数量经济学系列丛书

R语言统计学基础

吕小康 编著

清华大学出版社

数量经济学系列丛书

R语言统计学基础

吕小康 编著

清华大学出版社
北京

内 容 简 介

本书借鉴西方主流统计教材的模式，图例丰富，讲解清晰，使用实际数据进行统计分析，尤其注重对统计思维和软件技能的培养，是基于开源软件的新一代概率统计教材．本书可供研究型大学的经济学、社会学、心理学、政治学、管理学、教育学、医学、药学、生物学等专业作为本科阶段的统计入门教材及软件操作教程，也可供相关专业高年级本科生或研究生作为普通统计学教材之外的辅导教材，同时还可作为一本数据分析与R语言操作的入门教程．

图书在版编目（CIP）数据

R语言统计学基础/吕小康编著．—北京：清华大学出版社，2017
（数量经济学系列丛书）
ISBN 978-7-302-45592-9

Ⅰ．①R…　Ⅱ．①吕…　Ⅲ．①统计分析-统计程序　Ⅳ．①C819

中国版本图书馆CIP数据核字(2016)第283895号

责任编辑：张　伟
封面设计：常雪影
责任校对：王荣静
责任印制：杨　艳

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课 件 下 载：http://www.tup.com.cn，010-62770175-4506
印 装 者：三河市中晟雅豪印务有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　印　张：22　插　页：1　字　数：546千字
版　　次：2017年2月第1版　　印　次：2017年2月第1次印刷
印　　数：1～3000
定　　价：49.00元

产品编号：072559-01

前言 PREFACE

与传统的介绍相比，我更想说的是：这是一本有思想、有技术、说“人话”的新一代概率统计与数据分析入门教材．我希望读者在阅读此书之后，能够明白统计方法并不简单的只是一种硬邦邦冷、冰冰的“客观方法”，而是一种严谨且有弹性的思维方式；学习统计方法的过程是一种处处充满惊喜的智力探索过程：通过严格而系统的训练，逐一打开统计方法的黑箱．

为此，本书努力在以下几个方面体现自身特色，以充分拓展学生的数据想象力与分析力．① 贯彻统计思想重于统计计算的基本教学理念．本书的主要目的并不是培养专业统计研究人才，而是培养学生理性而健全的统计思维模式，以及使用基本统计方法解决本学科领域的实际问题的能力．② 以实证数据处理为中心阐述基本统计内容．本书的主要内容完全针对行为与社会科学中的实际研究情境设计，例子和习题同时具有可读性和知识性，注重从一手文献、大型社会调查中提取相关数据作为训练数据．③ 补充国内同类教材目前尚较为少见的重要内容．这主要包括抽样分布产生、实验数据的随机化检验、自助分布置信区间、效应值与统计功效等内容．④ 重视统计数据和统计结果的可视化呈现．全书利用 R 语言绘制了 100 多个统计图形，旨在培养学生的图形思维能力．⑤ 强调统计结果的合理表达，使普通读者能够更好地理解统计公式与计算结果在现实世界中的含义．

本书的第 1 章和第 2 章是传统概率论的内容，此部分内容需要一元函数微积分学的基础．第 3 章开始介绍统计学的内容．一般而言，统计学可分为两大块：描述性统计 (descriptive statistics) 和推论性统计 (inferential statistics)．描述性统计是有关数据采集、组织和呈现的统计学分支，主要涉及统计数字记录和归总、统计指标建立、统计图表制作等内容，其重点在于两个方面：① 数据的数字特征的概括，也就是集中趋势与离散趋势的概括；② 统计图表的制作与理解．本书对各种统计指标背后的构造思想进行深入的剖析，并结合 R 软件说明其应用．

