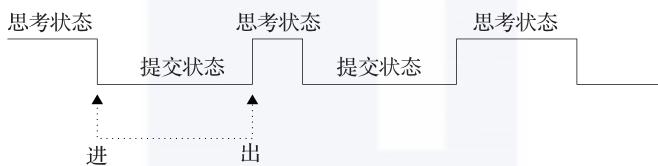


图 2-11 用户在思考和等待提交的作业返回之间交替

问题：如何定义交互式系统的响应时间？

答案：响应时间是图 2-10 和图 2-11 中“进”和“出”之间的作业时间。我们用 $E[R]$ 表示从“进”到“出”的平均时间，以区别于 $E[T]$ ， $E[T]$ 定义为

$$E[T] = E[R] + E[Z]$$

重点：虽然开放系统中的“响应时间”由随机变量 T 表示，对于闭合交互式系统，我们将 T 称为系统时间(或“系统中的时间”)并保留随机变量 R 作为响应时间。

目标：交互式系统的目地是找到一种方法，允许尽可能多的用户立即进入系统，这样他们都可以完成工作，同时保持 $E[R]$ 足够低。请注意，交互式系统与开放系统有很大不同，因为 N 的微小变化会对系统行为产生深远影响。

系统设计师提出的典型问题是：

- 考虑到原始系统，在保持 $E[R]$ 低于某个阈值的同时，能获得多高的 N ？也就是说， $E[R]$ 如何与 N 一起上升？
- 假设一个固定的多道程序设计水平 N 。鉴于我们可以对中央子系统进行更改(即使某些设备更快)，哪些更改将最大限度地提高 $E[R]$ ？

问题：假设我们正在对网站的性能进行建模。你是将网站建模为封闭的交互式系统还是开放系统？

答案：没有定论。有两种类型的研究论文。一方面，一旦用户点击链接(提交作业)，

通常在点击另一个链接之前等待接收结果。因此，用户表现得就像网站是封闭系统一样。另一方面，网站可能拥有大量用户，每个用户在使用网站时耗时都非常短暂。在这方面，网站可能更像开放系统。

Schroeder 等^[165]提出了“部分开放”系统的观点。在这里，用户是从外部到达的，就像在开放系统中一样，但是到达时会向系统发出 k 个请求，而每个请求只能在前一个请求完成时才发出(正如在封闭系统中)。

2.6.2 批处理系统

批处理系统的例子如图 2-12 所示。批处理系统看起来像一个思考时间为零的交互式系统。但是，批处理系统的目标有些不同。在批处理系统中，通常一个系统在一夜之间要运行许多作业。一个作业完成后，另一个作业就会启动。因此，中央子系统中总有 N 个作业。MPL 通常是预先确定的并且是固定的。例如，MPL 可能是适合内存的作业数。

目标：对于一个批处理系统，其目标是获得高吞吐量，以便尽可能多的作业在一夜之间完成。

系统设计人员提出的典型问题是：“我们如何改进中央子系统以最大化吞吐量？”

请注意，我们通常受某些固定最大 MPL 的限制(因为只有这么多作业适合内存，或其他原因)。因此，我们增加吞吐量的唯一方法是通过更改路由或通过加速某些设备来更改中央子系统。请注意，在批处理系统中，我们不关心响应时间，因为作业在一夜之间运行。

问题： X 在封闭系统中的含义是什么？

答案： X 是每秒“出”的作业数。请注意批处理系统的“进” = “出”。

2.6.3 封闭系统中的吞吐量

我们来看一些例子。

示例：单个服务器

图 2-13 显示了由单个服务器组成的封闭网络。

问题：图 2-13 中的吞吐量 X 是多少？

答案： $X = \mu$ 。

注意这与开放网络的情况非常不同，其中吞吐量与服务率无关！

问题：图 2-13 中的平均响应时间 $E[R]$ 是多少？

答案：对于封闭批处理系统， $E[R] = E[T]$ ，即系统

中的响应时间和时间是相同的。对于图 2-13， $E[T] = N/\mu$ ，因为每个“到达”等待 $N-1$ 个作业然后运行。

请注意， X 和 $E[R]$ 是反向相关的！

示例：串联服务器

现在考虑一个更复杂的封闭网络的例子，如图 2-14 所示。

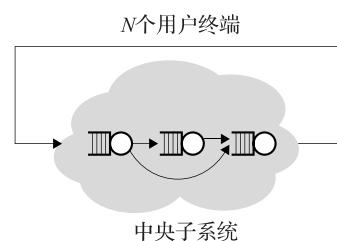


图 2-12 批处理系统

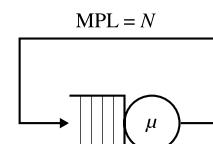


图 2-13 单服务器封闭网络

问题：吞吐量是多少？

答案：我们想说 $X = \min(\mu_1, \mu_2) \dots$

问题：为什么之前的答案不一定正确？

答案：如果我们知道较慢的服务器总是很忙，那么之前的答案是正确的，但情况不一定如此。想象一下 $N=1$ 。那么，速度较慢的服务器总是很忙。

问题：好的，但是当 $N=2$ 时会发生什么。现在看来，慢速服务器上始终至少有一个作业，不是吗？

答案：不，较慢的服务器仍然不忙。我们在这里忽略了一个事实，有时慢速服务器比快速服务器更快，因为这些服务率只是均值！那么我们实际上是否需要考虑作业规模分布才能得到确切答案？作业规模分布是否真的会影响答案？

我们很快就会回答这些问题……现在，让我们总结一下开放网络和封闭网络的行为之间的差异，以及为什么我们需要同时考虑两者。

2.7 封闭网络和开放网络之间的差异

开放系统

- 吞吐量 X 与 μ_i 无关。
- X 不受 μ_i 加倍的影响。
- 吞吐量和响应时间无关。

封闭系统

- X 取决于 μ_i 。
- 如果我们在保持 N 常数的同时将所有 μ_i 加倍，则 X 会改变。
- 实际上我们将在第 6 章中看到，对于封闭式系统，
更高的吞吐量 \Leftrightarrow 更短的平均响应时间

关于建模的问题

这是最后一个问题。几年前，我接到了 IBM 一些人的电话。他们试图将刀片服务器建模为单服务器队列。他们知道服务器的到达率 λ ，以作业/秒为单位。然而，他们想知道如何获得平均作业规模 $E[S]$ 。

问题：如何在单一服务器系统的实践中获得 $E[S]$ ？

答案：乍一看，你可能会因为 $E[S]$ 是隔离作业所需的平均时间而推理，你应该只将一个作业发送到系统中并测量其响应时间，重复该实验一百次得到一个均值。这在理论上是有道理的，但在实践中不能很好地工作，因为与系统加载一段时间的情况相比，缓存条件和其他因素对于单个作业的场景非常不同。

更好的方法是考虑 $E[S] = 1/\mu$ ，因此只需考虑服务器的服务率就足够了。为了获得 μ ，假设有一个开放系统，我们可以使 λ 越来越高，这将提高完成率，直到完成率达到某个值的水平，这将是速率 μ 。

更好的想法是将我们的服务器置于一个封闭系统中，没有思考时间。这样就可以保证服务器始终处于工作状态。现在，如果测量服务器上的完成率（每秒完成的作业），那么这将为我们提供服务器的 μ 。 $E[S]$ 则是 μ 的倒数。

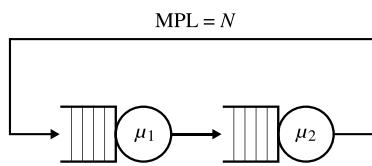


图 2-14 串联服务器封闭网络

2.8 阅读材料

特别有助于理解封闭排队网络的是 Lazowska(第 58~59 页)^[117] 和 Menascé(第 84~87 页)^[125]。这些都是精彩的书籍。

令人惊讶的是，文献中很少有人讨论封闭系统与开放系统的比较。例如，考虑一个具有负载 ρ 的封闭交互式单服务器系统，而不是具有负载 ρ 的相应开放系统。它们的平均响应时间如何相互比较？服务时间的变化如何影响封闭系统与开放系统？在[186]和[24]以及习题 7.2、习题 7.5、习题 13.7 和习题 13.8 中考虑了这些问题和许多其他问题。另一个问题是服务器上的调度策略（服务顺序）如何影响封闭系统与开放系统。这个问题直到 2006 年才真正讨论过^[165]。有关开放与封闭主题的最新讨论，我们推荐 Y. C. Tay 的书^[173]。

在本章中，我们多次提到确保到达速率小于服务率 ($\lambda < \mu$) 对于稳定性是必要的。这种情况也足以确保我们在本书中考虑的网络的稳定性。但是，对于更复杂的排队网络，这通常不是满足稳定性的充分条件。为了理解原因，我们推荐 Maury Bramson 的论文（参见[29]）。

2.9 习题

- 2.1 最高外部到达率。对于图 2-3 中给出的具有概率路由的队列网络，假设每个服务器的平均速率为 10 个作业/秒；也就是说， $\mu_i = 10$, $\forall i$ 。假设 $r_2 = r_3 = 1$, $p_{12} = p_{2,出} = 0.8$, $p_{23} = p_{13} = 0.2$, $p_{1,出} = 0$, 并且 $p_{31} = 1$ 。为保持系统稳定， r_1 的最大允许值是多少？
- 2.2 降速。(a) 作业到达服务器，以 FCFS 顺序为它们提供服务：



平均到达率是 $\lambda = 1/2$ 个作业/秒。作业规模（服务时间）根据随机变量 S 独立同分布，其中

$$S = \begin{cases} 1 & \text{概率为 } 3/4 \\ 2 & \text{否则} \end{cases}$$

测量了平均响应时间： $E[T] = 29/12$ 。

根据这些信息，计算平均降速 $E[\text{降速}]$ ，其中作业 j 的降速定义为 $\text{降速}(j) = T(j)/S(j)$ ，其中 $T(j)$ 是作业 j 的响应时间， $S(j)$ 是作业 j 的规模。

- (b) 如果(a)部分的服务顺序是最短作业优先(SJF)，那么同样的技术是否适用于计算平均降速？
- 2.3 调度顺序。(a) 对于单个服务器 CPU，其中作业根据某个进程到达，让 SRPT 表示抢占式调度策略，该策略始终以当前最短剩余处理时间为作业服务（假设其知道此信息）。对于由到达时间和每个作业的规模组成的任何到达序列，SRPT 调度将到达序列上的平均响应时间最小化。证明或反驳这一说法。
- (b) 作业降速的定义是作业的响应时间除以服务要求。(i) 许多人认为平均降速比平均响应时间更重要。你觉得这是为什么？(ii) 根据直觉，SRPT 调度策略应尽量减少平均降速。证明或反驳这个假设。

| 第二部分 |

Performance Modeling and Design of Computer Systems: Queueing Theory in Action

必要的概率背景知识

概率是分析建模的重要部分。第二部分提供了本书所需的所有概率知识。第 3 章简要介绍了本科概率知识。第 4 章回顾了两种生成随机变量的方法，这对于模拟队列非常重要。最后，第 5 章讨论了更高级的主题，例如样本路径、随机变量序列的收敛以及不同类型的均值(例如时间均值和整体均值)。这些概念很重要，在本书中都有提及。然而，这些概念也很难，读者可能希望在第一次阅读时略过第 5 章，然后在第 8 章和第 9 章学习完马尔可夫链之后返回该章以进行更深入的阅读。

第3章 |

Performance Modeling and Design of Computer Systems: Queueing Theory in Action

概率知识复习

在本书中，我们假设读者了解本科概率的知识，包括离散和连续随机变量。Sheldon Ross 的书 *Introduction to Probability Models*^[150] 的前 3 章(约 180 页)提供了对这些主题的相关讨论，我们鼓励读者阅读。在本章中，通过一些简单的说明性示例简要回顾了本书中我们需要掌握的具体概率概念。我们首先讨论事件的概率，然后转向随机变量。通过这些示例，加上本章末尾的习题，应该足以作为本书中本科概率知识的回顾。

