前　　言

Beauty is our business.
完美是我们的追求。

——Edsgar Wybe Dijkstra
埃德斯加 • 狄克斯特拉
（荷兰计算机科学家，1972 年图灵奖获得者）

背景

计算机和计算机网络是 20 世纪人类的伟大创造。当我们跨进 21 世纪的大门，蓦然回首，发现没有哪个发明像计算机和计算机网络这样对整个人类文明和社会进步发生了和发生着如此巨大、如此深刻的影响。计算机网络和计算机系统已经成为现代信息社会最重要的基础设施之一，其应用已遍布社会的各个领域，成为国家发展和社会进步的基本需求，是知识经济的基本载体和支撑环境。

性能评价（performance evaluation）是计算机网络和计算机系统研究与应用的重要理论基础和支撑技术，是通信和计算机科学领域的重要研究方向，也是一门理论与实践紧密相连、内容丰富、体系完整的学科。在国外，性能评价技术的研究和应用都十分广泛，《Performance Evaluation》就是一本专门介绍计算机网络和计算机系统性能评价的国际权威期刊，许多 IEEE 的会刊（transactions）也都有专门的 Performance Evaluation 专栏，还有许多关于性能评价的国际学术年会定期召开。另外，性能评价也是国外计算机、通信、信息科学等专业大学高年级学生和研究生的必修课程之一。而在国内，性能评价技术的教学与科研都亟待加强。

作者从事性能评价方面的研究工作已近 20 年，并在清华大学计算机科学与技术系讲授研究生课程“计算机网络和计算机系统的性能评价”，该课程以其理论高度和实践指导意义受到了清华研究生的欢迎。在讲课的过程中，作者对讲义进行了不断的充实与完善，本书就是在讲义的基础上写成的。书中取材新颖，并且包含了作者多年来在此领域的大量科学研究成果，是作者经验的总结。

性能评价方法

计算机网络和计算机系统性能评价的目的主要有三个：选择、改进和设计。具体而言，是指在众多的系统（方案）中选择一个最适合需要的系统（方案），即在一定的价格范围内选择性能最好的系统，达到较好的性能/价格比；对已有系统的性能缺陷和瓶颈进行改进和提高其运行效率；对未来设计的系统进行性能预测，在性能成本方面实现最佳设计或配置。

计算机网络和计算机系统的性能一般包括以下两个大的方面：一个方面是它的可靠性

或可利用性，亦即，计算机系统能正常工作的时间，其指标可以是能够持续工作的时间长度，如平均无故障时间，也可以是在一段时间内，能正常工作的时间所占的百分比。另一方面是它的处理能力或效率。这又可分为三类指标：一类指标是各种吞吐率，如系统在单位时间内能处理正常作业的个数。另一类指标是各种响应的时间，即从系统得到输入到给出输出之间的时间。再一类指标是各种利用率，即在给定的时间区间中，各种部件（包括硬设备和软系统）被使用的时间与整个时间之比。当然不同的系统对性能指标的描述有所不同，例如网络常用的性能评估指标为：信道传输速率、信道吞吐量和容量、信道利用率、传输延迟、响应时间和负载能力等。

计算机网络和计算机系统的性能取决于多种因素，最基本的因素在于系统的配置（即指系统构成所包括的各种软件、硬件的成分、数量、能力和系统结构、处理和调度策略等）和系统负载（即指工作负载和工作方式，例如交互方式、批处理方式等）。

性能评价的主要任务就是研究系统配置、系统负载、性能指标之间的相互关系。

性能评价的方法大致可以分为两类：

(1) 测量方法

通过一定的测量设备或一定的测量程序可以直接从计算机网络和计算机系统测得各项性能指标或与之密切相关的度量，然后由它们经过一些简单的运算求出相应的性能指标。这是最直接也是最基本的方法，其他方法在一定程度上也要依赖于它。但是这种方法只能适用于已经存在并运行的系统，而且比较费时间。测量方案和测量手段是测量方法的关键。