推论性统计学主要涉及如何从样本数据推论到总体数据的工作．通常而言，我们不可能针对研究对象的全体即总体做研究，而只可能根据总体的某个子集即样本做研究，并且希望将根据样本得到的信息，来归纳和推论总体的信息．本书的所有推论统计观点都基于频率学派的研究，这涉及第 4~7 章．其中，第 4 章讨论的抽样分布是推论统计学习的重点和难点，是社会科学研究中反事实框架的一个具体形式．第 5 章和第 6 章分别介绍参数估计和假设检验的内容．同时，本书还介绍了随机化实验中常用的推论框架：随机化分布，并介绍了最

近几十年发展迅速的自助法置信区间及 R 语言实现, 以拓展学生的统计视野与软件技能. 第 7 章主要介绍线性模型的基础内容, 主要包括线性回归和方差分析两大部分. 限于篇幅与自身学识, 本书并未涉及频率学派统计学之外的贝叶斯统计学的基本观点.

本书文字内容基于 Texlive 2015 平台写作而成, 统计分析和图形绘制基于 Rstudio 平台完成. 本书并不刻意回避英文, 涉及的概率统计人物均不做翻译, 直接以英语出现. 关键术语均注明英文原文, 以便检索. 例题和练习中的变量名称也多用英文, 这是为了与 R 中的变量命名原则相匹配. 本书所涉及的所有数据可从以下网址下载:

http://pan.baidu.com/s/1c20ZuWK

为节省篇幅, 软件安装指南、部分 R 语言基础统计操作、推荐阅读书目、练习题详细答案等拓展性内容及书中未完全涉及的统计内容以 PDF 形式存放于清华大学出版社官方网站, 请读者自行下载或向本人来信索取.

本书多数章节的内容在出版之前已作为内部讲义在南开大学周恩来政府管理学院各专业试用. 由衷感谢各界本科生和研究生同学对本书内容与表述方式提出的改进建议. 尤其要感谢 (以下排名不分先后) 我的助教、博士生张慧娟和王丛, 我的硕士生曹松峰、贾婷, 2015 级南开大学应用心理学全体学术硕士, 以及我指导过的本科生付英涛、陈丹忆、李亚静、张光耀、柳婷、荣杨、彭芷晴、付鑫鹏、穆蔚琦、杨旋、刘奕男、孙超然、隋晓阳等同学. 他们协助我校订了讲义中的文字、公式、例题和习题, 同时还帮助我撰写了部分章节的 LateX 文档与 R 语言操作说明, 并共同设计了部分练习题. 在此特别要向这些热心好学的学生致以诚挚的谢意!

感谢南开大学社会心理学系及周恩来政府管理学院诸位师长和同事对我的宽容, 使我能够自由地探索和实践自己的教学思想. 感谢张阔副教授的信任, 使我得以全程尝试用 R 软件进行心理统计课程教学的机会. 还要感谢教材例题与习题中 “神出鬼没”、备受调侃的 “柴教授” 的原型, 我的同门师弟柴民权博士. 他虽已是兰州大学管理学院的教师, 仍不改逗萌本色, 为本书贡献了 “柴教授” 的著名绰号, 以其独特方式证明他的持久影响力.

撰写此书虽已尽全力, 成书在即仍旧诚惶诚恐. 既恐出现纰漏, 贻笑大方; 更恐误人子弟, 罪莫大焉. 相关建议或批评, 可直接发至本人邮箱 xkdog@126.com 交流探讨. 如需更多国内外教学资料、统计习题、R 语言代码和考试试题, 也可直接发信索取, 我可承诺做到知无不言、全面分享.

最后, 用我很喜欢的一句英文谚语作为结尾吧:

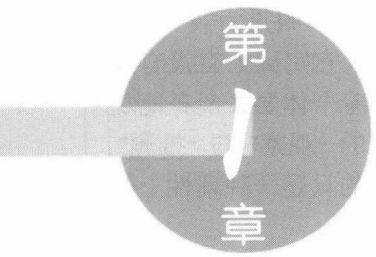
Throw your hat over the fence!(直译: 先把你的帽子扔过墙!)

这样你就有了翻墙而上的勇气.