3.1 样本空间和事件

概率通常根据某些实验来定义。实验的样本空间 Ω 是实验的所有可能结果的集合。

定义 3.1 事件 E 是样本空间 Ω 的任何子集。

例如，在投掷两个骰子的实验中，每个结果(也称为样本点)由 (i, j) 表示，其中 i 是第一个投掷，而 j 是第二个投掷。有 36 个样本点。我们可以考虑这个事件

$$E = \{(1,3) \text{ or } (2,2) \text{ or } (3,1)\}$$

骰子掷出的总和是 4。

通常，样本空间可以是离散的，这意味着结果的数量是有限的，或者至少可数无限的，或连续的，这意味着结果的数量是不可数的。

人们可以谈论事件的并集和交集，因为它们也是集合(例如， $E \cup F$ 、 $E \cap F$ 和 E^c ，其中 E 和 F 是事件，而 E^c (E 的补)，表示在 Ω 中但不在 E 中的点集)。

问题：对于骰子投掷实验，请考虑图 3-1 中 Ω 上定义的事件 E_1 和 E_2 。你认为 E_1 和 E_2 是独立的吗？

答案：不，它们不是独立的。我们稍后在定义独立性时会再讨论这一点。我们说 E_1 和 E_2 是相互排斥的。

定义 3.2 如果 $E_1 \cap E_2 = \emptyset$ ，则 E_1 和 E_2 相互排斥。

定义 3.3 如果 E_1, E_2, \dots, E_n 是事件，使得 $E_i \cap E_j = \emptyset, \forall i, j$ ，以及 $\bigcup_{i=1}^n E_i = F$ ，则我们说事件 E_1, E_2, \dots, E_n 划分集合 F 。

		E_1	E_2		
		(1,1)	(1,2)	(1,3)	(1,4)
		(2,1)	(2,2)	(2,3)	(2,4)
		(3,1)	(3,2)	(3,3)	(3,4)
		(4,1)	(4,2)	(4,3)	(4,4)
		(5,1)	(5,2)	(5,3)	(5,4)
		(6,1)	(6,2)	(6,3)	(6,4)
				(1,5)	(1,6)
				(2,5)	(2,6)
				(3,5)	(3,6)
				(4,5)	(4,6)
				(5,5)	(5,6)
				(6,5)	(6,6)

图 3-1 示例空间 Ω 中两个事件的插图

3.2 事件定义的概率

概率是根据事件定义的。

$P\{E\}$ = 事件 E 正在发生的概率

我们可以将每个样本点视为具有一定的发生概率，并且事件 E 发生的概率是 E 中样本点的概率的总和。例如，在两个骰子示例中，每个样本点(有序的一对数字)的概率为 $1/36$ 。

重要的是， Ω (其中 Ω 是样本空间)的概率定义为 1。

定义 3.4 两个事件的并的概率定义如下：

$$P\{E \cup F\} = P\{E\} + P\{F\} - P\{E \cap F\}$$

如果我们将事件视为如图 3-2 所示的集合，这应该是有意义的。观察到 $P\{E \cap F\}$ 项的减法是必要的，这样这些样本点不会被计算两次。

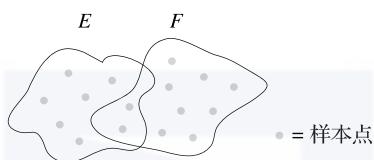


图 3-2 维恩图

定理 3.5 $P\{E \cup F\} \leq P\{E\} + P\{F\}$ 。

证明：这从定义 3.4 可得。 ■

问题：定理 3.5 什么时候是相等的？

答案：当 E 和 F 相互排斥时。

问题：假设你的实验包括投掷飞镖，它落在区间 $[0, 1]$ 内任何地方的概率都是相等的。飞镖落在 0.3 处的概率是多少？

答案：精确地降落在 0.3 处的概率定义为 0。为此，假设概率严格大于 0，例如 $\epsilon > 0$ 。那么，降落在 0.5 处的概率也将是 ϵ ，降落在任何理性点上的概率也将是 ϵ 。但是这些不同的降落结果是相互排斥的事件，因此它们的概率会加起来。因此，降落在间隔 $[0, 1]$ 中的概率将大于 1，这是不允许的，因为 $P\{\Omega\} = 1$ 。尽管精确地降落在 0.3 的概率被定义为 0，但是降落在间隔 $[0, 0.3]$ 中的概率被定义为 0.3。

3.3 事件的条件概率

定义 3.6 给定事件 F ，事件 E 的条件概率写为 $P\{E|F\}$ 并由下式给出，我们假设 $P\{F\} > 0$ ：

$$P\{E|F\} = \frac{P\{E \cap F\}}{P\{F\}} \quad (3.1)$$

$P\{E|F\}$ 应该被认为是假设我们已经将样本空间缩小到 F 中的点时，事件 E 发生的概率。为了看到这一点，请考虑图 3-3，其中 $P\{E\} = 8/42$ 和 $P\{F\} = 10/42$ 。

如果我们想象将空间缩小到 F 中的 10 个点，那么假设我们在集合 F 中，则在集合 E 中的概率应该是 $2/10$ 。即

$$P\{E|F\} = \frac{2}{10} = \frac{\frac{2}{42}}{\frac{10}{42}} = \frac{P\{E \cap F\}}{P\{F\}}$$

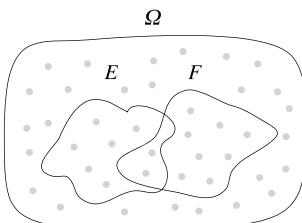


图 3-3 有 42 个样本点的样本空间

示例：表 3-1 显示了我每天的三明治选择。我们将“本周的上半周”定义为周一至周三(含周三)，将“本周的下半周”定义为周四至周日(含周日)。

表 3-1 我的三明治选择

周一	周二	周三	周四	周五	周六	周日
果冻	奶酪	火鸡	奶酪	火鸡	奶酪	无

问题： $P\{\text{奶酪} | \text{下半周}\}$ 是什么？

答案：这是问我下半周吃奶酪三明治的天数。答案显然是 4 天中的 2 天，也就是 $2/4$ 。或者，我们可以使用式(3.1)表示如下：

$$P\{\text{奶酪} | \text{下半周}\} = \frac{P\{\text{奶酪} \& \text{ 下半周}\}}{P\{\text{下半周}\}} = \frac{\frac{2}{7}}{\frac{4}{7}} = \frac{2}{4}$$

3.4 独立事件和有条件独立事件

定义 3.7 事件 E 和 F 是独立的，如果

$$P\{E \cap F\} = P\{E\} \cdot P\{F\}$$

问题：如果 E 和 F 是独立的，那么 $P\{E|F\}$ 是多少？

答案：假设 $P\{F\} > 0$ ，我们有

$$P\{E|F\} = \frac{P\{E \cap F\}}{P\{F\}} \stackrel{\text{独立}}{=} \frac{P\{E\} \cdot P\{F\}}{P\{F\}} = P\{E\}$$

也就是说， $P\{E\}$ 不受 F 是否为真的影响。

问题：两个互斥(非空)事件是否可以独立？

答案：不。在这种情况下， $P\{E|F\} = 0 \neq P\{E\}$ 。

问题：假设一个人投掷两个骰子。这些事件中的哪一对是独立的？

1) E_1 = “第一次投掷为 6”， E_2 = “第二次投掷为 6”

2) E_1 = “投掷的总和为 7”， E_2 = “第二次投掷为 4”

答案：它们都是独立的！

问题：假设我们已定义： E_1 = “投掷总和为 8”， E_2 = “第二次投掷为 4”。它们现在

独立吗？

答案：不。

在问题中经常出现的另一种独立概念(例如，参见习题 3.20)是有条件的独立性。

定义 3.8 给定事件 G ，两个事件 E 和 F 被认为是条件独立的，其中 $P\{G\} > 0$ ，如果

$$P\{E \cap F | G\} = P\{E | G\} \cdot P\{F | G\}$$

独立并不意味着有条件的独立，反之亦然。

3.5 总概率定律

观察到集合 E 可以表示为

$$E = (E \cap F) \cup (E \cap F^c)$$

也就是说， E 是集合 $E \cap F$ 和集合 $E \cap F^c$ 的并集，因为 E 中的任一点要么在 F 中，要么不在 F 中。

现在观察 $E \cap F$ 和 $E \cap F^c$ 是互斥的。因此

$$P\{E\} = P\{E \cap F\} + P\{E \cap F^c\} = P\{E | F\}P\{F\} + P\{E | F^c\}P\{F^c\}$$

其中 $P\{F^c\} = 1 - P\{F\}$ 。

定理 3.9 是一个泛化。

定理 3.9(总概率定律) 设 F_1, F_2, \dots, F_n 划分状态空间 Ω 。那么

$$P\{E\} = \sum_{i=1}^n P\{E \cap F_i\} = \sum_{i=1}^n P\{E | F_i\} \cdot P\{F_i\}$$

证明

$$E = \bigcup_{i=1}^n (E \cap F_i)$$

现在，因为事件 $E \cap F_i, i=1, \dots, n$ 是互斥的，我们有

$$P\{E\} = \sum_{i=1}^n P\{E \cap F_i\} = \sum_{i=1}^n P\{E | F_i\} \cdot P\{F_i\}$$

问题：假设我们对某种类型的事务失败的概率感兴趣。我们知道如果存在缓存故障，那么事务将有 $5/6$ 的失败概率。我们也知道如果网络出现故障，那么事务将有 $1/4$ 的失败概率。假设网络出现故障的概率是 $1/100$ ，并且缓存出现故障的概率是 $1/100$ 。这足以告诉我们事务失败的概率吗？

答案：很容易写出(但是错误的)

$$\begin{aligned} P\{\text{事务失败}\} &= P\{\text{事务失败} | \text{缓存故障}\} \cdot \frac{1}{100} + P\{\text{事务失败} | \text{网络故障}\} \cdot \frac{1}{100} \\ &= \frac{5}{6} \cdot \frac{1}{100} + \frac{1}{4} \cdot \frac{1}{100} \end{aligned}$$

问题：这个解有什么问题？

答案：我们所关注的两个事件——网络故障和缓存故障——不一定划分空间。这些事件的概率总和显然小于 1。此外，两个故障都可能存在非零概率。

人们需要非常小心，事件对整个样本空间既互斥又求和。

3.6 贝叶斯定律

有时，我们需要知道 $P\{F|E\}$ ，但是我们知道的是相反的方向： $P\{E|F\}$ 。是否有可能从 $P\{E|F\}$ 获得 $P\{F|E\}$ ？结果是，假设我们也知道 $P\{E\}$ 和 $P\{F\}$ 。

定理 3.10(贝叶斯定律)

$$P\{F|E\} = \frac{P\{E|F\} \cdot P\{F\}}{P\{E\}}$$

证明：

$$P\{F|E\} = \frac{P\{E \cap F\}}{P\{E\}} = \frac{P\{E|F\} \cdot P\{F\}}{P\{E\}}$$

总概率定律可以与贝叶斯定律结合如下：设 F_1, F_2, \dots, F_n 划分 Ω 。那么我们可以写为 $P\{E\} = \sum_{j=1}^n P\{E|F_j\} \cdot P\{F_j\}$ 。这产生以下结果：

定理 3.11(扩展贝叶斯定律) 设 F_1, F_2, \dots, F_n 划分 Ω 。那么

$$P\{F|E\} = \frac{P\{E|F\} \cdot P\{F\}}{P\{E\}} = \frac{P\{E|F\} \cdot P\{F\}}{\sum_{j=1}^n P\{E|F_j\} P\{F_j\}}$$