(2) 模型方法

首先对要评价的计算机网络或计算机系统建立一个适当的模型，然后求出模型的性能指标，以便对系统进行性能评价。模型中一般包括许多参数，这些参数的确定往往依赖于对实际系统的测量结果或对系统参数的估计。与测量方法相比，模型方法有两个优点：一是它不仅可以应用于已有的系统的性能评价，而且也可以应用于尚未存在系统的性能预测；二是它的工作量一般比测量方法要小，费用比测量手段的要少。

模型方法又可分为模拟方法和分析方法两种。模拟方法是用一个程序动态地模拟一个系统及其负载。一般首先使用一个模拟语言来为系统建立模型，然后在模拟时，通过用负载驱动系统模型从而得出模型的性能指标。模拟方法可以详细地刻画系统，得出较精确的性能指标，但是构造和使用模型时的费用较高。

分析方法则是应用数学理论与方法来研究和描述性能与系统、负载之间的关系。为了数学上描述与计算的方便，往往要对系统模型进行一些简化和假设，因而这种模型刻画系统的详细程度较低，得出的性能指标精度也较低。但是这种方法理论基础强，可以明显地刻画各种因素之间的关系，而且构造和使用模型时的费用也较低。

随着计算机技术的发展，系统的庞大和复杂化使得系统性能评价问题变得越来越复杂并日益引起人们的重视。提供有效的数学理论工具、直观的模型描述方法和有效的模型分析方法以及实用的辅助分析软件，是系统性能评价所面临的迫切需要解决的问题，这也正是本书所要介绍的性能评价技术的目标与核心。

内容安排

本书从结构上按照性能评价的模型技术与方法分为三个部分：

第一部分是排队论模型与性能评价。

用排队论数学理论来解决系统的描述问题是传统的分析方法，其数学求解的基础是马尔可夫随机过程。

第一章和第二章给出了阅读本书所需要的概率论和随机过程的一些基本概念和知识。第三章和第四章分别介绍了排队模型和排队网络模型，二者是性能评价的基础理论和经典模型。第五章介绍了非乘积解排队网络和近似分析方法，近似模型使得分析模型在数学上和计算上更加容易处理。第六章介绍了自相似传输模型，这是国际上网络研究的最新进展，仍有大量的问题有待解决，是计算机网络性能评价所面临的一个重要挑战。

第二部分是随机 Petri 网与性能评价。

20 世纪 80 年代初，随机 Petri 网的提出为系统的性能评价又提供了一个新的数学描述工具。系统性能评价方法，尤其是排队论分析方法的发展和所遇到的问题，包括并行系统的资源共享描述和非乘积解的问题，给 Petri 网的应用领域的拓宽和发展带来了勃勃生机。随机 Petri 网的提出，以及后来各种随机网的不断发展，使系统性能评价成 Petri 网最成功的应用领域之一。

作者从事不同级别 Petri 网和各种随机 Petri 网的研究近 20 年，并在国际上首先提出了随机高级 Petri 网(stochastic high-level Petri net, SHLPN)及其分析技术，以及在性能评价中的应用。这部分的主要内容是作者近 20 年研究工作的总结，是作者对当前随机 Petri 网理论、技术和应用的研究经验的综合。这部分的目标是使读者能够基本掌握随机 Petri 网的理论、模型方法、分析技术和应用思路，为读者的系统性能评价学习、工作和研究课题提供一条有效途径。

第七章至第九章介绍了各种随机 Petri 网理论和分析技术，包括：随机 Petri 网(SPN)、广义随机 Petri 网(GSPN)、随机回报网(SRN)、随机高级 Petri 网(SHLPN)和确定与随机 Petri 网(DSPN)。着重描述了可达集、可达分析和稳定状态概率的算法。介绍了系统的性能特性分析和算法。这几章是随机 Petri 网的基础知识，是后续章节的基础。