吕小康

于南开大学津南校区

2016 年 8 月 31 日



概率基础

高中阶段似已掌握概率的最基本运算, 大学阶段的学习应如何在其基础上更深一步? 某种程度上讲, 大学阶段的学习是一种 "知其然且知之其所以然" 的过程, 既要通过一定的数学练习培养起严格审慎的概率思维方式, 又要深入理解各种数学定义和公式背后的基本思想与实用意图, 并了解数学工具的前提假定与应用局限, 从而避免概率思维的误解与滥用. 本章将先回顾中学阶段的一些基础知识和数学符号, 并通过对概率概念及其计算方式的进一步阐释, 引出条件概率、事件的独立性等基础的概率论概念.

1.1 基础知识回顾

这一部分将回顾高中阶段的集合论术语与概率计算技巧.

1.1.1 基本术语与符号表达

概率论可以说是一门研究随机现象之数学模型的学科. 所谓随机现象(random phenomenon), 就是在一定条件下并不总是出现相同结果的现象. 对随机现象进行观察、记录、实验的过程, 称为随机试验(random experiment), 而其中的每一次观测则称为 trial(由于中文缺少单复数形式, 故翻译仍为试验, 但一个 experiment 可包含若干次 trials).

从 "几何" 意义上讲, 某一随机现象的所有可能结果的集合, 称为样本空间(sample space), 用大写希腊字母 Ω 表示; 而每一个不可再分解的试验结果, 称为样本点(sample point), 用小写希腊字母 ω 表示, 通常会加上数字下标, 如 $\omega_1, \omega_2, \cdots, \omega_n$ 表示不同的样本点. 如此, 随机事件(random event, 简称事件) 可以定义为某些样本点的集合, 或样本空间的某个子集(subset). 每一个样本点对应一个基本事件. 样本空间的最大子集, 即 Ω 本身, 称为必然事件(sure event); 样本空间的最小子集, 即空集$\varnothing$(empty set), 称为不可能事件(impossible event).

实际使用中, 随机事件可能有不同的表达方式: 直接用语言描述, 同一事件可能有不同的描述; 也可以用样本空间子集的形式表示, 此时需要理解它所表达的实际含义. 同时应当注意, 这里的 "试验" 与科学中的试验或实验并不是一回事, 这里所称的事件与日常语言中的事件也不是一回事. 概率论中的 "事件" 与 "试验", 应当连在一起作为一对相互联系的概念进行理解. 日常用语中的 "事件", 通常是指已经发生的情况, 如 "非典" 事件、"9·11" 事件, 等等. 而概率论中的事件, 仅仅是关于某种状况的一种陈述, 它可能已经发生过, 也可能没有发生过; 可能发生, 也可能不发生; "发生" 与否, 需等待 "试验" 的结果才能确定. 概率

论中称 "两个事件 A 与 B 共同发生或同时发生", 并不是真的要求你能够 "眼见为实" 地看到它出现, 而只是在说: "A 与 B 存在同时出现的逻辑上的可能", 至于它实际上有没有发生过, 并不是关注的重点. 事件的产生总依赖于试验, 这也不一定意味着个体要去亲身地观察和实验, 而可以只是一种逻辑上的思考与想象, 可以仅是一种理论上的 "观察" 与 "推测". 也就是说, 试验虽然可能涉及真实的、科学意义上的观测过程, 但更多的只是一种理性上的思考过程而已.

直观上讲, 用来表示随机事件结果的变量称为*随机变量*(random variable), 常用大写字母, 如 X、Y、Z 表示. 这其实是将具体的现象抽象化和符号化的过程. 后面会用更加数学化的语言来重新定义随机变量, 但不妨先做这一简单理解. 事件之间的关系和运算有很多种, 这里仅列出最常见的几种及其符号表示 (表 1.1), 以便参阅.