示例：一种测试用于诊断罕见疾病。该测试只有 95% 的准确率。这意味着在患有该疾病的人中，它将以 95% 的概率报告“阳性”（否则为阴性），并且在没有该疾病的人中，它将报告“阴性”的概率为 95%（否则为阳性）。假设每 10 000 个孩子中就有 1 个患上这种疾病。

问题：妈妈带她的孩子接受测试。鉴于测试结果是阳性的，妈妈应该有多担心？

答案：

$$\begin{aligned} & P\{\text{孩子患病} | \text{测试阳性}\} \\ &= \frac{P\{\text{测试阳性} | \text{患病}\} \cdot P\{\text{患病}\}}{P\{\text{测试阳性} | \text{患病}\} \cdot P\{\text{患病}\} + P\{\text{测试阳性} | \text{健康}\} \cdot P\{\text{健康}\}} \\ &= \frac{0.95 \cdot \frac{1}{10000}}{0.95 \cdot \frac{1}{10000} + 0.05 \cdot \frac{9999}{10000}} = 0.0019 \end{aligned}$$

因此，孩子患病的概率大约是 2/1000。

3.7 离散随机变量与连续随机变量

考虑一个实验，例如掷两个骰子。假设我们对两次掷出的总和感兴趣。该总和可以在 2 到 12 之间，每个事件具有不同的概率。与该实验相关联的随机变量 X 是表示实验值的方式（在这种情况下是掷出的总和）。具体来说，当我们写下 X 时，可以理解 X 有许多实例，范围从 2 到 12，并且不同的实例以不同的概率出现（例如， $P\{X=3\}=2/36$ ）。

定义 3.12 一个随机变量是实验结果的实值函数。

定义 3.13 离散随机变量最多可以取无数个可能值，而连续随机变量可以取一组不可数的可能值。

问题：哪些随机变量是离散的，哪些是连续的？

- 1) 两个骰子掷出的总和
- 2) 时间 t 之前的网站访问人数
- 3) 到下一次访问网站的时间
- 4) HTTP 请求的 CPU 要求

答案：其中第一个只能采用有限数量的值——介于 2 和 12 之间，因此它显然是一个离散的随机变量。网站的到达次数可以取 0, 1, 2, 3, …，即可数集，因此这也是离散的。通常，时间被建模为连续数量，即使我们通过计算机测量时间的能力存在非零粒度。因此，上面的数量 3 和 4 是连续的随机变量。

我们使用大写字母来表示随机变量。例如，我们可以将 X 定义为等于两个骰子之和的随机变量。那么

$$P\{X = 7\} = P\{(1,6) \text{ or } (2,5) \text{ or } (3,4) \text{ or } \dots \text{ or } (6,1)\} = \frac{1}{6}$$

重点：因为“实验的结果”只是一个事件，我们所了解的关于事件的所有定理也适用于随机变量。特别是，总概率法则成立。例如，如果 N 表示到时间 t 的网站访问次数，则 $N > 10$ 是一个事件。那么我们可以使用条件来得到

$$P\{N > 10\} = P\{N > 10 \mid \text{工作日}\} \cdot \frac{5}{7} + P\{N > 10 \mid \text{周末}\} \cdot \frac{2}{7}$$

一旦我们接下来研究常见随机变量的示例，所有这些将变得更加具体。

3.8 概率和密度

3.8.1 离散：概率质量函数

离散随机变量具有可数的值，每个值都有一定的概率。

定义 3.14 设 X 为离散随机变量，则 X 的概率质量函数 $p_X(\cdot)$ 定义如下：

$$p_X(a) = P\{X = a\}, \quad \sum_x p_X(x) = 1$$

X 的累积分布函数定义为

$$F_X(a) = P\{X \leq a\} = \sum_{x \leq a} p_X(x)$$

我们也写成

$$\bar{F}_X(a) = P\{X > a\} = \sum_{x > a} p_X(x) = 1 - F_X(a)$$

常见的离散分布包括伯努利分布、二项分布、几何分布和泊松分布，所有这些都将在下面讨论。

$\text{Bernoulli}(p)$ 表示单个硬币翻转的结果，其中硬币有概率 p 出现正面(我们将此事件映射到值 1)和概率 $1-p$ 出现反面(我们将此事件映射到值 0)。如果 X 是从 $\text{Bernoulli}(p)$ 分布中得到的随机变量，那么我们写： $X \sim \text{Bernoulli}(p)$ ，并定义 X 如下：

$$X = \begin{cases} 1 & \text{概率为 } p \\ 0 & \text{否则} \end{cases}$$

随机变量 X 的概率质量函数定义如下：

$$p_X(1) = p$$

$$p_X(0) = 1 - p$$

该概率质量函数如图 3-4 所示。

$\text{Binomial}(n, p)$ 建立在 $\text{Bernoulli}(p)$ 之上。给出一个正面出现概率为 p 的硬币（成功），我们将硬币翻转 n 次（这些是独立的翻转）。如果 $X \sim \text{Binomial}(n, p)$ ，则 X 表示翻转 $\text{Bernoulli}(p)$ 硬币 n 次时的正面数（成功）。观察到 X 可以采用离散值：0, 1, 2, …, n 。

随机变量 X 的概率质量函数定义如下：

$$p_X(i) = P\{X = i\} = \binom{n}{i} p^i (1 - p)^{n-i}, \quad i = 0, 1, 2, \dots, n$$

概率质量函数如图 3-5 所示。

$\text{Geometric}(p)$ 也建立在 $\text{Bernoulli}(p)$ 之上。我们再次拥有一枚正面出现概率为 p 的硬币（成功）。我们现在翻转它直到我们取得成功，这些是独立的实验，每个都是 $\text{Bernoulli}(p)$ 。如果 $X \sim \text{Geometric}(p)$ ，则 X 代表我们获得成功之前的翻转次数。

随机变量 X 的概率质量函数定义如下：

$$p_X(i) = P\{X = i\} = (1 - p)^{i-1} p, \quad i = 1, 2, 3, \dots$$

概率质量函数如图 3-6 所示。

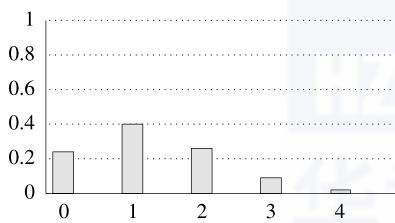


图 3-5 $\text{Binomial}(n, p)$ 分布的概率质量函数，
 $n=4$ 且 $p=0.3$

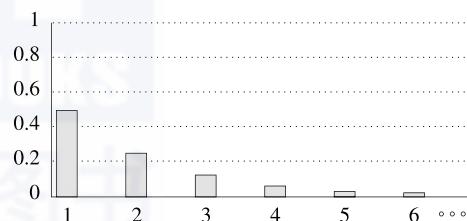


图 3-6 $\text{Geometric}(p)$ 分布的概率质量函数，
 $p=0.5$

问题：让我们回顾一下。假设你有一个存放了 n 个磁盘的房间。每个磁盘每年独立报废的概率为 p 。以下数量是如何分布的？

- 1) 第一年报废的磁盘数量
- 2) 特定磁盘报废之前的年数
- 3) 一年后特定磁盘的状态

答案：分布分别是：1) $\text{Binomial}(n, p)$ ，2) $\text{Geometric}(p)$ ，3) $\text{Bernoulli}(p)$ 。

$\text{Poisson}(\lambda)$ 是计算机系统分析中非常常见的另一种离散分布。我们通过概率质量函数定义 $\text{Poisson}(\lambda)$ 。尽管概率质量函数目前似乎没有任何意义，但我们在第 11 章中展示该分布的许多有趣属性。当观察非常大量独立源的混合时，泊松分布自然发生，每个独立源具有非常小的个体概率。因此，它可以是每单位时间到达网站或路由器的数量的合理近似值。

如果是 $X \sim \text{Poisson}(\lambda)$ ，那么

$$p_X(i) = \frac{e^{-\lambda} \lambda^i}{i!}, \quad i = 0, 1, 2, \dots$$

Poisson(λ)分布的概率质量函数如图 3-7 所示。

你可能已经注意到泊松分布与二项分布看起来并没有什么不同。事实证明，如习题 3.12 所示，如果 n 很大而 p 很小，那么 Binomial(n, p) 实际上非常接近 Poisson(np)。

3.8.2 连续：概率密度函数

连续随机变量能取无数个值。连续随机变量的范围可以被认为是实数轴上的间隔或间隔的集合。一个连续随机变量 X ，等于任何特定值的概率为零。我们根据密度函数定义一个连续随机变量的概率。

定义 3.15 连续随机变量 X 的概率密度函数是非负函数 $f_X(\cdot)$ ，其中

$$P\{a \leq X \leq b\} = \int_a^b f_X(x) dx \quad \text{其中 } \int_{-\infty}^{\infty} f_X(x) dx = 1$$

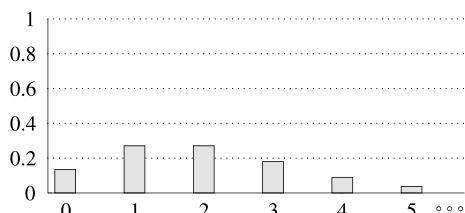


图 3-7 Poisson(λ) 分布的概率质量函数， $\lambda=2$

定义 3.15 如图 3-8 所示。

问题：对于所有 x ， $f_X(x)$ 必须 < 1 吗？

答案：不， $f_X(x) \neq P\{X=x\}$ 。

要解释密度函数 $f(\cdot)$ ，想想

$$f_X(x) dx \doteq P\{x \leq X \leq x + dx\}$$

问题：哪些是有效的概率密度函数？

$$f_X(x) = \begin{cases} 0.5x^{-0.5} & \text{如果 } 0 < x < 1 \\ 0 & \text{否则} \end{cases}$$

$$f_X(x) = \begin{cases} 2x^{-2} & \text{如果 } 0 < x < 1 \\ 0 & \text{否则} \end{cases}$$

$$f_X(x) = \begin{cases} x^{-2} & \text{如果 } 1 < x < \infty \\ 0 & \text{否则} \end{cases}$$

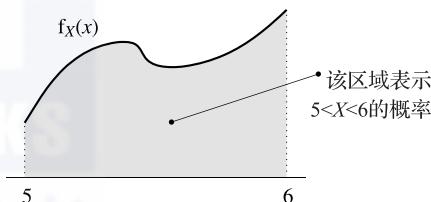


图 3-8 曲线下的面积表示随机变量 X 在 5 和 6 之间的概率，即 $\int_5^6 f_X(x) dx$

答案：只有第一个和第三个概率密度函数积分到 1，所以只有它们有效。

定义 3.16 连续随机变量 X 的累积分布函数是由函数 $F(\cdot)$ 定义的

$$F_X(a) = P\{-\infty < X \leq a\} = \int_{-\infty}^a f_X(x) dx$$

我们还可以写成：

$$\bar{F}(a) = 1 - F_X(a) = P\{X > a\}$$

问题：我们知道如何从 $f_X(x)$ 获得 $F_X(x)$ 。那么我们如何从 $F_X(x)$ 得到 $f_X(x)$ ？

答案：根据微积分的基本定理

$$f_X(x) = \frac{d}{dx} \int_{-\infty}^x F_X(t) dt = \frac{d}{dx} F_X(x)$$

有许多常见的连续分布。下面我们简单地定义几个：均匀分布、指数分布和帕累托

分布。

$\text{Uniform}(a, b)$, 通常写成 $U(a, b)$, 模拟 a 和 b 之间的长度 δ 的任何间隔都可能是一样的。具体来说, 如果 $X \sim U(a, b)$, 那么

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{如果 } a \leq x \leq b \\ 0 & \text{否则} \end{cases}$$

问题: 对于 $X \sim U(a, b)$, 如何求 $F_X(x)$?