第十章中讨论了随机 Petri 网的模型方法与模型的分解和压缩技术。随机 Petri 网模型性能评价的一个主要问题是模型状态空间的爆炸，状态的数量会随着模型的规模和复杂性的增加而指数性地增长，使实际系统的性能评价不可能。这一点会严重地阻碍 SPN 模型的实际应用，也是当前 SPN 研究的热点问题。这部分所介绍的技术可以应用到各种复杂、大型系统的性能模型和性能评价中，这些技术是掌握 SPN 分析方法的关键，也是深入进行 SPN 研究的基础。

第十一章着重介绍了随机 Petri 网在计算机网络和计算机系统性能评价中的应用，这部分精选了作者近年来在计算机网络和计算机系统性能评价领域的最新研究成果，包括通信网络协议的性能评价、计算机和软件系统的模型与性能评价、ATM 网络的性能模型与性能评价、多服务器多队列系统的模型与性能评价等。详细地介绍了应用随机 Petri 网对各种计算机和网络系统的对象描述方法和性能模型方法，深入讨论了系统模型的有效的分析

和化简技术,以及问题的求解方法,系统性能参数(例如系统的吐吞量、资源的利用率和用户平均响应时间等)的实际计算和分析。

第三部分是模拟技术与性能评价。

模拟技术也是系统性能评价的一种重要方法。模拟是通过在数字计算机上运行系统的模型,并且分析运行的输出结果,从而获得真实系统的某些性能预测。模拟技术通常是在数学分析模型不可求解或难于求解的情况下,或者其他分析方法的结果(由于假设条件的影响)与实际情况相差比较大时才使用,是一种最后的方法。

第十二章对模拟模型技术进行了概述。介绍了模拟模型技术的基本概念、模型方法以及模拟的过程与模型的建立。第十三章介绍了模拟程序软件,包括直接使用编程语言的模拟方法和 smpl 模拟软件的使用方法,以及模拟的输出分析。其中介绍了实际模型例子的分析软件程序,读者很容易套用这些例子编写自己的模拟软件程序。另外还介绍了作者使用 smpl 对 ATM 网络进行性能评价的实例。

本书特点与读者对象

本书具有以下鲜明的特点:

(1) 将排队理论模型、随机 Petri 网模型技术、模拟技术有机地融于一本书中,内容丰富全面,结构合理,层次分明,体系完整。

(2) 内容新颖,具有一定的理论高度和学术价值。书中包括了计算机网络和计算机系统性能评价领域的最新研究进展,并且精选了作者在此领域多年来的大量科学研究成果。

(3) 容易理解,面向应用。书中的内容基本上是自包含的,由浅人深,介绍了充分的必备知识。在介绍数学概念时强调直观性和物理背景,注意阐明定理和结论的意义和作用,强调方法和模型在实际计算机网络和计算机系统中的应用,而尽量避免繁杂的数学理论推导。对于每一种形式定义和推导,都给出模型例子进行引导。软件工具的介绍也方便了读者对实际问题的模型分析与应用。

本书十分适合我国系统性能评价领域的教学、科研工作和工程应用参考,既可以供计算机、通信、电子、自动化、信息等相关专业的教师、研究生和大学高年级学生作教材或教学参考书,也可以供专业科学研究人员和工程技术人员使用。

致谢

国家自然科学基金委员会(编号:69873012 等)和国家重点基础研究发展规划项目(编号:G1999032707)对作者的研究工作给予的连续资助,在此表示深深的谢意。

我的学生,北京科技大学信息工程学院的单志广博士在本书的写作过程中做了大量细致辛苦的编写工作,对此表示衷心的感谢!