表 1.1 概率论中的事件符号及其含义

符号表示	集合论意义	概率论意义
$A \subset B$	A 包含在 B 中	若 A 发生, 则 B 一定发生; 事件 A 蕴涵事件 B
$A = B$	A 与 B 相等	A 与 B 同时发生或同时不发生
$A \cap B$	交集 (intersection)	A 与 B 同时发生
$A \cup B$	并集 (union)	A 与 B 至少有一个发生
$A \cap B = \varnothing$	A 与 B 不相交 (disjoint)	A 与 B 互不相容 (互斥, mutually exclusive)
A^c 或 $\bar{A}$	A 的补集 (complement), $A + A^c = \Omega$	A 与 A^c 为对立事件
$A - B$	差集 (difference)	A 发生而 B 不发生

若样本空间 Ω 可划分为一系列两两互不相容的事件 $A_1, A_2, \cdots, A_n$, 且 $A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$, 即 $\bigcup_{i=1}^{n} A_i = \Omega$, 则称 $A_1, A_2, \cdots, A_n$ 为 Ω 的一个分割 (partition), 或称 $A_1, A_2, \cdots, A_n$ 是一个完备 (exhaustive) 事件组. 若干个事件的交集 $A_1 \cap A_2 \cap \cdots \cap A_n$ 则可记为 $\bigcap_{i=1}^{n} A_i$.

后面还会遇到其他类型的事件关系, 如 A 与 B 相互独立, 这在概率论及其实际应用中占有很重要的地位, 稍后再行展开.

在公式表达经常会遇到求和号 $\sum$ 和连乘号 $\prod$, 其基本形式如下:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

$$\prod_{i=1}^{n} x_i = x_1 x_2 \cdots x_n$$

在不至于引起歧义时, 为了追求方便, 上下标有时也略去不写.

1.1.2 基本计数原理与技巧

概率的计算通常离不开排列组合等相关的*计数技巧*(counting technique) 与计数原理.

基本计数原理有两个: *加法原理*(addition principle) 和*乘法原理*(multiplication principle). 加法原理的要义是: 做一件事情, 完成它有 n 类办法, 在第 1 类办法中有 m_1 种不同的方法, 在第 2 类办法中有 m_2 种不同的方法, $\cdots$, 在第 n 类办法中有 m_n 种不同的方法, 那

么完成这件事情共有 $m_1+m_2+\cdots+m_n$ 种不同的方法. 乘法原理的基本要义是: 如果完成一个事件可以分解为 n 个独立的步骤, 每个步骤均有 m 种实现方式, 那么, 完成这一事件总共可以有 $m\times n$ 种方法. 通常用一句话概括这两个原理的用法: 分类问题用加法, 分步问题用乘法.

排列组合是高中数学训练的一个重点. 这里不再重复, 仅列出常用概念的记号、定义与公式, 以便回顾.

定义 1.1 (阶乘)　阶乘(factorial), 即阶乘式的乘法, 定义如下:

$$n!=n\times(n-1)\times(n-2)\times\cdots\times3\times2\times1 \tag{1.1}$$

特别地, 规定 $0!=1$.

有时还可能遇到双阶乘(double factorial), 其定义为

$$n!!=\begin{cases}n\times(n-2)\times\cdots\times4\times2, & n\text{为偶数};\\ n\times(n-2)\times\cdots\times3\times1, & n\text{为奇数}\end{cases} \tag{1.2}$$

仍规定 $0!!=1$

在 R中, 计算阶乘的命令为 `factorial()`. 例如, 求 10! 的命令为

```
factorial(10)
```

答案为 3628800.

定义 1.2 (排列)　排列(permutation) 是指从 n 个不同元素中无放回 (without replacement) 地抽取 $r(r\leqslant n)$ 个元素所排成的一列 (考虑元素的先后次序). 此排列的总数记为 ${}_nP_r$, 又记为 P_n^r 或 A_n^r(A 是排列的另一英文 Arrangement 的首字母). 排列的计算方式如下:

$${}_n\mathrm{P}_r=\frac{n!}{(n-r)!} \tag{1.3}$$

特别地, 有 ${}_n\mathrm{P}_n=n!$.