答案:

$$F_X(x) = \int_a^x \frac{1}{b-a} dx = \frac{x-a}{b-a}$$

图 3-9 以图形方式描绘了 $f_X(x)$ 和 $F_X(x)$ 。

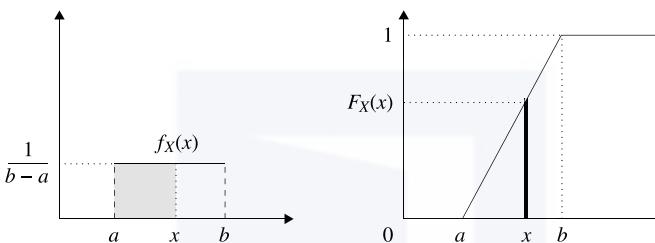


图 3-9 $U(a, b)$ 的概率密度函数 $f(x)$ 和累积分布函数 $F(x)$ 。左图中的阴影区域的面积等于右图中加粗段的高度

$\text{Exp}(\lambda)$ 表示指数分布, 其概率密度函数以指数方式下降。我们说随机变量 X 是以速率 λ 指数分布的, 写成 $X \sim \text{Exp}(\lambda)$, 如果

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

概率密度函数的图形如图 3-10 所示。累积分布函数 $F_X(x) = P\{X \leq x\}$, 由下式给出

$$F_X(x) = \int_{-\infty}^x f(y) dy = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$\bar{F}_X(x) = 1 - F_X(x) = e^{-\lambda x}, \quad x \geq 0$$

观察到 $f_X(x)$ 和 $\bar{F}(x)$ 都会以常数因子 $e^{-\lambda}$ 下降, 每个单位增加 x 。这一事实对于证明指数分布的“无记忆”特性非常重要, 如第 11 章所述。

$\text{Pareto}(\alpha)$ 是具有幂律尾部的分布, 意味着其密度以 $1/x$ 而非指数形式衰减为多项式, 如 $\text{Exp}(\lambda)$ 。参数 α 通常被称为“尾部参数”。通常假设 $0 < \alpha < 2$ 。正如我们后面所看到的, 这个 α 范围给出了 Pareto 无穷方差(方差在 3.9 节中定义)。如果 $X \sim \text{Pareto}(\alpha)$, 那么

$$f_X(x) = \begin{cases} \alpha x^{-\alpha-1} & x \geq 1 \\ 0 & \text{否则} \end{cases}$$

$$F_X(x) = 1 - x^{-\alpha}$$

$$\bar{F}_X(x) = x^{-\alpha}$$

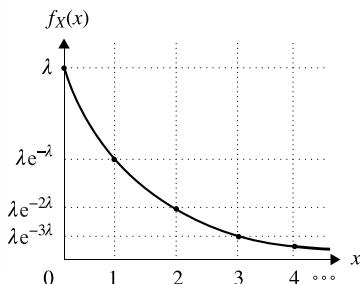


图 3-10 指数概率密度函数

虽然帕累托分布具有滑坡形状, 如指数的形状, 但它的尾部减少得更慢(比较两个分布的 $\bar{F}(x)$)。帕累托分布可以说有“重尾”或“肥尾”, 其中较低的 α 对应较胖的尾。这将在第 20 章中介绍。

我们没有提到的一个重要的连续分布是正态分布，又名高斯分布。关于正态分布的讨论需要首先理解期望和方差，因此我们将其推迟到3.14节。

3.9 期望和方差

分布的均值(也称为期望)紧跟在分布的概率质量函数(在连续分布的情况下为密度函数)之后。对于随机变量 X ，我们用 $E[X]$ 来表示它的期望。这在下表中定义。

离散情况	连续情况
$E[X] = \sum_x x \cdot p_X(x)$	$E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) dx$

对于离散随机变量 X 的情况，其期望可以被视为可能结果的总和，每个结果由其概率加权。

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

要了解它，请考虑以下示例。

示例：平均午餐费用

我午餐的平均费用是多少？

星期一	星期二	星期三	星期四	星期五	星期六	星期日
\$ 7	\$ 7	\$ 5	\$ 5	\$ 5	\$ 0	\$ 2

$$\text{平均费用} = \frac{7 + 7 + 5 + 5 + 5 + 0 + 2}{7}$$

|||

$$E[\text{费用}] = \frac{2}{7}(7) + \frac{3}{7}(5) + \frac{1}{7}(2) + \frac{1}{7}(0)$$

每个可能的值(7、5、2和0)按其概率加权。

问题：如果 $X \sim \text{Bernoulli}(p)$ ，那么 $E[X]$ 是多少？

答案： $E[X] = 0 \cdot (1 - p) + 1 \cdot (p) = p$ 。

问题：假设一枚硬币出现正面的概率为 $1/3$ 。在期望中，需要投掷硬币多少次才能获得正面？

答案：这只是 $E[X]$ ，其中 $X \sim \text{Geometric}(p)$ ， $p = 1/3$ 。假设 $X \sim \text{Geometric}(p)$ ，我们有

$$\begin{aligned} E[X] &= \sum_{n=1}^{\infty} n (1 - p)^{n-1} p = p \cdot \sum_{n=1}^{\infty} n \cdot q^{n-1}, \quad q = (1 - p) \\ &= p \cdot \sum_{n=1}^{\infty} \frac{d}{dq}(q^n) = p \cdot \frac{d}{dq} \sum_{n=1}^{\infty} q^n = p \cdot \frac{d}{dq} \left(\frac{q}{1-q} \right) = \frac{p}{(1-q)^2} = \frac{1}{p} \end{aligned}$$

因此，当 $p = 1/3$ 时，预期的翻转次数为3。

问题：如果 $X \sim \text{Poisson}(\lambda)$ ，那么 $E[X]$ 是多少？

答案：

$$\begin{aligned} E[X] &= \sum_{i=0}^{\infty} i \frac{e^{-\lambda} \lambda^i}{i!} = \sum_{i=1}^{\infty} i \frac{e^{-\lambda} \lambda^i}{i!} = \lambda e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} \\ &= \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} e^{\lambda} = \lambda \end{aligned}$$

问题：如果 $X \sim \text{Exp}(\lambda)$ ，那么 $E[X]$ 是什么？

答案：

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^{\infty} x \lambda e^{-\lambda x} dx = \frac{1}{\lambda} \quad (\text{分步积分})$$

观察到泊松分布的 λ 参数也是它的期望，对于指数分布， λ 参数是期望的倒数。我们将 λ 称为指数的“速率”。例如，如果直到下次到达的时间是指数分布的，每秒到达率为 3，则到下一次到达的预期时间为 $1/3$ 秒。

我们还可以考虑随机变量 X 的更高阶矩。随机变量 X 的 i 阶矩由 $E[X^i]$ 表示，定义如下：

离散情况	连续情况
$E[X^i] = \sum_x x^i \cdot p_X(x)$	$E[X^i] = \int_{-\infty}^{\infty} x^i \cdot f_X(x) dx$

更一般地说，我们可以讨论随机变量 X 的函数 $g(\cdot)$ 的期望。对于离散的随机变量 X ，定义如下：

$$E[g(X)] = \sum_x g(x) \cdot p_X(x)$$

对于连续的随机变量 X 定义如下：

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

问题：假设 X 定义如下：

$$X = \begin{cases} 0 & \text{概率为 } 0.2 \\ 1 & \text{概率为 } 0.5 \\ 2 & \text{概率为 } 0.3 \end{cases}$$

$E[X]$ 是多少， $E[2X^2 + 3]$ 是多少？

答案：

$$E[X] = (0)(0.2) + (1)(0.5) + (2)(0.3)$$

$$E[2X^2 + 3] = (2 \cdot 0^2 + 3)(0.2) + (2 \cdot 1^2 + 3)(0.5) + (2 \cdot 2^2 + 3)(0.3)$$

你可能已经注意到 $E[2X^2 + 3] = 2E[X^2] + 3$ 。这不是巧合，而是由于 3.13 节中讨论的期望的线性性。

定义 3.17 随机变量 X 的方差写为 $\text{Var}(X)$ ，是 X 与其期望的预期平方差(即我们期待 X 与其期望 $E[X]$ 之差的平方数)。定义如下：

$$\text{Var}(X) = E[(X - E[X])^2]$$

并且可以等效地表达为：

$$\text{Var}(X) = E[X^2] - (E[X])^2$$

(当我们在3.13节讨论期望的线性性之后,就会很明显地推导出为什么这些表达式是等价的。)

问题: 如果 $X \sim \text{Bernoulli}(p)$, 那么 $\text{Var}(X)$ 是什么?

答案:

$$E[X] = p$$

$$\text{Var}(X) = E[(X - p)^2] = p(1 - p)^2 + (1 - p)(0 - p)^2 = p(1 - p)$$

问题: $X \sim \text{Uniform}(a, b)$ 的方差是多少?

答案:

$$E[X] = \int_a^b x \frac{1}{b-a} dx = \frac{1}{b-a} \cdot \frac{(b^2 - a^2)}{2} = \frac{a+b}{2}$$

$$\text{Var}(X) = E\left[\left(X - \frac{a+b}{2}\right)^2\right] = \int_a^b \left(X - \frac{a+b}{2}\right)^2 \cdot \frac{1}{b-a} dx = \frac{(b-a)^2}{12}$$

表3-2显示了概率质量函数(或概率密度函数)以及许多常见分布的期望和方差。

表3-2 离散和连续分布

分布	概率质量函数 $p_X(x)$	期望	方差
Bernoulli(p)	$p_X(0) = 1 - p; p_X(1) = p$	p	$p(1 - p)$
Binomial(n, p)	$p_X(x) = \binom{n}{x} p^x (1 - p)^{n-x}, x = 0, 1, \dots, n$	np	$np(1 - p)$
Geometric(p)	$p_X(x) = (1 - p)^{x-1} p, x = 1, 2, \dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	$p_X(x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}, x = 0, 1, 2, \dots$	λ	λ
分布	概率密度函数 $f_X(x)$	期望	方差
Exp(λ)	$f_X(x) = \lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Uniform(a, b)	$f_X(x) = \frac{1}{b-a}, \text{如果 } a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Pareto(α), $0 < \alpha < 2$	$f_X(x) = \alpha x^{-\alpha-1}, \text{如果 } x > 1$	$\begin{cases} \infty & \text{如果 } \alpha \leq 1 \\ \frac{\alpha}{\alpha-1} & \text{如果 } \alpha > 1 \end{cases}$	∞
Normal(μ, σ^2)	$f_X(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, -\infty < x < \infty$	μ	σ^2

3.10 联合概率和独立性

我们经常感兴趣的是两个或多个随机变量同时出现某种情况的概率。例如, 我们可能想知道两个磁盘在某个时间间隔内同时崩溃的概率。两个磁盘的行为可能相关或不相关。作为另一个示例, 计算机系统性能通常根据能量延迟乘积^[68]来衡量, 即系统使用的能量和用户经历的延迟的乘积。能量和延迟通常彼此相关, 并且可以想象这两个随机变量之间的联合分布。在本节和下一节中, 我们将介绍正式表达这些想法所需的定义。

定义3.18 离散随机变量的 X 和 Y 之间的联合概率质量函数定义为

$$p_{X,Y}(x, y) = P\{X = x \& Y = y\}$$

这通常写为 $P\{X=x, Y=y\}$ 。类似地, $f_{X,Y}(x, y)$ 表示连续随机变量的 X 和 Y 之间的联合概率密度函数, 其中

$$\int_c^d \int_a^b f_{X,Y}(x, y) dx dy = P\{a < X < b \quad \& \quad c < Y < d\}$$

问题: $f_X(x)$ 和 $f_{X,Y}(x, y)$ 之间的关系是什么?