本书写作期间,我的家人和朋友给了我极大鼓舞和帮助,作者谨以此书献给我的家人和朋友。

林　闯

2001 年 1 月

北京清华园

目　　录

第一部分　排队论模型与性能评价

第三部分 模拟技术与性能评价

第一部分

排队论模型与性能评价

第一章 概率论基础

在开始正式讨论计算机网络和计算机系统的性能评价的模型和方法之前，我们首先给出最必要的概率论理论和随机过程的一些基本概念和知识，这些数学基础对于后续章节的分析讨论是很重要的，可供读者阅读和学习时参考。

1.1 概率的定义

概率关系着对事件的数量分配。一个事件 A 的概率 $P(A)$ 是对应事件 A 要发生可能性的数量分配。通常，我们考虑执行一个试验并获得一个结果。事件 A 是一个或一组特定的结果，且将某一概率分配给这个事件。

一般很难确切掌握概率的概念。不同的应用理论有着不同的概率表现方法。事实上，存在许多不同的概率定义。这里我们给出三种定义。

1. 公理化定义

一个概率论的形式化方法是从一定数量的定义概率度量的公理出发，经过推导规则达到概率的有效计算。公理是必须被接收的简单断言。一旦公理被接收，每一条规则就可能证明。

在公理和规则中要使用下列来自集合论的概念。必然事件 Ω 是在每一次试验中都发生的事件，它包括所有可能结果的全集或称"样本空间"。两个事件 A 和 B 的并集 $A\cup B$ 是当 A 或者 B 或者两者都发生时发生的事件。交集 $A\cap B$，也可写做 AB，是当事件 A 和 B 都发生时发生的事件。如果一个事件的发生不包括另一个事件的发生，那么这两个事件是互斥的。事件 $\overline{A}$ 是当 A 不发生时发生的事件。这些概念借助温尼(Venn)图更容易理解，如图 1.1.1 所示。在每一个图中，阴影部分对应图下面的表达式。图 1.1.1(c)和(d)对应着 A 和 B 不是互斥的情况，即某些结果是作为事件 A 和 B 两者的一部分而被定义的。图 1.1.1(e)和(f)对应着 A 和 B 是互斥的情况，在这种情况下 A 和 B 的交是空集。

通常用来定义概率的公理集合，包括如下：

(1) 对于每一个事件 A，有 $0\leqslant P(A)\leqslant 1$。

(2) $P(\Omega)=1$。

(3) 如果 A 和 B 是互斥的，则 $P(A\cup B)=P(A)+P(B)$。

公理 3 可以扩展到许多事件。例如，如果 A，B 和 C 是互斥的，$P(A\cup B\cup C)=P(A)+P(B)+P(C)$。应当注意到，这条公理并没有规定单个结果或事件的概率应该如何分配。基于上述公理可以推导出许多规则。下面是一些重要的规则：

$$P(\overline{A}) = 1 - P(A)$$

如果 A 和 B 是互斥的，则 $P(A\cap B)=0$

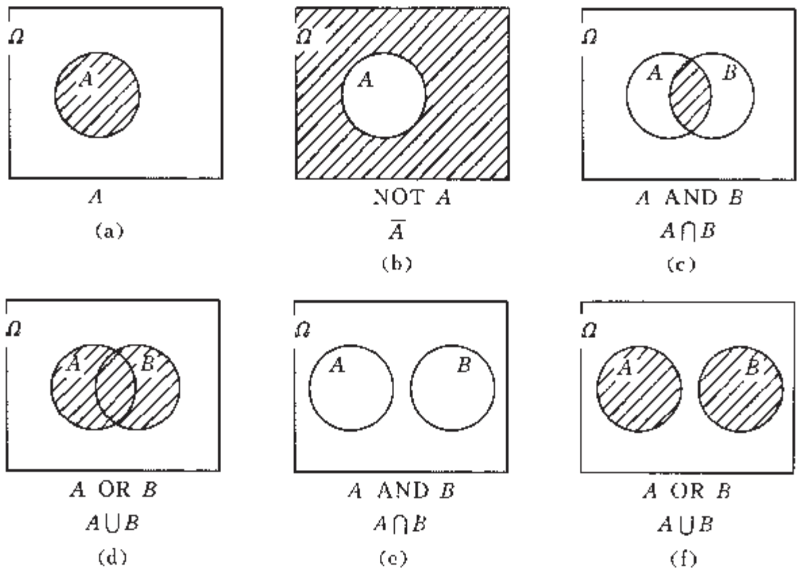

图 1.1.1　温尼(Venn)图

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$\begin{aligned} P(A \cup B \cup C) =& P(A) + P(B) + P(C) - P(A \cap B) \\ & - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \end{aligned}$$