本书采用 ${}_n\mathrm{P}_r$ 这一记号, 这与国外多数教材比较匹配, 与一般科学计算器上的记号也是相符的. 但国内似以 P_n^r 或 A_n^r 为主导. 通常行文中, "排列" 既可能指元素的一种排序方式, 又可能指所能可能排列的总数, 读者需要根据上下文来理解.

定义 1.3 (组合)　组合(combination) 是指从 n 个不同元素中无放回地抽取 $r(r\leqslant n)$ 个元素并成一组 (不考虑元素的先后次序), 记为 ${}_n\mathrm{C}_r$ 或 C_n^r 或 $\binom{n}{r}$. 或者说, 组合数其实考虑的是 n 个不同元素中无放回地抽取 $r(r\leqslant n)$ 个元素, 可以构成的不同子集的个数. 组合的计算方式如下:

$${}_n\mathrm{C}_r=\frac{{}_n\mathrm{P}_r}{r!}=\frac{n!}{(n-r)!r!} \tag{1.4}$$

特别地, 规定 ${}_n\mathrm{C}_0={}_n\mathrm{C}_n=1$.

排列组合的运算技巧非常丰富, 也总出现在各种数学竞赛的题目中. 然而对概率论和统计学的学习而言, 这些技巧并不处于核心地位. 故这里不再详细展开.

R中计算组合的命令为 `choose(n,k)`, 给出的是 ${}_n\mathrm{C}_k$ 的值. 例如, 求 ${}_{10}\mathrm{C}_5$ 的命令为

```
choose(10, 5)
```

答案为 252.

R并未提供直接计算排列的命令, 但注意到排列与组合间的倍数关系, 这可转换为组合与阶乘的乘积来求. 例如, 求 ${}_{10}P_5$ 时, 可利用关系式 ${}_{10}P_5 = {}_{10}C_5 \times 5!$, 输入如下命令:

```
choose(10, 5) * factorial(5)
```

答案为 30 240.

1.2 概率的计算方式与公理化定义

概率(probability) 是什么? 这其实是一个很难回答的问题, 答案也不统一. 这里不妨先引用一段美国统计学家福尔克斯 (Leroy Folks) 的话[①]:

科学理论是建立在没有定义(或定义得不好) 的名词上的. 定义质量, 力和加速度的尝试都是不满意的, 然而依据建立在这些名词上的理论, 飞机在飞, 火车在行驶, 卫星在围绕地球运行. 电子有时描述为粒子, 有时为波, 有时既是粒子也是波, 即使这个词没有确切定义, 而晶体管技术仍在前进. 概率存在类似的情况. 虽然概率这个词没有明确的定义, 但统计方法和概率模型却证明它们自身很有用.

理想中总希望对每个概念都进行精确定义, 然而并不总能做到这一点. 作为概率论的核心, "概率" 这一概念本身就是模糊不清的. 然而正如上面的引文所言: 这并不影响概率论的魅力与应用. 正确地理解这一点, 是大学阶段概率统计学习的重要前提. 在形式化地给出概率的公理化定义之前, 这里先简要介绍概率的几种计算方式 (或称实现方式). 它们在概率的公理化定义出现之前就已经存在, 并且更为直观.

1.2.1 古典概率

通常人们最为熟悉的概率就是*古典概率*(classical probability), 其问题形如: "从装有 10 个红球和 5 个白球的盒子中随机取出一球, 请问该球为红色的概率多大?" 答案显然为 $10/15 = 2/3$. 这里的 "随机取出一球"(或描述为 "任取一球"), 实际是指 "每个小球被取中的可能性相同", 同时盒子中的小球个数也是有限的 (这样能保证分母为有限的整数). 这些隐含的意思对于概率计算非常关键. 实际上, 试验结果的 "有限性" 与 "等可能性" 正是古典概率计算的先决条件.