答案: 应用总概率定律, 我们有

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy \quad \text{和} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx$$

同样

$$p_X(x) = \sum_y p_{X,Y}(x, y) \quad \text{和} \quad p_Y(y) = \sum_x p_{X,Y}(x, y)$$

与我们将两个事件 E 和 F 定义为独立的方式类似, 我们同样可以将两个随机变量定义为独立的。

定义 3.19 如果

$$p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y)$$

我们说离散随机变量的 X 和 Y 是独立的, 写成 $X \perp Y$ 。同样, 如果

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y), \quad \forall x, y$$

我们说连续随机变量的 X 和 Y 是独立的。

定理 3.20 如果 $X \perp Y$, 那么 $E[XY] = E[X] \cdot E[Y]$ 。

证明:

$$\begin{aligned} E[XY] &= \sum_x \sum_y xy \cdot P\{X = x, Y = y\} \\ &= \sum_x \sum_y xy \cdot P\{X = x\} P\{Y = y\} \quad (\text{根据 } \perp \text{ 的定义}) \\ &= \sum_x x P\{X = x\} \cdot \sum_y y P\{Y = y\} = E[X]E[Y] \end{aligned}$$

同样的论证适用于连续的随机变量。 ■

同样的证据表明, 如果 $X \perp Y$, 那么, 对于任意函数 g 和 f , 有

$$E[g(X)f(Y)] = E[g(X)] \cdot E[f(Y)]$$

问题: 如果 $E[XY] = E[X]E[Y]$, 这是否意味着 $X \perp Y$?

答案: 不, 见习题 3.10。

3.11 条件概率和期望

正如我们研究事件的条件概率, 即假设一个事件发生时, 另一个事件发生的概率我们也可以将其扩展到随机变量中的条件概率。从离散情况开始, 然后到连续情况。以下示例将有助于激发这一想法。

示例: 头发颜色

假设我们将班级中的人分为金发人(颜色值 1)、红发人(颜色值 2)、棕发人(颜色值 3)和黑发人(颜色值 4)。假设 5 名学生是金发, 2 名是红发, 17 名是棕发, 14 名是黑发。设

X 是随机变量，其值为头发颜色。那么 X 的概率质量函数如下所示：

$$p_X(1) = P\{\text{金发}\} = 5/38$$

$$p_X(2) = P\{\text{红发}\} = 2/38$$

$$p_X(3) = P\{\text{棕发}\} = 17/38$$

$$p_X(4) = P\{\text{黑发}\} = 14/38$$

现在我们说如果头发的颜色是金色或红色，那么这个人的头发是浅色的。如果头发的颜色是棕色或黑色，那么这个人的头发是深色的。设 A 表示人的头发颜色是浅色的事件。

$$P\{A\} = 7/38 \quad \text{且} \quad P\{A^c\} = 31/38$$

定义 3.21 设 X 是一个离散的随机变量，在可数空间上定义概率质量函数 $p_X(\cdot)$ 。令 A 为一个事件，满足 $P\{A\} > 0$ 。那么 $p_{X|A}(\cdot)$ 是给定事件 A 的 X 的条件概率质量函数。我们定义

$$p_{X|A}(x) = P\{X = x | A\} = \frac{P\{(X = x) \cap A\}}{P\{A\}}$$

更正式的说法是，如果 Ω 表示样本空间而 ω 表示样本空间中的样本点， $\{\omega : X(\omega) = x\}$ 是导致 X 具有值 x 的样本点集，那么

$$p_{X|A}(x) = P\{X = x | A\} = \frac{P\{\{\omega : X(\omega) = x\} \cap A\}}{P\{A\}}$$

因此，条件概率涉及缩小概率空间。例如

$$p_{X|A}(\text{金发}) = \frac{P\{(X = \text{金发}) \cap A\}}{P\{A\}} = \frac{\frac{5}{38}}{\frac{7}{38}} = \frac{5}{7}$$

同样， $p_{X|A}(\text{红发}) = 2/7$ 。

另一个示例

$$p_{X|A}(\text{棕发}) = \frac{P\{(X = \text{棕发}) \cap A\}}{P\{A\}} = \frac{0}{\frac{7}{38}} = 0$$

同样， $p_{X|A}(\text{黑发}) = 0$ 。

问题：如果我们对所有 x 求和 $p_{X|A}(x)$ ，我们得到了什么？

答案：

$$\sum_x p_{X|A}(x) = \sum_x \frac{P\{(X = x) \cap A\}}{P\{A\}} = \frac{P\{A\}}{P\{A\}} = 1$$

因此 $p_{X|A}(x)$ 是有效的概率质量函数。

定义 3.22 对于离散随机变量 X ，给定事件 A 的条件期望 X 如下：

$$E[X | A] = \sum_x x p_{X|A}(x) = \sum_x x \cdot \frac{P\{(X = x) \cap A\}}{P\{A\}}$$

问题：对于我们的示例，将金发视为值 1，而红发视为值 2， $E[X | A]$ 是什么？

答案：

$$E[X | A] = 1 \cdot \frac{5}{7} + 2 \cdot \frac{2}{7} = \frac{9}{7}$$

我们还可以考虑事件 A 是随机变量的实例的情况。例如， A 可能是事件 $Y=y$ 。那么在给定事件 $Y=y$ 的情况下书写 X 的条件概率质量函数是常见的

$$p_{X|Y}(x|y) = P\{X=x|Y=y\} = \frac{P\{X=x \& Y=y\}}{P\{Y=y\}} = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

和

$$E[X|Y=y] = \sum_x x \cdot p_{X|Y}(x|y)$$

随机变量条件化示例

取值为 {0, 1, 2} 的两个离散随机变量 X 和 Y 具有表 3-3 给出的联合概率质量函数。

表 3-3 联合概率质量函数 $p_{X,Y}(x, y)$

	$X=0$	$X=1$	$X=2$
$Y=2$	0	$\frac{1}{6}$	$\frac{1}{8}$
$Y=1$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{8}$
$Y=0$	$\frac{1}{6}$	$\frac{1}{8}$	0

问题：计算条件期望 $E[X|Y=2]$ 。

答案：

$$\begin{aligned} E[X|Y=2] &= \sum_x x \cdot p_{X|Y}(x|2) = \sum_x x \cdot P\{X=x|Y=2\} \\ &= 0 \cdot \frac{P\{X=0 \& Y=2\}}{P\{Y=2\}} + 1 \cdot \frac{P\{X=1 \& Y=2\}}{P\{Y=2\}} \\ &\quad + 2 \cdot \frac{P\{X=2 \& Y=2\}}{P\{Y=2\}} = 1 \cdot \frac{\frac{1}{6}}{\frac{7}{24}} + 2 \cdot \frac{\frac{1}{8}}{\frac{7}{24}} = \frac{10}{7} \end{aligned}$$

对于连续的实值随机变量 X ，给定事件 A 的 X 的条件概率密度函数类似于离散情况的条件概率密度函数，除了 A 现在是实数轴的子集，其中我们定义 $P\{X \in A\}$ 是 X 在子集 A 内取值的概率。

定义 3.23 设 X 是连续随机变量，在实数上定义了概率密度函数 $f_X(\cdot)$ 。设 A 是实数轴的子集，其中 $P\{X \in A\} > 0$ 。那么 $f_{X|A}(\cdot)$ 是给定事件 A 的 X 的条件概率密度函数。我们定义

$$f_{X|A}(x) = \begin{cases} \frac{f_X(x)}{P\{X \in A\}} & \text{如果 } x \in A \\ 0 & \text{否则} \end{cases}$$

与离散情况一样，条件概率密度函数在条件集 A 之外为零。在 A 中，条件概率密度函数具有与无条件形状完全相同的形状，除了由常数因子 $\frac{1}{P\{X \in A\}}$ 缩放，以便 $f_{X|A}(x)$ 积分到 1。

定义 3.24 设 X 为连续随机变量，在实数上定义概率密度函数 $f_X(\cdot)$ 。设 A 是实

数轴的子集，其中 $P\{X \in A\} > 0$ 。 X 在给定 A 下的条件期望写成 $E[X | A]$ ，有以下定义：

$$E[X | A] = \int_{-\infty}^{\infty} xf_{X|A}(x) dx = \int_A xf_{X|A}(x) dx = \frac{1}{P\{X \in A\}} \int_A xf_X(x) dx$$

示例：匹兹堡超级计算中心

匹兹堡超级计算中心(PSC)为来自全国各地的科学家运行大型并行作业。为了适当地向用户收费，根据所需的 CPU 小时数将作业分组到不同的箱中，每个箱具有不同的价格。假设作业持续时间是指数分布的，平均为 1000 个处理器小时。进一步假设所有需要少于 500 个处理器小时的作业被发送到箱 1，并且所有剩余的作业被发送到箱 2。

问题：考虑以下问题：

- (a) $P\{\text{作业被发送到箱 1}\}$ 是多少？
- (b) $P\{\text{作业持续时间} < 200 | \text{作业发送到箱 1}\}$ 是多少？
- (c) 持续时间 X 的条件密度 $f_{X|Y}(t)$ 是多少(其中 Y 是作业被发送到箱 1 的事件)？
- (d) $E[\text{作业持续时间} | \text{作业在箱 1}]$ 是多少？

答案：首先回忆一下对于 $X \sim \text{Exp}\left(\frac{1}{1000}\right)$ 我们有

$$f_X(t) = \begin{cases} \frac{1}{1000} e^{-\frac{t}{1000}} & t > 0 \\ 0 & \text{否则} \end{cases}$$

$$F_X(t) = P\{X \leq t\} = 1 - e^{-\frac{1}{1000}t}$$

那么

$$(a) F_X(500) = 1 - e^{-\frac{500}{1000}} = 1 - e^{-\frac{1}{2}} \approx 0.39$$

$$(b) \frac{F_X(200)}{F_X(500)} = \frac{1 - e^{-\frac{1}{5}}}{1 - e^{-\frac{1}{2}}} \approx 0.46$$

$$(c) f_{X|Y}(t) = \begin{cases} \frac{f_X(t)}{F(500)} = \frac{\frac{1}{1000} e^{-\frac{t}{1000}}}{1 - e^{-\frac{1}{2}}} & t < 500 \\ 0 & \text{否则} \end{cases}$$

$$(d) E[\text{作业持续时间} | \text{作业在箱 1}] = \int_{-\infty}^{\infty} tf_{X|Y}(t) dt = \int_0^{500} t \frac{\frac{1}{1000} e^{-\frac{t}{1000}}}{1 - e^{-\frac{1}{2}}} dt \approx 229$$

问题：为什么 1 号箱的预期作业规模小于 250?