例 1.1.1　抛骰子

考虑抛一个骰子。存在六种可能的结果。其中必然事件是六面骰子的任何一面在上面总要发生。事件{偶数}和{小于 3}的并是事件{1 或 2 或 4 或 6},这两个事件的交是事件{2}。事件{偶数}和{奇数}是互斥的。如果我们假定六种结果的每一个都是同样可能的,并且分配给每一个结果的概率为 1/6,很容易看到上述三个公理都被满足了。我们能应用概率规则获得如下结果:

$P\{偶数\} = P(2) + P(4) + P(6) = 1/2$

$P\{小于 3\} = P(1) + P(3) = 1/3$

$P\{\{偶数\} \cup \{小于 3\}\} = P\{偶数\} + P\{小于 3\} - P(2) = 1/2 + 1/3 - 1/6 = 2/3$ □

2. 相对频率定义

相对频率方式使用下列的概率定义。将一个实验做许多次,每一次称为一个试验。对于每一个试验,观察事件 A 是否发生。事件 A 的概率 P(A)定义为如下极限:

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n} \tag{1.1.1}$$

其中 n 是试验的次数,n_A 是 A 发生的次数。

例 1.1.2　投硬币

一个人多次投掷硬币。在经过很多次投掷后,硬币正面出现的次数与总共投掷的次数的比率在 0.5 左右,其正面和反面出现的概率相同。 □

3. 古典定义

概率的古典定义如下：

$$P(A)=\frac{N_A}{N} \tag{1.1.2}$$

其中 N 是可能结果的总个数，假设全部结果出现的可能性相同，N_A 是事件 A 在其中发生的结果的个数。

例 1.1.3 抛两个骰子

我们抛两个骰子并且确定总和为 7 的概率 p。你可能会想到有多个不同的总和，包括 $(2, 3, \cdots, 12)$，共 11 个，于是得出概率为 1/11。但这是不正确的。我们需要考虑等价可能的结果。为此目的，我们必须考虑到每一个骰子面的组合，并且要区分第一个骰子和第二个骰子。例如，结果(3, 4)和(4, 3)必须相区别。使用这种方法，总共有 36 种相同可能的结果，而且有 6 种结果(1, 6)，(2, 5)，(3, 4)，(4, 3)，(5, 2)和(6, 1)是我们所要求的。因此，$p=6/36=1/6$。 □

1.2 条件概率和独立性

我们经常希望知道在某些事件发生的条件下的概率。条件的影响是删除样本空间的一些结果。

形式化地说，假定事件 B 已经发生时事件 A 发生的条件概率 $P(A|B)$ 可以定义为如下比率式：

$$P(A \mid B)=\frac{P(AB)}{P(B)} \tag{1.2.1}$$

这里我们假定 $P(B)$ 不为零。

如果 $P(AB)=P(A)P(B)$，事件 A 和 B 叫做相互独立的事件。进一步，如果事件 A 和 B 是相互独立的，我们就有 $P(A|B)=P(A)$ 和 $P(B|A)=P(B)$。

例 1.2.1 抛两个骰子

在抛两个骰子时，如果我们知道至少一个骰于的面是偶数，得到总和为 8 的概率是多少？

解法 1：

我们可按如下方法推理：由于一个骰子是偶数且和为 8，另一个骰子必定是偶数。因此，有三个相同可能出现的合适结果——(2, 6)，(4, 4)和(6, 2)，而我们要考虑的总结果数为{36－(两个骰子都是奇数)}＝36－3×3＝27。最后的概率是 3/27＝1/9。