定义 1.4 (古典概型) 概率论中把满足下列条件的概率模型称为古典概型:

(1) 试验的所有可能结果是有限的;

(2) 试验的每一个结果出现的可能性相同;

(3) 事件 A 的概率 [记为 $P(A)$] 定义为

$$P(A) = \frac{\text{事件}A\text{包含的可能结果个数}}{\text{所有可能结果的个数}} \tag{1.5}$$

$$= \frac{\text{事件 } A \text{ 中包含的样本点数}}{\text{样本空间中的样本点数}} \tag{1.6}$$

① 福尔克斯: 统计思想 [M]. 魏宗舒、吕乃刚, 译. 上海: 上海翻译出版社, 1987: 55.

由古典概型中计算出来的概率就是所谓的古典概率.

之所以称为古典, 是因为这种概率被经典数学家们, 如 Blaise Pascal (1623—1662)、Pierre-Simon Laplace (1749—1827) 等人研究得最早、最透彻. 其中 "事件 A 包含的可能结果" 通常也称为*有利结果*(favorable outcomes), 这里的 "有利" 不是对谁有利的意思, 而只表示这是此时关注的结果 (outcomes of interest). 由于假定了每个结果发生的可能性相等, 古典概型又称为等可能概型, 但严格来说等可能概型并不仅仅局限于古典概型, 后面我们将明了这一点.

要注意的是, 并不是一个事件有 n 个可能结果, 这 n 个结果的可能性就是相同的. 这方面有一个最简单的例子: 假设我们每天都要出门上班 (上课), 此时只有两种可能: 出车祸, 不出车祸. 如果这两个结果是等可能的, 我们还能放心地出门吗? 实际上, 出车祸的可能远低于不出车祸的可能. 概率模型中的 "等可能" 只是一种内在的模型假定, 它不一定是真实的事实. 只有当事实能够与这一前提假定相一致, 才能运用等可能概型来做计算; 一旦事实不能满足这种假定, 就不能使用 "等可能" 的古典概型来计算概率. 这一点务必牢记.

这里仅以几例回顾相关计数技巧在古典概型中的应用.

例 1.1 (生日问题)　宿舍中 6 个人, 求 6 人生日 (只考虑月和日) 各不相同的概率 (假设 1 年有 365 天).

解　设 6 人依次排队 "选择" 生日. 第 1 人 "选择" 生日时, 共有 365 天可选; 第 2 人再选时, 为避开第 1 人的生日, 共有 $365-1=364$ 天可选. 依次类推, 并各自相乘, 即得有利结果数. 这实际上是 365 选 6 的排列数. 而所有可能结果数显然为 365^6. 故所求概率为 ${}_{365}\mathrm{P}_6/365^6$.

R中计算的命令为

```
choose(365,6) * factorial(6) / 365^6
```

答案是 0.959 5.

实际上利用 R可以很快算出任意人数宿舍中每个人的生日各不相同 (或至少有两人生日相同) 的概率, 这将留作练习供大家思考.

例 1.2　n 张奖券中有 r 张有奖, 共有 k 个人购买, 每人一张, 其中至少有一个人中奖的概率为多少 ($k<r$)?

此题直接去计算有利结果不太可行, 因为情况众多, 难以一一列举, 故取其对立事件, 即 "一个人都没有中奖" (设为事件 A), 并计算其概率; 则 $P(A^c)=1-P(A)$ 即是所求概率. 而要保证一个人都不中奖, 最简单的方法, 莫过于奖池的奖券都是不带奖的, 即所有摸出的奖券 (k) 都从没有奖的奖券 ($n-r$) 中抽取. 而所有可能结果显然为 $\binom{n}{k}$. 故所求概率为

$$P(A^c)=1-P(A)=1-\frac{\binom{n-r}{k}}{\binom{n}{k}}$$

例 1.3 (抽签问题)　袋中有 a 个白球、b 个黑球, 无放回地每次从中取出一球, 求第 $k(k\leqslant a+b)$ 次取到黑球的概率.

解 这里提供两种解法. 令 $X=\{$第 k 次取到黑球$\}$.

(1) 排列法. 假设先对每个小球进行编号, 然后把每个小球看成不同的小球, 并把取出的小球依次排入 $a+b$ 个方框中. 则样本空间的基本事件数为 $a+b$ 个球在 $a+b$ 个方框上的全排列, 此排列数为 $(a+b)!$.

第 k 次取到黑球相当于首先在第 k 个方框上放入黑球, 这共有 b 种排法; 然后在剩余的 $a+b-1$ 个方框内排剩下的 $a+b-1$ 个球 (不论黑白), 这共有 $(a+b-1)!$ 种排法. 由于这是分步问题, 使用乘法原理, 有利结果 X 的事件数为 $b(a+b-1)!$ 种. 故所求概率为

$$P(X)=\frac{b(a+b-1)!}{(a+b)!}=\frac{b}{a+b}$$

(2) 组合法. 排列法假设每个小球都有区别. 实际上, 更直观地, 每个小球除了颜色之外并无任何不同, 用组合法应当更合常理. 将小球取出后, 仍将其依次放入 $a+b$ 个方框中. 此时, b 个黑球在 $a+b$ 个方框的所有不同放法的组合数为 $\binom{a+b}{b}$ (因为 b 个黑球之间没有区别, 也就不需要考虑其先后排序, 因此是组合数), 而这个数实际上就是 b 个黑球被取出的不同取法数, 即样本空间中的基本事件总数.

第 k 次取到黑球相当于首先在第 k 个方框上必须放入黑球, 其余的 $b-1$ 个黑球可以任意地选择剩下的 $a+b-1$ 个方框中的任意 $b-1$ 个方框放入, 此时的组合, 共有 $\binom{a+b-1}{b-1}$ 种. 这也就是有利结果数. 故所求概率为

$$P(X)=\frac{\binom{a+b-1}{b-1}}{\binom{a+b}{b}}=\frac{\dfrac{(a+b-1)!}{(b-1)!a!}}{\dfrac{(a+b)!}{b!a!}}=\frac{b}{a+b}$$

此题的重要性在于它解决了抽签或抓阄法的公平性问题. 由于 a、b 只是提前确定好的球数, 即常数, 而 k 相当于抽签顺序. 此例中我们看到第 k 次取到黑球 (类似于标有中奖的签) 的概率与 k 本身无关, 而仅与黑白球事先的比例相关.

同时也可看到, 同样的题目, 在计算概率时, 所使用的样本空间可以不同 (这里分别用排列法和组合法计算了各自的样本空间), 由此得到的有利结果的样本点也不尽相同. 但只要样本空间的建立是合理的, 便可在各自空间下得到相同的概率. 在学习完独立性概念后, 还将对此题进行更深入的讨论.

例 1.4 (三门问题) **三门问题**(Monty Hall problem) 又直译为蒙提 · 霍尔问题, 出自美国的电视游戏节目 Let's Make a Deal, 问题名字来自该节目的主持人蒙提 · 霍尔 (Monty Hall). 情境大致如下: 你会看见三扇关闭了的门, 其中一扇的后面有一辆汽车, 选中后面有车的那扇门可赢得该汽车, 另外两扇门后面则各藏有一只山羊. 当参赛者选定了一扇门, 但未去开启它的时候, 知道门后藏有什么的节目主持人会开启剩下两扇门的其中一扇, 露出其中一只山羊. 主持人其后会问你要不要选择另一扇仍然关上的门. 问题是: 换另一扇门会否增加参赛者赢得汽车的概率? 这里假定所有人的偏好都是想赢得汽车而不是山羊.

解 答案是会, 虽然这似乎违反直觉 (直觉应当回答换与不换一个样, 都是 50% 的概率). 如果一开始就选对了汽车所在的门, 那么你改变选择就会输, 输的概率是 1/3; 如果你

一开始选错了 (概率是 2/3)，那么你改变选择就能赢, 主持人是否打开一扇后面有羊的门, 不会改变这一点. 所以, 换一扇门而赢得汽车的概率是 2/3.