答案：考虑指数概率密度函数的形状。现在将其截断为 500，并将所有内容扩展为使其积分到 1 所需的常量。较小的值仍有更多权重，因此预期值小于中点。

问题：如果作业持续时间服从 $\text{Uniform}(0, 2000)$ 分布，问题(d)的答案如何改变，期望仍然是 1000 吗？

答案：从逻辑上讲，鉴于作业位于箱 1 并且分布均匀，我们应该发现预期的作业持续时间为 250 小时。这是一个代数论证：

$$E[\text{作业持续时间} | \text{作业在箱 1}] = \int_{-\infty}^{\infty} tf_{X|Y}(t) dt = \int_0^{500} t \frac{\frac{1}{2000}}{\frac{500}{2000}} dt = 250$$

3.12 基于条件化的概率和期望

回想一下总概率定律，如果 F_1, \dots, F_n 分割样本空间 Ω ，那么

$$P\{E\} = \sum_{i=1}^n P\{E|F_i\}P\{F_i\}$$

这扩展到随机变量，因为“ $X=k$ ”是一个事件。

离散随机变量的总概率表示

$$P\{X=k\} = \sum_y P\{X=k|Y=y\}P\{Y=y\}$$

(可以对连续随机变量进行类似的陈述。)

这是一个强大的工具！它允许我们将问题分解为许多更简单的问题。像往常一样，诀窍是知道要对什么进行条件化。

示例：哪个指数先发生？

导出 $P\{X_1 < X_2\}$ ，其中 $X_1 \perp X_2$ ， $X_1 \sim \text{Exp}(\lambda_1)$ ， $X_2 \sim \text{Exp}(\lambda_2)$ 。

问题：你以什么作为条件？

答案：我们选择以 X_2 的值为条件，我们使用 $f_2(\cdot)$ 来表示 X_2 的概率密度函数：

$$\begin{aligned} P\{X_1 < X_2\} &= \int_{-\infty}^{\infty} P\{X_1 < X_2 | X_2 = k\} \cdot f_2(k) dk \\ &= \int_0^{\infty} P\{X_1 < k\} \cdot \lambda_2 e^{-\lambda_2 k} dk = \int_0^{\infty} (1 - e^{-\lambda_1 k}) (\lambda_2 e^{-\lambda_2 k}) dk = \frac{\lambda_1}{\lambda_1 + \lambda_2} \end{aligned}$$

定理 3.25 对于离散随机变量，

$$E[X] = \sum_y E[X|Y=y]P\{Y=y\}$$

对于连续随机变量，

$$E[X] = \int E[X|Y=y]f_Y(y) dy$$

证明：我们提供离散情况的证明：

$$\begin{aligned} E[X] &= \sum_x x P\{X=x\} = \sum_x x \sum_y P\{X=x|Y=y\}P\{Y=y\} \\ &= \sum_x \sum_y x P\{X=x|Y=y\}P\{Y=y\} = \sum_y \sum_x x P\{X=x|Y=y\}P\{Y=y\} \\ &= \sum_y P\{Y=y\} \sum_x x P\{X=x|Y=y\} = \sum_y P\{Y=y\} E[X|Y=y] \end{aligned} \quad \blacksquare$$

这个证明概括为

$$E[g(X)] = \sum_y E[g(X)|Y=y]P\{Y=y\}$$

当我们需要计算 X 或更高阶矩的方差时，这是非常重要的。

示例：几何分布

假设我们想要使用条件化来轻松计算参数 p 的几何分布的期望。也就是说，我们寻找 $E[N]$ ，其中 N 是获得第一个正面所需的翻转次数。

问题：我们对什么进行条件化？

答案：我们对第一次翻转的值 Y 进行条件化，如下所示：

$$\begin{aligned}E[N] &= E[N|Y=1]P\{Y=1\} + E[N|Y=0]P\{Y=0\} = 1p + (1+E[N])(1-p) \\pE[N] &= p + (1-p) \\E[N] &= \frac{1}{p}\end{aligned}$$

注意这个推导比我们对几何分布的期望的原始推导更简单！

3.13 期望的线性性质

下面是一个强有力的概率论定理。

定理 3.26(期望的线性性质) 对于随机变量 X 和 Y ,

$$E[X+Y] = E[X] + E[Y]$$

问题：定理 3.26 是否需要 $X \perp Y$?

答案：不！回想一下，我们确实需要独立性来简化 $E[XY]$ ，而不是 $E[X+Y]$ 。

证明：这是 X 和 Y 连续的证明。离散情况类似：只需用 $p_{X,Y}(x, y)$ 替换 $f_{X,Y}(x, y)$ 。

$$\begin{aligned}E[X+Y] &= \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} (x+y) f_{X,Y}(x,y) dx dy \\&= \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} xf_{X,Y}(x,y) dx dy + \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} yf_{X,Y}(x,y) dx dy \\&= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} xf_{X,Y}(x,y) dy dx + \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} yf_{X,Y}(x,y) dx dy \\&= \int_{x=-\infty}^{\infty} x \int_{y=-\infty}^{\infty} f_{X,Y}(x,y) dy dx + \int_{y=-\infty}^{\infty} y \int_{x=-\infty}^{\infty} f_{X,Y}(x,y) dx dy \\&= \int_{x=-\infty}^{\infty} xf_X(x) dx + \int_{y=-\infty}^{\infty} yf_Y(y) dy \\&= E[X] + E[Y]\end{aligned}$$

此标识可以简化许多证明。请考虑以下示例。

示例：Binomial 分布

$X \sim \text{Binomial}(n, p)$ 。 $E[X]$ 是多少？

问题：如果我们只使用 Binomial 分布的定义，我们对 $E[X]$ 有什么表达式？

答案： $E[X] = \sum_{i=0}^n i \binom{n}{i} p^i (1-p)^{n-i}$ 。这个表达式似乎令人生畏。

问题：我们可以将 $\text{Binomial}(n, p)$ 视为随机变量的总和吗？

答案：令

$$X = n \text{ 次试验中成功的次数} = X_1 + X_2 + \dots + X_n$$

其中

$$\begin{aligned}X_i &= \begin{cases} 1 & \text{如果第 } i \text{ 次试验成功} \\ 0 & \text{否则} \end{cases} \\E[X_i] &= p\end{aligned}$$

那么

$$E[X] = E[X_1] + E[X_2] + \dots + E[X_n] = nE[X_i] = np$$

这个结果应该是有意义的，因为 n 个硬币翻转，每个翻转到正面的概率为 p ，应该导致平均 np 个正面。

上面的 X_i 被称为指标随机变量，因为它们取值 0 或 1。在前面的示例中， X_i 是独立相同分布的。然而，即使试验不是独立的，我们也会有

$$E[X] = E[X_1] + \dots + E[X_n]$$

以下示例清楚地说明了这一点。

示例：帽子

在派对上， n 个人将帽子扔到圆圈的中间。每个人闭上眼睛，随机挑一个帽子。设 X 代表取回自己帽子的人数。我们的目标是确定 $E[X]$ 。

问题：我们如何将 X 表示为指标随机变量的总和？

答案： $X = I_1 + I_2 + \dots + I_n$ ，其中

$$I_i = \begin{cases} 1 & \text{如果第 } i \text{ 个人拿到了自己的帽子} \\ 0 & \text{否则} \end{cases}$$

观察虽然 I_i 具有相同的分布(通过对称)，但它们并不是彼此独立的！尽管如此，我们仍然可以使用期望线性来表示

$$E[X] = E[I_1] + E[I_2] + \dots + E[I_n] = nE[I_i] = n\left(\frac{1}{n} \cdot 1 + \frac{n-1}{n} \cdot 0\right) = 1$$

观察到期望线性也可以用来表示

$$E[X^2 + Y^2] = E[X^2] + E[Y^2]$$

尽管如此，这并不意味着期望线性适用于方差。为此，我们需要一个独立性假设，如下面的定理。

定理 3.27 设 X 和 Y 是随机变量，其中 $X \perp Y$ 。那么

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

证明：

$$\begin{aligned} \text{Var}(X + Y) &= E[(X + Y)^2] - (E[(X + Y)])^2 \\ &= E[X^2] + E[Y^2] + 2E[XY] - (E[X])^2 - (E[Y])^2 - 2E[X]E[Y] \\ &= \text{Var}(X) + \text{Var}(Y) + \underbrace{2E[XY] - 2E[X]E[Y]}_{\text{如果 } X \perp Y \text{ 则等于 } 0} \end{aligned}$$

3.14 正态分布

一个非常重要且无处不在的连续分布是正态分布。

定义 3.28 连续随机变量 X 被称为 $\text{Normal}(\mu, \sigma^2)$ 或 $\text{Gaussian}(\mu, \sigma^2)$ ，如果它有概率密度函数 $f_X(x)$ 形为

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \quad -\infty < x < \infty$$

其中 $\sigma > 0$ 。参数 μ 称为期望，参数 σ 称为标准差。

定义 3.29 $\text{Normal}(0, 1)$ 随机变量 Y 被认为是标准正态。它的累积分布函数表示为

$$\Phi(y) = F_Y(y) = P\{Y \leq y\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y e^{-t^2/2} dt$$

Normal(μ, σ^2)概率密度函数具有“钟形”形状，并且在 μ 周围明显对称，如图3-11所示。事实上，定义3.28中的 $f_X(x)$ 实际上是一个密度函数，可以通过将其变换为极坐标来看出这一事实（相信我，你不想看到具体的细节^[176]）。

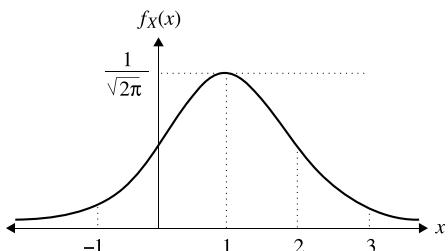


图3-11 Normal(1, 1)概率密度函数

定义3.30 设 $X \sim \text{Normal}(\mu, \sigma^2)$ ，那么 $E[X] = \mu$ 和 $\text{Var}(X) = \sigma^2$ 。

证明：因为 $f_X(x)$ 围绕 μ 对称，很明显 $E[X] = \mu$ 。

$$\begin{aligned}\text{Var}(X) &= \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x - \mu)^2 e^{-\frac{1}{2}((x-\mu)/\sigma)^2} dx \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-y^2/2} dy \quad (\text{通过改变变量 } y = (x - \mu)/\sigma \text{ 和 } dx = \sigma dy) \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y \cdot (ye^{-y^2/2}) dy = \frac{\sigma^2}{\sqrt{2\pi}} (-ye^{-y^2/2}) \Big|_{-\infty}^{\infty} + \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy \quad \text{通过分步积分} \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy = \sigma^2\end{aligned}$$

最后一行是通过使用事实

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy = 1$$

获得的。因为被积函数是标准正态的密度函数。 ■

3.14.1 线性变换特性

正态分布有一个非常特殊的属性，称为“线性变换特性”，它表示如果 X 是一个正态的随机变量，并且你采用 X 的线性函数，那么新的随机变量也具有正态分布。请注意，对于我们看到的其他分布，例如指数分布，此属性不适用。

定义3.31(线性变换特性) 设 $X \sim \text{Normal}(\mu, \sigma^2)$ 。设 $Y = aX + b$ ，其中 $a > 0$ 和 b 为标量。那么 $Y \sim \text{Normal}(a\mu + b, a^2\sigma^2)$ 。

证明：很容易证明 $E[Y] = aE[X] + b = a\mu + b$ 和 $\text{Var}(Y) = a^2\text{Var}(X) = a^2\sigma^2$ 。剩下的就是表明 $f_Y(y)$ 是正态分布的。我们将 Y 的累积分布函数与 X 的累积分布函数关联如下：

$$F_Y(y) = P\{Y \leq y\} = P\{aX + b \leq y\} = P\left\{X \leq \frac{y - b}{a}\right\} = F_X\left(\frac{y - b}{a}\right)$$

我们现在就按 y 给两边求导：

$$\frac{d}{dy}F_Y(y) = \frac{d}{dy} \int_{-\infty}^y f_Y(t) dt = f_Y(y)$$

$$\frac{d}{dy}F_X\left(\frac{y-b}{a}\right) = \frac{d}{dy} \int_{-\infty}^{\frac{y-b}{a}} f_X(t) dt = f_X\left(\frac{y-b}{a}\right) \cdot \frac{d}{dy}\left(\frac{y-b}{a}\right) = f_X\left(\frac{y-b}{a}\right) \cdot \frac{1}{a}$$