解法 2：

在本例中，A＝{和为 8}，B＝{至少一个骰子是偶数}。$P(AB)$ 是包含所有总和为 8 并且至少一个骰子是偶数的概率。我们已经知道有 3 个这样的结果。因此 $P(AB)=3/36=1/12$。同时可以知道 $P(B)=3/4$。使用上式，可得：

$$P(A \mid B)=\frac{P(AB)}{P(B)}=\frac{1/12}{3/4}=\frac{1}{9}$$

这个结果同解法 1 的推导结果相同。 □

1.3 贝叶斯定理

贝叶斯(Bayes)定理是概率论中最重要的结论之一。首先我们必须说明全概率公式。给定一组互斥事件 $E_1, E_2, \cdots, E_n$,这些事件的并集包括所有可能的结果,同时给定一个任意事件 A,那么全概率公式可以表示为:

$$P(A) = \sum_{i=1}^{n} P(A \mid E_i)P(E_i) \tag{1.3.1}$$

贝叶斯定理可以陈述如下:

$$P(E_i \mid A) = \frac{P(A \mid E_i)P(E_i)}{P(A)} = \frac{P(A \mid E_i)P(E_i)}{\sum\limits_{i=1}^{n}(A \mid E_i)P(E_i)} \tag{1.3.2}$$

贝叶斯定理在实际中用于解决下面类型的问题。假设 $E_1, E_2, \cdots$ 构成样本空间的一个划分,并且假设已知它们的概率 $P(E_i), i=1, 2, \cdots$。假设事件 $E_1, E_2, \cdots$ 在随机试验中不能或没有直接观察到,只能观察到与这些事件直接相联系的某事件 A。事件 A 必与事件 $E_1, E_2, \cdots$ 之一相伴随而出现,并且已知条件概率 $P(A\ E_i)$。假设现在在试验中出现了事件 A,要求在此条件下,对事件 $E_1, E_2, \cdots$ 出现的可能性作出判断,即求出它们关于 A 的条件概率 $P(E_i|A), i=1, 2, \cdots$。通常 $P(E_i)$ 可由以往的数据分析得到,称为 E_i 的"先验概率",而 $P(E_i|A)$ 是在得到有利于 E_i 的证据之后该事件真正发生的概率,称之为"后验概率"。所以贝叶斯公式又叫做"后验概率公式"。

例 1.3.1 假定我们在噪音很大的传输线路上正在传输一组 0 和 1 的信号。$S0$ 和 $S1$ 分别代表在一个给定的时间一个 0 和一个 1 被传送的事件,而 $R0$ 和 $R1$ 分别是在给定时间一个 0 和一个 1 被接收的事件。假定我们知道传输源的发送概率,$P(S1)=p, P(S0)=1-p$。现在观察这条传输线,当一个 0 和一个 1 被传送时,由 $P(R0|S1)=p_a$ 和 $P(R1|S0)=p_b$ 决定发生错误的频率。如果收到一个 0,我们可以计算发生错误的条件概率,即接收到一个 0 而发送的是一个 1 的条件概率,使用贝叶斯定理有:

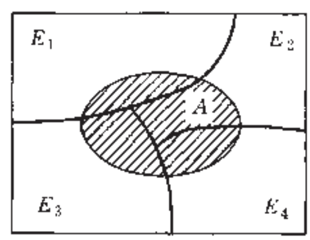

图 1.3.1 全概率和贝叶斯定理的图形表示

$$P(S1 \mid R0) = \frac{P(R0 \mid S1)P(S1)}{P(R0 \mid S0)P(S0) + P(R0 \mid S1)P(S1)}$$

$$= \frac{p_a p}{p_a p + (1-p_b)(1-p)}$$

全概率和贝叶斯定理的图形表示见图 1.3.1。 □

小 结

· 对于真实世界中许多系统的建模,特别是通信网络和计算机体系结构,概率的描述

和推导是非常重要的。

- 概率理论有自然、直观的解释和简单的数学公理。
- 相关的事件可由条件概率描述。
- 全概率公式允许对问题进行分解表达,这是随机模型的一个极为重要的工具。
- 相互独立事件交集的概率是每个事件概率的乘积。