这个问题曾引起很大的争议, 许多成名的数学家, 数学和统计学博士, 以及此类问题的爱好者均不愿意相信 2/3 这一结果. 相关的探讨有许多, 大家可以搜索 "三门一羊" 或 "Monty Hall Problem", 可以搜到许多解释. 这里不详细展开, 仅作为一个引例, 以期激发同学们对概率问题的兴趣. 会编程的同学可以使用软件来模拟这一问题, 得出经验上的论证.

实际上, 还有一些心理学家以实验的方式研究这一问题解决过程中的认知表征问题, 可以为心理学专业的同学提供一个研究的样板: 心理学不仅关注人怎样做出正确的推理, 还关注人为什么会做出错误的推理①.

1.2.2　经验概率

古典概型的概率基于等可能这一前提, 但生活中的许多事件并不总是 "等可能" 的. 例如工厂生产的产品, 有正品、有次品, 正常的生产线中总是次品占极小的比例; 又如某学科成绩的优秀率, 也总是一小部分学生得优, 而多数不得优. 诸如此类的问题就不适宜使用古典概型, 且无法通过预先计算的方式去推理其概率, 而必须加以试验, 等试验结果出来后, 才可能知道其 "概率" 为多少. 如次品率, 通常是以大量随机抽取的样品中的次品数, 除以抽取的样品总数得出. 这实际上是个 "频率"(relative frequency). 直观来讲, 当试验的次数足够多时, 使用频率来作为 "概率" 是合理的, 此时的频率应当会稳定于某个理论中的真实 "概率" 值. 这种概率就称为经验概率(empirical probability), 意指必须经过试验才能确定的概率.

定义 1.5 (经验概率)　经验概率又称概率的统计定义, 其基本思想如下:

(1) 用来确定某一事件 A 的是否出现的试验可大量重复 (理论上常视为无限次) 地进行;

(2) 在 n 次重复试验中, 记 m 为事件 A 出现的数次, 则记其频率为 $f_n(A)=\dfrac{m}{n}$;

(3) 经由大量的观测发现, 当 n 充分大时, $f_n(A)$ 会稳定于某个常数 p 附近. 此时就称 p 为频率的稳定值, 即经验概率值.

注意这里第 (3) 点并不能写成如下形式: $\lim\limits_{n\to\infty} f_n(A)=p$. 这是因为 $f_n(A)$ 并不总是随着 n 的增大而无限接近于常数 p, 而有可能会大于 p, 等于 p, 或小于 p. 例如抛一枚均匀的硬币, 理论上的正面 (Head, 简记 H) 向上与反面 (Tail, 简记 T) 向上的可能性是一样的. 但在做试验时, 抛 100 次时可能正好出现 H:T=50:50 的情况, 而第 101 次时就可能出现 50:51 的可能; 随着次数的增多, 频率反而偏离了常数值. 因此, 它不能用简单的极限语言来描述, 而只能笼统地说: 随着 n 变大, $f_n(A)$ 稳定于 p 的可能性应当更大一些.

有同学可能会想到, 这似可用如下改进形式的极限语言表达:

$$\lim_{n\to\infty} P(|f_n(A)-p|\geqslant\epsilon)=0$$

① 这方面感兴趣的同学可以参考王宝玺、向玲、张庆林 2006 年发表于《心理发展与教育》的《表征影响三门问题解决的实验研究》及华裔经济学家 Keith Chen(2008) 发表于 *Journal of Personality and Social Psychology* 的 *How Choice Affects and Reflects Preferences: Revisiting the Free-Choice Paradigm* 一文, 可以对这一问题有更深入的了解. 如果大家以后进一步了解杰出的心理学家 Amos Tversky(1937—1996) 和 Daniel Kahneman(1934—2002 年获诺贝尔经济学奖) 的工作, 可以发现研究人如何犯错竟然也可以获诺贝尔奖. 他们的主要文献可参见《不确定状况下的判断: 启发式和偏差》(2013, 中国人民大学出版社) 和《思考, 快与慢》(2012, 中信出版集团股份有限公司) 两本书.