因此我们已经证明了

$$f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right)$$

对这个求值，我们有

$$\begin{aligned} f_Y(y) &= \frac{1}{a} f_X\left(\frac{y-b}{a}\right) = \frac{1}{a \sqrt{2\pi}\sigma} e^{-(\frac{y-b}{a}-\mu)^2/2\sigma^2} = \frac{1}{\sqrt{2\pi}(a\sigma)} e^{-(y-b-a\mu)^2/2a^2\sigma^2} \\ &= \frac{1}{\sqrt{2\pi}(a\sigma)} e^{-(y-(b+a\mu))^2/2a^2\sigma^2} \end{aligned}$$

所以 $f_Y(y)$ 是一个正态的概率密度函数，期望为 $a\mu+b$ ，方差为 $a^2\sigma^2$ 。 ■

不幸的是，我们不知道如何将正态的密度从 0 到 y 的积分进行符号化。为了计算正态分布的累积分布函数，我们必须使用数字积分结果表来表示 $\Phi(y)$ ，例如[200]中给出的结果^②。接下来给出该表的子集以供参考：

y	0.5	1.0	1.5	2.0	2.5	3.0
$\Phi(y)$	0.6915	0.8413	0.9332	0.9772	0.9938	0.9987

问题：看一下你看到的表，例如， $\Phi(1)=0.8413$ 。这告诉我们标准正态分布在离期望一个标准差范围内的概率是多少？

答案：我们得到 $P\{Y<1\}=0.84$ 。我们想知道 $P\{-1<Y<1\}$ 。

$$\begin{aligned} P\{-1 < Y < 1\} &= P\{Y < 1\} - P\{Y < -1\} = P\{Y < 1\} - P\{Y > 1\} \quad (\text{根据对称性}) \\ &= P\{Y < 1\} - (1 - P\{Y < 1\}) = 2P\{Y < 1\} - 1 = 2\Phi(1) - 1 \\ &\doteq 2 \cdot 0.84 - 1 = 0.68 \end{aligned}$$

所以概率为 68%，我们在期望的一个标准差内。

同样，我们可以使用相同的参数来表示概率为 95%，我们在期望的两个标准差内，并且概率为 99.7%，我们在期望的三个标准偏差内，等等。

问题：以前的结果表示标准正态。如果我们没有标准正态怎么办？

答案：我们可以使用线性变换特性将非标准正态转换为标准正态。下面是它的工作原理：

$$X \sim \text{Normal}(\mu, \sigma^2) \Leftrightarrow Y = \frac{X - \mu}{\sigma} \sim \text{Normal}(0, 1)$$

所以

$$P\{X < k\} = P\left\{\frac{X - \mu}{\sigma} < \frac{k - \mu}{\sigma}\right\} = P\left\{Y < \frac{k - \mu}{\sigma}\right\} = \Phi\left(\frac{k - \mu}{\sigma}\right)$$

定理 3.32 如果 $X \sim \text{Normal}(\mu, \sigma^2)$ ，那么 X 偏离其期望小于 k 个标准差的概率与标准正态偏离其期望小于 k 个标准差的概率相同。

证明：设 $Y \sim \text{Normal}(0, 1)$ 。那么

② 实际上，再也没有人会用这个表格了，因为有一些近似方法可让你将表格中的值计算到小数点后七位，如 [131] 所示。

$$P\{-k\sigma < X - \mu < k\sigma\} = P\left\{-k < \frac{X - \mu}{\sigma} < k\right\} = P\{-k < Y < k\}$$

定理 3.32 说明了为什么在标准差方面比在绝对值方面更容易思考。

问题：智商测试的支持者将告诉你，人类智商(IQ)已经显示为正态分布，期望为 100，标准差为 15。有多少人的智商高于 130(“天赋临界点”)？

答案：我们正在寻找智商高于期望两个标准偏差的人群。这与标准正态超过其期望两个标准偏差以上的概率相同，即 $1 - \Phi(2) = 0.023$ 。因此，只有约 2% 的人智商高于 130。

3.14.2 中心极限定理

考虑采样该州所有个体的高度并取均值。我们很快定义的中心极限定理(CLT)表示这个均值倾向于正态分布。即使我们采用大量独立同分布随机变量的均值也是如此，其中随机变量来自明显非正态的分布，比如统一分布。正是这个特性使正态分布如此重要！

我们现在更正式地说明这一点。设 $X_1, X_2, X_3, \dots, X_n$ 是独立同分布的随机变量，其中有一些期望 μ 和方差 σ^2 。注意：我们不假设这些是正态分布的随机变量。事实上，我们甚至没有假设它们必然是连续的——它们可能是离散的随机变量

令

$$S_n = X_1 + X_2 + \dots + X_n \quad (3.2)$$

问题： S_n 的期望和标准差是多少？

答案： $E[S_n] = n\mu$ 和 $\text{Var}(S_n) = n\sigma^2$ 。因此标准差是 $\sigma\sqrt{n}$ 。

令

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

问题： Z_n 的期望和标准差是多少？

答案： Z_n 的期望为 0，标准差为 1。

定理 3.33(中心极限定理(CLT)) 设 X_1, X_2, \dots, X_n 是独立同分布随机变量的序列，具有共同的期望 μ 和有限方差 σ^2 ，并且定义

$$Z_n = \frac{X_1 + \dots + X_n - n\mu}{\sigma\sqrt{n}}$$

那么 Z_n 的累积分布函数收敛到标准正态累积分布函数。也就是，对于每个 z

$$\lim_{n \rightarrow \infty} P\{Z_n \leq z\} = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-x^2/2} dx$$

证明：证明将使用变换，因此将 CLT 的证明推迟到第 25 章，见习题 25.15。 ■

问题：式(3.2)中 S_n 的分布是什么？

答案：通过线性变换特性， $S_n \sim \text{Normal}(n\mu, n\sigma^2)$ 。

中心极限定理是非常普遍的，它解释了许多导致正态分布的自然现象。CLT 适用于独立同分布随机变量的任何总和的这一事实允许我们证明，当 n 升高时， $\text{Binomial}(n, p)$ 分布，即独立同分布 $\text{Bernoulli}(p)$ 随机变量的总和，收敛于正态分布。当我们在第 11 章中更深入地研究泊松分布时，我们将看到泊松(λ)分布也可以被视为独立同分布随机变量的总和；因此泊松(λ)分布也很好地用正态分布近似，期望为 λ ，方差为 λ 。

我们现在说明使用正态分布近似复杂和的分布。

示例：总和的正态近似

想象一下，我们正在尝试传输信号。在传输过程中，有 100 个源独立地产生低噪声。每个源产生一定量的噪声，均匀分布在 $a = -1$ 和 $b = 1$ 之间。如果噪声总量大于 10 或小于 -10，那么它会破坏信号。但是，如果噪声总量的绝对值小于 10，那么这不是问题。

问题：100 个信号的总噪声绝对值小于 10 的近似概率是多少？

答案：让 X_i 成为来自源 i 的噪声。观察到 $\mu_{X_i} = 0$ 。观察到 $\sigma_{X_i}^2 = (b-a)^2/12 = 1/3$ 和 $\sigma_{X_i} = 1/\sqrt{3}$ 。设 $S_{100} = X_1 + X_2 + \dots + X_{100}$ 。

$$\begin{aligned} P\{-10 < S_{100} < 10\} &= P\left\{\frac{-10}{\sqrt{100/3}} < \frac{S_{100} - 0}{\sqrt{100/3}} < \frac{10}{\sqrt{100/3}}\right\} \approx 2\Phi\left(\frac{10}{\sqrt{33.33}}\right) - 1 \\ &= 2(0.9572) - 1 = 0.9144 \end{aligned}$$

因此，信号被破坏的近似概率小于 10%。在实践中，这种近似值非常好。

3.15 随机变量的随机数的和

在许多应用程序中，人们经常需要添加一些独立同分布随机变量，其中这些变量的数量本身就是一个随机变量。具体来说，我们在谈论以下表达式中的数量 S 。设 X_1, X_2, X_3, \dots 是独立同分布随机变量。令

$$S = \sum_{i=1}^N X_i, \quad N \perp X_i$$

其中 N 是非负整数值随机变量。

我们现在回顾一下如何推导出像 $E[S]$ 和 $E[S^2]$ 这样的量，本书中将需要这些量。

问题：为什么我们不能直接应用期望线性性？

答案：线性方程仅适用于 N 为常数的情况。

问题：这会给你任何想法吗？

答案：让我们对 N 的值进行条件化，然后应用期望线性性。

$$\begin{aligned} E[S] &= E\left[\sum_{i=1}^N X_i\right] = \sum_n E\left[\sum_{i=1}^N X_i \mid N = n\right] \cdot P\{N = n\} = \sum_n E\left[\sum_{i=1}^n X_i\right] \cdot P\{N = n\} \\ &= \sum_n nE[X] \cdot P\{N = n\} = E[X] \cdot E[N] \end{aligned} \tag{3.3}$$

问题：我们可以使用相同的技巧获得 $E[S^2]$ 吗？

答案：对 N 进行条件化的困难在于我们最终得到了一个我们需要平方的大数目，并且如何做到这一点并不明显。考虑以下公式：

$$E[S^2] = \sum_n E[S^2 \mid N = n] \cdot P\{N = n\} = \sum_n E\left[\left(\sum_{i=1}^n X_i\right)^2\right] \cdot P\{N = n\}$$

一个更好的想法是先导出 $\text{Var}(S \mid N = n)$ ，然后用它来得到 $E[S^2 \mid N = n]$ 。按照定理 3.27，有

$$\text{Var}(S \mid N = n) = n\text{Var}(X)$$

也请注意

$$\begin{aligned} n\text{Var}(X) &= \text{Var}(S \mid N = n) = E[S^2 \mid N = n] - (E[S \mid N = n])^2 \\ &= E[S^2 \mid N = n] - (nE[X])^2 \end{aligned}$$

从前面的表达式，我们得到

$$E[S^2 | N = n] = n \text{Var}(X) + n^2 (E[X])^2$$

它遵循

$$\begin{aligned} E[S^2] &= \sum_n E[S^2 | N = n] \cdot P\{N = n\} = \sum_n (n \text{Var}(X) + n^2 (E[X])^2) P\{N = n\} \\ &= E[N] \text{Var}(X) + E[N^2] (E[X])^2 \end{aligned}$$

此外

$$\begin{aligned} \text{Var}(S) &= E[S^2] - (E[S])^2 = E[N] \text{Var}(X) + E[N^2] (E[X])^2 - (E[N] E[X])^2 \\ &= E[N] \text{Var}(X) + \text{Var}(N) (E[X])^2 \end{aligned}$$

我们已经证明了定理 3.34：

定理 3.34 设 X_1, X_2, X_3, \dots 是独立同分布随机变量。令

$$S = \sum_{i=1}^N X_i, \quad N \perp X_i$$

那么

$$\begin{aligned} E[S] &= E[N] E[X] \\ E[S^2] &= E[N] \text{Var}(X) + E[N^2] (E[X])^2 \\ \text{Var}(S) &= E[N] \text{Var}(X) + \text{Var}(N) (E[X])^2 \end{aligned}$$

方差技巧非常酷。你可能想知道我们如何获得三阶矩 $E[S^3]$ ，因为方差技巧在那里不起作用。答案是使用变换分析(生成函数)，这将很容易提供 S 的任何阶矩。本主题将在第 25 章中介绍。