参考文献

1. Stallings W. High-speed networks: TCP/IP and ATM design principles. Prentice-Hall International, Inc., 1988.
2. Harrison P G, Patel N M. Performance modelling of communication networks and computer architectures. Addison-Wesley, 1993
3. Trivedi K S. Probability & statistics with reliability, queuing, and computer science applications. Prentice-Hall, 1982
4. Hock N C. Queueing modelling fundamentals. John Wiley & Sons, 1997

习　题

1.1 抛两个骰子,令 X、Y 和 Z 分别是第 1 个骰子显示的点数、第 2 个骰子显示的点数和两个骰子显示点数之和。从事件概率 $P(X\leqslant 1, Z\leqslant 2)$ 和 $P(X\leqslant 1)P(Z\leqslant 2)$ 说明事件 X 和 Z 不是相互独立的。

1.2 对子在一个样本空间 Ω 的 n 个事件 $E_1, E_2, \cdots, E_n$,证明:

$$P(E_1\cup\cdots\cup E_n)=\sum_{i=1}^{n}P(E_i)-\sum_{i_1<i_2}P(E_{i_1}E_{i_2})+\cdots+(-1)^{r+1}\sum_{i_1<\cdots<i_r}P(E_{i_1}\cdots E_{i_r})$$
$$+\cdots+(-1)^{n+1}P(E_{i_1}\cdots E_{i_n})$$

1.3 假定每一个箱子能容纳 N 个球,有多少种方法将 N 个球放进 M 个箱子中?

1.4 (空盒问题)有 N 个盒子和 n 个质点。现在把 n 个质点随机地分配到 N 个盒子中去,假设每个质点落入各盒的概率相等。试求事件 $A_m=\{$恰好有 m 个空盒$\}$的概率。

1.5 (贝特朗[Bertrand]奇论)在半径为 r 的圆(C)中任意引一条弦,试求弦长大于圆内接正三角形之边长 $r\sqrt{3}$ 的概率 p。

1.6 在一组人群中,占比例 p 的人有一种疾病。开发了一种测试方法,使用这种方法进行测试,一个人有这种疾病而获得正结果的概率为 v,一个人没有这种疾病而获得负结果的概率为 c。显然如果 v 和 c 的值都接近 1,则这种测试方法就是好方法。如果一个人使用这种方法进行测试的结果为正,那么这个人有这种疾病的概率是多少?

1.7 有这样一个游戏,在三扇关闭的门中,奖品藏在其中的一个后面,其余的门后面都是假的奖品。参加游戏的观众从三扇门中选择一个,并站在这个门的前面。主持人(知道奖品藏在哪里)打开其他两扇门中的一个,向该观众显示一个假的奖品,并且问他是否想改变其最初的选择。问:为了使选中奖品的概率达到最大,这位观众应该坚持最初

的选择，还是选择另外一扇门，还是这两者根本没有区别？

1.8 （仆里耶[Polya]坛子）假设一开始坛子中盛有 N 个球，其中 N_1 个白球和 N_2 个黑球（$N_1+N_2=N$）。现在从坛中接连随机地取 n 次球，每次取出一个，记下它的颜色，并且在下次取球之前把该球连同另外 r 个和它同样颜色的球一起放入坛中。试求在 n 次取球过程中，白球恰好出现 m 次的概率（仆里耶坛子模型在医学上常用来模拟疾病的传染）。

第二章　随机过程概述

2.1　随机变量

随机变量是一个试验所有可能的输出事件集合(样本空间 Ω)到实数的一个映射,即一个随机变量将每一个事件与一个实数联系起来,给每一个事件结果分配一个数值。随机变量映射输出事件到实数是给事件一个数值解释。