3.16 习题

- 3.1 期望脑筋急转弯。一位朋友告诉我，在他上学的第一年，他从未参加过不到 90 名学生的课程。他说几乎所有的朋友都有同样的经历。然而，院长坚持认为平均新生班级规模为 30 名学生。怎么会这样？用一个简单的数字示例解释发生了什么。
- 3.2 书呆子 Ned。书呆子 Ned 每天都会邀请一位新认识的女孩约会。女孩同意的概率为 1/100，女孩不同意的概率为 99/100。Ned 找到女朋友需要超过 100 天的概率是多少？
- 3.3 方差。使用期望线性来证明 $\text{Var}(X) = E[X^2] - E[X]^2$ 。
- 3.4 用于条件的链式规则。设 E_1, E_2, \dots, E_n 为 n 个事件，每个事件具有正概率和非零的交集。证明

$$P\left(\bigcap_{i=1}^n E_i\right) = P\{E_1\} \cdot P\{E_2 | E_1\} \cdot P\{E_3 | E_1 \cap E_2\} \cdots P\left\{E_n | \bigcap_{i=1}^{n-1} E_i\right\}$$

- 3.5 评估风险。Queueville 航空公司知道预订航班的人中平均有 5% 不会登机。(他们通过假设每个人独立的没有登机的概率为 5% 来对这些信息进行建模。)因此，他们的政策是为可以容纳 50 名乘客的航班出售 52 张机票。每个出现的乘客都有可能获得座位的概率是多少？
- 3.6 有条件的实践。对于表 3-3 中的联合概率质量函数，计算 $E[X | Y \neq 1]$ 。
- 3.7 方差如何扩展。考虑以下两个随机变量：

$$X = \begin{cases} 3 & \text{概率为 } \frac{1}{3} \\ 2 & \text{概率为 } \frac{1}{3} \\ 1 & \text{概率为 } \frac{1}{3} \end{cases} \quad Y = \begin{cases} 30 & \text{概率为 } \frac{1}{3} \\ 20 & \text{概率为 } \frac{1}{3} \\ 10 & \text{概率为 } \frac{1}{3} \end{cases}$$

- (a) Y 是 X 的缩放版本。 X 和 Y 是否具有相同的方差?
(b) 直观地, 如果我们认为 X 表示以秒为单位的测量值, 而 Y 表示以十分之几秒为单位的测量值, 那么我们希望感觉 X 和 Y 具有相同的方差。计算机系统中的一个常见度量是平方变异系数, 其中 X 的平方变异系数写为 C_X^2 , 定义为 $C_X^2 = (\text{Var}(X))/(E[X]^2)$ 。这可以视为标准化方差。 C_X^2 和 C_Y^2 如何比较?

3.8 了解方差和风险。设 c 为整数, 其中 $c > 1$ 。我们给出了随机变量 X 的 c 个独立实例: 称这些为 X_1, X_2, \dots, X_c 。

- (a) 哪个方差较小: $\text{Var}(X_1 + X_2 + \dots + X_c)$ 还是 $\text{Var}(cX)$? 计算每一个。
(b) 共同基金的卖点是它们比购买单一股票风险更小。解释这个说法。

3.9 恒等式。设 A 和 B 是独立的随机变量。证明或证伪以下断言:

$$E[A/B] = E[A]/E[B]$$

3.10 乘积的期望。证明或证伪以下断言: 如果

$$E[XY] = E[X] \cdot E[Y]$$

那么 X 和 Y 是独立的随机变量

3.11 Binomial 的方差。设 $X \sim \text{Binomial}(n, p)$ 。使用定理 3.27 轻松导出 $\text{Var}(X)$ 。
3.12 Binomial 的泊松近似。证明当 n 很大且 p 很小时, $\text{Binomial}(n, p)$ 分布能很好地用 $\text{Poisson}(np)$ 分布近似。[提示: 从 $\text{Binomial}(n, p)$ 分布的概率质量函数开始。设置 $p = \lambda/n$ 。展开所有项。获取限制并显示你获得 $\text{Poisson}(\lambda)$ 分布, 其中 $\lambda = np$.]

3.13 概率界限。你被告知数据库中的平均文件大小为 6K。
(a) 解释为什么会有(从期望的定义)不到一半的文件可以具有 $> 12K$ 的大小。
(b) 现在, 你将获得最小文件大小为 3K 的附加信息。获得大小 $> 12K$ 的文件百分比的更严格上限。
3.14 服务质量。如果处理请求的时间超过 7 秒, 公司将支付罚款。处理请求包括两个任务: (a) 检索文件, 这需要一些指数分布期望为 5 的时间 X ; (b) 解析文件, 这需要一些时间 Y , Y 独立于 X 并且服从 $\text{Uniform}(1, 3)$ 分布, 期望为 2。鉴于处理请求的平均时间显然是 7 秒, 公司认为罚款是不公平的, 因为它必须对一半的请求支付罚款。这是正确的吗? 罚款必须支付的实际时间比例是多少? 与 $1/2$ 有多少不同?

3.15 正相关。我们说事件 A 和 B 是正相关的, 如果

$$P\{A|B\} > P\{A\} \quad (3.4)$$

证明或证伪: 式(3.4)可推出

$$P\{B|A\} > P\{B\} \quad (3.5)$$

3.16 协方差。由 $\text{cov}(X, Y)$ 表示的任意两个随机变量 X 和 Y 的协方差由下式定义:

$$\text{cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

- (a) 证明 $\text{cov}(X, Y) = E[XY] - E[X]E[Y]$ 。
(b) 如果 $X \perp Y$, 关于 $\text{cov}(X, Y)$ 我们能说些什么?
(c) 设 X 和 Y 为指标随机变量, 其中

$$X = \begin{cases} 1 & \text{如果事件 } A \text{ 发生} \\ 0 & \text{否则} \end{cases}$$
$$Y = \begin{cases} 1 & \text{如果事件 } B \text{ 发生} \\ 0 & \text{否则} \end{cases}$$

证明如果事件 A 和 B 正相关(见习题 3.15), 那么 $\text{cov}(X, Y) > 0$, 而如果 A 和 B 负相关, 那么 $\text{cov}(X, Y) < 0$ 。注意: 这个概念可以扩展到一般随机变量 X 和 Y , 而不仅仅是指标随机变量。

3.17 正态近似。比尔·盖茨邀请了 1000 位朋友参加晚宴。每个人都被要求做出贡献。贡献是独立同分布泊松分布随机变量, 期望为 1000 美元。比尔希望筹集到 100 万美元。
你的工作是计算比尔的筹集款 $< 999,000$ 美元的概率。

- (a) 使用本章的正态近似进行计算。
(b) 现在为这个概率写一个精确的表达式，然后使用计算器或小程序来评估表达式。
- 3.18 **联合分布。**你的助教，Eric 和 Timmy，已同意在下午 2 点到 3 点之间举行会议，以设计下一个家庭作业。他们相当繁忙，并且不太确定何时可以到达，因此假设他们每个人的到达时间是独立的并且在一小时内均匀分布。每个人同意为其他助教等待 15 分钟，之后他将离开。Eric 和 Timmy 能够见面的概率是多少？
- 3.19 **天气预报的贝叶斯推理。**为了举行一个干燥的户外婚礼，John 和 Mary 决定在沙漠中结婚，那里平均每年下雨天数是 10 天。不幸的是，天气预报员预报明天(即 John 和 Mary 的婚礼当天)下雨。假设天气预报员不完全准确：如果第二天下雨，预报员会有 90% 的概率预测下雨。如果第二天干燥，预报员仍然有 10% 的概率会(错误地)预测下雨。鉴于这些信息，在 John 和 Mary 的婚礼期间下雨的概率是多少？
- 3.20 **贝叶斯卫生保健测试推理。**一家制药公司开发了一种针对 H1N1 流感病毒的潜在疫苗。在对疫苗进行任何测试之前，开发人员假设，他们的疫苗有 0.5 的概率是有效的，有 0.5 的概率为无效。开发人员对疫苗进行了初步的实验室测试。该初始实验室测试仅部分指示疫苗的有效性，准确度为 0.6。具体而言，如果疫苗有效，则该实验室测试将以 0.6 的概率返回“成功”，而如果疫苗无效，则该实验室测试将以 0.6 的概率返回“失败”。
(a) 实验室测试返回“成功”的概率是多少？
(b) 鉴于实验室检测结果“成功”，疫苗有效的概率是多少？
(c) 开发人员决定增加第二个实验(这个实验在人类身上)，它比初始实验室测试更具指示性，准确度为 0.8。具体而言，如果疫苗有效，那么人类测试将以 0.8 的概率返回“成功”。如果疫苗无效，那么人类测试将以 0.8 的概率返回“失败”。鉴于实验室测试和人体测试都取得了“成功”，疫苗有效的概率是多少？添加这个额外的测试有什么用？假设两种测试(人体测试和实验室测试)在条件上独立于疫苗有效或无效。
- 3.21 **约会成本：通过条件化获得期望和方差。**一个男人，为了寻找妻子，尝试了两种方法：慷慨和吝啬。当这个男人尝试慷慨的方法时，他最终会在他的约会对象身上花费 1000 美元，这个人最终与他分手的概率为 0.95，但与他结婚的概率为 0.05。当这个男人尝试吝啬的方法时，他会在约会对象身上花费 50 美元，其最终会和他分手。到目前为止，在他的生命中，这个人只经历过失败，所以他无法分辨哪种方法更好。因此，他决定随意选择一种方法(慷慨或吝啬)。
(a) 假设这个男人今天开始寻找，他找到妻子的预期费用是多少？
(b) 计算男人最终花在找妻子上的钱的方差。
- 3.22 **几何分布的方差。**设 $X \sim \text{Geometric}(p)$ 。证明 $\text{Var}(X) = (1-p)/p^2$ 。[提示：使用条件化。]
- 3.23 **芯片的良品与次品。**一家芯片供应商生产 95% 的良品和 5% 的次品。良品每天失败的概率为 0.0001。次品每天失败概率为 0.01。你买随机芯片。设 T 是芯片出现故障的时间。计算 $E[T]$ 和 $\text{Var}(T)$ 。
- 3.24 **期望的替代定义[⊖]。**(a) 设 X 为非负的、离散的整数值随机变量。证明
- $$E[X] = \sum_{x=0}^{\infty} P\{X > x\}$$
- (b) 设 X 为非负的、连续的随机变量。证明
- $$E[X] = \int_{x=0}^{\infty} P\{X > x\} dx$$
- (c) 设 X 为非负的连续随机变量。这个数量有更好的名字吗？
- $$\int_{x=0}^{\infty} xP\{X > x\} dx$$
- 3.25 **基于条件化的期望。**Stacy 的容错系统只有在连续 $k=10$ 次失败时才会崩溃。如果每分钟发生一次

[⊖] 提示：本习题的结果将在整本书中调用。

故障的概率为 $p=1/10$ ，那么在 Stacy 的系统崩溃之前的预期分钟数是多少(通常以 k 和 p 表示)?
[提示：写一个递归关系。]

- 3.26 Napster——由 RIAA 带给你。作为我兄弟的礼物，我决定创作一个他最喜欢的乐队的所有歌曲的集合。我需要下载乐队的 50 首歌曲。不幸的是，每当我输入乐队名称时，我都会收到乐队的随机歌曲。设 D 表示获得所有 50 首歌曲所需的下载次数。
- (a) $E[D]$ 是多少？给出一个封闭形式的近似值。
- (b) $\text{Var}(D)$ 是多少？(这里不需要封闭形式。)
- 3.27 分数矩。鉴于正态分布的丑陋，我很高兴地说它在我的研究中从未出现过……直到几天前！事情是这样的：我有一个随机变量 $X \sim \text{Exp}(1)$ ，我需要计算 $E[X^{\frac{1}{2}}]$ 。为什么我需要正态分布来完成计算？我最终得到了什么答案？以下是一些提示：首先应用分部积分，然后对变量进行正确的更改。如果你做得对，应该得到标准正态分布。请记住，指数范围从 0 到 ∞ ，而正常范围从 $-\infty$ 到 ∞ 。