随机变量 R 在实数子集合 X 中的概率定义如下:

$$P\{R \in X\} = P(A)$$

其中,$R(a)\in X$,当且仅当 $a\in A$ 并且 A 是一个事件。上式的左端 $R\in X$ 是 $R(w)\in X$ 的缩写,其中 $w\in\Omega$。

例 2.1.1　独立地抛两枚硬币,随机变量 X 是硬币正面在上的次数之和,因此 $X\in\{0, 1, 2\}$。硬币正面在上用 h 表示,硬币反面在上用 t 表示。概率表达如下:

$$P\{X = 0\} = P(\{(t,t)\}) = 1/4$$
$$P\{X = 1\} = P(\{(h,t),(t,h)\}) = 1/2$$
$$P\{X = 2\} = P(\{(h,h)\}) = 1/4$$

□

注意对应 $X=0,1,2$ 的事件是互斥事件,所以概率分配的和为 1。

如果随机变量取值的集合 X 是有限的或可列无限个数的不同值,则随机变量就是离散的;如果随机变量取值的集合 X 是不可数的无穷个数的不同值,则随机变量就是连续的。

2.1.1　分布和密度函数

一个连续的随机变量 X 可以使用它的分布函数 $F(x)$ 或者它的密度函数 $f(x)$ 描述:

分布函数:

$$F(x) = P\{X \leqslant x\} \tag{2.1.1}$$

特别地,$F(-\infty)=0$, $F(\infty)=1$。

密度函数:

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x}F(x) \tag{2.1.2}$$

$$F(x) = \int_{-\infty}^{x} f(y)\mathrm{d}y \tag{2.1.3}$$

其中,$\int_{-\infty}^{\infty} f(y)\mathrm{d}y = 1$。

对于离散型随机变量,它的概率分布可描述如下:

$$P_X(k) = P\{X = k\} \tag{2.1.4}$$

并且 $\sum_{\text{all } k} P_X(k) = 1$。

我们经常关心一个随机变量的某些特性,而不是它的整体分布。数学期望就是一个重要的特性:

连续情况

$$E[X] = \mu_X = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x \tag{2.1.5}$$

离散情况

$$E[X] = \mu_X = \sum_{\text{all } k} k P\{x = k\} \tag{2.1.6}$$

随机变量其他有用的特性测量:

二阶矩:

连续情况

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x) \mathrm{d}x \tag{2.1.7}$$

离散情况

$$E[X^2] = \sum_{\text{all } k} k^2 P\{x = k\} \tag{2.1.8}$$

方差:

$$\mathrm{Var}[X] = E[(X - \mu_X)^2] = E[X^2] - \mu_X^2 \tag{2.1.9}$$

标准差(均方差):

$$\sigma_X = \sqrt{\mathrm{Var}[X]} \tag{2.1.10}$$

方差和标准差是用来度量随机变量 X 与其均值 $E(X)$ 的偏离程度的。对于一个常数 a,我们有:

$$E[aX] = aE[X] \tag{2.1.11}$$

$$\mathrm{Var}[aX] = a^2 \mathrm{Var}[X] \tag{2.1.12}$$

数学期望称为一阶统计特征,二阶矩和方差是二阶统计特征。更高阶统计特征可从概率密度函数推导出来。

对于任意两个随机变量 X 和 Y,我们有:

$$E[X + Y] = E[X] + E[Y] \tag{2.1.13}$$

2.1.2 多维随机变量

对于多于两个的多维随机变量,我们经常关心一个变量的变化是否会影响到其他变量。我们要定义一些重要的相关程度的测量。一般多维随机变量的统计特征要求定义它们的联合概率密度函数或联合概率分布函数:

分布函数:

$$F(x_1, x_2, \cdots, x_n) = P\{X_1 \leqslant x_1, X_2 \leqslant x_2, \cdots, X_n \leqslant x_n\} \tag{2.1.14}$$

密度函数: