数据分析的统计学入门
OpenIntro Statistics

Fourth Edition

(原著第 4 版，翻译初版草稿)

(作者)

David Diez[1]

Mine Çetinkaya-Rundel[2]

Christopher D Barr[3]

(编译)

王世尧[4]

李雪琦[5]

任俊景[6]等[7]

[1] 数据科学家，OpenIntro 创始人
[2] 杜克大学副教授，RStudio 专业讲师
[3] Varadero Capital 投资分析师
[4] 国际货币基金组织，特许金融分析师
[5] 世界银行，跨越数据银河公众号联合创始人
[6] 武汉大学，Galadata 数据银河社负责人
[7] 特别感谢：边逸凡，卢籽文，王敬轩，王怡雯，吴逸飞，武悦，叶子阳在本书第 5 至 6 章初稿编纂种的重要贡献。

原著前言

在找统计学的入门课程？选 OpenIntro Statistics 准没错。该书作为应用统计学的初级读物，其内容严谨、清晰、简明、好懂。编写之初，我们参照的是在校本科生的水平，但没想到现在它在高中甚至研究生的课程中都颇受欢迎。

我们首先希望读者能通过这本书收获统计学基本思维和方法，并在此之上了解以下三点：

- 统计学领域有着非常广泛的实践应用
- 并不是只有「数学大咖」才能玩转数据
- 数据难免混乱，统计学工具也常常不完美。但是，只要你理解你手中工具的「所能及」和「所不能及」，你就能更好地使用它们去了解这个世界。

教材大纲

本书中章节如下：

1. **走进数据**：数据结构，变量，和一些基本的数据收集技巧。
2. **总结数据**：数据摘要，图表，和「使用随机过程进行统计推断」的初体验
3. **概率**：概率的基本原理
4. **随机变量的分布**：正态分布模型和其他核心分布模型
5. **统计推断基础**：结合估算人口比例的背景，讲解统计推断的总体思路
6. **基于分类数据的推断**：使用正太和卡方分布对比例和表格进行推断
7. **基于数值数据的推断**：使用学生分布对一个或两个样本的均值进行推断，对比两个样本时的统计功效，以及对比多个均值的 ANOVA 方法
8. **走近线性回归**：针对数值型因变量和一个解释变量的回归，该章大部分内容在第 1 章节中已经覆盖
9. **多元和逻辑回归**：针对数值型因变量和类别型因变量以及多个解释变量的回归

OpenIntro Statistics 的内容设计考虑了大家对灵活选择学习话题的需求。例如，如果使用这本书的主要目的是尽快了解多元回归，那么你可以只阅读以下必要章节：

- 第 1 章，和第 2 章的 2.1 和 2.2 环节：这些会帮你打下扎实的基础，让你了解数据结构和书中使用的统计学概念。
- 第 4 章的 4.1 环节：让你对正态分布有良好的认知。
- 第 5 章：学习统计推断使用的核心工具集。
- 第 7 章的 7.1 环节：为理解学生分布打基础。
- 第 8 章：了解一元回归的大致概念和原理。

示例和指导练习

示例的作用是帮助你更好了解统计方法是怎么被应用到实践中的。其内容格式如下：

示例 0.1

E

这是一个示例，如果在示例中提出了一个问题，你可以在哪里找到答案？

答案：你可以在这里，示例的解答部分{footnote}/译者注{footnote}，找到答案！

当我们觉得读者掌握的知识已经足够去解答某个问题了，我们就会把示例变成指导练习，其格式如下：

指导练习 0.2

G

指导练习的答案一般会附在原书当页脚注中。[1]

除了文中的指导练习，每个环节和章节的结尾也会有一些练习题。在附录 A 中可以找到奇数编号的环节/章节末练习题解答。[2]

额外资源

相关的视频概览，PPT，统计软件实验室，本书用到的数据集，以及更多资源可以通过以下链接获取：

openintro.org/os

我们还通过添加附录 B 来让数据中的数据更容易获取：附录 B 是第四版新添加的，它为正文中使用的每个数据集提供了额外的背景信息。通过官网 openintro.org/data 页面，还可以找到这些数据集的在线指南，以及一个配套的 R 语言包。

我们非常欢迎大家通过官网 openintro.org/os/typos 页面提供反馈，包括任何拼写错误。

对于高中的学习者，请考虑使用 Advanced High School Statistics 这本教材，这是一个 OpenIntro Statistics 的高中版本。这是 Leah Dorazio 老师为高中学生和美国 AP 统计学课程在本书基础上量身定制的教材。

[1] 这里就可以找到指导练习 0.2 的答案。

[2] 译者注：我们会优先翻译正文内容，环节末/章节末的练习题在初稿中暂时会跳过。

致谢

这个项目得以实现，多亏了作者及作者名单外的那些参与编纂者的热情和奉献。此外，我们还要感谢 OpenIntro 团队长期以来的投入，感谢在 2009 年本书首次发布以来，数百名对本书内容提供了宝贵反馈的学生和老师。

我们也想感谢很多帮助我们审阅本版书稿的老师，他们是：Laura cion, Matthew E. Aiello-Lammens, Jonathan Akin, Stacey C. Behrensmeyer, Juan Gomez, Jo Hardin, Nicholas Horton, Danish Khan, Peter H.M. Klaren, Jesse Mostipak, Jon C. New, Mario Orsi, Steve Phelps, 和 David Rocko。正是他们的宝贵反馈，让本书的文本内容得到了极大的完善。

第 1 章

走近数据 Introduction to data

一直以来，科学家们试图通过认真观察所得和严谨方法来解决问题。这些认真观察所得，也就是**数据 Data** 组成了统计学研究的「脊梁骨」。它们往往是通过类似田野记录[1]，调查问卷和试验等方式收集上来。统计学是一门研究如何更好地进行数据收集、分析，以及有效地从数据中得出结论的科学。在这第一章中，我们既关注数据的属性，也关注如何收集数据[2]。

跨越数据银河

系列推文合集

更多视频，演示文稿，和其他相关资源，请访问：
http://www.openintro.org/os

[1] 译者注：田野调查是一个术语，指的是所有实地参与现场的调查研究工作，也称「田野研究」，通过调查所得记录被笔者译为「田野记录」。

[2] 译者注：理解是数据有自己的生命周期，而这个周期的第一步就是数据收集（产生），所以它才被作者提高到了和「数据本质属性」同样的高度，放在了第一章。

1.1 案例分析：使用颅内支架来预防脑中风

第 1.1 环节主要介绍一个统计学的经典挑战：评估一项医疗手段的有效性。在该环节中使用的术语（实际上可以说本章中涉及的所有术语），会在下文被重复提起。这一部分的目的就是让大家能够对统计学扮演的角色有一个大概的感觉。

在这个环节我们会走近一项医疗试验。这项试验的目的是研究使用颅内支架[1]来预防脑中风[2]的有效性。支架经常被用于心脏病的防范和术后康复：医生把支架搭建在血管内，保证血流的通畅。而不少医生都希望能够通过类似在心血管里面搭支架的逻辑，在脑血管内也搭支架，从而让有脑中风风险的患者也受益。在了解这些背景后，我们开始走近数据。首先我们写下作为研究者要回答的最主要的问题：

使用颅内支架能够降低脑中风的风险吗？

专家们进行了一项试验，一共涉及 451 位存在脑中风风险的患者。接着他们把这些志愿参加试验的人分成两组，每个志愿者都被随机分配到了其中一组中：

试验组 Treatment Group：试验组的患者接受了颅内支架的治疗手段，并遵医嘱服药、严控风险指标、采取更健康的生活方式。

对照组 Control Group：对照组患者不接受颅内支架，只是遵照相同的医嘱进行调养。

在这个过程中，专家们随机分配了 224 位患者到试验组，227 位患者到对照组。对照组的作用就是能够作为一个基准参照，好让我们能够看到并衡量试验组接受的颅内支架治疗的效果。专家们选取了两个时间点：在开始试验的 30 天后，和开始试验的 365 天后[3]。其中 5 位患者的结果在图 1.1 中进行了列举。患者的试验结果被登记为两种情况：出现中风 stroke，或者无异常 no event；出现中风意味着，在试验开始到观测时间点的这段期间，患者出现了至少一次脑中风。

我们如果花大量时间去逐个观测每位患者的得病和治疗情况，或许也可以回答最初的那个「最主要的问题」，但这无疑会是一个耗时长且痛苦的过程。而如果我们开展一项统计学研究（和上述试验一样），就可以一次性去分析所有人的数据。下图 1.2 把原始数据用一种更「有帮助」的方式进行了总结。在下面这张表中，我们可以很快了解到在整个试验中发生了什么。例如，如果要计算试验组中的患者有多少在 30 天内出现了中风情况，我们就直接看左半部分，找到「试验组」和「出现中风」的交叉点，从而了解到这样的患者有 33 人。

---

[1] 译者注：请结合第 4 项脚注看，正是因为病因是血管收缩变窄，所以才需要用支架把血管撑开，保持畅通。

[2] 译者注：中风是传统的中医名称，在这里应该指的是由于脑血管收缩阻塞（或者血管突然破裂）导致血液无法正常流入大脑的一种疾病。

[3] 译者注：选取这两个时间大概是因为：30 天的时间点体现了短期效果，365 天的时间点体现了治疗的长期效果。

病人	组别	0-30 天	0-365 天
1	实验组	无异常	无异常
2	实验组	出现中风	出现中风
3	实验组	无异常	无异常
...			
450	对照组	无异常	无异常
451	对照组	无异常	无异常

图 1.1：脑中风研究中五位患者结果的节选表格

	0-30 天		0-365 天	
	出现中风	无异常	出现中风	无异常
实验组	33	191	45	179
对照组	13	214	28	199
总数	46	405	73	378

图 1.2：脑中风研究的描述性统计量

指导练习 1.1

在试验组的 224 人中，45 人在一年内出现了脑中风。使用这两个数字，计算有百分之多少的是试验组患者在一年内出现了脑中风？（注：指导练习的答案可以在脚注中找到）[1]

我们可以通过表格计算出一些具有概括性的统计量，**概括性统计量 Summary Statistic** 是指那些能够通过一个数字概括很多数据的统计量。例如，上述研究的主要结果就可以用两个统计量描述，即试验组和对照组患者中出现中风的比例：

(1) 试验（接受支架）组在 1 年内出现脑中风的患者比例：45/224 = 0.20 = 20%

(2) 对照组在 1 年内出现脑中风的患者比例：28/227 = 0.12 = 12%

这两个统计量很有用，因为通过它们，我们直接看到了两组之间的差别，而且这个差别多少有些让人意外：试验组，也就是接受了治疗的患者，比对照组的患者中风的比例多了 8%！该信息很关键。首先，它和医生们设想的（颅内支架能够减少中风）相反；其次，它自然地引出了一个统计学问题：这些数据是否足以说明，试验组和对照组的治疗效果「真的」有所不同？

[1] 脑中风的患者比例：45/224 = 0.20 = 20%

别小看第二个问题，它很容易被人们忽视。假设我们扔一枚质地均匀的 1 元硬币，那么正面（字）和反面（花）出现的概率应该各是 50%。可实际上，如果我们真的扔这样的一枚硬币 100 次，我们大约不会正好观察到 50 次正面朝上。这种偏差其实普遍存在。几乎只要是数据产生过程[1]，都难逃这种实际观测与概率不等的偏差。所以刚刚提到的颅内支架试验中，那 8%的差值很有可能可以归因于自然偏差。只是对于同样大小的样本，偏差越大，我们就越难说服自己说这样的偏差只是偶然。所以我们真正需要正视的问题是：多大的偏差才足够大？大到让我们认为这并不是「偶然」，而去正视试验组和对照组的治疗效果确实不同。

这个问题有点深奥了，是不是？其实当我们还没有掌握完备的统计学理论和工具，来帮助解答该问题的时候，我们也可以直接下结论：在以上研究中，我们观察到了足够有说服力的数据证据，证明颅内支架不利于治疗脑中风。

注意！开头短短的这句「在以上研究中」和后面的结论同等甚至更加重要。千万不要轻易把这个结论给普遍化，从而说它对所有患者和所有类型的颅内支架都适用。这个试验考察的患者具有很明显的特质：他们都是自愿加入试验的。而「自愿加入试验」的脑中风患者可能无法代表「所有」的脑中风患者。此外，这次试验只使用了一种颅内支架（由波士顿科技出品的 Wingspan 品牌），而现实中还有很多别的支架。尽管不能随便把结论普遍化，这个试验至少教会了我们重要的一课，也就是统计学研究的结果很可能和预期不一致，而我们应该准备好去迎接这种超出预期。

[1] 译者注：数据产生过程听起来有点唬人，其实就是任何可以产生数据的统计行为，比如扔硬币 100 次，然后记下每次的结果；再比如调查走访，然后记下每个人的回答……

1.2 数据基础

对很多数据研究来说，有效组织数据和描述数据往往是第一步。本环节介绍了使用数据矩阵来组织数据的概念，同时也会介绍一些术语，用以描述本书中不同形式的数据。

1.2.1 观测值，变量和数据矩阵

图 1.3 展示了一个随机抽样得出的贷款数据库的第 1,2,3 和第 50 行。这 50 行数据我们把它记作 loan50 数据集[1]，loan50 就是数据集的名字。

这个数据集的每行都代表了一笔贷款。数据集中一行信息的正式名称是一个**案例 Case** 或者一个**观测值单元 Observation Unit**（简称观测值，英文 observation）。而其中的一列代表了不同笔贷款的同一个特征，称为**变量 Variables**。例如，第一行代表了一笔价值是 7500 刀，利息率是 7.34%的贷款，这笔贷款的借款人在马里兰州，年薪是大约 7 万美元。

G

指导练习 1.2

图 1.3 中第一笔贷款的级别是？它的借款方的住房类型是？[2]

在实践中，有一件事至关重要：那就是通过询问和确认，保证你能够理解数据的方方面面。例如，我们一定要知道数据集中每个变量的含义是什么，并且搞清楚计量单元[3]是什么。对于 loan50 数据集变量的描述列举在图 1.4 中。

图 1.3 中的数据就是一个**数据矩阵 Data Matrix**。数据矩阵是一种常见和有效的数据整理形式，尤其是在用电子表格收集数据的时候。数据矩阵的每一行都对应着一个特定的对象，每一列则对应着一个变量。

记录数据时，尽可能使用数据矩阵，除非你有充分的理由去采用其他形式。数据矩阵的结构让我们可以把新的案例添加成行，把新的变量添加成列。

---

[1] 译者注：译者喜欢把任何二维数据表都称作「数据库」，但如果从 dataset 的字面翻译，数据集确实是个更好的译名。而数据库可能对应的是数据集的集合，也意味着层次更多，维度更多，内容更复杂的数据集合。

[2] 该笔贷款的级别是 A 级，借款方的住房类型是租房。

[3] 译者注：这个计量单元，或者叫做统计单位非常重要。我们先来举一个我亲身经历过的例子，我曾经在世界银行做一些乡村家庭数据的分析，然后其中涉及到两个数据集的合并，一个数据集的一行代表一户人家，另一个数据集的一行代表家庭中的一个人。而最开始我直接把它们合并，结果自然是得出了非常荒谬的结论。然后才意识到，他们两个数据集的计量单元 unit of measurement 不同，应该先把个人数据进行汇总，处理成为以家庭（户）为单位的数据，然后再进行合并。所以计量单元就代表了数据收集的基本单元，代表了每行数据的代表对象，看数据集，一定要去理解计量单元，明确一行数据描述的是一个什么样的对象。

	loan_amount 贷款额	interest_rate 利率	term 期数	grade 级别	state 州	total_income 总收入	homeownership 住房情况
1	7500	7.34	36	A	马里兰	70000	租房
2	25000	9.43	60	B	俄亥俄	254000	抵押贷款
3	14500	6.08	36	A	马里兰	80000	抵押贷款
...	...	...	...	...	...	...	
50	3000	7.96	36	A	加利福尼亚	34000	租房

图 1.3：loan50 数据集中的四行节选

变量（括号内为译名）	描述
loan_amount（贷款额）	借款方到手的贷款金额，单位：美元
interest_rate（利率）	贷款年利息率，单位：百分比
term（期数）	贷款的时长，单位：月
grade（级别）	贷款级别，取值是从A到G，代表了贷款的质量和违约风险
state（州）	借款方居住地所在州
total_income（总收入）	借款方的总收入，包括除主收入外的其他收入
homeownership（住房情况）	借款方居住房屋是全款购置，或抵押贷款，或仍在租房

图 1.4：loan50 数据集的变量和变量描述

指导练习 1.3

G 我们平时所学课程中的作业、小结和考试成绩，通常是以数据矩阵的形式记录在成绩簿里，你会如何用数据矩阵的形式整理成绩数据呢？（比如会有哪些变量，观测值单元是什么等）[1]

指导练习 1.4

G 我们来看一个美国州郡[2]数据库，假设集中包含了乡镇名、所属省份、2017 年人口、2010 到 2017 年人口变化、贫困率以及六个更多的特征，这些数据要如何通过数据矩阵进行整理呢？[3]

图 1.5 展示了指导练习 1.4 中所描述的数据库，图 1.6 展示了所有变量的描述。

[1] 本题答案不唯一，一个比较常见的方法是把每个学生的信息记录在每行中，然后对每项练习，作业和考核去添加各自的列。这样安排的好处是可以通过观察一行数字的变化来了解学生的历史成绩。此外，还应该添加一些列记录学生个人信息，例如一列记录学生名字。

[2] 美国的郡（county）的行政区划是在市（city）上面的，属于州（state）的下一级；例如译者所在的州就是弗吉尼亚州（Virginia），然后是阿灵顿郡（Arlington），接着才是水晶市（Crystal City）。

[3] 每个郡可以当作一个观测值，然后对每个郡来说有 11 项信息被记录。一个有 3142 行和 11 列的数据表可以存下这些数据，每行就代表了一个郡，然后每列/变量代表了某个类别所有郡的信息。

	name	state	pop	pop_change	poverty	homeownership	multi_unit	unemp_rate	metro	median_edu	median_hh_income
1	Autauga	阿拉巴马	55504	1.48	13.7	77.5	7.2	3.86	有大都市区	上过大学	55317
2	Baldwin	阿拉巴马	2E+05	9.19	11.8	76.7	22.6	3.99	有大都市区	上过大学	52562
3	Barbour	阿拉巴马	25270	-6.22	27.2	68	11.1	5.9	无	高中毕业	33368
4	Bibb	阿拉巴马	22668	0.73	15.2	82.9	6.6	4.39	有大都市区	高中毕业	43404
5	Blount	阿拉巴马	58013	0.68	15.6	82	3.7	4.02	有大都市区	高中毕业	47412
6	Bullock	阿拉巴马	10309	-2.28	28.5	76.9	9.9	4.93	无	高中毕业	29655
7	Butler	阿拉巴马	19825	-2.69	24.4	69	13.7	5.49	无	高中毕业	36326
8	Calhoun	阿拉巴马	1E+05	-1.51	18.6	70.7	14.3	4.93	有大都市区	上过大学	43686
9	Chambers	阿拉巴马	33713	-1.2	18.8	71.4	8.7	4.08	无	高中毕业	37342
10	Cherokee	阿拉巴马	25857	-0.6	16.1	77.5	4.3	4.05	无	高中毕业	40041
...	...	...	...	...	...	...	...	...	...	...	...
3142	Weston	怀俄明	6927	-2.93	14.4	77.9	6.5	3.98	无	上过大学	59605

图 1.5：从 county 数据集中节选的 11 行

变量	描述
name	郡名
state	郡所在州名（或华盛顿特区）
pop	2017年人口
pop_change	2010年至2017年人口变化率。例如，第一行的1.48就指对于这个郡，从2010年到2017年人口增长了1.48%
poverty	贫困人口百分比
homeownership	房屋所有者（或与房屋所有者同居所，例如孩子和父母一起住在父母房子里）占总人口百分比
multi_unit	公寓楼占所有房屋百分比
unemp_rate	失业率，单位：百分比
metro	郡中是否有大都市区
median_edu	受教育程度中位数，取值是：高中未毕业，高中毕业，上过大学，和本科毕业
median_hh_income	郡中所有家庭收入的中位数，家庭收入的定义是全部15岁及以上家庭成员的收入总和

图 1.6：county 数据集的变量和变量描述

1.2.2 变量的类型

我们来仔细看一下上面数据集中失业率（unemp_rate），人口（pop），州（state），还有教育程度中位数（median_edu）这几个变量。这几个变量间显然存在着本质上的不同，但是却又共享了一些共同点。[1]

首先考虑失业率变量，它是个典型的**数值型 Numerical** 变量，因为它可以从大范围的数字中取值，并且允许对取值进行有意义的加减乘除以及均值计算。要注意的是，不是所有记录数字的变量都可以被归为数值型变量。例如电话号码：尽管是由几位数字构成，但是对电话号码进行加减运算，或者取几个电话号码的均值显然毫无意义。

接着我们看人口变量，它也是个数值型变量，不过它似乎和失业率又有些细微差别。这个差别就在于，人口变量的取值只能取非负的整数[2]（0/1/2...）。正因如此，我们把数值型变量再细分一下，而把人口归入**离散数值型 Discrete** 变量类别中，因为它只能从数轴上跳动取值。与之对应，能够从数轴上连续取值的变量叫做**连续数值型 Continuous** 变量，失业率就属于此类。

回到数据表中来看州这个变量。美国有 51 个州/特区，所以州这个变量就可以有 51 种取值：AL（阿拉巴马 Alabama），AK（阿拉斯加 Alaska）……一直到 WY（怀俄明 Wyoming）。因为这个变量的取值是分成这 51 类的，所以它可以被称作**分类 Categorical** 变量，分类变量能够取到的那些类别也叫分类变量的**值 Levels**。

最后，我们再来看教育程度的中位数这个变量。这个变量描述了各郡居民教育程度的中位数，有高中以下（below_hs）、高中毕业（hs_diploma）、上过大学（some_college）、本科毕业（bachelors）几个取值。那么我们很容易判断它是一个分类变量，但它似乎融合了一些数值型变量的特征：即取值虽然不能加减，但却可以排序比较（本科毕业 > 上过大学 > 高中毕业 > 高中以下）。我们把这样的变量分类归为分类变量下面的子类别：叫做**序数 Ordinal** 变量。而相对的，如果只是一个普通的分类变量，其取值排序无意义，这样的变量就归入另一个子类别：**名义 Nominal** 变量中。为了方便教学，我们把本书中所涉及的所有序数变量都当作名义分类变量处理。

---

[1] 本质不同指的变量记录的信息内容不同，而共同点指的是例如失业率和人口的记录都以数字形式呈现，而州和教育程度中位数都以文字形式呈现。

[2] 译者注：有时候我们会在一些报告上看到带有小数点后几位数字的人口数据，这是因为那些数据往往是以万/亿为单位的。例如说我国在 2020 年底约有总人口 14.2 亿人，这里尽管加了小数点，但由于带上了单位，所以理解起来也很自然。而严格意义上来说，无论什么国家，省份，城市，人口都一定得是非负的整数。

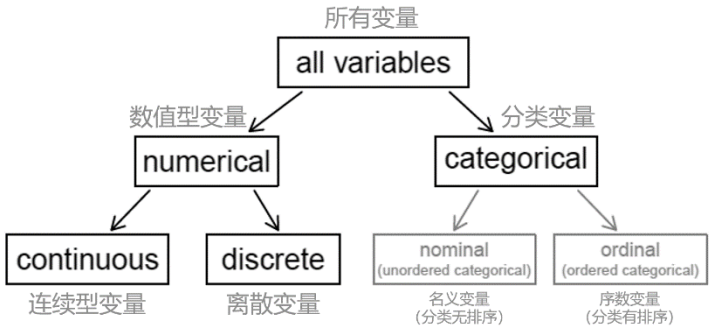


图 1.7：各类型变量细分

示例 1.5

我们来考虑某门统计学课学生的数据，对每个学生都记录了以下三个变量：兄弟姐妹的个数、该学生身高和该学生之前是否上过任何统计学课程。请判断这三个变量分别是连续数值型变量、离散数值型变量或是分类变量。

E

答案：兄弟姐妹的个数和身高都是数值型变量，其中前者是计数（只能取整数），所以是离散数值型的；而身高的值可以连续变动，所以是连续数值型变量。之前是否上过任何统计学课程这个变量只有两个取值：上过和没上过，所以属于分类变量。

指导练习 1.6

G

一项试验正在评估一种新药治疗偏头痛的有效性，其中组别（group）变量用来区分试验组和对照组，偏头痛次数（num_migraines）变量记录了三个月内，患者出现偏头痛的次数，请判断以上两个变量是数值型还是分类变量。[1]

1.2.3 变量间的关系

很多分析产生的背景都是：研究者想要探寻某两个或者多个变量之间的关系。一个社会科学的研究者可能试图回答以下问题：

(1) 如果某地的「房屋所有者占总人口的比例」低于全国平均值，那么该地公寓楼数量是多于还是低于全国平均水平？

(2) 如果某地的人口增长速度高于全国平均值，那么该地的家庭收入的中位数是会高于还是低于全国平均值？

(3) 是不是受教育水平中位数越高，家庭收入的中位数也越高？

[1] 指导练习 1.6 答案：组别变量只有两个有意义的取值，所以是分类变量。对偏头痛次数进行算术运算是有意义的，所以它是一个数值型变量；更具体来说，这是一个计数（只能取整数），所以它还是一个离散数值型变量。

想要回答这些问题，我们就要收集数据，例如之前图 1.5 展示的美国各郡的数据集就是个例子。通过其中一些概括性统计量，我们就可以为回答上述问题找找思路。此外，我们还可以借助一些图表来从视觉上探索数据。

散点图是一种用来展示两个数值型变量间[1]关系的图表，图 1.8 展示了「房屋拥有者占总人口比例」和「公寓楼占全部楼房比例」之间的关系，图上的每个点都代表着一个郡。例如，被红色圈出来的点对应了图 1.5 数据集中的 413 号郡：佐治亚州的查塔胡奇郡。该郡的公寓楼占比为 39.4%，房屋拥有者占比为 31.3%。这张散点图体现了两个变量间的一种关系：公寓楼比例越高的郡，自己拥有房屋的居民比例越低（说明大家都在租公寓，而不自己买房）。我们可以想想这种关系背后的一些原因，然后针对每个可能的原因进行研究，进而分析出最合理的解释。

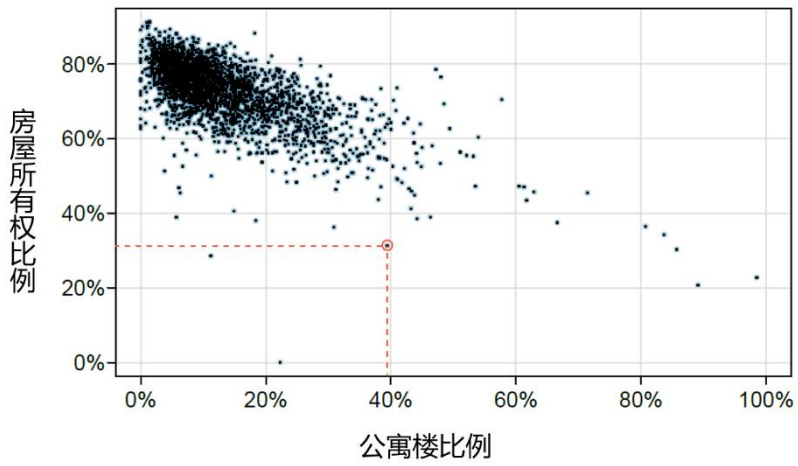


图 1.8：一张散点图：体现了美国各郡「房屋拥有者占总人口比例」和「公寓楼占全部楼房比例」关系的散点图。图中圈出的红点代表了佐治亚州的查塔胡奇郡，该郡的公寓楼占比为 39.4%，房屋所有权比例为 31.3%

通过这张图，我们可以说：「房屋拥有者占总人口比例」和「公寓楼占全部楼房比例」之间是相关的，因为在图中可以观察到明显的从左上至右下的分布趋势。当两个变量相互关联的时候，我们可以称他们为**相关变量 associated variables**。相关变量在英文中既可以用 associated variables 表示，也可以叫 dependent variables。

指导练习 1.7

G

请回到图 1.3 中的 loan50 数据集（如下），然后想想看哪些变量间可能有关联，请试着提出两个猜想。[2]

[1] 译者注：散点图的横轴和纵轴是两条数轴，所以只有可以进行算术运算的数值型变量放上去才有意义

[2] 例如如下的两个问题：(1) 贷款额和总收入之间的关系是什么？(2) 如果某借款人的收入大于平均值，他们贷款的利息率是会倾向于高于或是低于平均利息率？

示例 1.8

请观察图 1.9，其中涉及：美国各郡从 2010 年至 2017 年「7 年来人口变化率」以及「家庭收入中位数」。你觉得的这两个变量是有关联的吗?

E

答案：从图上可以看出，如果家庭收入的中位数越高，那么该郡的 7 年来人口变化率也越高。尽管这个趋势并不适用于图上所有点（郡），但是总体趋势确实如此。这么说来，这两个变量之间确实有某种联系，他们也就是一组相关变量。

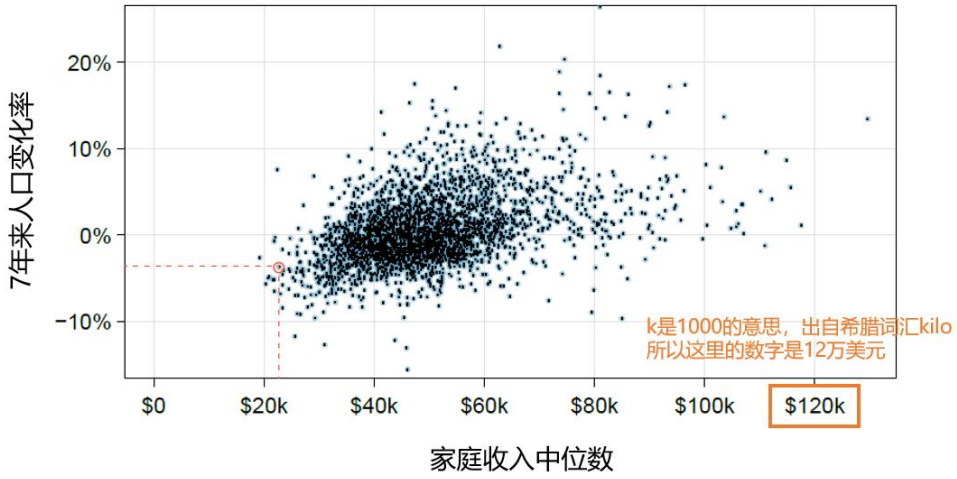


图 1.9：关于「7 年来人口变化率」和「家庭收入中位数」的散点图。肯塔基州的奥斯利郡在图中左下角被标出，它自 2010 年来人口缩减了 3.63%，家庭收入的中位数是 22736 美元

因为图 1.8 中的散点呈现一种向下的趋势（即有更多公寓楼的郡也伴随着更低的房屋拥有者比例），所以我们说图中的两个变量呈**负相关 negatively associated**。与之对应，我们可以在图 1.9 中看到两个**正相关 positive association** 的变量（即家庭收入中位数更高的郡也倾向于有更高的人口增长率）。

如果两个变量之间没有关联，那么我们说他们之间是**独立的 independent**。通俗些讲，两个独立变量之间「没有故事发生」。

相关或独立，只能选一个

一组变量相互之间要么是有关联的，要么是互相独立的。没有一组变量互相既是相关的，又是互相独立的。

1.2.4 解释变量和响应变量

当研究两个变量之间的关系的时候，除了单单讨论它们相互关联或相互独立，我们有时也想了解其中一个变量的变化是否会引起另一个变量的变化。之前我们对美国郡县 county 数据集提过这样一个问题：如果某地的人口增长速度高于全国平均值，那么该地的家庭收入的中位数是会高于还是低于全国平均值？（见环节 1.2.3 开篇第一段）」现在我们换个问题:

如果一个郡的家庭收入中位数增加了，这会导致人口的增长吗?

在这个问题中，我们不再只关心相关性，而是在思考是否一个变量会影响（或者说改变）另一个变量。而如果我们如此假设，假设它们二者中一个（家庭收入中位数）会影响另一个（7 年人口变化率），那么施加影响的变量「家庭收入中位数」就叫**解释变量 explanatory variable**，同时「7 年来人口变化率」就叫**响应变量 response variable**[1]。

解释变量和响应变量

当我们怀疑一个变量会从「因果上」影响另一个变量的时候，我们就把造成影响的前者记为解释变量，把受到影响的后者记为响应变量。它们之间的关系如下：

解释变量 ----------（可能影响）> 响应变量

有一点请铭记：这样的变量标记方式万万不代表他们两者间一定存在因果关系。想要确定因果，需要更严谨的试验设计、数据收集和之后的统计评估。一组变量相互之间要么是有关联的，要么是互相独立的。没有一组变量互相既是相关的，又是互相独立的。

1.2.5 观察性研究和试验性研究

数据收集的过程主要可以分成两类：观察性研究和试验性研究。

为了解释清楚两者，我们首先要明确数据是如何产生的（how the data arise）。如果研究者不人为地对数据产生过程进行干预和控制，那么这就更像一个**观察性研究 observational study**。例如，研究者可能直接通过调查问卷，或者既有的医疗/企业资料中去搜集信息。再具体点，以疾病研究为例，研究者可能选择跟踪调查一群有相似特征的患者，试图发现规律，进而提出疾病由何而生的假设。在这些情景中，研究者很少干预数据产生过程，从而直接收集到具有因果说服力的数据。所以总而言之，观察性研究往往可以说明相关性，但却无法推断因果性。

[1] 译者注：更加通俗的名称是「因变量」和「自变量」，不过原书作者没有采用这种命名方式是因为「因变量」的英文也是 dependent variable，所以容易和上文提到的「相关变量 associated variables」的另一种叫法混淆。而为了和原书保持一致，我们在本书的翻译中也广泛使用「解释变量」和「响应变量」。

如果研究者需要深入调查因果性的时候，他们会考虑设计**试验 experiment**。常见的思路是先确定要研究的解释变量和响应变量，然后控制其他变量保持不变。例如，还是在医疗领域，我们想研究一种药物是否会降低心脏病患者的死亡率。那么我们会把药物的使用与否当作解释变量，患者的死亡率当作响应变量。然后试验流程往往是先召集一批患者，并对他们分组，然后让每组患者都接受除了药物外相同的治疗，随后观察比较两组患者的死亡率。如果在分组的时候遵循了随机的原则，那么这个试验也被称为**随机试验 randomized experiment**。随机试验需要保证分组流程的随机性，比如我们可以掷硬币，根据结果正反面来把患者分到两组中。如果是让患者自己选择，或者根据年龄分组，那么这就不够「随机」。我们已经学过了试验组和对照组的差别，试验组的患者会试用这种新药，而对照组的患者往往会在不知情的情况下接受**安慰剂 placebo**。让患者不清楚自己接受的是药物还是安慰剂是为了减少患者自身心理因素的影响（以免接受了药物治疗的患者会倾向于给自己积极的心理暗示）。在章节 1.1 的案例（研究颅内支架和脑中风）中，对照组就没有接受安慰剂，而这有可能会造成试验结论存在偏差。

所以可以看得出来，试验性研究更关注因果，所以会需要更严谨的设计，更多的投入，和更加精密的统计工具。尽管如此，观察性研究很多时候也已足以说明问题，只是我们要记得提醒自己：相关性不等于因果性。

相关≠因果

相关很多时候都不意味着因果，而因果性的结论往往依赖严密的随机试验。

1.3 抽样原则和策略

想要开展统计研究，务必要明确需要回答哪些问题。这是应该最优先考虑的，先于数据收集、数据分析和数据可视化。如果能够明确一个具体的研究问题，将会有助于识别研究所属的领域，所涉及的案例，和所依赖的变量。明确了问题之后，考虑数据从何而来才变得至关重要。我们需要依赖可靠的数据收集手段，去达成既定的研究目标。

1.3.1 总体和样本

我们来分析下面三个研究问题：

(1) 大西洋剑鱼是一种肉质鲜美的海鱼，而我们知道汞是一种对人体有毒的物质。那么大西洋剑鱼体内的平均汞含量是多少呢?

(2) 过去五年，杜克大学的学生完成本科学位平均需要花多长时间?

(3) 已知一种新药刚被研发，那么这种药到底能不能减少严重心脏病患者的死亡人数?

以上，每个研究问题其实都指代了一个**总体 population**。在第一个问题中，总体是所有的大西洋剑鱼，其中每条剑鱼代表着一个个体。很多时候，要把总体中的每个个体都调查一遍成本太高，因此，我们往往只抽取一部分样本 (以下简称抽样)。**样本 sample** 是由总体中的被抽取的一部分个体构成的，也可以说是总体的一个子集。例如，我们可以从大西洋里捕捉 60 条剑鱼，那么这 60 条剑鱼就构成了一个样本，我们可以通过这个样本来估计总体（即大西洋所有剑鱼）的相关数据。

G

指导练习 1.9

对于以上提出的第二个和第三个问题，他们的总体和个体分别是什么？[1]

1.3.2 轶事证据

对于之前的三个问题，我们来看看以下三个可能的回答：

(1) 有新闻报道称一名男子吃了剑鱼之后出现了食物中毒，由此说明剑鱼中的汞含量一定很高。

(2) 我遇到过两个杜克大学的学生，他们都花了 7 年多的时间才毕业，所以在杜克大学完成学位的所需要的时间比别的学校都要长。

(3) 我朋友的父亲心脏病发作后服用了一种新药，结果还是去世了，说明新药不管用。

[1] 对于第二个问题，首先要明确，我们讨论的应该只是那些完成了学位的学生，而未能完成学位的学生花在项目上的时间应该不予考虑。所以，研究的总体应该是所有杜克大学过去五年内本科毕业的学生，而个体则是每一个这样的学生。对于第三个问题，任何一名严重心脏病的患者都是该研究的一个个体，而总体则是所有患者构成的群体集。

以上三个回复都是基于具体数据的，但是它们都存在了两个问题：首先，它们的数据都只包含了一到两个个体案例；其次且更重要的是，这些个体案例不一定就能代表总体。这种偶然得到的数据被称为**轶事证据 anecdotal evidence**。

轶事证据

在使用这种轶事证据时一定要注意：我们不用全盘否定他们，因为这些证据或许也是真实可信的。但即便如此，它们可能也只代表了某些特殊情况，而不一定能反映总体。

图 1.10：2010 年 2 月，有媒体上的相关专家用一场暴风雪来反驳全球变暖。喜剧演员乔恩·斯图尔特指出，这只是「一」个国家中，「一」个地区的，「一」场暴风雪

轶事证据通常代表一些特殊情况。因为具有某些骇人听闻的特点，所以它们更容易被人记住。例如，是「七年逾期毕业的学生」还是「四年正常毕业的学生」更让人印象深刻？通常来说都是前者。但是，如果我们正在做某一问题的调查统计，就应该着眼于一个具有普遍代表性的样本。

1.3.3 从总体中抽取样本

如果我们在杜克大学抽取一部分学生，来研究该校学生过去五年内的毕业所需时间，那么该校所有过去五年内的毕业生就构成了总体，被抽中的毕业生就构成了样本。 通常来说，我们需要在总体中随机取样。最基本的随机取样原理就像抽奖一样，例如，在抽取杜克大学毕业生的时候，可以把他们所有人的名字都写在纸条上，然后放在盒子里，从中抽出 100 个。这被抽中的 100 个毕业生就代表了一个随机样本。随机取样能够减少统计偏差。

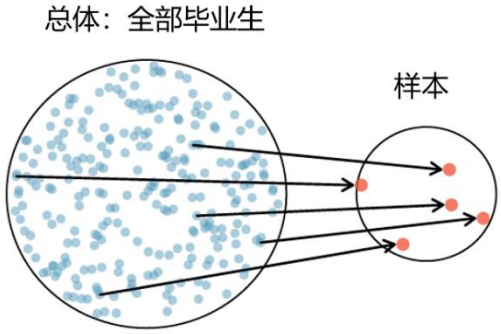


图 1.11：该图中，从总体（全部毕业生）中随机抽出了五位学生组成了一个样本

示例 1.10

如果在这个 (杜克大学学位完成时间) 研究中，我们请一位营养学专业的女生主观选择一些毕业生，然后由他们组成样本进行研究。你觉得这个她可能会选什么样的学生？你觉得她选出来的学生能代表全部毕业生吗？

E

答案：她选择的样本中，很有可能健康领域的毕业生会占更大的比例。当然她也有可能随机挑选，从而使样本有普遍代表性。不过，当我们采取这种「主观挑选」的方式去抽样，我们自然面临着选出**有偏 biased** 样本的风险。即使我们没有存心，或者说故意去增大某一群体的比例，但是还是可能产生「我们自己意识不到但确实存在」的偏差。

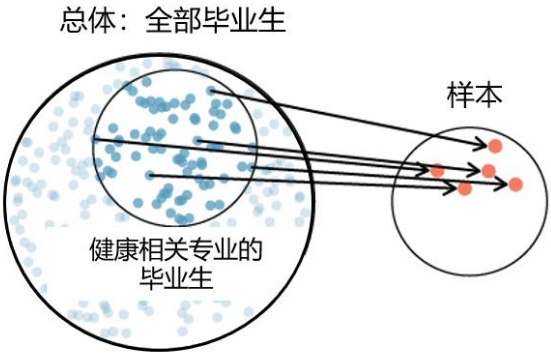


图 1.12：在取样时，一个营养学专业的学生可能会不经意间选择更多健康相关专业的毕业生，从而使得样本中各专业的占比不合理

如果把抽取毕业生样本的过程交给某一个人全权决定，即使他/她完全无意，最终得到的取样结果，也很有可能会由于这个人的喜好而产生**偏差 bias**。随机取样就可以避免这个问题。最基本的随机取样就像抽奖一样，取出的样本被称为**简单随机样本 simple random sample**。在抽取简单随机样本的时候，总体中的每个体被抽到的概率都是相等的，并且个体之间没有隐含关系。

随机取样能够很大程度上减少偏差，但即使进行了随机取样，偏差也有可能通过其他方式产生。例如，过低的**应答率 response rate** 也可能让样本产生偏差。如果我们从总体中随机挑 100 个人发放调查问卷，但只收到了其中 30 个人的回复（应答率为 30%）。那么尽管抽样的过程是随机的，我们也很难确定最终这 30 个人的样本是否还能代表最初挑选的 100 人，进而代表总体。由于应答率过低而产生的偏差叫做**无应答偏倚 non-response bias**。

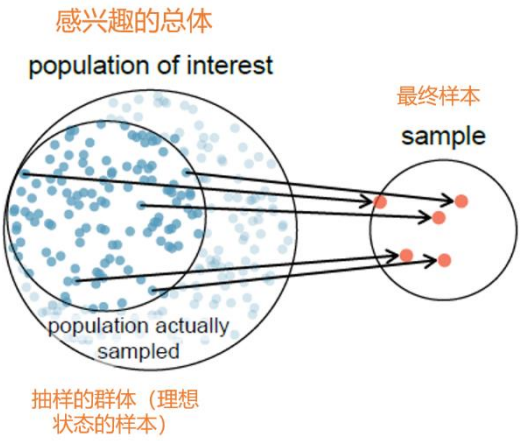


图 1.13：现实中，由于总有人（出于客观原因）无法或者（主观原因）不愿意接受调查，所以真正进入样本的可能只是总体中特定的「愿意提交答案的」个体[1]。这个问题几乎不可能被彻底解决

除了低应答率之外，**方便抽样 convenience sampling** 也有可能产生偏差。方便抽样是「随意」而非「随机」[2]选择被调查者。例如，调查者在纽约的布朗克斯街头随意拦截路人展开调查，那么最终的样本就不能代表全部纽约市民，因为越是经常路过布朗克斯街的人就越容易被调查到，反之就越不容易被调查到。我们通常也很难判断：方便抽样得到的样本到底代表了总体中的哪一部分个体。

指导练习 1.11

G

很多线上的商品，店家或者公司都有评分系统，而我们也经常会参考这些评分做决定。那么如果对于某一商品的差评率有 50%，你是否认同这句话：买了该商品的人中有半数是对商品不满意的？[3]

[1] 译者注：举个例子，如果在网上公开投放调查问卷，那么客观上没条件上网的人就一定被排除在样本之外；还有比如调查收入水平时，有些富豪可能主观上不愿意被调查。

[2] 译者注：随机和随意的区别在于：随机是严格按照某一符合随机分布的科学系统（例如抽纸条，掷硬币等等）；而随意是指由客观大环境和主观心情交叉决定，并不服从某种科学分布。

[3] 该题无固定答案，只需要注意：任何线上的评分都是基于「出于某种动机想主动提供反馈的人」所提供的评分统计出来的。就本练习而言，根据不一定站得住脚的生活经验，我们发现人们在产品达不到期望的时候总不忘了抱怨，而在产品达到或超出预期的时候总吝于赞扬。正因如此，我们才会质疑在淘宝这样的平台上，50%的差评率是否存在偏差，而不能代表买了产品的人中有半数都不满意。当然，这些结论是从生活经验出发，我们也愿意对各种看法保持开放态度。

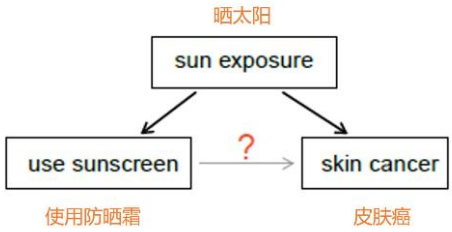
1.3.4 观察性研究

如果不对数据的产生过程进行干预，那么收集上来的数据就被称为**观察性数据 observational data**。例如，之前在章节 1.2 举例时用到的个人贷款数据（loan50 数据集）和美国州郡数据（county 数据集）都属于观察性数据。我们前面讲了，基于试验来进行因果推断是更合理的。对应的，使用观察性数据进行因果推断往往是有风险的，所以不被推荐。因此，观察性数据往往仅被用来去发现相关性，进而做出假设，而这些假设需要通过试验加以验证。

指导练习 1.12

G 假设一项观察性研究的课题是《防晒霜和皮肤癌》。然后通过观测，这项研究发现如果一个人用的防晒霜越多，那么他/她就越有可能得皮肤癌。这是否意味着：防晒霜会导致皮肤癌呢？[1]

一些之前的研究表明，使用防晒霜不仅不会导致皮肤癌，实际上还能降低患皮肤癌的风险。那么上述练习中发现的「防晒霜和皮肤癌之间的正相关」又是怎么一回事呢？这就要引入被忽略变量的概念。「防晒霜的使用」和「患皮肤癌的概率」间的正相关，可以用一个被忽略的变量解释。这个缺失的变量就是：晒太阳（或者专业点：日晒）[2]。如果一个人总是整天地晒太阳，那么他/她很有可能使用更多的防晒霜，同时有更大的可能患皮肤癌。如果只是做简单的观察性研究，「晒太阳」这个因素很可能就未被考虑，从而得出一些与以往研究相悖的结论。



像「晒太阳」这样的变量，就是统计学中的**混淆变量 confounding variable**。混淆变量在英文中有时候也叫 lurking variable（直译潜在变量）， confounding factor（直译混淆因子），或者 confounder（直译混杂因素）。混淆变量需要同时和解释变量与响应变量都相关。在观察性研究中，如果想得出因果结论，方法之一就是尽可能穷尽所有的混淆变量。不过这实际上很难做到，因为无法保证所有混淆变量都被考虑到、且进行了统计测量。

指导练习 1.13

G

[1] 当然不，请阅读本练习后紧接着的一段。

[2] 译者注：这里由于是初次举例，所以原著也是尽可能精简语言，方便理解。否则，日晒应该还可以细分为日晒的时间和程度等等。

大家是否还记得，图 1.8 中我们展示了「房屋拥有者占总人口比例」和「公寓楼占全部房屋比例」之间的负相关关系。尽管趋势存在，直接判断这两个变量间存在因果关系是不合理的。那么你能找到一个混淆变量解释这种负相关关系吗？[1]

观察性研究又可以分成两种：前瞻性研究和回顾性研究。**前瞻性研究 prospective study** 以现在为起点，随着事件的推进去追踪记录个体信息。例如，医疗领域中的癌症研究者可能就会长达数年跟踪研究一群患者，从而探索哪些行为会影响人们得癌症的风险。具体点的一个例子叫 The Nurses' Health Study（护士健康研究）。它始于 1976 年，并在 1989 年时进一步扩大研究。与前瞻性研究对应，**回顾性研究 retrospective study** 则是收集之前事件中数据。例如，同样是医疗领域中的癌症研究，回顾性研究者会侧重审查既往病例。现实中，数据往往是由一些前瞻性变量和一些回顾性的变量共同组成的。

1.3.5 四种抽样方法

几乎所有的统计方法都要基于随机性的概念之上。如果观察性数据在收集的时候不遵循随机原则，那么后续的统计学估计和误差分析就会变得不可靠。既然随机性如此重要，我们就在此讨论四种随机的抽样方法：简单抽样，分层抽样，整群（或聚类）抽样，和多阶段抽样。图 1.14 和图 1.15 就针对这些方法提供了图解说明：

简单随机抽样 Simple random sampling 大概是最符合直觉的随机抽样形式。我们举例来看：假设我们要分析中国超级足球联赛（简称中超）的球员工资，我们可以就采用简单随机抽样的方式。我们把所有中超球员的名字一个一个写到单独的纸条上，然后把所有纸条放在一个大纸箱里面摇匀，最后从里面抽出 96 张纸条。这就形成了一个「简单随机抽样」的样本。这是因为总体中每个球员被选入样本的概率都相等，同时一个球员被抽选到样本中，不会影响（或者提供额外信息）其他球员是否会被抽到。

分层抽样 Stratified sampling 是一种使用了「分而治之」思路的抽样策略。首先我们把总体划分成多个群组，每个群组都被称为一个**层 strata**。在层的选择上，我们可以把有相似特征的一系列个体集中到一起作为一个层。分好层之后，我们会引入刚刚讲过的简单随机抽样的方式，再从每个层中抽出若干个个体。还以刚刚的中超球员薪资水平为例，中超有 16 支球队，我们可以把其中每支当做一个层，因为每个球队的球员工资水平相对来说是接近的（不得不承认，有的球队就是比别的球队有钱）。接着我们从每个球队中随机地选择 6 名球员，形成共计 96 名球员的样本。

[1] 本题无固定答案。我们可以举出的一个例子是「人口密度」。如果一个郡人口稠密（且房屋有限），那么自然这个郡就有相当一部分居民需要住在多单元的公寓楼中。同时，如果人口密度大，那么对应的房屋需求也就大，房价也就更高，从而让在当地买房变得更加困难。所以大家都租房子（而不是买房）可能不是由于公寓楼太多导致的，而是由于人太多导致的。

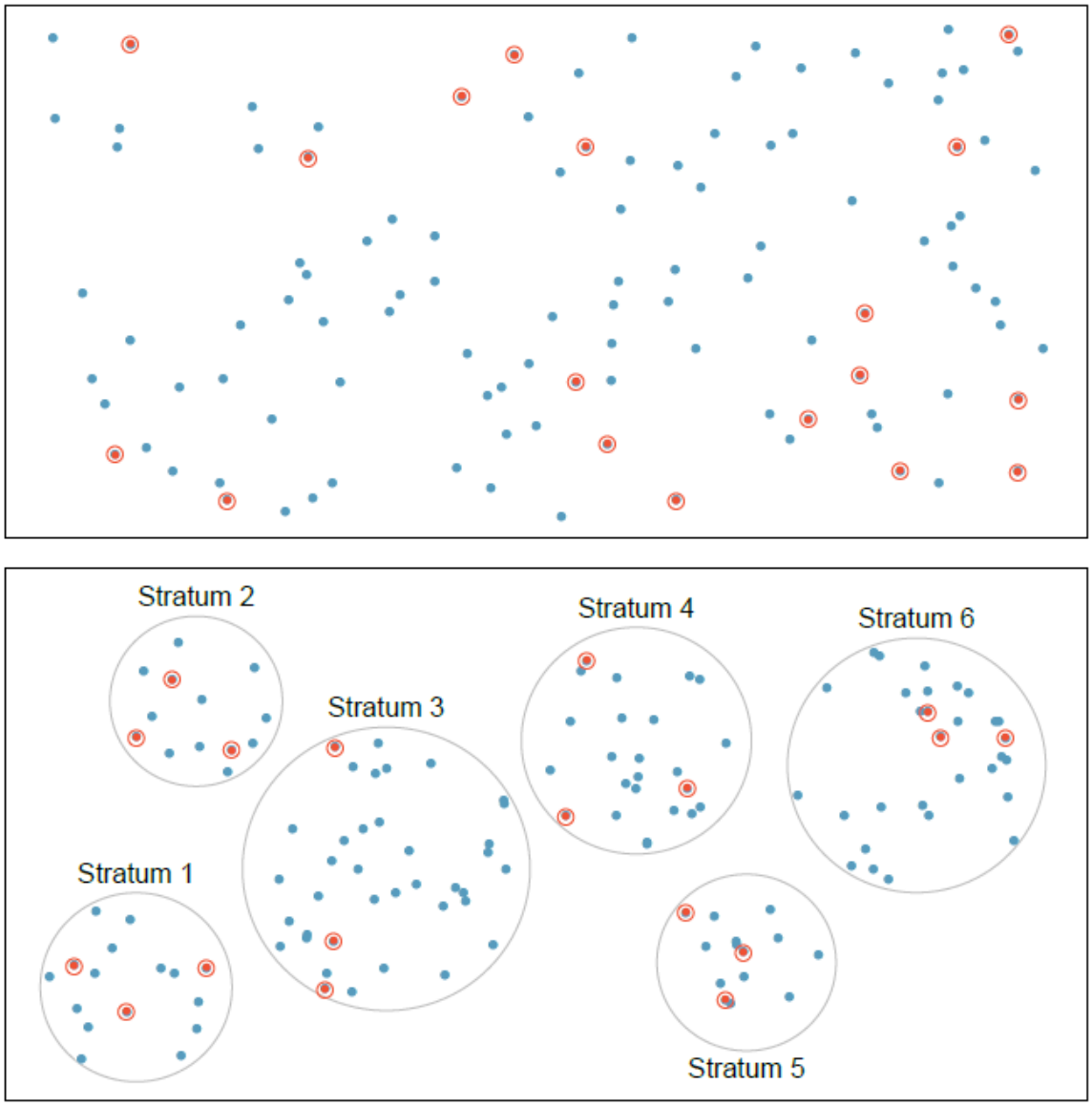


图 1.14：简单随机抽样（上方）和分层抽样（下方）的图解。在上面的图中，展示了使用简单随机抽样方法抽取 18 个个体。在下面的图中，我们可以看到分层抽样的概念设计：Stratum 是「层」的意思，总体首先被不重叠地划分成多个「子总体」，每个子总体也就被称为一个「层」。接着在每层中，我们使用简单随机抽样的方法抽取若干（图中是三个）个体

当我们感兴趣的信息在同一层的个体间相似的时候（例如同个球队球员的工资），分层抽样是非常好用的。不过分层抽样的缺点在于：使用分层抽样得到的数据进行数据分析的时候，比使用简单随机抽样数据要麻烦。由于本书是统计学的初级教材，所以并未涉及到分析分层抽样数据的工具。而如果大家需要分析它们，还请做延伸阅读和学习。

示例 1.14

在分层抽样中，为什么同层个体间越相似越好？

E

答案：因为如果一层的相似度高，那么随机抽取出的那些个体就更能稳定地反映该层的信息，带来更准确的「层级别」的统计量估计。而因为总体的统计量是对层级别的统计量的进一步汇总，这个过程中，如果能够在每层做到更精确，最后汇总得到的总体估计也就自然更精确。

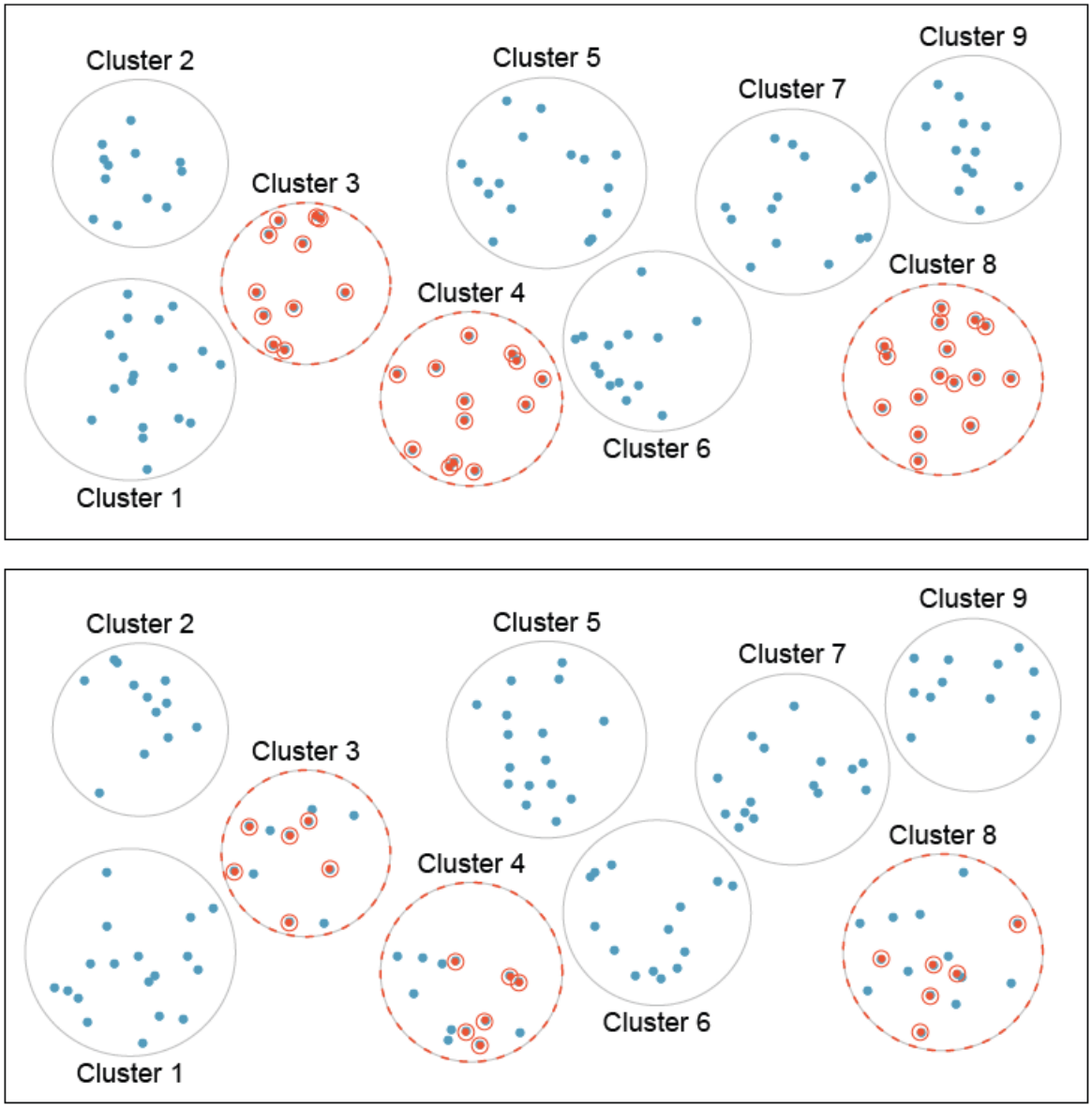


图 1.15：整群抽样（上方）和多阶段抽样（下方）的图解。在上面的图中，展示了整群抽样的概念：首先把所有数据划分成 9 个群集，然后从中取出 3 个群集，并抽取每个群集中所有的个体组成样本。在下面的图中，我们可以看到多阶段抽样的示例：同样是首先划分成 9 个群集，同样是抽取其中 3 个，唯一不同点在于多阶段抽样中我们对每个群集只抽取其中部分个体（而不是全部）加入样本。

在**整群抽样 cluster sampling** 中，我们也是把总体分成很多组，不过这些组叫做**群集 cluster**。然后与分层抽样（遍历所有层）不同，整群抽样中我们只会抽取部分群集，然后把每个被抽到的群集中的所有个体都纳入样本。**多阶段抽样 multistage sampling** 和整群抽样有点像，不过多阶段抽样不会把每个群集中所有个体都保留，而是像分层抽样那样从抽到的每个群集中再进行简单随机抽取，这样选择个体加入样本。

因为不需要遍历，有时候整群抽样或者多阶段抽样的性价比更高。同时，我们刚刚提到了分层抽样希望每层个体间尽可能相似，而层与层之间可以有明显不同。与之对应的是，整群/多阶段抽样希望差异发生在群集内部，而群集与群集之间应大体相同。举个例子：如果要进行社区抽样调查，而我们把一个社区当成一层或者一个群集。如果每个社区内居民很相似（收入，习惯等等），而社区间居民差别很大，分层抽样就会是不错的选择；而如果同社区内居民状况差异很大，那么整群和多阶段抽样将会是不二之选。

示例 1.15

假设我们想研究印度尼西亚热带农村中居民的疟疾感染率，并为此收集了以下信息：我们感兴趣的区域中有 30 个村子，村与村之间环境情况非常相似。然后我们现在需要对该区域中的 150 名村民做测试，你会考虑使用什么样的抽样方法呢？

E

答案：我们可以直接使用简单随机抽样的方式，不过这样就会让收集成本变得非常高（需要进入到很多村子中进行取样）。由于不了解每个村子内个体差别是否很大（如果很大，则无法使用分层抽样），所以分层抽样的方法也不一定可取。这种情况下，整群抽样或者多阶段抽样听起来更好一些。如果我们决定使用多阶段抽样，我们可以考虑从 30 个村子中先随机地挑选 15 个村子，然后从每个村子里再随机挑选 10 名村民。这样一来，我们可能可以解决成本过高的问题。多阶段抽样形成的样本一样也是可靠的，只是在分析它们的时候需要用到本书不曾讨论的一些高级计量技巧。

1.4 试验

如果研究者用科学方式对数据的产生过程进行有计划的干预，那么这样的研究就叫做**试验 experiment**。如果在干预的时候遵循了随机的原则（例如通过掷硬币的方式对患者个体分组），那么这个试验也被称为**随机试验 randomized experiment**。当我们想要说明两个变量间存在因果关系的时候，随机试验是很重要的。

1.4.1 统计试验的设计原则

随机试验通常要遵循以下四个原则：

对照原则 Controlling：除了干预措施外，研究者应该尽可能控制试验组和对照组的其他因素保持一致。这样才能够仅就「施加干预与否」形成鲜明对照。举例来说，如果研究的是某药物的治疗效果，那么病人服药的方式就属于「其他因素」，应该被控制一致。因为有的病人可能会一口吞下药片，或者只借助一点点水服药，而有的病人可能吃一片药要喝一整杯水。为了控制「喝水的量」这个因素，医生可以要求每个病人在服药的时候都喝下 350 毫升的水。

随机原则 Randomization：对于那些无法控制的因素，研究者可以考虑通过随机分组的方式来尽可能消除影响。例如，有的人可能因为饮食习惯而更容易得某种病，而我们恰巧要研究这种病的非饮食成因。这时候，随机分组就可以汰除掉饮食习惯的影响，而关注被研究的特定致病因子。不仅如此，如果对照组和试验组真的是随机分配的，我们也可以避免无关要素带来的结论偏差。

重复原则 Replication：研究者观察的个体越多，那么对解释变量和响应变量间因果性的估计也就会越准确。在一个单独的研究中，**重复 replicate** 的体现就是收集足够大的样本。不仅如此，重复还有另一层含义：我们也经常看到一群科学家通过重复之前做过的研究来核实已知的结论。

区组原则 Blocking：除了一些不能直观计量的因素外，研究者有时候会怀疑（或者已经确定）已收集的数据中有些变量[1]会影响响应变量。这时候，我们可以先根据这个变量对个体进行分**区 blocks**。例如，一个药物对心脏病疗效的试验，我们可以先把参与试验的病人根据发病风险分成低风险区和高风险区，然后再对每个区的病人进行随机分组。接着，我们把低风险区随机抽取的一半病人和高风险区随机抽取的一半病人分入试验组，剩下的分入对照组，如图 1.16 所示。这样的一种战略保证了对照组和试验组中有同样数量的低风险心脏病患者和高风险心脏病患者。

[1] 译者注：一般研究中，我们会首先确定感兴趣的解释变量和响应变量，然后试图找到二者之间的因果关系。这里要注意「感兴趣的解释变量」和「解释变量」之间的区别。例如要研究「快餐」能否引起「脱发」，那么「快餐」就是感兴趣的解释变量，但同时其他可能的解释变量还有：「不运动」，「熬夜」，「压力大」等等。

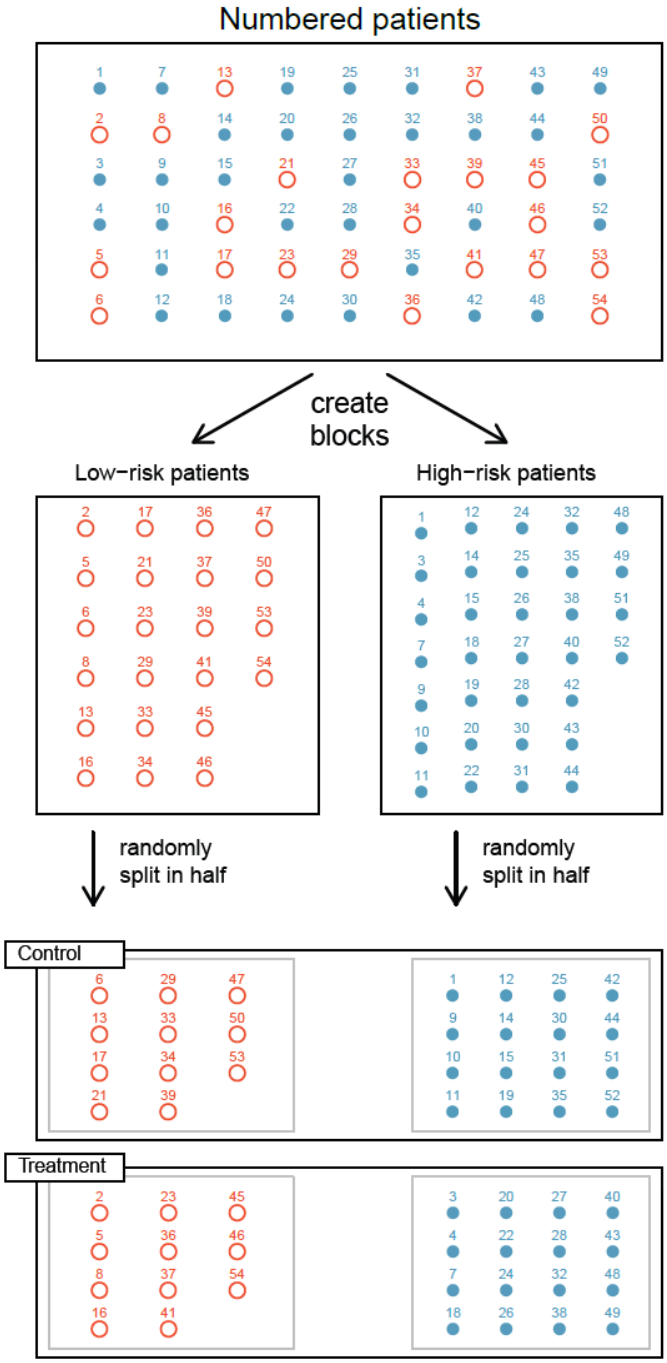


图 1.16：使用风险变量进行分区：所有患者首先被分成低风险区和高风险区，之后在每个区内再随机分成试验组和对照组。这样可以保证试验组和对照组中的两种风险患者相对比例是相同的。

对于任何一项试验研究来说，以上提到的四个原则的前三项都适用。本书也为满足这些的试验数据分析提供了统计学工具。区组原则相对来说是一个更加高级的设计方法，所以我们可能也需要超出本书的统计学知识来分析采用了分区原则后收集得到的数据。[1]

[1] 译者注：原书作者一直在提醒大家：越复杂和高级的数据收集手段，也对应着越复杂和高级的计量统计工具。

1.4.2 减少试验中的人为偏差

随机试验可以说是数据收集的黄金法则，但是它却不是无偏因果推断的充分条件。以人为调查对象的研究就是产生「无意识偏差」的特别好的例子。我们还是考虑大家想必很熟悉的「新药和心脏病」的例子。在这个例子中，研究者需要知道新药是否能降低心脏病患者的死亡率。

那么我们的研究员首先设计了一个随机试验。之所以选择试验，是因为他们不仅仅对相关性感兴趣，更想试图得出一些关于药效的因果的结论。参与试验的志愿者们（也是心脏病患者）被随机地分成两组。其中**试验组 treatment group** 会接受新药的治疗，**对照组 control group** 则不尝试新药。

现在，请把自己想象成一个参与了这项试验研究的志愿者。假设你在试验组里，就会有人拿着一种看起来很高级的全新药物给你使用，而你自然也会给自己一种积极的心理暗示，期待这种药物会起效。与之对应，假设你在对照组中，你只是无所事事地在等待中度过试验阶段，那么你自然就觉得来参与这次试验对你没啥影响，同时还可能不断给自己消极的心理暗示，希望参与试验不要增加自己患病死亡的风险。这一正一负的心理状态形成鲜明对比，从而让试验本身就带来了两种效应：一是药物的效应，二是心理和情绪波动的效应。

作为药物领域的研究，设计本试验的研究员显然对后者没什么兴趣，更何况它的存在还可能造成试验结果的不准确[1]。为了消除这种偏差，研究者意识到不应该让病人知道他们被分到了哪个组。而如果研究者能够做到这点，而让病人对他们的分组状况不知情，那么这个试验就被称为**盲法试验 blind**。

以临床医疗为例，盲法试验的困难在于，如果参与者发现自己没有接受「特殊治疗」（比如服一种新药或者注射一种新疫苗），那么他/她就能轻易意识到自己是在对照组中。这个问题的解决方案是，给所有对照组中的病人提供一种「假的」但是和试验组方式一致的「特殊治疗手段」。在医疗领域我们把这个「假的」药剂叫做**安慰剂 placebo**。灵活选择合适的安慰剂是盲法试验成功的关键。一个很经典的安慰剂的例子就是，在让试验组的病人服下被测试的药丸的时候，给对照组的病人服下外包装一样的糖丸。这样无论是服用了真药还是安慰剂的病人往往都会倾向于猜测自己服下的是真药，从而给自己施加一样积极的心理暗示，避免了心理因素给试验结果带来的影响。

这种积极的心理暗示还会带来**安慰剂效应 placebo effect**， 或者又称假药效应，伪药效应。即尽管服下的是糖丸或者类似的没有真实疗效药物，病人的病情或多或少还会真的有所改善。

---

[1] 译者注：这里的偏差可能是个比较抽象的概念，译者可以试着解释一下：首先我们站在上帝视角，来设定药物真的没有效果。但是如果上述心理因素确实存在，并且真的足够明显，这就会导致在现实中我们观察到了试验组的病人死亡率确实显著低于对照组。那么站在现实视角，不知道药物到底有没有效果的我们，在观察到了试验组死亡率较低这个事实之后，是否就会倾向于做出「药物确实有效」的结论？而这一结论会和事实相悖，其原因就是我们没考虑到其实是「新药可能能治好我的病」的积极心理预期导致了更低的死亡率。

我们说了给病人分组的时候要**设盲 blinding**，那么医生呢？其实不仅病人的心理因素会影响试验结果，医生的心理因素和对应行为同样会影响试验，造成偏差。大家设想：作为一个参与试验的医生，他/她会不会容易对已知试验组的病人更感兴趣，然后给予他们更多的关注和照拂？毕竟这些病人的病情发展很可能直接和药物疗效挂钩。这样一来，试验组和对照组间「医生的关注和照拂」这个变量就没能得到很好地控制。为了防止这种因素带来的偏差（而且我们发现这种偏差有时候真的会对试验结果造成不可忽视的影响），现代的研究都会采用一种**双盲 double-blind** 的设计，让无论医生还是病人都无从得知他们到底是在哪个组中。

指导练习 1.16

G 请回顾在章节 1.1 中的关于颅内支架和脑中风的例子。这个例子中的研究设计能被称为「试验」吗？这项研究有「设盲」吗？这项研究有采用「双盲」的设计吗？[1]

指导练习 1.17

G 在章节 1.1 的案例中，研究者们有可能引入安慰剂吗？如果可以，那么这个安慰剂要怎么设计？[2]

在阅读了指导练习 1.17 的内容后，你可能会对使用「假手术」做对照的伦理逻辑有所质疑。甚至于在我们介绍试验的时候，你可能也想过既然一项手术有可能对病人有好处，那么为什么不让所有病人都接受手术？这些问题恐怕不是该书作为一本统计学教材所能回答的。不过我们可以明确的是：如果采用「假手术」的对照手段，虽然确实可以制造安慰剂效应，但无疑会带来额外的风险；而如果什么都不做，尽管病人可能会意识到自己身处对照组中，但是也维持了病人原本的个人风险水平。

关于试验（尤其是临床试验）和安慰剂，其实总是有很多不同观点交流碰撞，而且我们也很难明确地说谁对谁错。例如，就是因为假手术会带来额外的风险，那么使用假手术做对照的行为就是不道德吗？要知道，如果没有引入安慰剂，那么所有接受了试验的病人很可能只是因为心理效应而有所好转。这样，即使新医疗手段实际并无效果，我们也可能会得出「值得推广」的结论。而且，在推广这种无效（或许价格还很高昂）的治疗手段中，很可能浪费掉很多时间，人力等资源。这些资源本来是可以用在已知有效的治疗手段上的。所以如果不采用假手术，会不会不仅让试验变成了无用功，甚至于还有副作用。错误的结论最终耽误了那些本来会选择其他治疗手段的病人。这会不会是更大的不道德呢？

---

[1] 首先，因为该例子中病人被分成两组，一组进行试验，另一组做对照，所以可以被称得上是试验。其次，根据描述，病人们是可以直观区分出自己到底是在试验组还是在对照组，所以这个试验没有「设盲」。最后，因为试验没有「设盲」，更谈不上「双盲了」。

[2] 这个问题要复杂一些，因为在这个例子中试验组的病人不是在试吃新药（用外包装类似的糖丸就可以做安慰剂），而是接受了颅内支架手术。那么这个问题的本质其实就是：我们有办法让病人「觉得」自己接受了手术吗？事实上这也是有可能的，甚至有些试验中已经在采用一种被称为**假手术 sham surgery** 的手段。在假手术中，病人也经历一场手术，只是术中并不对病人完全施加试验手段（例如颅内支架）。而因为经历了手术，对照组的病人也会获得安慰剂效应，从而减少心理因素的影响。

第 2 章

总结数据 Summarizing data

本章内容会着重关注概括性统计量的计算方法和制图。我们会使用统计软件来生成本章节中的一些统计表和统计图。由于这可能是你第一次接触它们，我们会放慢脚步，一步一步来展示。理解本章内容，对于学习后续章节至关重要。

跨越数据银河

系列推文合集

更多视频，演示文稿，和其他相关资源，请访问：
http://www.openintro.org/os

2.1 研究数值型数据

本环节中，我们会探索「概括」数值型变量的方法。例如，我们之前举过个人贷款数据集 loan50，其中贷款额变量就是一个数值型变量。贷款额之所以可以被归为数值型，是因为讨论两笔贷款额度之间的数学差异是有意义的。从这个角度出发，地区代码和邮政编码虽然也是数字，但是却不是数值型变量，因为对他们进行数学运算毫无意义。因此地区代码和邮政编码就是分类变量。

在接下来的两个环节中，我们会使用章节 1.2 中引入的两个数据集：个人贷款 （loan50） 和美国郡县 （county） 数据。如果你想回忆一下这两个数据集里都有哪些变量，请参考前面的图 1.3 和图 1.5。

2.1.1 配对数据和散点图

散点图 scatterplot 对两个数值型变量提供了一种直观的、可以看到每个观测值的可视化方式。在前面的图 1.8 中，我们就使用了散点图，来研究 county 数据集中「房屋拥有者比例」和「公寓楼比例」之间的关系。那么下面这张图就是在比较 loan50 数据集中，「借款方总收入（total_income）」和借了多少钱「贷款额 （loan_amount）」之间的关系。在任何散点图中，一个点都代表了一个观测值。而因为在 loan50 数据集中有 50 行观测值，所以在图 2.1 中也就有 50 个点。

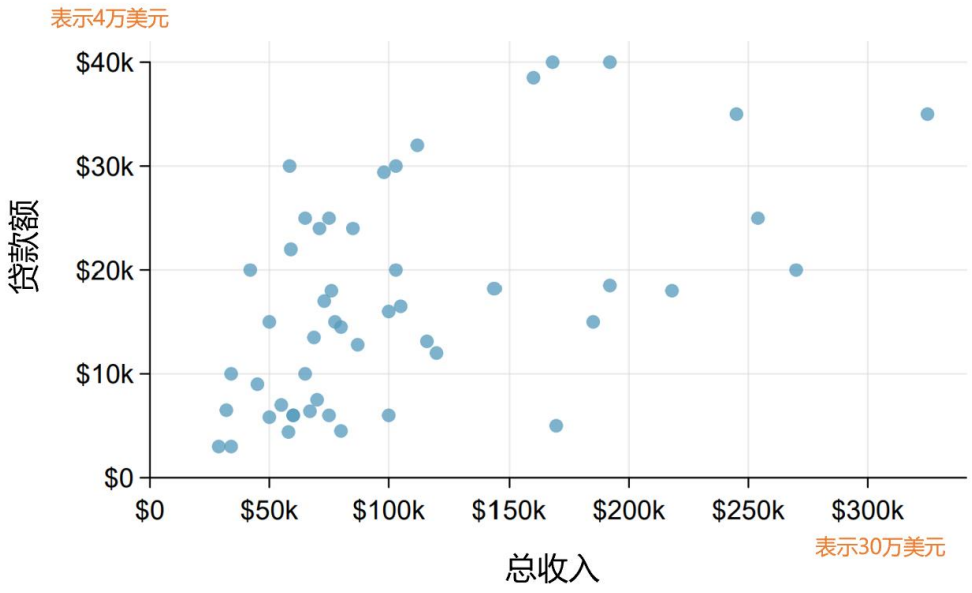


图 2.1：loan50 数据集中对比「总收入」和「贷款额」的散点图

观察图 2.1，不难发现在图的左边，收入在 10 万美元（$100k）以下的人不在少数。而收入在 25 万美元以上的人就屈指可数了。

示例 2.1

图 2.2 展示了一张比较各郡「家庭收入中位数」和「贫困率」的散点图。从图上看，这两个变量间的关系有什么特点吗？

答案：这两个变量间很明显存在非线性的关系，这可以从图上的虚线看出。这张图和我们之前看过的很多散点图都有所不同。之前书中给出的散点图，都没有如此清晰地展示出两个变量间的非线性曲线关系。

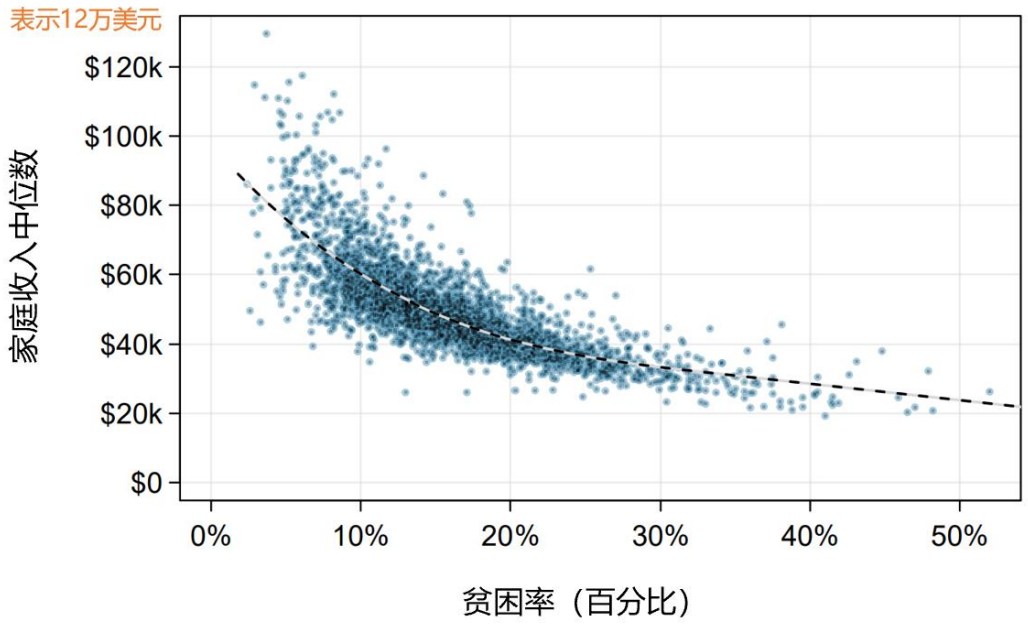


图 2.2：关于 county 数据集各郡「家庭收入中位数」和「贫困率」的散点图。我们已经找到了和数据拟合的统计模型，并在图上用一条虚线标识了出来

指导练习 2.2

散点图对于揭示数据信息有什么作用？[1]

指导练习 2.3

你能否描述两个变量，它们间的散点图呈现倒 U 型（或者说马蹄铁型：∩）？[2]

[1] 无标准答案。散点图因其能快速发现变量间的关系，而在数据可视化中占有一席之地。而且无论两个变量间的关系是简单或是复杂，散点图总能帮助我们从视觉上直观感受数据，进而提出关系猜测。

[2] 可以想象这样两个变量：纵轴是对你的「好处」，而横轴是一种「适量才好的东西」。比如「健康」和「喝水」就可能符合题设描述：我们需要水，喝一些水无疑是健康的，但是如果喝过量的水，那么就对健康无益处了。原书的这个例子真的让译者脑洞大开：玩游戏的效用和时间？恋爱的愉悦度和谈恋爱的次数？老板的赞赏和工作中付出的努力？果然什么都是适度才好呀！

2.1.2 均值和点图

我们书中一开始就以散点图为例展示数据制图，它的制作需要用到两个变量。但有时候，我们只想专注观察一个数值型变量，那么我们就可以把二维的散点图去掉一个维度，使用一种很基本的一维点图。**点图 dot plot** 可以理解成单个变量的「散点图」，请以下图为例，观察一下最基本的点图的特征。图 2.3 是 loan50 数据集中贷款利率这个变量的点图展示。然后我们把其中取值一样的点不再重叠而是堆积起来，就得到了如图 2.4 所示的一张堆积点图。

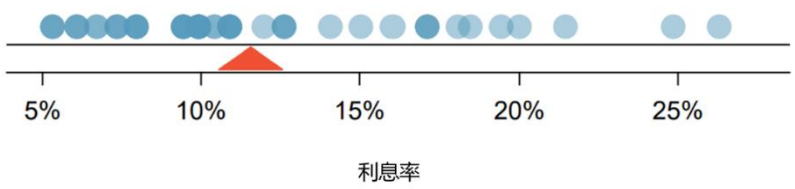


图 2.3：loan50 数据集中「利息率」变量的点图，该分布的均值在图上以红色三角标识

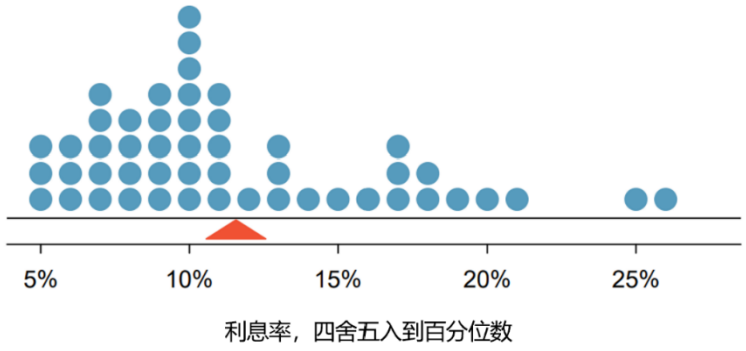


图 2.4：loan50 数据集中「利息率」变量的堆积点图，该图中利息率被四舍五入到最近的百分位数，分布均值同样用红色三角标出

均值 mean，也就是我们常说的**平均数 average**，是一种衡量数据分布中心的常见统计量。如果要计算上述数据集中利息率的均值，我们可以把所有利息率加起来，然后除以观测值的总数。

$$\bar{x} = \frac{10.90\% + 9.92\% + 26.30\% + \cdots + 6.08\%}{50} = 11.57\%$$

样本的均值我们一般使用这个头顶带横杠的$\bar{x}$（可以使用 LaTeX 公式输入）的符号表示，英文记作：x-bar（bar 就是短棍的意思，指上方横杠）。字母 x 此处指代利息率这个变量，它头顶的横杠代表着均值。通过上式可以看出，数据集中 50 笔贷款的平均利息率是 11.57%。为了帮助大家理解均值的概念，我们可以把它想象成数据分布的「平衡点」。通过图 2.3 和图 2.4 可以看到，代表均值的红色三角就像天平的支点一样，让整个数据左右两部分保持平衡。

均值

样本均值可以通过把所有观测值的取值加总，再除以观测值个数的方式计算：

$$\bar{x} = \frac{x_1 + x_2 + \ ... \ + x_n}{n}$$

其中x_1，x_2，……，x_n代表了数据集中的 n 个观测值对应的变量取值。

G

指导练习 2.4

观察计算利息率均值的公式，结合 loan50 数据集，你能回答 x_1 代表了什么，x_2 代表了什么吗？接着你能推理到普遍情况，猜测 x_i 代表了什么吗？[1]

G

指导练习 2.5

在 loan50 数据集中，n 的取值是多少？[2]

刚刚讨论的 loan50 贷款数据集是从一个更大的总体（一个名为 Lending Club 的美国 P2P 贷款平台的所有贷款）中取的样本。如果条件允许，我们也可以像计算样本均值那样来计算总体的均值。不过需要注意，在算式中我们需要用另一个符号（而不是 $\bar{x}$）来表示总体的均值：μ。这个符号是希腊字母，读作「miu」。它往往被用来指代总体中所有个体某信息的平均数。因为一个数据集中有很多变量，比如说这些变量分别是 x/y/z……。有时为了区分不同变量的总体均值，我们会在字母 μ 后面加上一个下标，例如用 μ_x 来代表变量 x 的总体均值。现实中，像计算样本均值那样去精确统计总体均值（即把每个个体的信息都收集之后取平均）成本往往太高。所以一般统计学家们会采取一种折中的手段：通过某变量 x 的样本均值 $\bar{x}$ 来估计其总体均值 μ_x。

E

示例 2.6

如果了解总体中的所有贷款利息率的平均值（一般感兴趣的研究问题都是针对总体的），我们可以通过样本的信息来进行估计。基于样本中的 50 笔贷款的信息，你觉得谁会是总体贷款利息率 μ_x 的一个合理估计？

答案：样本的均值，即刚刚计算出的 11.57%，可以作为总体均值 μ_x 的估计。尽管它并不完美，但至少在这个示例中，样本均值　是一个估计总体均值的最好选择。

[1] x_1代表了数据集中第 1 笔贷款（但不是编号为 1 的贷款，而是算式中第 1 项）的利息率，也就是 10.90%，x_2代表了第 2 笔贷款的利息率，也就是 9.92%……那么以此类推，　就代表了第 i 笔贷款（i 取值在 1 到 50 之间）的利息率。例如，如果 i=4，那么x_i就对应了x_4，即第 4 笔贷款。

[2] n的取值就是样本大小：50。

从第 5 章开始，我们将涉及到一些工具，用来评判如何才能让点估计（用某个样本统计量来估计总体统计量，例如样本均值这样）更精确。想必不难想象，样本越大，点估计就越准确，即越有机会接近总体的实际值。

示例 2.7

均值在统计中非常好用，因为无论数据分布如何，这个统计量都可以作为一个「标准化」了的指标，便于我们快速理解和比较。那么你能举两个例子，来展示均值在数据比较上的作用吗?

答案:

E

1. 我们想要知道某种新药在预防哮喘上会不会比传统药物更有效。于是我们设计了一个包含 1500 名病人的试验，其中 500 人用新药，余下 1000 人作为对照组使用传统药物，最后统计结果如下：使用新药的病人中，总共记录了 200 次哮喘发作。而使用传统药物的病人中，总共记录了 300 次哮喘发作。如果只是比较 200 和 300 这两个数字，很容易导致我们得出新药有效的结论。但实际上，两组人数是不同的，所以我们不能简单地拿哮喘发作总数作比较，而应该看平均每人的发作次数：

新药组：200/500 = 0.4 次；传统药物组：300/1000 = 0.3 次，从数据可以看出，传统药物的使用者平均哮喘发作次数更少，所以新药效果并没有那么理想。

2. 老埃去年在美国搞了辆食物餐车卖墨西哥鸡肉卷，最近三个月生意渐渐稳定了，他过去三个月总共赚了$11,000，大约工作了 625 个小时。因为自生意稳定仅三个月，所以这个赚钱总额并不能很好地评估他的收益。那么我们可以帮他算一个小时平均收入，即 11,000/625 = 17.60 美刀每小时（硕士毕业入职世界银行第一年一般可以拿到每小时 25 美金左右)。算出这个平均时薪，老埃等于是可以用一个标准化的指标来进行比较，比如比比之前工作的时薪，或者和其他餐车运营者来一较高低。

示例 2.8

假设我们想要计算美国人的平均收入，那么我们就可以考虑使用之前用作案例的美国郡县数据集：代号 county。已知这个数据集有 3142 个郡或县，并统计了每个郡县的个人平均收入，我们是不是可以直接对这 3142 个平均值再取一步简单平均，从而得到一个对美国国民平均收入的估计呢？[1]

答案：这样做其实不合理，因为 county 数据集里面每个郡的人数各不相等。如果我们只是取简单平均，即把各郡的人均收入加起来再除以总郡数，就相当于我们把一个有几千人的小郡和一个有几百万人的大郡一视同仁了。所以，比较合理的做法是通过每个郡的人均收入和人口数量计算出每个郡的总收入，再把所有郡的总收入加总，最后除以所有郡的总人口。以 county 数据集为例，如果我们是采用这种加总再除以总人口的方式计算，可以得到人均收入是$30,861；而如果是直接把 3142 个郡的人均收入简单平均，就会得到$26,093 的数字，一人就少了 4000 多美元。

E

其实有一定数学基础的小伙伴不难反应过来，这相当于对各郡的人均收入再做一个**加权平均 weighted mean**。OpenIntro 的官网制作了一个针对加权平均的补充材料：openintro.org/d?le=stat wtd mean。

---

[1] 译者注：人均这个词除了 average 之外，还有个更加形象准确的表达：per capita，人均收入就是：per capita income。

2.1.3 分布和直方图

点图的特点在于：它在图上把每个观测值的具体取值都用一个点标注了出来。这样固然可以让我们直接看到取值，但是大家也可以想想，如果数据集中观测值的个数非常多会变成什么样子？那样的话，有限长度的线段上，密密麻麻「挤」满了点，或者即使用堆积点图，也会「堆」满了点，变得难以观察理解。所以更多时候，我们不会选择把每个点的取值都展示出来，而是把每个取值想成是属于一个组（英文习惯叫 bin，直译的话就是箱）。例如，针对 loan50 数据集，我们制作了一张表，见图 2.5。它统计了分别有几笔贷款落在不同的利率区间（5.0%到 7.5%，7.5%到 10.0%，等等）。需要注意，在这种分类下，一定要明确对于落在两组交界处的观测值的归属。例如我们就人为规定，如果是位于交界处，就自动归入数字较小的那个组中（10.0%归入 7.5%到 10.0%的组）。这些被分组之后的个数统计数字，通过一个个长方形的柱子展示排列，得到了如图 2.6 所示的**histogram 直方图**。乍一看，似乎还有点像之前图 2.4 绘制的堆积点状图，不过直方图分组更明确，更集中。

Interest Rate	5.0% - 7.5%	7.5% - 10.0%	10.0% - 12.5%	12.5% - 15.0%	···	25.0% - 27.5%
Count	11	15	8	4	···	1

图 2.5：分组后的「利息率」数据各组个数统计

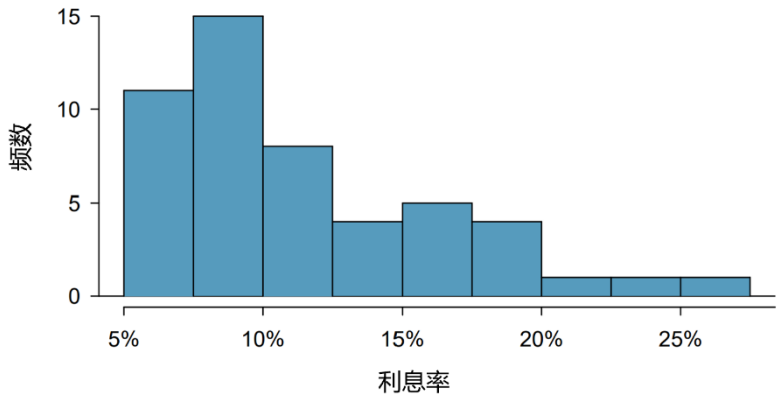


图 2.6：「利息率」数据分布直方图，可以看出数据明显呈现右偏趋势

直方图可以帮助我们了解**数据密度 data density**。如果一个组的柱子越高，也就代表着有更多的数据落在这个组区间内。例如我们通过上图可以看到，利息率在 5%到 10%之间的贷款数目远多于利息率在 20%和 25%之间。这些柱子的高低起伏直观地标明了数据是如何跟随利息率变化而或密或疏地分布的。

直方图对于帮助我们了解数据分布的形状非常有帮助。图 2.6 的直方图传递了如下信息：大多数贷款的利息率都在 15%以下，而利息率在 20%以上的贷款屈指可数。当数据像这样有着越向右数

目越少的趋势的时候，我们就形象地说数据（这里就是「利息率」变量的数据）在右侧有一条长长的尾巴，对应英文是：has a longer right tail。这种数据分布用一个专业术语描述就是**右偏 right skewed**。

说完右偏，大家应该能想象**左偏 left skewed** 数据长什么样子了：从直方图来看，左偏数据将会是越靠近左侧的柱子越低，并且左半部分整体显著低于右半部分。这样，数据就在左边拖着一条长长的尾巴，被称为左偏数据。而如果通过直方图画出的某变量分布在左右两边大差不差，整体比较均衡，那么我们就会说该变量的分布是**对称 symmetric** 的，既不左偏，也不右偏。

通过长尾来识别偏度

当数据分布朝着一个方向逐渐变稀疏（体现在直方图上就是柱子越来越低直到趋近于横轴）的时候，我们就称其为**长尾 long tail** 分布。在左侧拖着长尾叫左偏分布，在右侧拖着长尾的叫右偏分布[1]，有时候右偏分布也被称为**正偏 positively skewed**，处于这种状态的数据特征也被称为**正偏态 positive skewness**。

指导练习 2.9

G

观察图 2.3 和图 2.4，你能从这两张图中看出数据的偏度吗？直方图和点图，哪种类型的图更便于观察数据**偏度 skewness** 呢？[2]

指导练习 2.10

G

除了均值外，有哪些信息是你只能从点图（没办法从直方图中）获取的？[3]

除了可以观察数据分布的偏度，直方图也可以帮助我们来识别数据的众数信息。**峰 mode**[4]指的是分布中一座明显突出的「山峰」。在上面的「利息率」直方图中，可以看到只有左边一座明显突出的「山峰」。

在数学课上，我们都学过 mode 一词是众数的意思，具体来说就是在数据中出现频次最多的那个数。但是在真实的统计案例中，我们面对的数据往往不是像{1,2,2,2,3,4}这样的完美集合，而是具体的统计信息。所以在高精度的统计中，也很有可能对于某个感兴趣的变量，整个数据集所有观测值的取值都各不相同。所以很多时候，数学上众数的定义显然不适合统计实践。

图 2.7 展示了三个直方图，分别有一、二和三个明显突出的「山峰」。这就对应了三种分布，分

---

[1] 译者注：现实中很多数据都存在偏态，而右偏数据的一个经典例子就是收入。大多数人的收入都在一个合理的区间中，但是有少数富人收入却非常高。所以如果画一个直方图，就可以随着横轴收入从零到非常高的水平，柱子是先快速变高，然后接着不断变低，并且最后在最右侧还会有几个非常低的柱子（代表收入很高但是人数很少），即右侧拖着一条长长的尾巴。

[2] 其实硬要说的话，这两张图也可以看出数据是右偏的。不过显然一维点图是最不易于看出偏度的。相比之下，堆积点图要好一些，直方图最方便。

[3] 每笔贷款利息率的具体数值。这里排除均值其实主要是因为在前面点图的案例中，本书用红色三角标注出了均值。其实如果只是标准的点图或者直方图，都是不能直接看出均值信息的（毕竟均值是个基于计算所得的统计量）。

[4] 其实原著此处 mode 似乎是想表示直方图里的一座座「山峰」，例如假设数据直方图左侧有个高的「山峰」，右侧有个低点的「山峰」，似乎作者是把这两个都算做了 modes。

别叫做：**单峰 unimodal**、**双峰 bimodal** 和**多峰 multimodal**。

任何一种多于两个峰的分布都属于多峰。在图 2.7 所示的单峰（最左图）分布中，我们看到除了明显突出的（频数为 15 的）山峰以外，还有一个次高的没那么突出的（频数为 10 的）山峰，但是对于这个次高的山峰，我们并不能将它称作统计学意义上的「峰」，因为它之比相邻柱子的频数只高出很少的几个观测值。

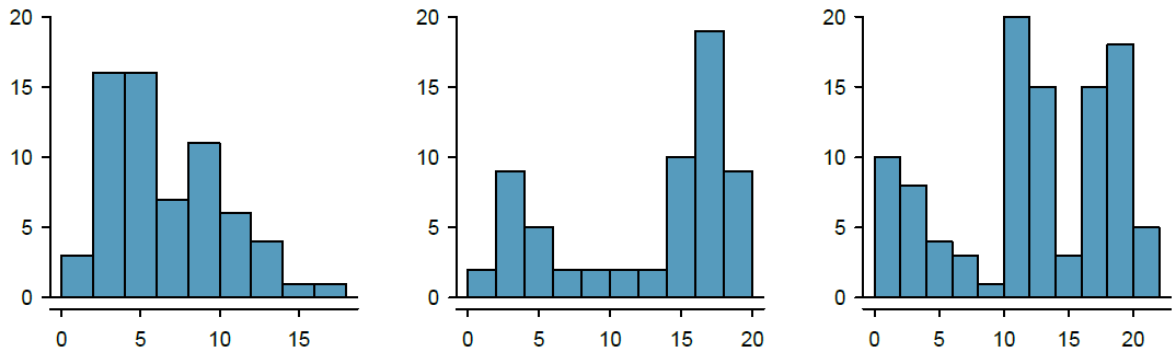


图 2.7：针对「明显突起」的山峰个数，从左到右分别对应了：单峰分布、双峰分布和多峰分布。对于最左边的这张图，因为我们只考虑「明显突起」的山峰个数，而非任何山峰都考虑，所以它只算单峰分布

示例 2.11

E 针对图 2.6 展示的利率分布，请判断它是单峰的、双峰的还是多峰的。

答案：单峰分布。

指导练习 2.12

G 如果我们对某小学 1-3 年级所有的学生和老师进行身高测量，再把测量结果整理成数据集，那么你认为这个数据集中有几个「峰」的可能性最大？[1]

很多时候，我们其实并不需要针对「峰」的个数给出一个确定的答案，这也是为什么本书中没有对「明显突起」给出很严谨的定义。更重要的是，观察峰的过程，其实是我们更好地了解数据的过程。

2.1.4 方差和标准差

我们之前讲到，均值是用来描述一组数据的中心的统计量，而数据间的差异也同样重要。我们

[1] 有两个峰的可能性最大。一个峰是学生身高形成的，另一个峰是老师身高形成的，也就是说这个数据很可能呈现双峰分布。

接下来会讲到两个描述数据间差异的统计量：方差和标准差。这两个统计量在数据分析中都非常有用，不过它们的计算公式稍微复杂了点。

我们把观测值和均值之间的差异称作**偏差 deviation**。以下列举了利息率变量第一、第二、第三和第五十号观测值的偏差：

$$\begin{aligned} x_1 - \bar{x} &= 10.90 - 11.57 = -0.67 \\ x_2 - \bar{x} &= 9.92 - 11.57 = -1.65 \\ x_3 - \bar{x} &= 26.30 - 11.57 = 14.73 \\ &\vdots \\ x_{50} - \bar{x} &= 6.08 - 11.57 = -5.49 \end{aligned}$$

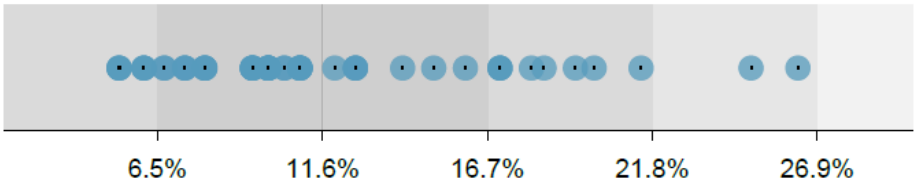
如果我们把每个观测值的偏差先取平方，再计算出这些平方的均值，就得到了样本的**方差 variance**，用s^2来表示。

$$\begin{aligned} s^2 &= \frac{(-0.67)^2 + (-1.65)^2 + (14.73)^2 + \cdots + (-5.49)^2}{50 - 1} \\ &= \frac{0.45 + 2.72 + 216.97 + \cdots + 30.14}{49} \\ &= 25.52 \end{aligned}$$

当我们在计算样本方差的时候，分母是 $n-1$ 而不是 n ，这会产生计算结果上的细微差别，但是用 $n-1$ 得到的结果会更准确[1]和有用一些。 我们给每个观测值的偏差取平方，这样会产生两个效果：首先，这么做让本身就比较大的值变得更大了，我们比较 $(-0.67)^2$、$(-1.65)^2$、$(14.73)^2$ 和$(-5.49)^2$ 就能看出；其次，这么做消除了负号，只能得到非负值。**标准差 standard deviation** 是由方差开方计算得到的：

$$s = \sqrt{25.52} = 5.05$$

在我们用符号 s^2 和 s 来表示方差和标准差的时候，也可以通过加上脚注来说明 s_x^2 和 s_x 是针对 $x_1, x_2, \ldots, x_n$ 这些观测值的。和均值同理，总体的方差和标准差也用专门的符号表示：σ^2表示总体的方差，表示总体的标准差。σ是希腊字母 sigma。



利息率，$\bar{x}$ = 11.57%，s_x = 5.05%

图 2.8：对于利息率这个变量，50 笔贷款中，34 笔的利息率都落在（离均值的）一个标准差以内，48 笔都落在（离均值的）两个标准差以内。通常情况下，70%的观测值都落在一个标准差以内，95%都在两个标准差以内，但也并非所有情况都是这样

[1] 译者注：至于到底为什么用 得到的结果更准确，可能因为背后的原理比较复杂，这里作为入门阶段就不详述了，建议自行搜索。

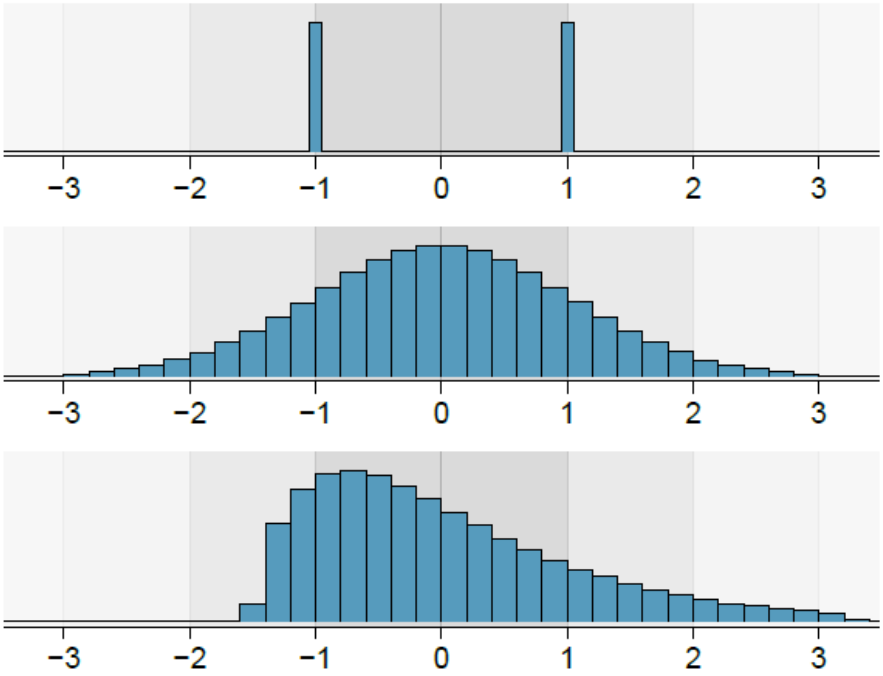


图 2.9：三种非常不同的分布，它们的均值和标准差完全相等，均值 $\mu = 0$，标准差 $\sigma = 1$

指导练习 2.13

之前我们介绍了分布形状的概念。一个好的分布形状描述应该包含分布的形态（单峰、多峰等等）和分布的偏态（左偏、右偏、对称等等）。以图 2.9 为例，你能解释下为什么这两个维度的描述缺一不可吗？[1]

2.1.5 箱形图，四分位数，和中位数

箱形图 boxplot 会在图上展示五个统计量来总结数据信息，同时异常值（离均值较远）也会以点的形式被标记出来。图 2.10 展示了一张箱形图，同时在箱形图的左边用浅蓝色的点绘制了一张竖直的点图，与箱形图形成对照。这两张图都是基于 loan50 数据集中的「利息率」变量绘制的。

[1] 在图 2.9 中，我们看到三种非常不同的分布。但是每个分布都是有一样的均值、方差和标准差的。使用形态描述，我们就可以区分（从上至下）第一个单峰分布和第二个双峰分布。使用偏态描述，我们就可以区分最后一个右偏分布和前两个对称分布。而直方图则是完美包含这两种描述以及更多信息的「完全体」。通过直方图我们可以把数据分布的故事讲得更加完整。不过即使不依赖直方图，我们也可以通过形态和偏态这两个维度来描述一个分布的基本特征。

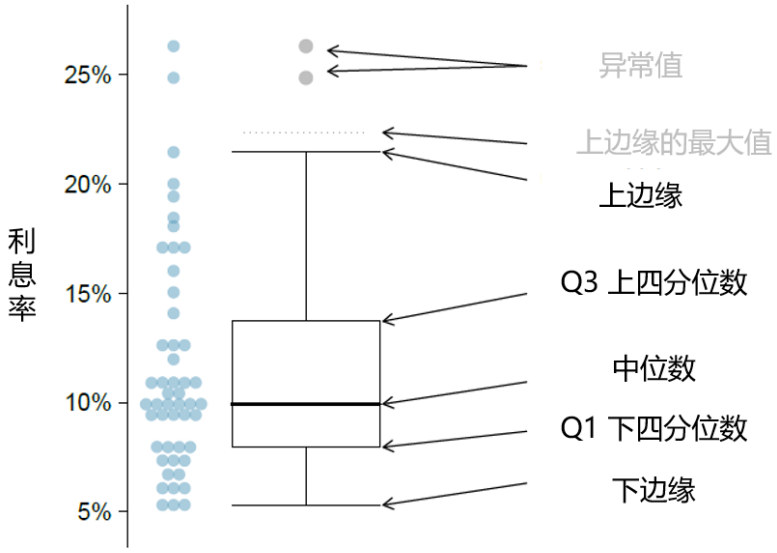


图 2.10：「利息率」变量的竖直点图和箱形图（带标签）

绘制箱形图的第一步是把**中位数 median** 用一条黑色的线段在正中画出来。这条线段会把所有数据点对半分成上下两部分。通过图 2.10 可以看到，右侧箱形图中间的中位数线段，把左侧点图的点对半分割。因为 loan50 数据集里面有 50 个观测值，所以上下两部分各有 25 个观测值落入其中。这种情况下，中位数的取值是最接近中间（第 25 位和第 26 位）的取值的平均数。而在本数据集中这两个取值正好是一样的，所以中位数取值就是：(9.93%+9.93%)/2 = 9.93%。如果数据集中观测值的个数是奇数，那么对数据从小到大顺序排列后，应该正好有一个值可以把数据分成两半，这种情况下我们就不用再取平均，把数据平分的那个值就是中位数（例如 7 个观测值中的第 4 个，恰好把数据等分成 1-3 和 5-7 两部分）。

中位数：正中间的那个「它」

我们把数据从小到大排列，中位数就是正中间的观测值。如果总共有偶数个点，那么会同时有两个数位于中间位置。这样中位数就取它们的算术平均值就好。

绘制箱形图的第二步是画一个长方形，代表了靠近中间的那 50%数据的范围。这个长方形又被称为「箱 box」，箱形图的取名也是由此而来。这个箱状图形的长度被称为**四分位距 interquartile range**。英文中经常简称其为 IQR。它和标准差的作用相似，可以衡量数据的离散程度。数据越离散，标准差就会越大，而一般来说四分位距也会越大，直观体现在图上就是中间的箱形区域很长。这个箱形区域的上下边界对应了**上四分位数 the third quartile** 和**下四分位数 the first quartile**。上四分位数代表了有 75%的数据小于这个值，而下四分位数则对应了 25%下方的数据。在箱形图中，我们往往使用 Q_3 和 Q_1 来标识上下四分位数。

四分位距（IQR）

四分位距是箱型图中箱的高度，它的计算公式是：

$$IQR = Q_3 - Q_1$$

Q_3 和 Q_1 分别对应了第 75 百分位数和第 25 百分位数。

G

指导练习 2.15

有百分之多少的数据落在 Q1 和中位数之间？又有多少落在中位数和 Q3 之间？[1]

接着我们看延伸到箱形区域外面的数据，我们用**须子 whisker**（直译，通俗一般就称上下边缘）来反映它们。图 2.10 的上边缘和下边缘都对应了两根须子，他们不代表数据的极大值和极小值，而是代表了非异常值的范围。约定俗成地，这两根须子到最近的四分位数的距离不会超过 1.5 倍的 IQR。比如图 2.10 中，上方的须子就是先找到所有取值不大于 Q_3+1.5 倍 IQR 的点，然后在满足条件的、取值最大的一个点处画一条线段。需要注意的是，上边缘的须子位置不一定非要正好在 Q_3+1.5 倍 IQR 处。如图所示，虚线代表了上边缘能取到的最大值，即 Q_3+1.5 倍 IQR。而实际上并没有数据点取到这个值。在上边的须子上方，还有两个灰色的点。这两个点就是因为取值大于 Q_3+1.5 倍的 IQR，所以被标记为异常值。

同理我们再来看下方边缘。由于没有任何点取值小于 Q_1-1.5 倍 IQR，所以该箱形图下边缘的下方就没有数据点。因此也就没有再用虚线画出 Q_1-1.5 倍 IQR 的位置（因为画上去就有点画蛇添足了）。

在上下须子边缘外的任何观测值，即**异常值 outliers** 都被用一个一个点来标记出来。其余的数据则无需再用点的形式画在箱型图上。这样做的目的是方便识别那些距离「数据大部队」较远的观测值，同时减轻中间部分核心信息的提取压力。在这张箱形图案例中，利息率是 24.85%和 26.30%的点被标为异常值，这两个数字也确实高得有些离谱了。

异常值往往很极端

我们一般也会说异常值相比其他数据来说是很极端的。分析异常值很有用，比如：

1. 有助于识别为什么分布有明显的偏度；
2. 有助于找到数据收集或者录入中的明显错误；
3. 有助于帮助我们发现数据一些有趣的特性（比如收入数据，亿万富翁们明显属于异常值，但是这就是收入数据的特点：有少数人拥有很多财富）。

[1] 分别各是 25%。

指导练习 2.16

请使用图 2.10，目测估计 loan50 数据集中利息率的 Q1，Q3 和 IQR。[1]

2.1.6 统计量的稳健性

你觉得，前面讨论的利息率数据会受到异常值 26.3%什么样的影响？如果实际上正确的利息率只有 15%，但是却被阴差阳错输入成了 26.3%，那会发生什么？而如果实际上的利息率比 26.3%还要高，比如 35%，那么这种输入错误又会对数据的概括统计量产生什么样的影响？我们把这些情形绘制成图表，如图 2.11 所示。同时，我们在图 2.12 中计算了每种情形下的一些相关统计量：

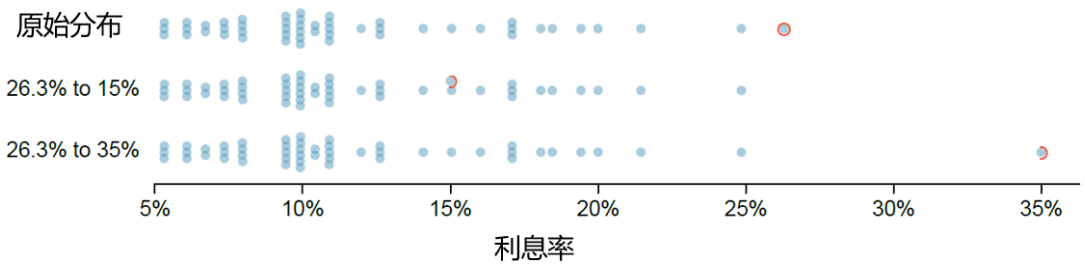


图 2.11：原始「利息率数据」和修改后的数据的对比点图

		稳健		不稳健	
	情形	中位数	IQR	均值	标准差
原始	利率数据变动	9.93%	5.76%	11.57%	5.05%
情形1	26.3% -- > 15%	9.93%	5.76%	11.34%	4.61%
情形2	26.3% -- > 35%	9.93%	5.76%	11.74%	5.68%

图 2.12：四个统计量的稳健性检验对比：在「利息率」数据的一个极端值改变之后，整个样本的中位数，IQR，均值和方差会发生什么样的变动

指导练习 2.17

两个问题：

(a) 均值和中位数，谁更容易受到极端值的影响？

(b) IQR 和标准差，谁更容易收到极端值的影响？[2]

[1] 目测可能 Q1 大约是 8%，Q3 为 14%，IQR 就是 6%。真实值是 Q1 为 7.96%，Q3 是 13.72%，IQR 是 5.76%。

[2] 均值更易受影响，也就是更不稳健；标准差更易受影响，相比 IQR 更不稳健。

根据上面的图片，我们不难发现，中位数和四分位距，即 IQR 这两个统计量更加**稳健 robust**。因为极端值的改变对它们变动的影响非常小。反过来看均值和标准差，在我们修改极端值的时候，它们都或多或少发生了波动。仅修改一个极端值就能够引起这两个统计量的变动，因此我们说均值和标准差对极端值的变化很敏感。在某些特殊情形下，这种敏感性尤其值得我们注意。

示例 2.18

中位数和四分位距，即 IQR，在图 2.12 中并没有改变，为什么会这样呢?

E

答案：因为中位数和和四分位距仅仅对四分位距区域内（即 Q_1 和 Q_3 之间）的数据敏感，而在我们修改极端值的时候，没有引起这些区域内数据的变动，因此也就导致了中位数和四分位距的取值相对稳定。

指导练习 2.19

G

loan50 数据集中的贷款额的分布是右偏的（有几笔大额贷款使得右侧拉出一条长长的尾巴）。在这种情况下，如果我们想知道比较典型的贷款额大致是多少，我们应该更看重均值还是中位数？[1]

2.1.7 转换数据（特别话题）

当数据呈现出很大偏度的时候，有时候就需要我们对数据进行转换。

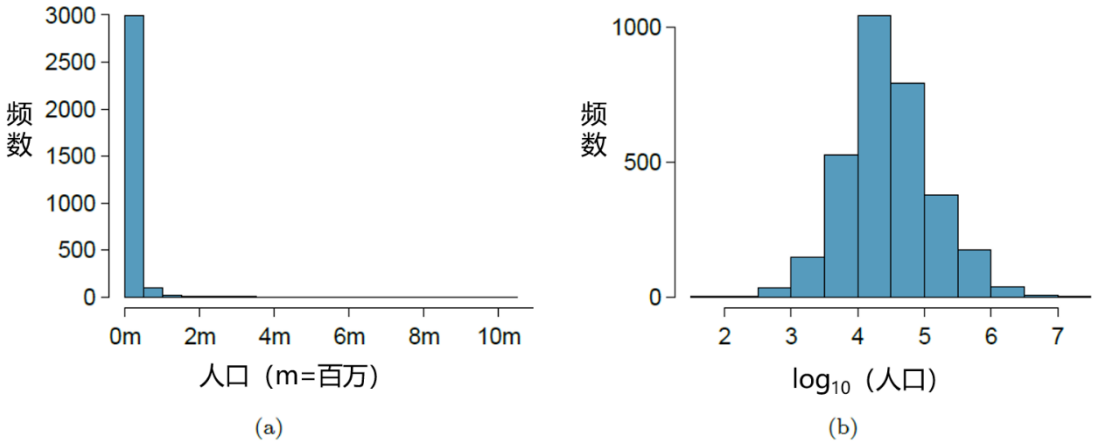


图 2.13：(a)基于美国各郡人口的直方图；(b)基于各郡人口取（以 10 为底的）对数之后的直方图。对于图(b)来说，x 轴表示 10 的次方数，例如：x 轴上的 4 表示 10^4=10,000

[1] 视情况而定。如果我们只是想知道一笔贷款一般是多少，看中位数会更准确些。但如果我们想进行一些计算，比如：如果要提供 1000 笔贷款，我们需要有多少资金？这个问题用均值计算会更好一些（因为我们要考虑到极端值的情况，而均值能够反应极端值带来的影响，而中位数不能）。

示例 2.20

E 在分析美国各郡人口数据时，图 2.13(a)呈现出极度右偏，这样会对我们的分析带来哪些不便呢？

答案：几乎所有的数据都落在了最左侧的箱中，这样我们就很难看出很多数据分布上有意思的细节。

在处理极度右偏数据（尤其是大部分数据都接近于零）的时候，我们可以对数据进行转换。**转换 Transformation** 是指使用函数对数据进行重新缩放。例如，如图 2.13(b)所示，是对美国各郡人口取以 10 为底的对数后得到的新的分布。这样得到的新数据是对称的，并且和源数据相比，任何极端值都显得不那么极端了。通过控制异常值和极端偏差，这样的转换通常可以让我们更轻松地针对数据构建统计模型。

除直方图外，我们也可以对散点图涉及的一个或两个变量进行转换。图 2.14(a)展示了美国各郡 2010 到 2017 年的人口变化，x 轴为 2010 年各郡人口（变化前人口）。从图 2.14(a)中，我们很难得到有价值的信息，因为人口变化这个变量呈现出很极端的偏差。然而，如果我们对这个变量取以 10 为底的对数，就得到了图 2.14(b)。从图 2.14(b)中，我们可以清楚的看到一些正相关性。而且，我们还可以进一步分析它们之间的线性回归关系，关于这部分内容我们会在第八章里讲到。

除了取对数之外，还有很多其他的数据转换方式。比如，取次方根和取倒数也都是很常见的方式。进行数据转换的目的一般包括：换种方式查看数据结构、减少数据分布偏差、帮助建模分析、把非线性数据转化成线性关系等等。

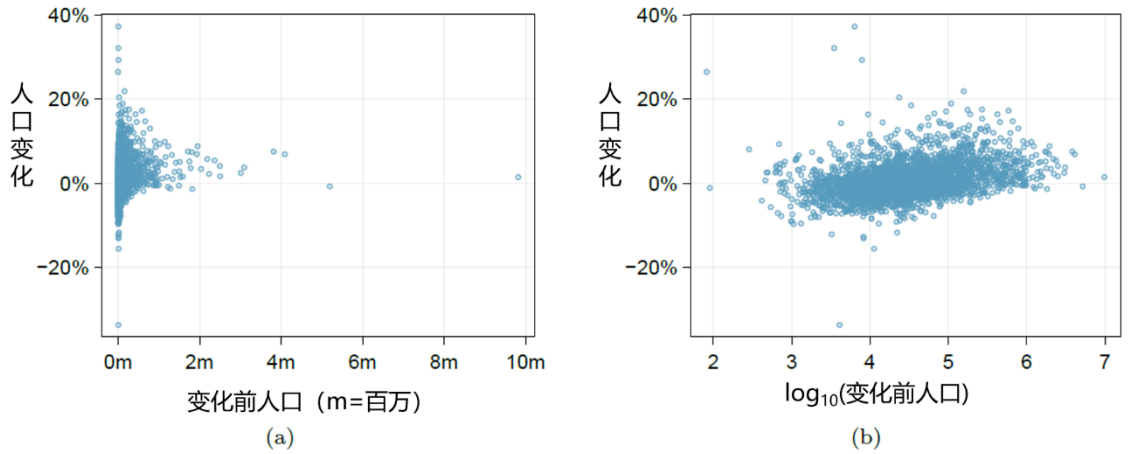


图 2.14：(a)基于人口变化百分比和变化前人口数量之间关系的散点图；(b)基于人口变化百分比和取对数后的变化前人口之间关系的散点图

2.1.8 制作数据地图（特别话题）

在 county 数据集中，我们可以对很多数值型变量制作点图、散点图或者箱型图，但无论以上哪种图其实都没办法展示数据全貌。因为这些数据都是基于每个郡县，或者说，是基于地理区域的数据。这种时候，我们就可以绘制一个**密度地图 intensity map** 来用不同颜色表示变量的大小变化。图 2.15 和图 2.16 展示了四张密度地图，其包含的变量信息依次是：贫困率，失业率，房屋拥有者比例，和家庭收入中位数。在地图右侧的图例标明了不同颜色对应的值的大小。尽管密度地图在获取特定郡县的数字时稍逊一筹，但是它却可以较好呈现变量在地域上分布的趋势，进而帮助我们去构思一些有趣的研究问题。

示例 2.21

从下方的密度地图中，你对贫困率和失业率的地理分布有什么有意思的发现吗？

答案：可以看出，贫困率在某几个地方显著地高。具体点儿说，在很靠南的位置，例如亚利桑那州和新墨西哥州的一些郡县的颜色明显比别的地方要深。此外，还有其他部分区域，例如密西西比州和肯塔基州的贫困率也较高。

E

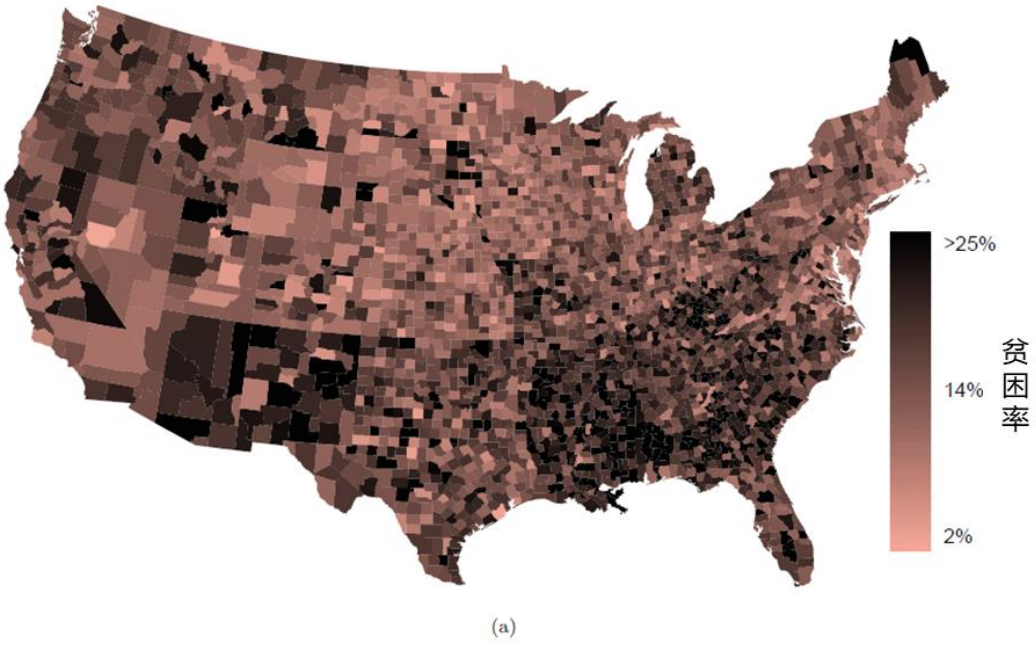
失业率的话也呈现类似的趋势，通过这种趋势的相似性，我们也可能看出「贫困率」和「失业率」这两个变量间的相关关系，而且这种关系很好说得通。此外，观察这两张图还能得到一个结论：从数值上来说，贫困人口的百分比是比失业人口的百分比要高的。这说明了在一些人工作的时候，他们的收入可能并不够高，以至于他们还是陷入了贫困的陷阱中。

指导练习 2.22

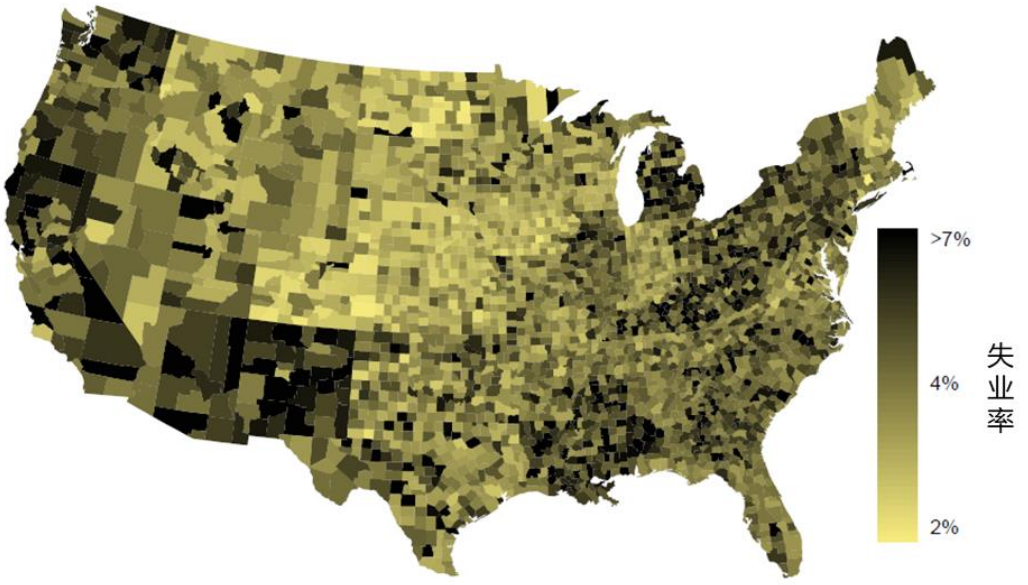
G

从图 2.16 可以看出家庭收入中位数的分布有哪些有趣的特点？[1]

[1] 本题无固定答案。可以看到大城市里的人们往往收入也更高（尽管也有部分例外情况），这些地方在图上呈现出更加深的颜色。所以我们或许可以通过寻找颜色较深的点，来分辨出美国的大城市都在哪里。

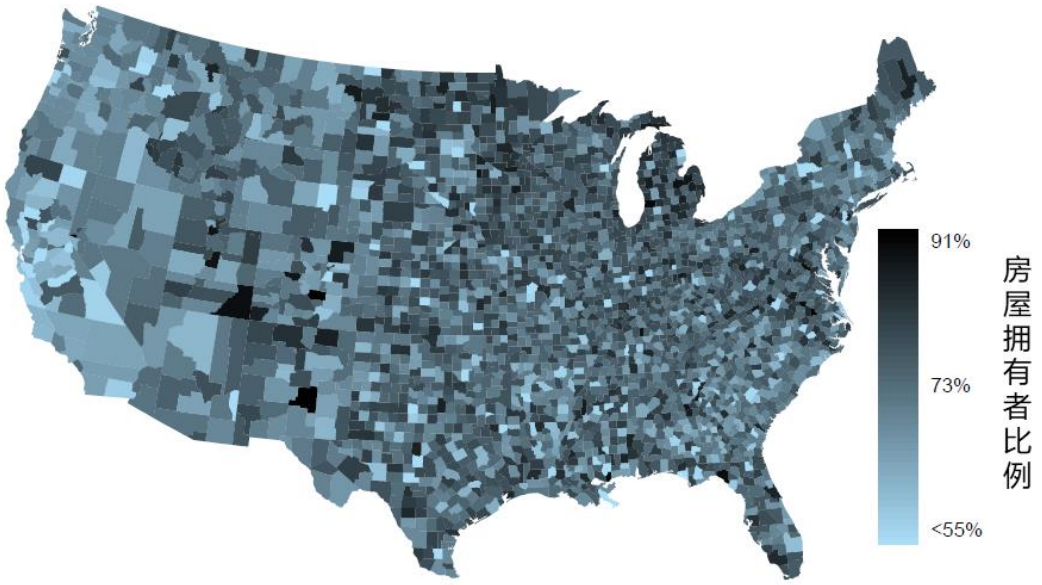


(a)

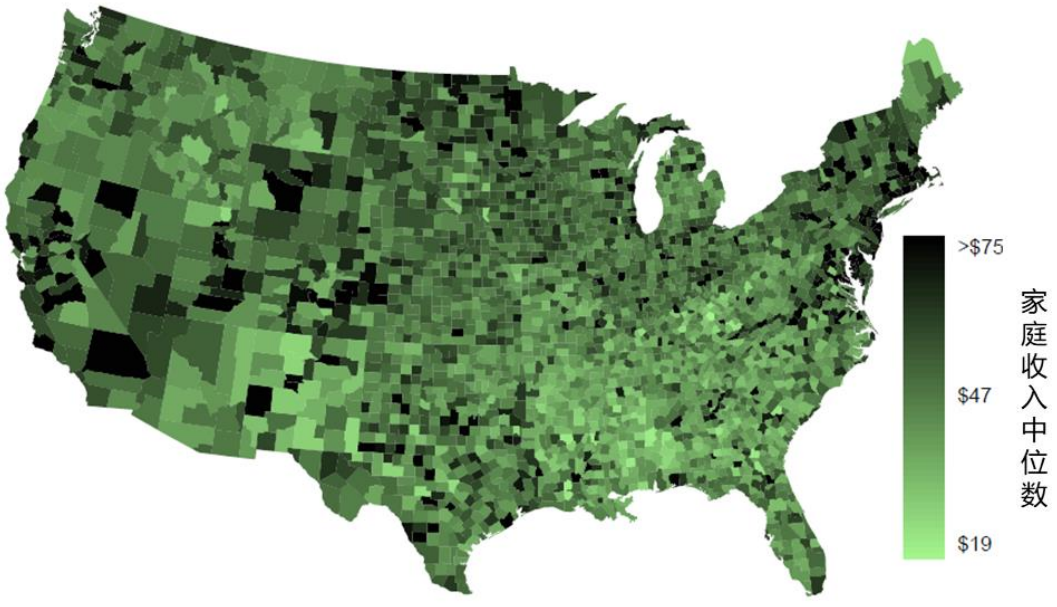


(b)

图 2.15：(a)贫困率（百分比）(b)失业率（百分比）



(a)



(b)

图 2.16：(a)房屋拥有者比例（百分比）(b)家庭收入中位数（$1000s）

2.2 研究分类数据

本环节中，我们会介绍一些统计分类变量的方法，包括使用表格和一些其他的基础工具。之前讨论的个人贷款数据集 loan50，它其实是从一个名为 loans 的更大的数据集中选出的 50 笔贷款的信息。这个名为 loans 的数据集中有 10000 笔贷款，它们的来源是一个名为 Lending Club 的美国 P2P 贷款平台。在本环节中，我们将以这个名为 loans（注意不是 loan50）的数据集为例，研究其中名为「住房情况」和「申请类型」的两个分类变量。在 loans 数据集中，「住房情况」这个变量可以取到的值包括：「租房」，「抵押贷款」（拥有房屋但是房贷还没还清），「自己拥有」。「申请类型」则分为「个人独立申请」和「联合申请」两种类型（联合申请代表贷款[1]并非由借款方独立申请，而是和其他伙伴一起发起的申请）。

2.2.1 列联表和柱形图

图 2.17 展示了「住房情况」和「申请类型」两个变量间一些统计数据。一张像这样统计两个分类变量的表被称作**列联表 contingency table**。每个表中的数字都代表：在特定的组合下，观测到的数据集中满足条件的情形总数。例如 3496 表示了在 loans 数据集中，借款者是租房同时进行个人独立申请的贷款共有 3496 笔。观察该列联表的最右侧和最下行，可以看到每行和每列的总数。**行总计 row totals** 计算的是同行内所有数字相加之和（例如，3496+3839+1170=8505），而**列总计 column totals** 计算的是同一列的数字之和。基于列联表的思维，我们可以把图 2.17 中的数字都替换成占总数的百分比。我们也可以单独制作一张表，其中只包含按照一个变量分类统计的信息。

		住房情况			
		租房	抵押贷款	自己拥有	行总计
申请类型	个人独立申请	3496	3839	1170	8505
	联合申请	362	950	183	1495
	列总计	3858	4789	1353	10000

图 2.17：「住房情况」和「申请类型」的列联表

[1] 译者注：这里的贷款是指 loans 数据集里的贷款，而非房屋抵押贷款，所以不要和「住房情况」变量的「抵押贷款」取值搞混。

住房情况	个数
租房	3858
抵押贷款	4789
自己拥有	1353
总数	10000

图 2.18：「住房情况」分类的频数统计表

当我们需要展示单个分类变量的频数分布的时候，**柱形图 bar plot** 是不二选择。图 2.19 的左图就展示了一张对「住房情况」变量绘制的柱形图。在右侧的图上，我们把频数转换成了百分比（例如，对于借款人是「租房」的贷款，其比例是：3858/10000 = 38.58%）。

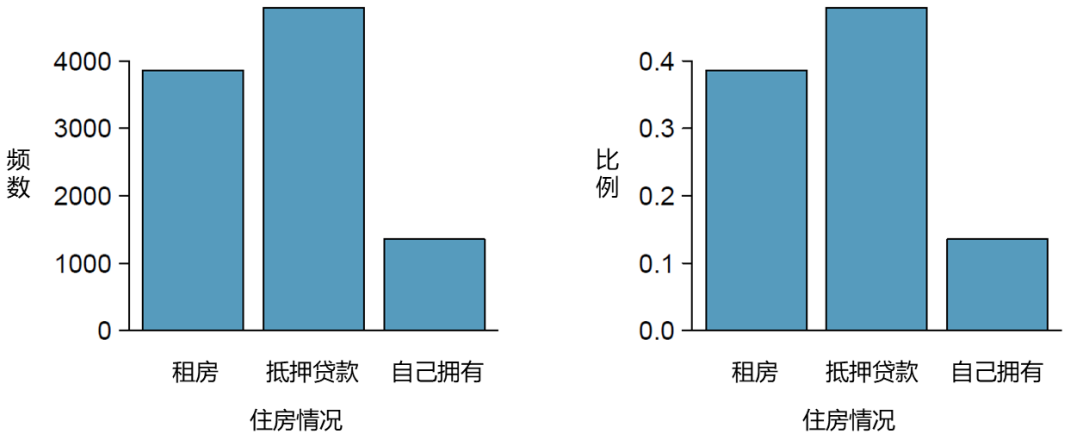


图 2.19：两张「住房情况」变量的柱形图，左侧展示的是频数，右侧展示的是每组比例

2.2.2 行和列的比例

有时候除了关心单个变量的取值比例，了解一个变量在另一个变量分类下的细分比例也很有用。为了实现这点，我们只需要对现有的列联表稍作调整。图 2.20 就是针对图 2.17 取**单行比例 row proportions** 计算得到的结果。其计算方式是把每个数字除以对应的行总计。比如，之前的 3496，作为借款方是租房同时进行个人独立申请的贷款数，在本张表中就被 3496/8505 = 0.411 代替。那么这个数字代表了什么呢？它代表着所有以个人名义独立申请的贷款中，有如此比例的贷款借款人是在租房。

	租房	抵押贷款	自己拥有	行总计
个人独立申请	0.411	0.451	0.138	1.000
联合申请	0.242	0.635	0.122	1.000
列总计	0.386	0.479	0.135	1.000

图 2.20：一张注明了单行比例的列联表，由于四舍五入的问题，其中有些比例数字相加并不完全相等（第二行联合申请三个数字相加只有 0.999）

除了可以制作单行比例的列联表之外，我们也可以依样画葫芦制作**单列比例 column proportion** 的表格。同单行比例原理一样，我们可以把图 2.17 频数表中的数字除以对应的列总计，以得到单列比例。图 2.21 所示的表格就是单列比例的列联表，其中左上角的 0.906 代表在所有借款方是租房居住的贷款中，绝大多数（有 90.6%）的贷款都是个人独立申请的。我们也通过这张表进行横向比较，发现相比于借款人住房情况是「抵押贷款」（80.2%）或者「自己拥有」（86.5%），借款人是租房的贷款中，个人独立申请的比例更高[1]。当借款人的住房类型不同时，个人独立申请占到的比例也不同，所以我们可以把这当做「住房情况」和「申请类型」两个变量相关的证据之一。此外，我们不仅通过列比例表来探寻相关性，也可以通过上面的行比例表发掘一样的信息。

	租房	抵押贷款	自己拥有	行总计
个人独立申请	0.906	0.802	0.865	0.851
联合申请	0.094	0.198	0.135	0.150
列总计	1.000	1.000	1.000	1.000

图 2.21：一张注明了单列比例的列联表，由于四舍五入的问题，其中有些比例数字相加并不完全相等（第二行联合申请三个数字相加只有 0.999）

指导练习 2.23

(a) 图 2.20 中的 0.451 代表着什么?

(b) 图 2.21 中的 0.802 代表着什么?[2]

指导练习 2.24

(a) 图 2.20 中的 0.122 代表着什么?

(b) 图 2.21 中的 0.135 代表着什么?[3]

[1] 译者注：首先无论借款人住房是哪种类型，个人独立申请的贷款比例都比较高。说明整个平台还是个人独立申请贷款是主流。其次，租房类型的个人申请比例更高可能是因为租房的借款人可能单身的更多，所以没有另一半来联名申请。

[2] 0.451 是个人独立申请者中以抵押贷款方式买房的比例；0.802 是以抵押贷款方式买房的申请者中，个人申请所占比例。

[3] 0.122 是所有联合申请的贷款中，申请者自己拥有房屋占的比例比例；0.135 所有借款方自己拥有房屋的贷款中，联合申请的申请者占的比例。

示例 2.25

数据科学家们会尝试用统计方法来过滤邮件中的的垃圾邮件。通过检索邮件中的「某些特征」，我们可能可以把邮件分成「垃圾邮件」和「非垃圾邮件」两类。这些所谓的「特征」指：邮件中是否不包含数字，或者是否有一些很小或者很大的数字；邮件内容有没有 HTML 格式的信息，尤其是字体加粗标注的超链接。现在有这样一个包含很多封邮件的名为 emails 的数据集，我们可以关注它其中的两个变量：「邮件格式」和「是否是垃圾邮件」。在图 2.22 展示的列联表中，可以看到这两个变量的统计信息。那么问题来了：如果想要依赖这张表格来得出一个分类依据，数据科学家们应该更加关注单行比例还是单列比例?

E

答案：从逻辑上来说，我们应该更关注每类格式中垃圾邮件的比例，而不是垃圾邮件中不同格式占有的比例。所以，应该去计算单列比例会更有帮助。

通过计算单列比例，我们可以得出一个结论：就是纯文本格式的邮件中垃圾邮件占比更大。纯文本格式邮件中，垃圾邮件的比例是 209/1195 = 17.5%，而含有 HTML 的邮件中，垃圾邮件比例是 158/2726 = 5.8%。当然，应该明确这个结论其实不能直接帮助我们进行垃圾邮件的判断，因为哪怕是纯文本邮件，也依然有 80%以上的比例不是垃圾邮件。而且根据常识，显然我们不能仅靠邮件格式来做垃圾邮件的判定。尽管如此，通过列联表得到的信息还是非常有用的，我们可以把它和其他维度的信息结合，通过更多变量的分析讨论，从而更合理、自信地进行垃圾邮件的自动判定。

	纯文本	HTML	行总计
垃圾邮件	209	158	367
非垃圾邮件	986	2568	3554
列总计	1195	2726	3921

图 2.22：「是否是垃圾邮件」和「邮件格式」的列联表

通过示例 2.25，我们想传递一个信息：就是单行比例表和单列比例表并不能完全等价。在我们确定使用一种类型的表格统计数据前，应该谨慎考虑选择哪种架构的表，尽管有时候这个选择并没有想象中的直观。

示例 2.26

回到图 2.20 和图 2.21 的研究情形中，在研究「住房情况」和「申请类型」变量的时候，我们能清晰地判定某种表（单行比例/单列比例）会更有用吗?

E

答案：并不能。讨论「住房情况」和「申请类型」变量，与垃圾邮件案例不同的是，我们无法清晰地判定哪个变量是响应变量，哪个又是解释变量。通常来说，在做比例列联表的时候我们会把解释变量当做条件，计算另一个变量的比例。例如，在垃圾邮件案例中，「邮件格式」显然是解释变量，所以我们就把它当做条件，计算「是否是垃圾邮件」的比例，也就是使用单列比例列联表。而当变量关系不明的时候，自然很难判断哪种结构的比例表更有用了。

2.2.3 涉及两个变量的柱形图

计算单行或者单列比例的列联表对于研究两个变量间的关系非常有帮助。而如果从可视化的角度来说，就不得不谈一谈**堆积柱形图 stacked bar plot**。

堆积柱形图对列联表进行了直观的可视化呈现。例如，在图 2.23(a)中，我们首先基于「住房情况」绘制一张柱形图，数据将呈现三列的样式。接着，我们把每列都依据「申请类型」分成两部分，分别用黄色和蓝色标示。

与堆积柱形图长相有些类似的另一种图叫**并列柱形图 side-by-side bar plot**，见图 2.23(b)。

我们介绍的最后一种柱形图叫**百分比堆积柱形图 percent stacked bar plot**，这是一种对堆积柱形图进一步标准化后绘制的图表。这种类型的图隐去了横向分布的不同柱子的高度信息，但是让柱子内的比例对比更加清楚。通过图 2.23(c)可以明显看到不同的借款人住房类型会对应不同的个人独立申请比例。

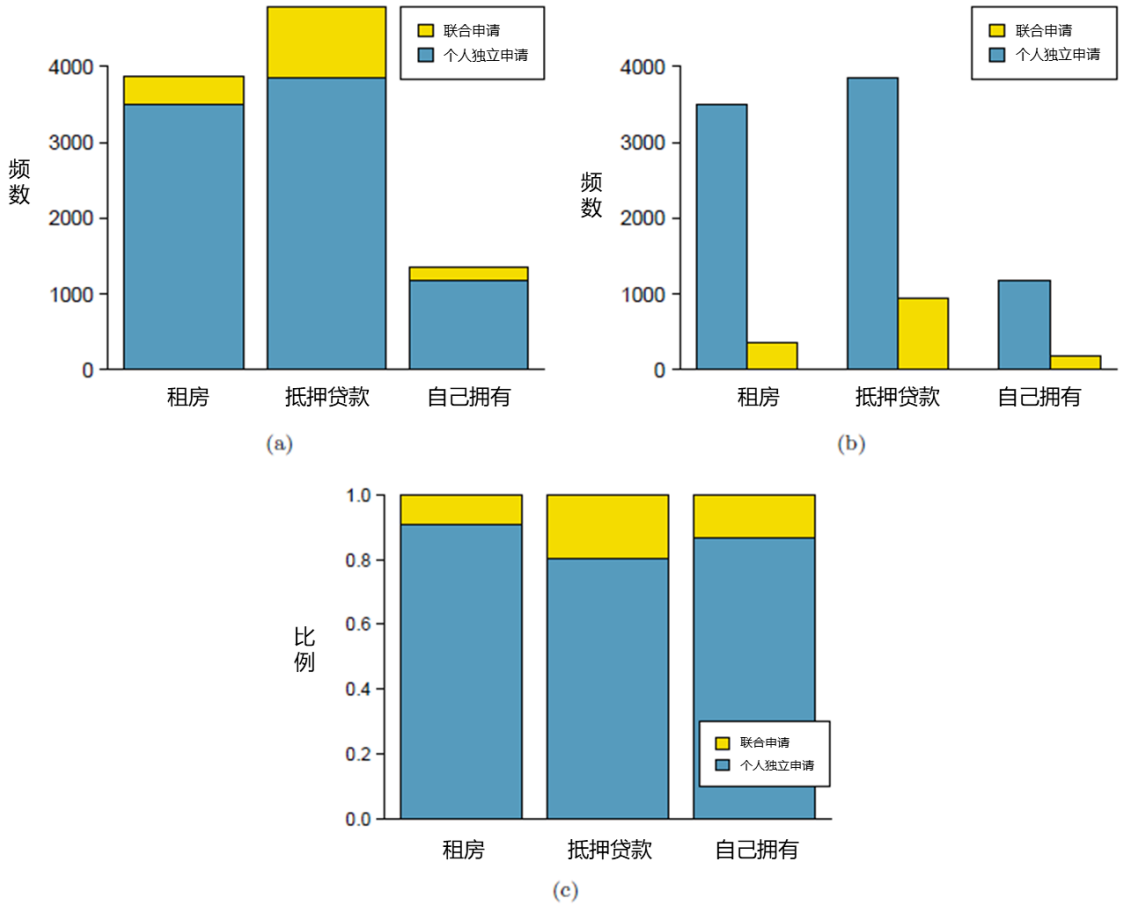


图 2.23：(a)「住房类型」的堆积柱形图，每个柱子按照「申请类型」分成两类；(b)并列柱形图；(c)百分比堆积柱形图

示例 2.27

请研究图 2.23 中的三张图表，你觉得分别在什么情况下每种图会更有用?

答案：堆积柱形图：在我们能够准确定义解释变量和响应变量的时候最有用，因为我们在绘制这张图的时候，需要先按照一个变量进行横向的列分类，然后把另一个变量用作柱子内的划分依据。

E 并列柱形图：如果使用并列柱形图，则比较难判定哪个变量是解释变量，哪个变量是响应变量。它的优势在于，很容易观察到（以上图为例）六个柱子每支单独的高度。但是，这也反映了它的缺陷：即要占用横向更多的空间，比如图 2.23(b)就显得有些局促。此外，如果当两个相邻的柱子差别的很大的时候，我们可能就不太容易观察到变量间的相关性。

百分比堆积柱形图：在横向分布的几根柱子的总高度差别很大（分布不均衡）的时候非常适用。例如上图中，住房类型是「自己拥有」的观测值总数大约只有「抵押贷款」类的三分之一（真的很多人贷款买房啊!），这就给判断比例关系增加了很多困难。这时候使用百分比堆积柱形图就能很清晰地看到比例，但是对应的「牺牲」就是我们不再能够直接从图上看出每根柱子代表的观测值频数。

2.2.4 马赛克图

如果在百分比堆积柱形图的基础上，想要看到每种分类的频数，**马赛克图 mosaic plot** 就是很不错的选择。它通过区域面积的大小来反映频数的信息。

那就让我们来一起绘制我们的第一张马赛克图吧！首先，我们把一个正方形的区域按照「住房类型」的三种分类划分成三列区域，如图 2.24(a)所示。每列都代表了一种借款人的住房类型，柱子的宽度反映了每种借款人住房类型对应的贷款数。可以从图上看出，借款人自己拥有房子的贷款数目要少于借款人抵押贷款买房的贷款数目。

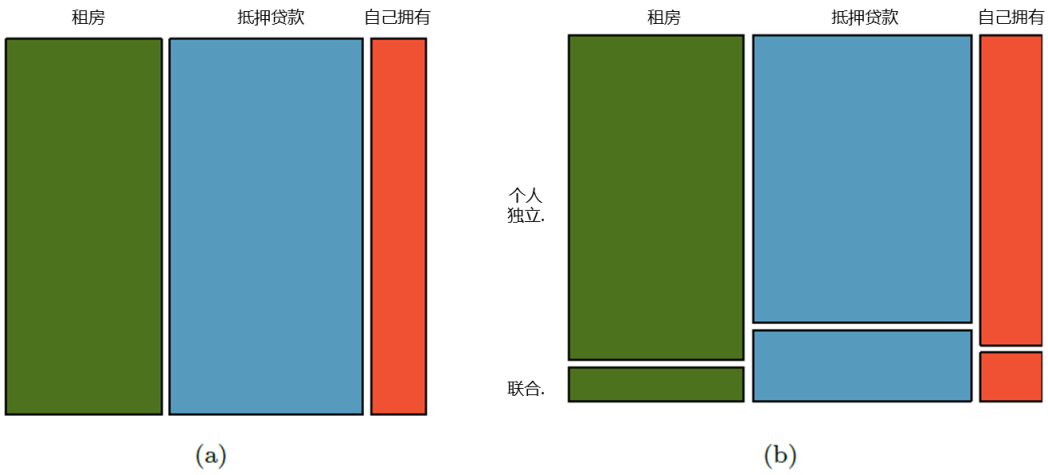


图 2.24：(a)单变量「住房类型 」马赛克图；(b)双变量马赛克图

接着，为了完成马赛克图，我们把图 2.24(a)的单变量马赛克图再用「申请类别」变量进一步分割，形成如图 2.24(b)样子的图表。在这张图中，每列的区域都按照个人独立申请和联合申请的贷款数比例分成两部分，上半部分对应个人独立申请，下半部分对应联合申请。我们除了可以用借款方的住房类型来分纵列，也可以用申请类型来分纵列，如下图 2.25 所示。和我们最开始讲柱形图一样，通常来说我们都是首先用解释变量来分列，然后再用响应变量把每个区域分成几行。

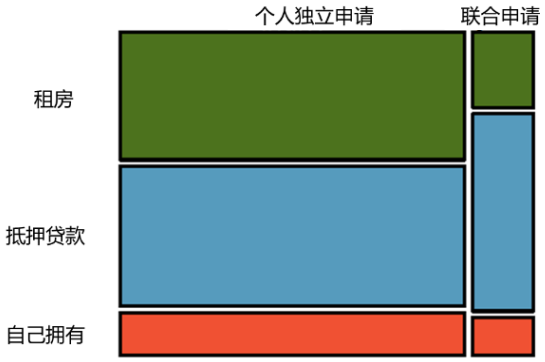


图 2.25：首先按「申请类型」分成两列，然后按照「住房类型」分成三行的马赛克图

2.2.5 本书唯一一幅饼状图

图 2.26 左侧展示了一张饼状图，右侧是一张和它展示同样信息的柱形图。饼状图在做分类概览的时候非常有用，不过，它的不足就是当我们想要进一步获取细节信息的时候，就会变得很困难。例如，考虑如下信息：借款人住房类型是「抵押贷款」对应的贷款数目比「租房」对应的数目更多。这个信息通过右侧的柱形图可以轻易获得，但是在左侧的饼图上却并不是那么明显，可能要多花好几秒钟盯着看才能反应过来。所以一般我们认为柱形图起到的作用是覆盖饼状图的，这也是为什么如果没有特殊的偏好和需求，我们在需要制作饼状图的时候，会倾向于用柱形图代替。

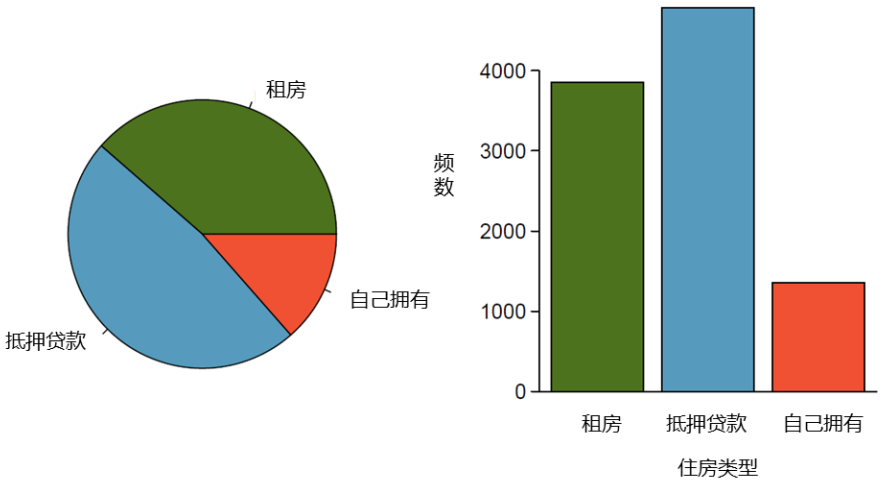


图 2.26：「住房类型」的饼状图和柱形图

2.2.6 比较多组间的数值型变量

通过分组比较数值变量，我们常常会有一些有意思的发现。而这里用到的知识我们前面都已经有所涉及：对数据进行分组和数值型变量的绘图。在此我们介绍两种很方便的比较方法：使用并排箱型图和空心直方图。

我们回到 county 数据集上，然后来比较一下各郡县的家庭收入中位数。不过这次，我们把所有郡县分成两组：2010 到 2017 年间人口有所增长的郡县，以及这期间人口没有增长的郡县。之所以如此分组，是因为我们想看看人口增长与否和收入增长之间有没有关系。不过要注意的是我们这里使用的是观察性数据（第 1 章环节 1.3.4 的概念），我们可能无法推断因果，而仅能窥探一下二者相关与否。

通过统计，一共有 1454 个郡县的人口在 2010 至 2017 年间发生了增长，同时也有 1672 个郡县人口没有增长（其中有一个人口持平，其余的发生了下降）。我们从有人口增长的郡县中随机抽出了 100 个，从没有人口增长的郡县中随机抽出了 50 个，接着把它们的家庭收入中位数列到了图 2.27 所示的表格中。大家可以通过这张表格感受下「家庭收入中位数」的源数据。

150个郡县的「家庭收入中位数」，千美元 ($1000s)

人口有增长						人口没有增长		
38.2	43.6	42.2	61.5	51.1	45.7	48.3	60.3	50.7
44.6	51.8	40.7	48.1	56.4	41.9	39.3	40.4	40.3
40.6	63.3	52.1	60.3	49.8	51.7	57	47.2	45.9
51.1	34.1	45.5	52.8	49.1	51	42.3	41.5	46.1
80.8	46.3	82.2	43.6	39.7	49.4	44.9	51.7	46.4
75.2	40.6	46.3	62.4	44.1	51.3	29.1	51.8	50.5
51.9	34.7	54	42.9	52.2	45.1	27	30.9	34.9
61	51.4	56.5	62	46	46.4	40.7	51.8	61.1
53.8	57.6	69.2	48.4	40.5	48.6	43.4	34.7	45.7
53.1	54.6	55	46.4	39.9	56.7	33.1	21	37
63	49.1	57.2	44.1	50	38.9	52	31.9	45.7
46.6	46.5	38.9	50.9	56	34.6	56.3	38.7	45.7
74.2	63	49.6	53.7	77.5	60	56.2	43	21.7
63.2	47.6	55.9	39.1	57.8	42.6	44.5	34.5	48.9
50.4	49	45.6	39	38.8	37.1	50.9	42.1	43.2
57.2	44.7	71.7	35.3	100.2		35.4	41.3	33.6
42.6	55.5	38.6	52.7	63		43.4	56.5	

图 2.27：这张表中，100 个人口有增长的郡县的家庭收入中位数（以千美元为单位）被列在左边。50 个人口没有增长的郡县的家庭收入中位数列在右侧

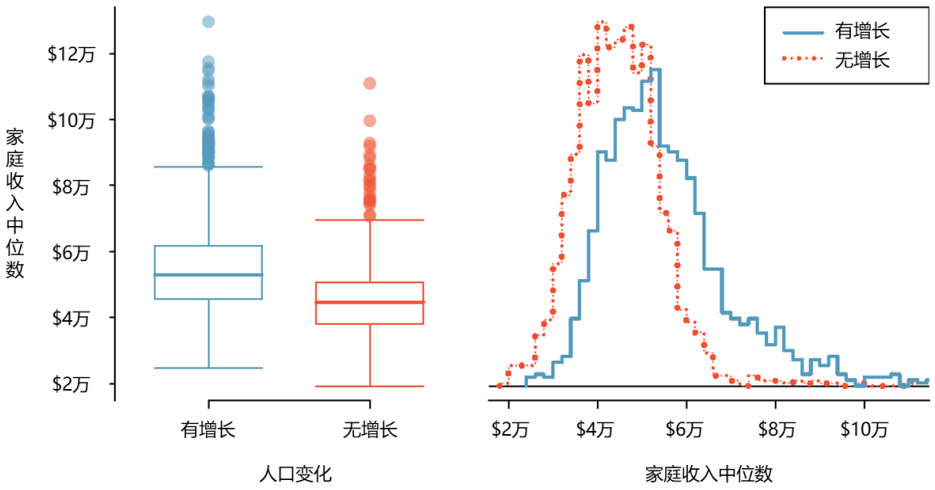


图 2.28：「家庭收入中位数」的并排箱型图（左）和空心直方图（右），每张图中的郡县被按照人口增长与否分成两部分

并排箱型图是一种跨组比较的很经典的工具。上面的图 2.28 中，可以看出人口发生了增长的郡县的收入中位数整体要比无增长的更高。注意绘制这种图的时候，尽量去保证左侧使用同一个坐标轴，让比较变得更容易。

右侧的空心直方图对于分组比较数值型变量也很有用。之所以用「空心」的就是为了避免重叠造成的信息不明。当然其实该图也可以用实心但是调整了透明度的直方图来代替。

指导练习 2.28

请使用图 2.28 来比较两个分组的郡县的家庭收入。关于每组收入的中心取值你有什么发现？关于每组收入的离散情况你有什么发现？两组的收入分布具有一致性吗？每组收入分组又有多少个「明显」的峰？[1]

指导练习 2.29

对于图 2.28 中的两幅图，你觉得每幅图中哪个部分最有用？[2]

[1] 答案可能不固定。参考答案是：人口发生增长的组别收入看起来也更高（左边箱型图中位数大约在$4 万 5，右边大约在$4 万）。同时人口增长组的数据也更离散（左侧箱型图的 IQR 更大）。两个组别的收入分布都有些右偏，同时也都是单峰的。此外，箱形图上也标注出了有很多离中心较远的点，不过这对于一个包含观测值总数有一百或者几百个的数据集来说并不让人感到意外。

[2] 答案可能不固定。参考答案是：箱型图中的中位数和 IQR 信息可能对于比较数据分布的中心以及离散情况很有用，而空心直方图可能对于观察分布形状，偏度和潜在的异常形态更有用。

2.3 案例分析：疟疾疫苗

示例 2.30

某位老师把教室里的学生分成了两组：坐在左边的的为一组，坐在右边的为一组。如果$\hat{p}_L$和$\hat{p}_R$分别代表左半边的学生和右半边的学生中拥有苹果产品的比例，那么$\hat{p}_L$和$\hat{p}_R$应该相等吗？

答案：$\hat{p}_L$和$\hat{p}_R$可能会比较接近，但是很可能不完全相等。

指导练习 2.31

如果你不认为一个学生「坐在教室的左边或者右边」和「他/她是否拥有苹果产品」有关系，那么这相当于对上述两个变量做了何种假设？[1]

2.3.1 统计结果的差异

本环节我们来考虑一项关于名为 PfSPZ 的新种疟疾疫苗的研究。该研究采用试验的设计，所有 20 名志愿者被随机分到两组中：其中 14 名志愿者接受了试验阶段的疫苗接种，其余 6 名志愿者则接受了安慰剂。19 周后，专家们让所有 20 名志愿者暴露于「对药物敏感的」疟疾毒株下（注：这里使用「对药物敏感的」病毒毒株是出于伦理道德考虑，从而确保感染可以被治疗）。试验的结果统计在下图 2.29 所示的表格中，其中试验组的 14 名志愿者有 9 人没有出现感染症状，而对照组的 6 人全部出现了感染症状。

		结果		
		感染	未感染	行总计
接种	疫苗	5	9	14
	安慰剂	6	0	6
	列总计	11	9	20

图 2.29：疟疾疫苗试验的概括性统计量

指导练习 2.32

该研究是观察性研究还是试验？研究类型对于研究结果有什么影响？[2]

[1] 相当于假设这两个变量间相互独立。

[2] 该研究是一个试验，因为志愿者被随机分到了一个试验组一个对照组。因为这是个试验，所以其结果可以被用于进行「接种疫苗」和「感染与否」的因果关系推断。

在该研究中，相比对照组，接种了疫苗的试验组只有很小一部分人出现了感染症状（35.7%对比 100%）。但是，由于该样本太小了，尽管我们可以尝试进行因果推断，却没有足够有说服力的证据说明疫苗是有效的。

示例 2.33

有时候，我们会要求数据科学家评估支撑统计结果的证据强度。当我们观察上面的感染率数据并试图评估结果的显著性的时候，脑子里会飘过什么念头？

E

答案：根据观察到的感染率数据（试验组的 35.7%对比对照组的 100%），说明该疫苗很有可能是有效的。但是，我们其实没办法完全确定这个数据是不是单次试验的偶然。因为即使背后的真相是疫苗无效，也就是按理说实验组和对照组感染率应该没有区别，我们也有可能观察到两组感染率结果不相同的情况（由于样本的选择和数据的波动）。此外，样本越小，试验的随机性也就越差，也就是说，偏差出现的可能也就越大。

示例 2.33 是一个小提醒，提醒大家通过哪怕是试验研究观察到的结果也不一定能完美反映两个变量的之间的关系。因为，现实中存在**随机噪声 random noise**，也就是自然存在的扰动项。其原理就和尽管我们知道掷硬币获得正反的概率相同，但是掷一千次硬币却几乎不会出现正反各五百次的情况。像上面图 2.29 中展示的数字，尽管两组感染率差别很大，由于样本很小（样本越小，随机噪声的影响也就越大），我们无法去判断这种差别是疫苗真的有效的体现还是说是一次试验的偶然。

统计学上为了去尽可能得出判断结论，就引入了置信水平、原假设和备择（备用选择）假设的概念。这些概念乍一听有些让人头大？别紧张，我们来慢慢展开：在统计学中，我们用 H_0 代表原假设，读作 H-nought；用 H_A 代表备择假设，读作 H-A。

H_0：**原假设 Independence model**。即假设「试验措施」变量和「结果」变量之间是独立的。这种假设判定它们之间并没有关系，而如果观察到了统计数据的差异（例如上例中 64.3%的感染率差距），那么该差异纯粹是由于偶然概率造成的。

H_A：**备择假设 Alternative model**。即假设「试验措施」变量和「结果」变量之间不是独立的，或者说，是相关的（可以看出 H_0 和 H_A 必有一个正确，它们互斥）。而如果观察到了统计数据的差异（例如上例中 64.3%的感染率差距），这种差异就说明「试验措施」起到了某种效果。

如果说原假设为真，也就是新疫苗其实对疟疾的感染率没有影响，那么意味着什么呢？它将意味着最后观察到的那 11 名出现感染症状的志愿者无论分到哪组，都会感染。也意味着剩余的 9 人，无论分组如何，都不会被感染。那么每组里面观察到多少感染病例，多少无感染病例，就完全取决于分组的概率了。比如，如果一不小心把 11 个「无论如何都会感染」的志愿者全分到疫苗组中，那么即使是试验组也会观察到 100%的感染率。

现在我们再考虑备择假设，也就是新疫苗能够预防疟疾感染。那么这又意味着什么呢？这将意味着，试验组因为接种了疫苗，所以总能观察到相对较少的感染比例。也就是说，我们再做几次试验，都应该观察到试验组的感染率比对照组的感染率要低。

在统计学研究的结论判断中，研究者往往会从上面两个假设中选择一个。而选择逻辑就是（下面这句话非常重要）：我们观察到的差异是否足够大，大到我们不认为原假设为真的前提下在一次随机试验中出现如此反常的差异，大到足以放弃原假设。如果差异确实足够大，并且数据也支持备择假设的判断，我们就会放弃原假设，选择备择假设，即判断疫苗是有效的。

2.3.2 模拟试验

如何判断观察到的差值是否足够大了呢？我们来模拟一下。我们首先假设疫苗是无效的，也就是意味着观察到的感染率数字差完全由随机分组的过程产生。那么我么尽可能多次地进行随机分组尝试和统计，看看上面出现的 64.3%的差值在众多次随机分组过程中是不是常见情况。如果随随便便一分组，都能观察到类似 64.3%这么大的差值，那也就很可能说明这个数字还挺常见的，也就是不够大到让我们觉得它是反常的。而如果分了很多次组，都无法再现如此大的差值，那也就很可能说明该数字确实反常，而这种反常就是疫苗有效的最直接支持证据。

图 2.29 中展示了 11 个最后发生感染的病人和 9 个没有感染的志愿者。为了模拟，我们倒回到分组前，并假设疫苗和是否感染无关。按照这种假设，最后 11 个发生感染的病人无论如何都会感染，而 9 个未发生感染的也无论如何都不会感染。这样一来，我们就可以用 20 张卡片替代志愿者，然后直接统计「由于分组导致的」统计结果的模拟情形。

对于这次模拟，我们就在 20 张替代志愿者的卡片中的 11 张上写下「会感染」，另外 9 张写下「不会感染」。然后我们来对这些卡片重新分成试验组和对照组，其中试验组随机抽取 14 张，对照组随机抽取 6 张。对于抽卡的结果，我们制作了如下图 2.30 所示的表。

		结果		
		感染	未感染	行总计
接种（模拟）	疫苗	7	7	14
	安慰剂	4	2	6
	列总计	11	9	20

图 2.30：模拟的结果，这里的疫苗组和安慰剂组的茶饮仅是由于分组的随机性导致

G

指导练习 2.34

在图 2.30 中两个组之间的感染率差值是多少？这和之前的 64.3%比怎么样？[1]

2.3.3 检验独立性

我们在上个指导练习中计算了在原假设下，模拟出来的两组感染率的差值。在刚刚的描述中，我们提到用随机抽卡的方式来进行模拟，而这个过程交给计算机来做会更加高效。关于如何制作相应的计算机模拟算法我们在此不再详述，而我们使用写好的算法继续进行模拟：

又一组模拟之后的差值计算：2/6 - 9/14 = -0.310

又一组模拟之后的差值计算：3/6-8/14 = -0.071

……

就这样不断模拟和记录，直到模拟的次数足够多，多到我们根据模拟的结果能够构建出一个「仅由于随机性导致」的差值的分布。图 2.31 就以堆积点图的形式展现了 100 次计算机模拟的结果。其中每个点在横轴上的数字都对应了两组间感染率的差值（对照组减去疫苗组）。

从图上可以看出，这个差值的分布大概是以数字 0 为中心的。由于我们的模拟是建立在原假设为真的基础上的，我们可以说在原假设情形下，对差值的期望「大约」应该是 0。这里的「大约」一词是因为在该研究中，我们的样本实在太小了（20 个）。

E

示例 2.35

根据图 2.31 的信息，你觉得模拟次数中，有多少机会观察到一个至少 64.3%差值？是很大机会，基本没有机会，还是完全观察不到？

答案：可以大约估摸出，一个至少 64.3%（大于等于）的差值出现的情形只占近 2%。这么低的一个概率说明这样的事件基本没有机会发生。

[1] 指导练习 2.34 答案：4/6 - 7/14 = 0.167，或者说对照组感染率要高 16.7%。这个差值和 64.3%数字比确实要小很多。

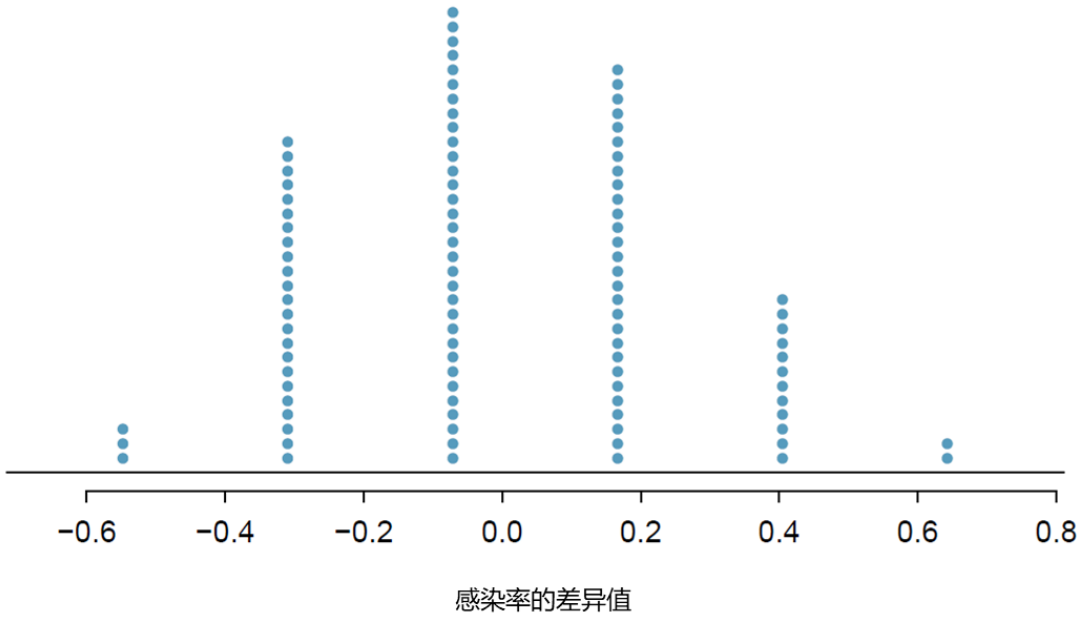


图 2.31：假设原假设 H_0 为真前提下，进行 100 次随机模拟过程的组间感染率差值的堆积点图，这些差值的产生是不受疫苗接种与否的影响的。图上只有右侧两个点的差值不小于 64.3%，即真实研究中得到的差值

那么既然在原假设为真的前提下，观察到 64.3%是一个几乎不会发生的事件，那么对此就有两种可能的解释：

H_0：**原假设 Independence model**。原假设依然为真，疫苗没有用，都是分组惹的祸。而我们真的是「凑凑凑凑」巧在一次随机过程中观察到了一个非常小概率的偶然事件。

H_A：**备择假设 Alternative model**。备择假设为真，即其实疫苗是有用的。

那么在讲了这么多之后，我们现在就有两个选择了：第一种选择：我们得出结论，说没有足够的证据证明原假设为伪，即没有足够证据证明疫苗有效。第二种选择：我们得出结论，说我们有足够的证据来拒绝原假设 H_0，所以判断疫苗是有效的。如果在一项正式的研究中，我们又真的观察到了如上所述的数据和推理。那么我们通常会去选第二项，即拒绝原假设。因为我们一般会拒绝接受如下想法：在一次纯随机过程中「凑凑凑凑」巧观察到了一个非常小概率的事件[1]。在这个疟疾疫苗的案例中，按照这个思路，在报告的结尾我们就会总结说我们有足够的证据，让我们相信新种疫苗确实能够起到预防疟疾的作用。

[1] 译者注：原著作者在这里做了一段注释，来延伸聊了一下轶事证据。其实主要是因为这里的判断，有可能有小伙伴会钻牛角尖：举个例子，我今天早上打开手机，看到某公众号推文说有人抽奖中了 100 万。抽奖中 100 万是个小概率事件吧？类似的还有很多小概率事件吧？而我们天天都在观察到。既如此，那为什么我们对于统计试验中观察到的小概率事件要特别重视，以至于做出拒绝原假设的结论呢？……这可能是因为，我们这里强调的是通过「随机过程」还观察到了小概率事件。在公众号推文看到别人中奖，真的是随机过程嘛（大概公众号就是专发这个的吧……）？而科学试验的前提就是随机过程，正因如此，这样的凑巧才特别值得重视。

统计学中有一个细分领域叫做统计推断，就是为了评估统计数据产生的差异是否是由于分组概率所致。在统计推断中，数据科学家们会沿用上面的逻辑来判断是原假设为真更合理还是备择假设为真更合理。其实现实中，也难免也会有错误推断的情况发生，就像随机过程中的小概率事件一样。我们不能保证总能通过统计学分析来准确选对「真相背后」的那个正确假设。尽管如此，统计推断却能够赋予我们「趁手的工具」，让我们能够评估和控制错误推断出现的概率。在本书第 5 章中，我们会对假设检验和结论选择进行一个正式的介绍。在接下来的第 3 章和第 4 章中，我们会来帮大家打一些概率论的基础，从而让大家在面对后续章节时有更严密的知识体系储备。

第 3 章

概率 Probability

概率理论毫无疑问构成了统计学的重要基础。尽管你可能对本章中的很多内容已经有了一些了解，但对大部分读者来说，这可能还是头一回去搭建一个正规的概率理论框架。

本章将会为未来的章节提供扎实的理论基础，同时也为你获取更深层次的统计学理解提供些许途径。不过也不用紧张，虽然该章节涉及的概率知识对统计学理论有奠基作用，你也无需对所有内容都精通。我们更希望你能形成一个属于自己的概率知识框架，熟悉相关的术语和逻辑，体会概率的理论之美，感悟概率的实践之趣。

跨越数据银河

系列推文合集

更多视频，演示文稿，和其他相关资源，请访问：
http://www.openintro.org/os

3.1 概率的定义

统计学是以概率为基础的，虽然概率并不是应用各种统计推断方法的必要条件，但它可以帮助你更深入地理解这些方法，并为今后的统计学习打下更好的基础。

3.1.1 一些简单示例

在我们讨论更专业的理论之前，让我们先来看看一些基本的例子，这些例子可能会帮你更熟悉我们即将讨论的专业知识。

示例 3.1

E 如果有一个六面体的骰子，六个面的编号分别为 1、2、3、4、5 和 6。掷骰子时，得到 1 的概率是多少?

答案: 如果骰子是公平的，那么得到 1 的机会和得到任何其他数字的概率一样大。由于有六个结果，那么这个概率是 1/6。

示例 3.2

E 下一次骰子掷到 1 或 2 的概率是多少?

答案：1 和 2 构成了 6 个等可能结果中的两个，所以得到这两个结果之一的概率是 2/6 = 1/3。

示例 3.3

E 下一次骰子得到 1、2、3、4、5、6 的概率是多少?

答案：100%。结果必须是这些数字之一。

示例 3.4

没有骰到 2 点的概率是多少?

E 答案:因为摇到 2 的概率是 1/6 或 16.6%,不掷到 2 的概率必须是 100%−16.6% = 83.4% 或 5/6。或者，我们可以注意到，没有摇到 2 和得到 1、3、4、5 或 6 是一样的，这构成了 6 个等可能结果中的 5 个，概率为 5/6。

示例 3.5

考虑掷两个骰子，如果第一个骰子有 1/6 概率是 1，第二个骰子有 1/6 概率是 1，得到两个 1 的概率是多少?

E

答案：如果第一个骰子是 1 的概率是 16.6%，第二个骰子也是 1 的概率是 1/6，那么两个骰子都是 1 的概率是(1/6) × (1/6)或者 1/36。

3.1.2 概率

我们使用概率的目的是为了构建工具，进而来描述和理解明显的随机性，我们通常把概率定义为产生某**结果 outcomes** 的**随机过程 random process**。

投掷一个骰子 → 1、2、3、4、5 或者 6

抛硬币 → 字 或者 花

掷骰子或抛硬币就可以被看作一个随机的过程，每个过程都会产生一个结果。

概率

我们这样定义**概率 probability**：某结果出现的概率即是如果我们观察一个随机过程无数次，该结果出现的次数所占的比例。

我们可以把概率定义为一个比例，它总是取 0 到 1（包括）之间的值，我们也常常把取值写成 0%和 100%之间的百分数。我们还可以通过一个掷骰子的例子来把概率和比例联系在一起：假设多次投掷一枚质地均匀的骰子，并把前 n 次投掷中数字 1 朝上的情况比例记作 $\hat{p}_n$。可以想象随着投掷次数的增加，$\hat{p}_n$ 将越来越趋向于 (收敛于) 得到数字 1 的概率，p = 1/6。图 3.1 显示了投掷 100,000 次时的比例趋势。不难发现，$\hat{p}_n$在 p 附近趋于稳定，这样趋势在统计学中可以被描述为**大数定律 Law of Large Numbers**。

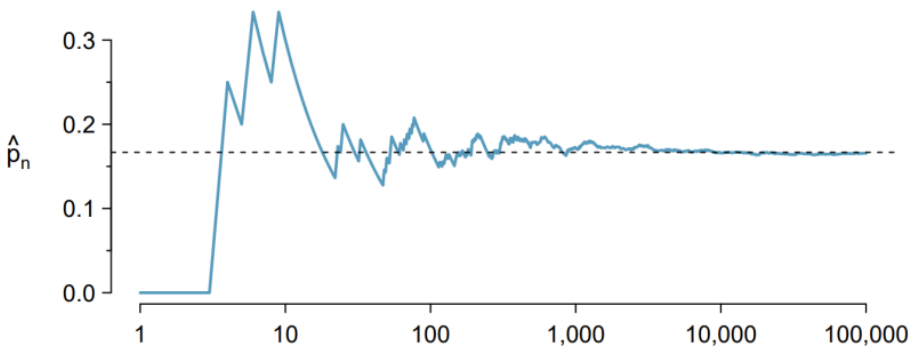


图 3.1：随着投掷次数的增加，数字 1 出现的比例的越来越趋近于其相应概率 1/6。

大数定律

随着观察次数的增加，出现特定结果的事件比例 $\hat{p}_n$ 收敛于该结果的发生的理论概率p。

有时候，尽管观察的次数已经很多，但观察到的比例也可能会偏离概率，就像在图 3.1 中横轴接近 1000 的部分多次出现 $\hat{p}_n$ 从 1/6 偏离，这似乎违背了大数定律。但实际上，这些偏差随着观察次数的进一步增加会逐渐减小。即大数定律的满足条件是需要观察次数真的足够「大」才行。

我们掷出 1 点的概率用字母 p 表示。我们也可以把这个概率写成：

$$P\big(\text{骰到 1 点}\big)$$

当我们越来越熟悉 P 这个符号的使用时，我们可以进一步简化。例如，如果很清楚这个过程是「掷骰子」，我们可以将 $P\big(\text{骰到 1 点}\big)$ 缩写为$P(1)$。

指导练习 3.6

比较经典的两个随机过程包括掷骰子和抛硬币。

(a) 想想看你能不能举出另一个随机过程；

(b) 说明你举出的随机过程的所有可能结果。例如，如果说掷骰子是一个随机过程，所有可能的结果分别是数字 1、2、……到 6。[1]

在本章中，我们判断一个事情是否是随机过程的标准可以宽泛些，这样主要是为了避免大家钻牛角尖。例如，在投掷骰子的时候，或许会有人提出每个人投掷的力度不同，或者投掷次数多了之后可能会习惯性按照某种特定方式投掷，导致事件发生的过程不再随机。如果是严密的科学试验，这样的考虑的确有潜在价值，但在本章介绍概率的时候，我们不必对判定某件事是随机过程的太过苛刻。有时候有的事件发生过程可能很复杂，影响因素也非常多，这时候我们可以简单的把它视为随机过程去理解。例如指导练习 3.6 答案中的第四个例子提到了室友的刷碗行为。诚然室友晚上是否会刷碗可能会受到各种因素影响，比如他/她心情是否预约，晚饭是谁做的，家中是否有洗碗机等等。然而，即使室友的刷碗行为的发生不是真正的随机，由于影响因素太多，事件是否会发生的过程细究过于复杂，所以这时将他/她的刷碗行为直接视作一个随机过程是被允许的。

[1] 这里有四个例子。(1) 某人在下个月是否会生病是一个明显的随机过程，结果可以是生病了和没生病。(2) 我们可以通过随机选取一个人并测量其身高来产生一个关于人的身高的随机过程。这个过程的结果将是一个正数。(3) 下周股票市场是上涨还是下跌似乎也是个随机过程，其结果可能是上升、下降和没有变化。另外，我们还可以用股市的百分比变化作为数字结果。(4) 你的室友今晚是否洗碗可能是一个随机过程，可能的结果是洗碗和不洗碗。

3.1.3 互斥结果

两个结果事件如果不能同时发生，则称之为**互斥 disjoint** 或**互不相容 mutually exclusive**。例如，如果我们掷一个骰子，掷出结果为 1 和结果为 2 是互斥的，因为它们不能同时出现。然而，结果为 1 和「掷出一个奇数」并非互斥，因为当掷出的结果是 1 时，这两个结果相当于都出现了。互斥和互不相容这两个术语是等价的，可以互换。此外他们对应的英文也在很多场景下有应用，希望大家也可以留心专门记忆一下。

计算互斥结果的概率很简单。当掷出一个骰子时，结果为 1 和结果为 2 是互斥的，我们通过将它们各自的概率相加，来计算结果 1 或结果 2 发生的概率（注意在概率公式书写中我们一般用英文单词 *or* 表示「或」）：

$$P(1\ or\ 2) = P(1) + P(2) = 1/6 + 1/6 = 1/3$$

那么掷骰子得出 1、2、3、4、5 或 6 的概率又是多少呢？在这里，所有结果也都是互斥的，所以我们将各个概率相加：

$$\begin{aligned} &P(1\ or\ 2\ or\ 3\ or\ 4\ or\ 5\ or\ 6) \\ &\quad = P(1) + P(2) + P(3) + P(4) + P(5) + P(6) \\ &\quad = 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1 \end{aligned}$$

加法法则 Addition Rule 保证了在结果互斥的情况下计算概率准确性，关于加法法则的具体竹。

互斥结果的加法法则

如果 A_1 和 A_2 代表两个互斥的结果，那么结果 A_1 或 A_2 发生的概率由以下公式给出：

$$P(A_1 or A_2) = P(A_1) + P(A_2)$$

如果某随机过程有许多互斥结果A_1, ..., A_k，那么在一次过程取得它们其中之一的概率为：

$$P(A_1) + P(A_2) + \cdots + P(A_k)$$

指导练习 3.7

我们对掷骰子得出 1、4 或 5 的概率非常感兴趣：

(a) 解释为什么 1、4、5 的结果是互斥的；

(b) 应用互斥结果的加法法则来确定$P(1 or 4 or 5)$。[1]

[1] (a) 随机过程是一个掷骰子的过程，这些结果中最多只能出现一个，这意味着它们的结果互斥。(b) $P(1\ or\ 4\ or\ 5) = P(1) +$

指导练习 3.8

在第 2 章的贷款数据中，房屋所有权的变量描述了借款人是租房、有抵押贷款还是拥有自己的房产。在 10,000 笔贷款中，3858 笔的借款人是租房，4789 笔的借款人是以抵押贷款形式购房，1353 笔的借款人拥有自己的房屋。

(a) 租金、抵押贷款和自有房产的结果是否互斥?

(b) 分别确定有抵押贷款和自有房屋贷款的比例。

(c) 使用互斥结果的加法法则来计算：从数据集中随机选择的贷款是给「有抵押贷款的人或是自有住房的人」的概率。[1]

数据科学家很少处理单个的结果，而是考虑结果的集合。让 A 代表掷骰子结果为 1 或 2 的事件，B 代表掷骰子结果为 4 或 6 的事件。我们把 A 写成结果的集合$1，2$，$B = 4，6$。这些集合通常被称为**事件 events**。因为 A 和 B 没有共同的元素，所以它们是互斥事件。图 3.2 表示 A 和 B。

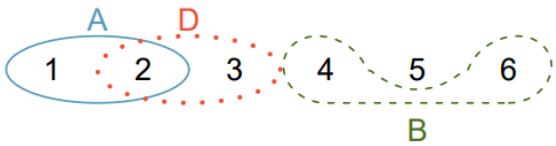


图 3.2：三个事件，A，B 和 D，由掷骰子的结果组成。A 和 B 是互斥的。

加法法则适用于互斥结果和互斥事件。互斥事件 A 或 B 之一发生的概率是其单独概率之和：

$$P(AorB) = P(A) + P(B) = 1/3 + 1/3 = 2/3$$

指导练习 3.9

(a) 利用加法法则验证事件 A 的概率$P(A)$是 1/3；

(b) 对事件 B 做同样的处理。[2]

指导练习 3.10

(a) 以图 3.2 为参考，事件 D 代表什么结果?

(b) 事件 B 和 D 是否不互斥?

(c) 事件 A 和 D 是否不互斥? [3]

指导练习 3.11

在指导练习 3.10 中，你确认图 3.2 中的 B 和 D 是不互斥的。计算事件 B 或事件 D 发生的概率。[4]

$P(4) + P(5) = 1/6 + 1/6 + 1/6 = 3/6 = 1/2$。

[1] (a) 是的。每笔贷款只被归入一个房屋所有权级别；(b) 抵押贷款：$4789/10000 = 0.479$。自有房产：$1353/10000 = 0.135$；(c) $P(\text{抵押贷款或自有房产}) = P(\text{抵押贷款}) + P(\text{自有房产}) = 0.479 + 0.135 = 0.614$。

[2] $P(A) = P(1\ or\ 2) = P(1) + P(2) = 1/6 + 1/6 = 2/6 = 1/3$；(b) 同样，$P(B) = 1/3$。

[3] (a) 代表了骰子是 2 和 3 的两种结果；(b) 是的，事件 B 和 D 是不互斥的，因为它们没有共同的结果；(c) 事件 A 和 D 有一个共同的结果，即 2，所以它们不是互斥的。

[4] 因为 B 和 D 是不互斥的事件，所以使用加法法则：$P(B\ or\ D) = P(B) + P(D) = 1/3 + 1/3 = 2/3$。

3.1.4 非互斥事件的概率

我们来考虑如下场景：在一副由 52 张牌组成的普通扑克牌中随机进行抽取。在该场景下我们来进行一些不相交的事件的计算。在此之前，我们先展示一下扑克牌中所有的牌的花色和数字，如图 3.3 所示。如果你不熟悉普通牌组中的牌，请看脚注[1]。

图 3.3：一副牌里面 52 张不重复的牌。

指导练习 3.12

(a) 随机选择的牌是钻石的概率是多少?

(b) 随机选择的牌是面子牌的概率是多少? [2]

当结果可以被归类为「属于」或者「不属于」两个或三个变量（或者属性，或者随机过程）时，**文氏图 Venn Diagram** 就很有用。图 3.4 中的文氏图用一个圆圈代表方片，另一个圆圈代表花牌。如果一张牌既是方片又是花牌（♦J，♦Q 和 ♦K），它就会落入两个圆圈的交叉点。如果它是方片，但不是花牌，它就仅仅是左侧椭圆的一部分，而不在右侧椭圆中（以此类推）。方片牌的总数由代表方片的圈的总数给出：$10 + 3 = 13$。图上我们也标记出了概率（例如：$10/52 = 0.1923$）。

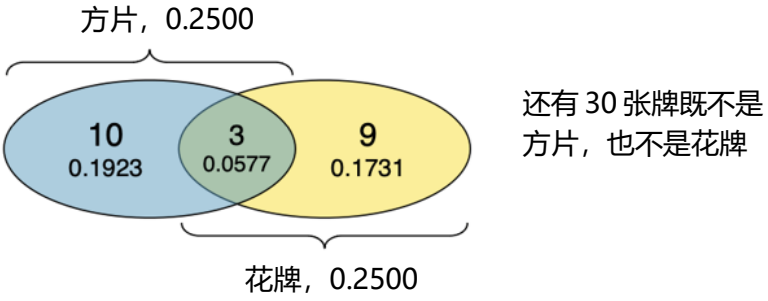


图 3.4：关于一副牌里方片和花牌的文氏图。

[1] 52 张牌分成 4 种**花色 suits**：梅花♣ club，方片♦ diamond，红桃♥ heart 和黑桃♠ spade。每种花色包含从 2 到 10 以及士兵 Jack，皇后 Queen，国王 King 和尖 Ace 共 13 张牌。所以每张牌都代表了唯一的花色和数字组合。然后士兵、皇后和国王对应的 J、Q 和 K 这三张牌又被称为**花牌 face cards**。

[2] 因为 B 和 D 是不互斥的事件，所以使用加法法则：$P(B \text{ or } D) = P(B) + P(D) = 1/3 + 1/3 = 2/3$。

我们用 A 表示随机选择的一张牌是方片的事件，B 代表它是一张花牌的事件。我们如何计算 $P(\text{A } or \text{ B})$？注意事件 A 和 B 并非互斥的（上面已经列举了：♦J，♦Q 和 ♦K 都属于这两类）。所以我们不能直接对非互斥的事件使用加法规则。这时，我们使用文氏图就可以较好辅助解决这个问题。我们首先将两个事件的概率相加……

$$P(A) + P(B) = P(\diamondsuit) + P(\text{花牌}) = 13/52 + 12/52$$

然而，在两个事件的交集中，有三张牌被计算了两次。我们必须纠正这种重复计算（注意，在书写公式的时候我们常常用 and 来代替「和」或者「与」逻辑）：

$$\begin{aligned} P(A \text{ } or \text{ } B) &= P(\diamondsuit \text{ } or \text{ 花牌}) \\ &= P(\diamondsuit) + P(\text{花牌}) - P(\diamondsuit \text{ } and \text{ 花牌}) = 13/52 + 12/52 - 3/52 \\ &= 22/52 \ = \ 11/26 \end{aligned}$$

上面的方程就是个很好的**普适加法法则 General Addition Rule** 的一个例子。

普适加法法则

如果 A 和 B 是任何两个事件，无论是否互斥，那么其中至少有一个会发生的概率是：

$$P(\text{A } or \text{ B}) = P(\text{A}) + P(\text{B}) - P(A \text{ \& } B)$$

其中 $P(A \text{ \& } B)$ 是两个事件同时发生的概率。

提示：「或」逻辑是包容性的。即当我们在统计学中写或 or 的时候，我们默认关系是「与/或」。除非特别说明，否则 A 或 B 发生意味着 A 发生、B 发生或者 A 和 B 同时都发生。

指导练习 3.13

(a) 如果 A 和 B 是两个互斥事件，请解释为什么 $P(A \text{ \& } B) = 0$？

(b) 根据上题，请证明对于互斥事件来说，普适加法法则可以简化为一般加法法则使用。[1]

指导练习 3.14

第 2 章的 loans 贷款数据集的 10,000 笔贷款中，1495 笔是联合申请（例如夫妇共同申请），4789 笔的借款人是以抵押贷款形式购房，950 笔则拥有以上两项特征。请根据这个背景绘制一幅文氏图。[2]

---

[1] 如果 A 和 B 是互斥事件，它们两者就永远无法同时发生，所以同时发生的概率为 0；(b) 如果 A 和 B 是互斥事件，由于上题已经证明了 $P(A \text{ } and \text{ } B) = 0$，所以我们只需要从普适加法法则的公式中删去这一项，就可以直接得到简化的一般加法法则。

[2] 图略，图上应展示贷款的数目信息和响应的概率 probabilities。其中左侧椭圆应包含 3839 + 950 = 4789 笔贷款，右侧椭圆应包括 950 + 545 = 1495 笔贷款，中间的交集正好是 950 笔贷款。对应的概率分别是 0.384, 0.095 和 0.055。

指导练习 3.15

(a) 根据你在上个指导练习中绘制的文氏图，请判断从 loans 数据集中随机抽取一笔贷款，它恰好「是联合申请的，且借款人是以抵押贷款方式购房」的概率？

(b) 如果还是随机抽一笔贷款，它至少具备上述两项特征之一的概率又是多少？[1]

3.1.5 概率分布

概率分布 probability distribution 是一个将所有互斥结果及其对应的概率结合在一起的表格。图 3.5 显示了两个独立掷出的骰子之和的概率分布。

骰子点数和	2	3	4	5	6	7	8	9	10	11	12
概率	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

图 3.5：两个骰子点数和的概率分布。

概率分布的规则

概率分布是一个同时带有可能出现的结果及其对应概率的列表，它满足三个规则：

(1) 列出的结果之间必须是互斥的；

(2) 每个发生的概率必须在 0 和 1 之间；

(3) 概率总和必须是 1。

指导练习 3.16

图 3.6 显示了美国家庭收入的三种分布情况，其中只有一个是正确的。你能判断出哪一个是正确的吗？其他两个有什么问题？[2]

收入范围	$0-25k	$25k-50k	$50k-100k	$100k+
(a)	0.18	0.39	0.33	0.16
(b)	0.38	-0.27	0.52	0.37
(c)	0.28	0.27	0.29	0.16

图 3.6：美国家庭收入的一些可能概率分布（指导练习 3.16）。

[1] (a) 可以从图上直接观察出答案：中间交集部分的概率对应了 0.095；(b) 该题的答案应该是把文氏图中三个区域的概率累加，得到 0.384 + 0.095 + 0.055 = 0.534（进行了一定的四舍五入）。

[2] 选项(a)的概率之和不为 1；选项(b)的第二个概率为负；目前就剩下(c)选项，检查发现不同收入区间列出的结果之间是互斥的，每个结果发生的概率在 0 到 1 之间，同时概率发生的总和为 0.28 + 0.27 + 0.29 + 0.16 = 1，均满足概率分布的要求，因此，选项(c)最可能是美国家庭收入的实际分布。

第 1 章中强调了通过制图以进行快速数据总结的重要性。概率分布也可以用柱状图来进行总结和描述。例如，美国家庭收入的概率分布情况如图 3.7 所示。两个骰子之和的概率分布表如图 3.5 所示，其分布柱状图如图 3.8 所示。

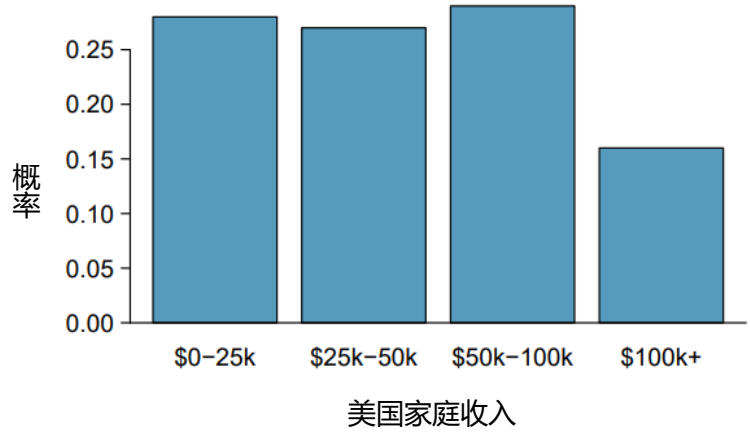


图 3.7：美国家庭收入的概率分布柱状图。

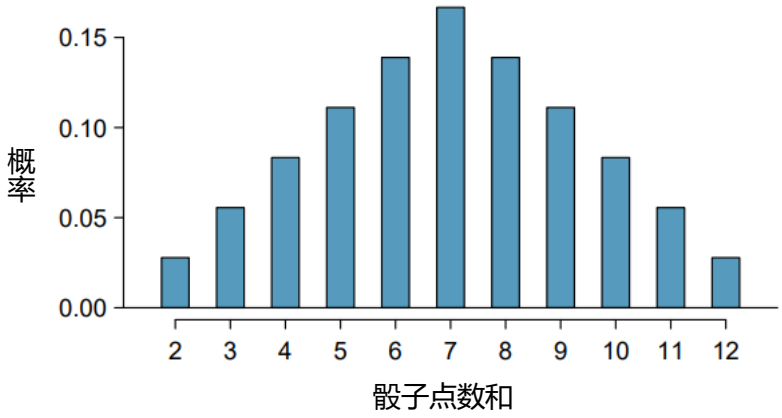


图 3.8：两个骰子点数和的的概率分布柱状图。

在这些柱状图中，柱子的高度代表结果发生的概率。若结果是数值型并且是离散的，柱状图往往是很好的可视化选择。柱状图从外观上看有点像之前介绍的直方图，事实上它们也的确有很多相通之处：横轴都可以取数据区间，柱子的高度代表了对应比率的大小等等。可以看到上方的图 3.8 就很像直方图。我们在此也预告一下：另一个柱状图的例子可以参考后面的图 3.18。

3.1.6 事件的补集

当我们掷出一个骰子的时候，可能得到的结果为1,2,3,4,5,6中的任何一个，这六种可能出现的结果组成的集合被称作掷骰子的**样本空间 sample space (*S*)**，我们经常用样本空间来辅助讨论某个事件没有发生的场景。

我们让$D = 2,3$代表掷骰子结果为 2 或者 3，那么，事件 D 的**补集 complement** 就是「掷骰子的样本空间内，所有除了 2 和 3 以外的可能结果」，我们记作$D^c = 1,4,5,6$，即 D^c 中包含了所有除了事件 D 结果外的可能结果。图 3.9 展示了 D，D^c 和样本空间 S 之间的关系。

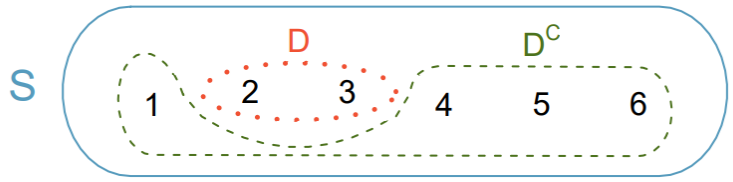


图 3.9：事件$D = 2,3$，其补集$D^c = 1,4,5,6$，S 代表了样本空间，包含了所有 6 种掷骰子可能出现的结果。

指导练习 3.17

(a) 在事件 D 不变的前提下，通过计算证明 $P(D^c) = P(1\ or\ 4\ or\ 5\ or\ 6)$；

(b) $\mathrm{P(D)} + \mathrm{P(D^c)}$ 等于多少？[1]

指导练习 3.18

事件$A = 2,3$，$B = 4,6$，这两个事件的划分如前文图 3.2：

(a) 写出 A^c 和 B^c 分别代表什么？

(b) 计算 $\mathrm{P}(A^c)$ 和 $\mathrm{P}(B^c)$；

(c) 计算 $\mathrm{P(A)} + \mathrm{P}(A^c)$ 和 $\mathrm{P(B)} + \mathrm{P}(B^c)$。[2]

对于一个事件 A 来说，它的补集 A^c 具有两个很重要的属性：(1) 任何一个不属于 A 的结果都一定属于 A^c；(2) A 和 A^c 互斥。其中，第一点属性又意味着：

$$P(A\ or\ A^c) = 1$$

即如果一个结果不在 A 中，那么它一定在 A^c 中。而第二点属性则适用于加法定律，即：

$$\mathrm{P(A}\ or\ A^c) = \mathrm{P(A)} + \mathrm{P}(A^c)$$

[1] (a) 由于每个可能的结果都互斥且都为 1/6，所以$P(\mathrm{D^c}) = 4/6 = 2/3$；(b) 同理，$P(D) = 1/6 + 1/6 = 1/3$，又因为 D 和 D^c 互斥，所以 $P(D) + P(\mathrm{D^c}) = 1$。

[2] (a) $A^c = 3,4,5,6$，$B^c = 1,2,3,5$；(b) 因为每种可能的结果都是互斥的，所以$P(A^c) = 2/3$，$P(B^c) = 2/3$；(c) 因为 A和 A^c 互斥，B 和 B^c 也互斥，所以$P(A) + P(A^c) = 1$，$P(B) + P(B^c) = 1$。

如果我们将以上两个等式结合起来，就能得到一个更有用的等式，来表达一个事件和其补集之间的关系：

补集

事件 A 的补集用 A^c 来表示，A^c 代表了所有不在 A 中的可能结果，A 和 A^c 的数学关系表现为：

$$\mathrm{P}(A) + \mathrm{P}(A^c) = 1，或\ \mathrm{P}(A) = 1 - \mathrm{P}(A^c)$$

对于简单的事件来说，我们可以通过直接计算的方法算出 $\mathrm{P}(\mathrm{A})$ 或者 $\mathrm{P}(A^c)$，但是当事件越来越复杂的时候，直接计算往往就不那么直观和容易了，这时候通过补集的角度切入往往能节省很多时间。

指导练习 3.19

我们让事件 A 代表：掷出两个骰子的点数之和小于 12。那么……

(a) A^c 代表了什么？

(b) 根据之前的图 3.5，计算 $\mathrm{P}(A^c)$；

(c) 计算 $\mathrm{P}(\mathrm{A})$。[1]

指导练习 3.20

我计算掷出两个骰子时，以下事件的概率：

(a) 点数之和不是 6；

(b) 点数之和大于等于 4，即计算事件$B = 4,5,\ldots,12$的概率；

(c) 点数之和小于等于 10，即计算事件$D = 2,3,\ldots,10$的概率。[2]

[1] (a) A^c 代表了两个骰子点数之和等于 12 的事件；(b) $\mathrm{P}(A^c) = 1/36$；(c) $\mathrm{P}(\mathrm{A}) = 1 - \mathrm{P}(A^c) = 1 - 1/36 = 35/36$。

[2] 首先 $P(6) = 5/36$，因此 $P(\text{点数和不是 } 6) = 1 - 5/36 = 31/36$；(b) 首先明确补集，即点数之和为 2 或 3，$P(2\ or\ 3) = 1/36 + 2/36 = 1/12$，因此 $P(B) = 1 - P(B^c) = 1 - 1/12 = 11/12$；(c) 我们依旧先明确补集，即 $P(D^c) = P(11\ or\ 12) = 2/36 + 1/36 = 1/12$，因此 $P(D) = 1 - P(D^c) = 11/12$。

3.1.7 独立性

就像我们在讨论「变量」和「观测」时强调它们的独立性一样，随机过程也可以是独立的。如果知道一个随机过程的结果并不能为另一个随机过程的结果提供有用的信息，那么这两个随机过程就被称作是**相互独立 independent** 的。例如，掷硬币和掷骰子就是两个独立事件，因为即使知道掷硬币结果为正面，我们也无法以此推断出掷骰子的结果。我们再来举个反例，有些股票价格经常出现一起涨或者一起跌的情况，那么它们就不是相互独立的。

示例 3.5 展示了一个很简单的两个相互独立事件的例子：掷两枚骰子。现在我们想得到两枚骰子都掷出 1 的概率。假如这两枚骰子一个是红色的，另一个是白色的，那么如果我们已知红色骰子掷出的结果为 1，也无从知道白色骰子的结果。在示例 3.5 中，我们曾经探究过这个问题，当时我们通过以下的逻辑计算出了两枚骰子都掷出 1 的概率：红色骰子掷出 1 的概率为 1/6，并且在红色骰子掷出 1 的情况下，白色骰子掷出 1 的概率为 1/6（见图 3.10），因为这两个事件是相互独立的，所以它们同时发生的概率可以由两个事件单独的概率相乘得到，即 1/6 × 1/6 = 1/36。这套逻辑可以被普遍适用于很多的独立事件。

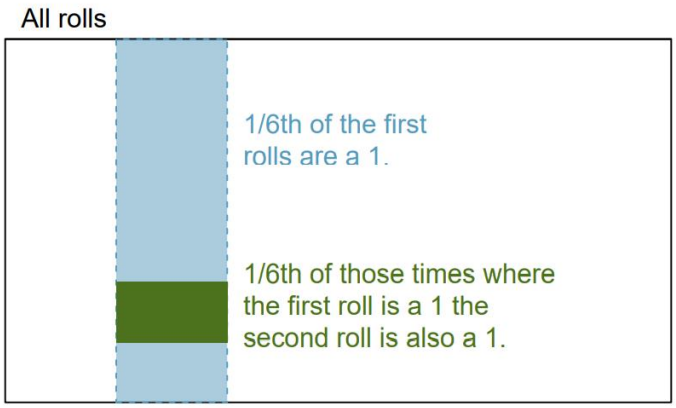


图 3.10：1/6 的情况下，掷出第一个骰子结果为 1，在此基础上，又有 1/6 的情况下掷出第二个骰子结果也为 1。

示例 3.21

如果除了前两个骰子外还有一个蓝色的骰子，那么三个骰子都掷出 1 的概率是多少?

答案：和示例 3.5 的逻辑相同，如果 1/36 的概率白色和红色的骰子都掷出 1，那么在此基础上还有 1/6 的情况下蓝色的骰子也掷出 1，即：

$$
\begin{aligned}
&P(\text{白色、红色和蓝色骰子都掷出 } 1) \\
&\quad = P(\text{白色骰子掷出 } 1) \times P(\text{红色骰子掷出 } 1) \times P(\text{蓝色骰子掷出 } 1) \\
&\quad = (1/6) \times (1/6) \times (1/6) = 1/216
\end{aligned}
$$

示例 3.21 展示的就是独立事件的乘法法则。

独立事件的乘法法则

如果 A 和 B 代表了两个不同的独立事件，那么 A 和 B 同时发生的概率可以通过它们的乘积计算得到：

$$P(A \text{ and } B) = P(A) \times P(B)$$

同理，如果 $A_1, \ldots, A_k$ 表示 k 个相互独立的事件，那么这些事件同时发生的概率为：

$$P(A_1) \times P(A_2) \times \cdots \times P(A_k)$$

指导练习 3.22

据统计，美国约 9%的人是左撇子。如果我们现在随机选两个人，因为两个人之于美国全体人口数来说太小了，所以我们可以认为这两个人相互独立，那么……

(a) 这两个人都是左撇子的概率是多少？

(b) 这两个人都是右撇子的概率是多少？[1]

指导练习 3.23

假设现在随机选出 5 个人：

(a) 他们全都是右撇子的概率是多少？

(b) 他们全都是左撇子的概率是多少？

(c) 他们不都是右撇子的概率是多少？[2]

假设惯用手和性别这两个变量相互独立，也就是说知道某人性别无法推断出他的惯用手是左手还是右手（反之亦然），那么根据乘法定律，我们就可以计算出一个随机选择的人是右撇子并且是女性[3]的概率：

$$P(\text{是右撇子并且是女性}) = P(\text{是右撇子}) \times P(\text{是女性})$$

$$= 0.91 \times 0.50 = 0.455$$

[1] (a) 因为第一个人和第二个人是相互独立的，我们使用乘法法则就可以轻易得出 $0.09 \times 0.09 = 0.0081$；(b) 我们可以假设即惯用左手也惯用右手的人群比例为 0（因为无论如何总应该有偏好），那么就可以根据惯用左手的比例来算出惯用右手的比例是 0.91，那么沿用上一小题中的思路，可以轻松算出随机抽到的两个人都是惯用右手的概率为 $0.91 \times 0.91 = 0.8281$。

[2] (a) 因为每个人都是相互独立的，所以根据独立事件的乘法法则，$P(\text{五个人都是右撇子}) = P(\text{第一个人是右撇子}) \times P(\text{第二个人是右撇子}) \times \cdots \times P(\text{第五个人是右撇子}) = 0.91 \times 0.91 \times 0.91 \times 0.91 \times 0.91 = 0.624$；(b) 同理，$P(\text{五人都是左撇子}) = 0.09 \times 0.09 \times 0.09 \times 0.09 \times 0.09 = 0.0000059$；(c) 利用补集的概念，$P(\text{五个人不都是右撇子}) = 1 - P(\text{五个人都是右撇子}) = 1 - 0.624 = 0.376$。

[3] 美国人口中女性的占比约为 50%，所以在这里我们用 0.5 表示随机抽取一个人为女性的概率，但该概率在其他国家可能不同。

指导练习 3.24

假设现在随机选出 3 个人：

(a) 第一个人是男性并且是右撇子的概率是多少？

(b) 前两个人都是男性并且是右撇子的概率是多少？

(c) 第三个人是女性并且是左撇子的概率是多少？

(d) 前两个人都是男性右撇子并且第三个人是女性左撇子的概率是多少？[1]

示例 3.25

如果我们洗牌后再随机抽出一张牌，那么抽出的这张牌是♥红桃和抽出的这张牌是 Ace 这两个事件相互独立吗？

答案：抽到红桃的概率为 1/4，抽到 Ace 的概率为 1/13，抽到红桃 A 的概率是 1/52，因为

$$P(\text{抽到 ♥ 红桃 Ace}) = P(\text{抽到 ♥ 红桃}) \times P(\text{抽到 Ace})$$

满足乘法法则，所以这两个事件是相互独立的。

[1] 只提供最终结果：(a) 0.455；(b) 0.207；(c) 0.045；(d) 0.0093。

3.2 条件概率

上一节中我们介绍了加法法则和乘法法则，他们都设计到了两个及以上事件间的概率计算。事实上，事件或变量间的关系往往是很有用的。例如汽车保险公司通过一个人的驾驶记录，能判断他发生交通事故的风险。这类有先后依属关系的概率属于条件概率的讨论范畴，条件概率也是我们本节要讨论的话题主角。

3.2.1 一利用列联表探索概率

我们来看一个名为 photo_classify 数据集的例子，该数据集的样本包括来自一个照片分享网站的 1822 张照片。数据科学家一直在努力改进一种分类器，这种分类器能判断照片是不是属于时尚类。科学家们用这 1822 张照片对分类器进行了一次测试。把每张照片进行两次分类，对应下表两个变量：一个变量名为机器判断（mach_learn，即 machine learning/ML，也就是通过机器学习根据一定算法系统预判照片是不是属于时尚类）；另一个变量名为真实结果（truth），它代表通过人工仔细分辨后判断的照片实际分类情况。虽然本书的之前章节一直使用中文变量名称，并通过直角引号（「」）来标识变量，但我们在这章的翻译中会做一些改变。由于例子中的英文变量名大多是由多个词汇缩写拼接而成，如果完整翻译会因为太长而导致段落更难理解，因此我们决定在引用变量名称及变量取值的时候直接使用原著中的英文名称。希望大家能够在最初接触变量和取值英文名称的时候，结合自己的英文知识或者线上字典来理解并记忆清楚他们的含义。必要时我们也会用括号或者脚注的形式来进行含义备注。

mach_learn 的取值分为 pred_fashion（机器判断为时尚类图片）和 pred_not（机器判断为非时尚类图片）; truth 的取值分为 fashion（人为判断为时尚类图片）和 not（人为判断为非时尚类图片）。图 3.11 总结出了结果。

		truth		
		fashion	not	Total
mach_learn	pred_fashion	197	22	219
	pred_not	112	1491	1603
	Total	309	1513	1822

图 3.11：photo_classify 数据集的总结列联表。

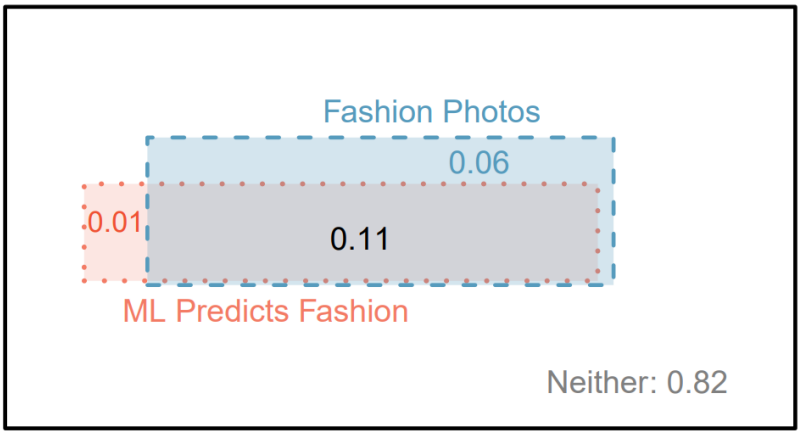


图 3.12：用方形盒子展示的一张关于 photo_classify 数据集的文氏关系图。

示例 3.26

已知照片实际上属于时尚类，那机器学习分类器能做出正确判断的概率有多大？

答案：我们可以利用数据估计概率：在 309 张时尚类照片中，机器学习算法正确判断出了 197 张。

$$P\left(\text{基于 truth 取值是 fashion，}\ \text{mach}_{\text{learn 取值是}\text{pred}_{\text{fashion}}}\right) = 197/309 = 0.0638$$

示例 3.27

我们从数据集中挑一张照片，已知机器学习算法判断照片不属于时尚类。那机器学习算法判断错误的概率是多少？

答案：已知机器学习分类器预测照片不属于时尚类，那么数据来自数据集第二行。在 1603 张照片中，实际上有 112 张照片属于时尚类，即这 112 张判断错误。

$$P\left(\text{基于}\text{mach}_{\text{learn 取值是}\text{pred}_{\text{not}}}\text{，truth 取值是 fashion}\right) = 112/1603 = 0.070$$

3.2.2 边际概率和联合概率

图 3.11 包含了照片分类中 mach_learn 和 truth 两个变量各自取值的计数情况。其中，从最后一行和最后一列的总计 (Total) 中可以得出样本的**边际概率 marginal probabilities**，也就是只考虑一个变量算出的概率。例如，仅仅基于 mach_learn 变量所得出的概率就是一个边际概率：

$$P\left(\text{基于}\text{mach}_{\text{learn 取值是}\text{pred}_{\text{fashion}}}\right) = 219/1822 = 0.12$$

基于两个或两个以上的变量或过程所得出的结果称为**联合概率 joint probability**，在联合概率的标记方法中，用逗号代替 *and* 是很常见的：

$$P\left(\text{基于}\text{mach}_{\text{learn 取值是}\text{pred}_{\text{fashion 且 truth 取值是 fashion}}}\right) = 197/1822 = 0.11$$

边际概率和联合概率

基于单变量的概率是*边际概率*；涉及多变量的概率往往是*联合概率*。

我们使用表中比例来统计照片分类样本的联合概率。将图 3.11 中每一项数字除以表中总数 1822，得到图 3.13 中的比例。图 3.14 展示出 mach_learn 和 truth 这两个变量的联合概率分布情况。

	truth: fashion	truth: not	Total
mach_learn: pred_fashion	0.1081	0.0121	0.1202
mach_learn: pred_not	0.0615	0.8183	0.8798
Total	0.1696	0.8304	1.00

图 3.13：关于 photo_classify 数据集的概率表。

Joint outcome	Probability
mach_learn is pred_fashion and truth is fashion	0.1081
mach_learn is pred_fashion and truth is not	0.0121
mach_learn is pred_not and truth is fashion	0.0615
mach_learn is pred_not and truth is not	0.8183
Total	1.0000

图 3.14：关于 photo_classify 数据集的联合概率分布。

指导练习 3.28

请验证图 3.14 是符合概率分布的定义的：各结果互斥，所有概率非负且概率和为 1。[1]

3.2.3 定义条件概率

即使结果可能不够精准，电脑分类器已经能预测一张照片是否属于时尚类了。而为了验证电脑判断的准确率，我们希望更充分利用变量的取值信息，来更准确地估计另一个变量取到特定值的概率。在本例中，我们沿用上一节的案例，继续讨论变量 mach_learn 和变量 truth 间的条件概率。

在数据集中随机取一张照片，它属于时尚类的概率约为 0.17。基于这个信息请思考以下问题：「如果通过机器学习已经预判照片属于时尚类，我们基本能断定照片确实是时尚类呢？」答案是可以的。为此，我们来关注符合条件的 219 个案例，根据数据来推断当机器学习预测照片属于时尚类时，照片有多大概率确实如此？

$$P\left(\text{基于}\mathrm{mach}_{\text{learn 取值是}}\mathrm{pred}_{\text{fashion 且 truth 取值是 fashion}}\right) = 197/219 = 0.900$$

[1] 四种结果的出现都是互斥的，因为按照定义四种结果是根据两个变量能取到的两种取值的组合得出，所以不会存在包含关系。所有概率非负且其和为 1，所以满足概率分布的标准定义条件。

条件概率 Conditional Probability 是基于一定条件计算出来的结果。本例中的条件是：机器判断照片属于时尚类。条件概率涉及两个重要概念，**结果 outcome of interest** 和**条件 condition**。可以把条件看作已知真实的信息，或者说是已知的结果。我们在写条件概率时，通常用竖线分隔左侧的结果和右侧的条件。

$$
\begin{aligned}
&P\left(\text{基于}\mathrm{mach}_{\text{learn 取值是}\mathrm{pred}_{\text{fashion}}}, \text{truth 取值是 fashion}\right) \\
&= P\left(\text{truth 取值 fashion} \,\middle|\, \mathrm{mach}_{\text{learn 取值是}\mathrm{pred}_{\text{fashion}}}\right) \\
&= 197/219 = 0.900
\end{aligned}
$$

概率表达式括号中用于分割的竖线，在口语中可以读作：基于。例如$P(A|B)$就是基于 B 发生的前提，A 发生的概率。

在上式中，基于机器学习算法预判照片属于时尚类的前提，我们可以计算出机器判断正确的概率，并用分数把它表示出来。

$$
\begin{aligned}
&P\left(\text{truth 取值是 fashion} \,\middle|\, \mathrm{mach}_{\text{learn 取值是}\mathrm{pred}_{\text{fashion}}}\right) \\
&= \frac{\text{\# truth 取值是 fashion 且 mach\_learn 取值是 pred\_fashion 的数量}}{\text{\# mach\_learn 取值是 pred\_fashion 的数量}} \\
&= 197 \,/\, 219 \;=\; 0.900
\end{aligned}
$$

计算条件概率时，我们只考虑符合条件的那些个体，然后计算这些个体中符合设定结果的概率。在上例中，我们只看 mach_learn 取值是 pred_fashion 的个体数量，再计算其中 truth 取值是 fashion 的个体数量，进而计算出概率。

更多时候，我们能接触到的是对应变量的边际概率和联合概率，而不是具体样本的统计频数。例如，我们通常会记录某种疾病的百分比发病率，而不是记录某几次实验中对应患病的病例数。因此，我们希望即使不使用个体数量信息，也能计算条件概率。接下来，我们就以上面计算机器判断正确的概率为例，来展开看看是如何拜托频数计算条件概率的。

我们只看符合条件的个体，也就是机器学习已经预判照片属于时尚类的那 219 个个体。我们要计算的条件概率是指在这些个体中人为判断照片属于时尚类（即 truth 取值为 fashion 的照片）的比例。假设我们只有图 3.13 中的信息，也就是说，只有概率数据。从表中信息可知，mach_learn 取值为 pred_fashion 的概率约为 12.0 %，mach_learn 和 truth 都判断照片为时尚类的概率约为 10.8%。那么随机取 1000 张照片，mach_learn 取值为 pred_fashion 的个体数量是 0.120 × 1000 = 120，mach_learn 和 truth 都判断照片为时尚类的个体数量是 0.108 × 1000 = 108。那么要计算的条件概率就可以表示为：

$$
\begin{aligned}
&P\left(\text{truth 取值是 fashion} \,\middle|\, \text{mach}_{\text{learn 取值是pred}_{\text{fashion}}}\right) \\
&= \frac{1000 \text{ 照片中 truth 取值是 fashion 且mach}_{\text{learn 取值是pred}_{\text{fashion 的数量}}}}{1000 \text{ 照片中mach}_{\text{learn 取值是pred}_{\text{fashion 的数量}}}} \\
&= 108 / 120 = 0.900
\end{aligned}
$$

严格来说，0.108 和 0.120 两个概率可以写作：

$$
P\left(\text{truth 取值是 fashion 且mach}_{\text{learn 取值是pred}_{\text{fashion}}}\right) = 0.108
$$

$$
P\left(\text{mach}_{\text{learn 取值是pred}_{\text{fashion}}}\right) = 0.120
$$

通过以上的计算，我们可以汇总出条件概率的通式：

条件概率

基于 B 的前提下 A 的条件概率可以表示为

$$
P(A \mid B) = P(A \,\&\, B)/P(B)
$$

注意：在熟悉了这样的标记法后，在接下来的行文中，我们会尽量使用 *and* 符号来连接两件事，表示与、且关系，但是有时候为了说明方便，我们也可能会穿插使用 $P(A, B)$ 的书写模式。总之大家要学会明确区分 P 括号内的事件和事件关系连接符。无论是 $P(A, B)$ 还是$P(A \,\&\, B)$，表达地都是一个意思。

指导练习 3.29

(a) 请用 P 和括号的方式表达「在所有人为判断出的时长类照片中，机器学习分类法也判断正确的概率」。注意！这里的条件是限定了变量 truth 的取值，不是变量 mach_learn 的取值；

(b) 计算出上一小题中的概率值。可以参考图 3.13。[1]

[1] (a) P(mach_learn 取值是 pred_fashion | truth 取值是 fashion)；(b) 根据条件概率展开式，可以知道我们需要用到的两个概率分别是联合概率：0.1081 和边际概率：0.1696，相除可知最终答案为：0.1081/0.1696 = 0.6374。

指导练习 3.30

(a) 请计算在已知照片属于时尚类的前提下，机器学习算法做出错误判断的概率；

(b) 利用上一小题的结果和指导练习 3.29 的结果，计算：

$$P\left(\text{mach}_{\text{learn 取值是pred}_{\text{fashion}}} \,\middle|\, \text{truth 取值是 fashion}\right) + P\left(\text{mach}_{\text{learn 取值是pred}_{\text{not}}} \,\middle|\, \text{truth 取值是 fashion}\right)$$

(c) 请结合直觉分析推理，为什么计算出的结果会是 1？[1]

3.2.4 1721 年波士顿天花事件

我们来看一个名为 smallpox 的历史天花病例数据集，该样本里面有 6224 个观测值个体，这些个体于 1721 年在波士顿感染了天花。当时的医生认为接种疫苗可以使人体在可控范围内接触病毒，从而刺激人体产生抗体并自动免疫。这样在产生免疫后感染真正有毒性的病毒时，就可以让免疫系统更快反应，降低死亡率。

数据集中每个个体指代一个人，对应两个变量：疫苗接种（inoculated）和结果（result）。变量 inoculated 取值可以是已接种疫苗（yes）或者未接种疫苗（no）；变量 result 取值可以是感染后存活（lived）或者感染后死亡（died）。图 3.15 和图 3.16 总结了这些数据。

		inoculated		
		yes	no	Total
result	lived	238	5136	5374
	died	6	844	850
	Total	244	5980	6224

图 3.15：关于 smallpox 数据集的列联表。

		inoculated		
		yes	no	Total
result	lived	0.0382	0.8252	0.8634
	died	0.0010	0.1356	0.1366
	Total	0.0392	0.9608	1.0000

图 3.16：关于 smallpox 数据集的比例分布表，由每个计数除以总数 6224 得出。

[1] (a) $P\left(\text{mach\_learn 取值是 pred\_not} \,\middle|\, \text{truth 取值是 fashion}\right) = 0.0615/0.1696 = 0.3626$；(b) 根据之前题目计算的答案（见脚注），可得加和为 1；(c) 之所以和为 1，是因为我们相当于基于同一个条件下，对另一个事件以及事件的补集进行了加总。因为当我们人为判断照片为时尚类的条件限定后，这些照片只可能存在两种情况：即要么被机器学习分类器也判定为时尚类，要么没有被机器学习分类器判定为时尚类。正因如此，所以上一个小题中让计算的其实就是在特定条件下事件和事件补集发生的概率和，自然应该就是 1。

指导练习 3.31

请使用规范标记法（即用 P 表示）写出“随机选取一个没有接种疫苗的人，他最后死于天花”的概率，并运用图 3.16 中的信息列式计算出这个概率。[1]

指导练习 3.32

计算一个人接种疫苗且死于天花的概率。这一结果和指导练习 3.31 中的结果有什么差异？[2]

指导练习 3.33

已知波士顿的人们自主选择是否接种疫苗，请思考下面几个问题。

(a) 这是一项观测性研究还是试验？

(b) 我们可以借助这些数据推断出接种疫苗和降低死亡率之间的因果关系吗？

(c) 有没有什么混淆变量会影响 inoculated 和 result 的取值？[3]

3.2.5 普适乘法法则

在 3.1.7 中我们已经了解过独立事件之间的乘法法则。本章我们将介绍一种**普适乘法法则**，这一法则对于并非相互独立的事件同样适用。

普适乘法法则

已知 A 和 B 指代两个事件或结果，则

$$P(A \,\&\, B) = P(A \mid B) \times P(B)$$

把 A 看作结果，B 看作条件，更便于理解。

以上普适乘法法则实际上只是把条件概率的公式进行了重新排列。

注：条件概率公式是 $P(A \mid B) = P(A \,\&\, B)/P(B)$。

[1] $P(\text{result 取值是 died} \mid \text{inoculated 取值是 no}) = \frac{P(\text{result 取值是 died 并且 inoculated 取值是 no})}{P(\text{inoculated 取值是 no})} = \frac{0.1356}{0.9608} = 0.1411$

[2] $P(\text{result 取值是 died} \mid \text{inoculated 取值是 yes}) = \frac{P(\text{result 取值是 died 并且 inoculated 取值是 yes})}{P(\text{inoculated 取值是 yes})} = \frac{0.0010}{0.0392} = 0.0255$。粗略估算下，接种疫苗群里的死亡率大约 1/40，而未接种疫苗群体的死亡率大约是 1/7。

[3] (a) 观测性研究；(b) 不能，我们不能基于观测性研究做任何因果推断；(c) 可能有多种混淆变量，比如是否有条件享受最前沿的医疗资源。

示例 3.34

以 smallpox 数据集为例来看。假设我们已知两点信息：（1）96.08%的居民未接种疫苗，（2）未接种疫苗的居民中有 85.88%感染后存活了下来。我们怎样计算一个居民未接种疫苗但感染后存活的概率?

答案：我们可以利用普适乘法法则来计算结果，然后用表 3.16 的数据来验证结果是否正确。我们要计算：

E

$$P(\text{result 取值 lived 且 inoculated 取值 no})$$

又已知：

$$P(\text{result 取值 lived} \mid \text{inoculated 取值 no}) = 0.8588 \qquad P(\text{inoculated 取值为 no}) = 0.9608$$

也就是有 96.08%的居民未接种疫苗，这些居民中有 85.88%感染后存活了下来：

$$P(\text{result 取值 lived 且 inoculated 取值 no}) = 0.8588 \times 0.9608 = 0.8251$$

这一计算符合普适乘法法则。计算结果与图 3.16 中联合概率 0.8252 一致（若不考虑四舍五入）。

指导练习 3.35

G

若已知 $P(\text{inoculated 取值为 yes}) = 0.0392$，$P(\text{result 取值 lived} \mid \text{inoculated 取值 yes}) = 0.9754$，计算一个人接种疫苗且感染后存活的概率。[1]

指导练习 3.36

G

如果接种疫苗的人们感染后存活率为 97.54%，那么接种疫苗的人们感染后死亡的概率是多少？[2]

条件概率之和

我们令 $A_1, \ldots, A_k$ 代表一个变量或某个过程可以取到的所有相互独立的结果。那么，基于事件 B 发生的前提下，我们可以得到：

$$P(A_1 \mid B) + \cdots + P(A_k \mid B) = 1$$

对于 A 和 A 的互补事件 A^c 来说，同样基于事件 B 发生的前提下，我们也可以得到：

$$P(A \mid B) = 1 - P(A^c \mid B)$$

指导练习 3.37

G

基于上面计算出的概率判断，接种疫苗是否能降低感染天花人群的死亡率？[3]

---

[1] 答案是 0.0382，可以在图 3.16 中得到验证。
[2] 因为总共只有两种可能的结果：存活和死亡，所以接种疫苗的人们感染后死亡的概率为 100% - 97.54% = 2.46%。
[3] 基于指导练习 3.33，因为这是一个观察性研究，所以并不能由此确定接种疫苗和感染死亡率之间的因果关系。

3.2.6 条件概率中的独立事件

如果两个事件相互独立，那么知道其中一个事件的结果对于我们估计另一个事件发生的概率没有什么帮助。我们可以用条件概率来证明这个论断在数学上的合理性。

指导练习 3.38

我们令 X 和 Y 分别代表掷出两个骰子的结果。

(a) X 取值为 1 的概率是多少?

(b) X 和 Y 的取值都为 1 的概率是多少?

(c) 利用条件概率的公式计算 $P(Y = 1 \mid X = 1)$。

(d) 计算 $P(Y = 1)$，结果与 $P(Y = 1 \mid X = 1)$ 相同吗？请说明原因。[1]

根据指导练习 3.38 可知，利用乘法法则计算独立事件的概率时，条件不会影响结果：

$$\begin{aligned} P(Y = 1 \mid X = 1) &= \frac{P(Y = 1 \ \& \ X = 1)}{\mathrm{P(X = 1)}} \\ &= \frac{P(Y = 1) \times P(X = 1)}{\mathrm{P(X = 1)}} \\ &= \mathrm{P}(Y = 1) \end{aligned}$$

指导练习 3.39

罗恩在赌场参加轮盘赌，并且他注意到前面五次小球都停在了黑色格子里。因为计算可知小球连续六次停在黑格里的概率很小（大约 1/64），所以他选择在红色区投注。他的推理过程有什么问题？[2]

3.2.7 普适乘法法则

树形图 Tree Diagram 可以把数据分析的结果和概率用树形结构展示出来。如果处理一组数据需要进行多个步骤，且每一个步骤都以前一个为基础时，树形图的用处最大。

据此，**smallpox** 这个数据集就很符合树形图的使用场景。首先，变量 inoculation 可以把样本群体分为 yes（已接种）和 no（未接种）两部分。然后，再看每部分人口存活率。图 3.17 中树形图展示出了这种结构。代表 inoculation 的第一个分支是**主要分支 primary branch**，后面的是**次要分支 secondary branch**。

[1] (a) 1/6；(b) 1/36；(c) $\frac{P(Y=1 \ \& \ X=1)}{\mathrm{P(X=1)}} = \frac{1/36}{1/6} = \frac{1}{6}$；(d) 结果和 (c) 相同，$P(Y = 1) = 1/6$，因为 X 和 Y 相互独立，所以 Y=1 发生的概率不受到 X 的影响。

[2] 他忽略了各次轮盘赌的结果相互独立。事实上，许多赌场会利用赌徒的这种心理，将博彩游戏之前的几次结果公开，使赌徒误信下一次赔率更小。这就是**赌徒谬误**。

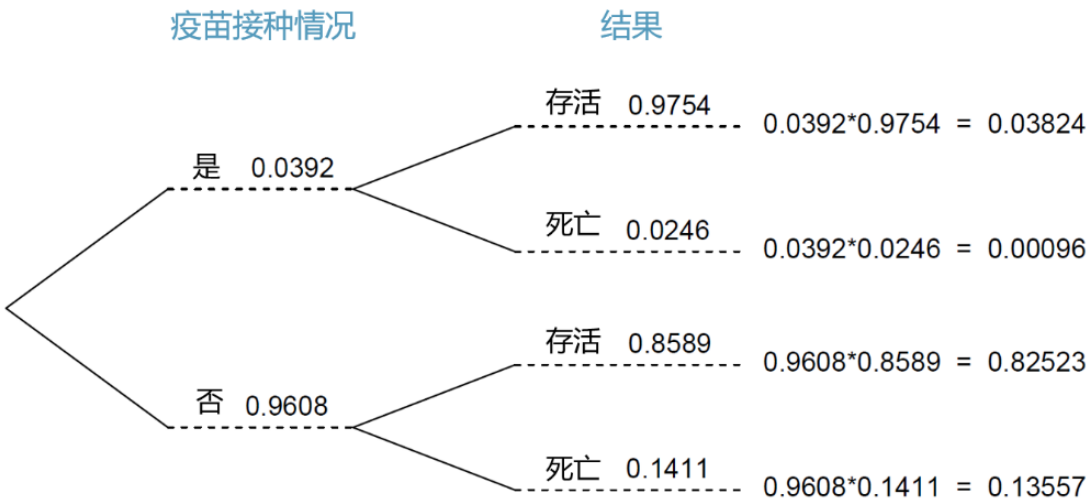


图 3.17：基于 smallpox 数据集的树形图。

如图 3.17，树形图注明了边际概率和条件概率。图 3.17 中，变量 inoculation 可以把 smallpox 数据集中的个体分为已接种和未接种两部分，边际概率分别为 0.0392 和 0.9608。基于上一级的条件，我们可以计算出次级分支的条件概率。例如，在图 3.17 中，最上面的分支代表在已接种的前提下（即 inoculation 变量取 yes）个体存活（即 result 变量取 lived）的概率。要计算这个分支上的联合概率，我们通常会把从左到右的所有数字相乘，结果写在分支最后。这个计算过程遵循普适乘法法则：

$$
\begin{aligned}
&P(\text{inoculated 取值 yes 且 result 取值 lived}) \\
&\quad = P(\text{inoculated 取值 yes}) \times P(\text{result 取值 lived} \mid \text{inoculated 取值 yes}) \\
&\quad = 0.0392 \times 0.9754 = 0.0382
\end{aligned}
$$

示例 3.40

我们来看一门统计学课程的期中和期末考试成绩数据。有 13%的学生在期中考试得到 A。在期中考试得到 A 的学生有 47%在期末考试也得到了 A，而在期中考试没有得到 A 的学生只有 11%在期末考试得到了 A。随机选取一个学生，已知他在期末考试得到了 A，那么他在期中考试也得到 A 的概率是多少?

答案：题目要求计算$P(\text{期中成绩为 A} \mid \text{期末成绩为 A})$。

E

而为了要计算出这个结果，我们需要知道$P(\text{期中成绩为 A}\,\&\,\text{期末成绩为 A})$和$P(\text{期末成绩为 A})$的值。但题干中并没有提供相关信息，且由已知信息不容易计算出来。在遇到这种涉及概率较多且解题暂时没有思路的情景，构建树形图就会很有帮助。

构建树形图时，先发生事件对应变量的边际概率通常用于构建主要分支。本例中，在主要分支上的边际概率对应期中考试结果，而次要分支上的条件概率对应在期中考试成绩已知的前提下期末考试结果。列出树形图后，可以计算出所需的概率值。

要计算边际概率$P(\text{期末成绩为 A})$，可以把次要分支上期末成绩为 A 对应的所有联合概率相加：

E

$$P(\text{期中成绩为 A}\,\&\,\text{期末成绩为 A}) = 0.0611$$

$$P(\text{期末成绩为 A})$$

$$= P(\text{期中成绩为其他}\,\&\,\text{期末成绩为 A}) + P(\text{期中成绩为 A}\,\&\,\text{期末成绩为 A})$$

$$= 0.0957 + 0.0611 = 0.1568$$

然后把两个概率相除可得：

$$P(\text{期中成绩为 A}\,|\,\text{期末成绩为 A}) = \frac{P(\text{期中成绩为 A}\,\&\,\text{期末成绩为 A})}{P(\text{期末成绩为 A})}$$

$$= \frac{0.0611}{0.1568} = 0.3897$$

所以随机选取一个学生，已知他在期末考试得到了 A，那么他期中考试成绩是 A 的概率约为 0.39。

指导练习 3.41

G

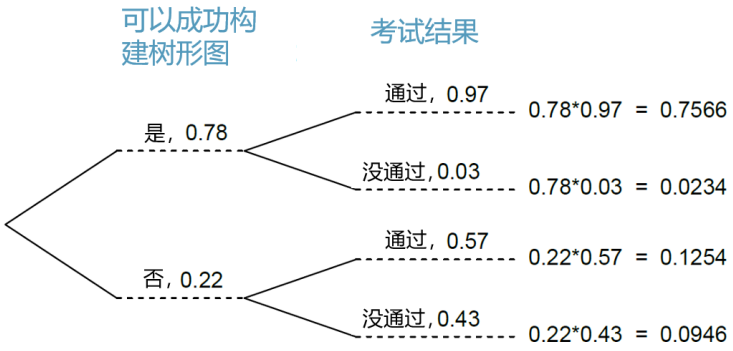
已知：学过统计学入门课程后，78%的学生可以成功构建出树形图。可以成功构建出树形图的学生中，有 97%通过了考试；而无法成功构建出树形图的学生中，只有 57%能通过考试。

(a) 根据题干信息构建树形图。

(b) 随机选取一个学生，他通过考试的概率是多少？

(c) 已知一个学生通过了考试，那么他能够构建树形图的概率是多少？[1]

[1] (a)如下图；(b)找出通过考试的两个联合概率并相加，得到$P(\text{通过考试}) = 0.76566 + 0.1254 = 0.8820$；
(c)$P(\text{可以成功构建树形图}\,|\,\text{通过考试}) = P(\text{可以成功构建树形图}\,\&\,\text{通过考试})/P(\text{通过考试}) = 0.7566/0.8820 = 0.8578$。



3.2.8 贝叶斯公式

很多时候我们已知条件概率：

$$P(\text{变量 1 取值情况} \mid \text{变量 2 取值情况})$$

但实际上需要取反，求条件概率：

$$P(\text{变量 2 取值情况} \mid \text{变量 1 取值情况})$$

已知第一个条件概率时，可以利用树形图求第二个条件概率。但有时利用树形图也求不出来。面对这种情况，有一个通用公式**贝叶斯公式 Bayer’s Theorem** 非常有用。

首先，我们来看一个使用贝叶斯公式取反求条件概率的例子，同时还要借助树形图来分析。

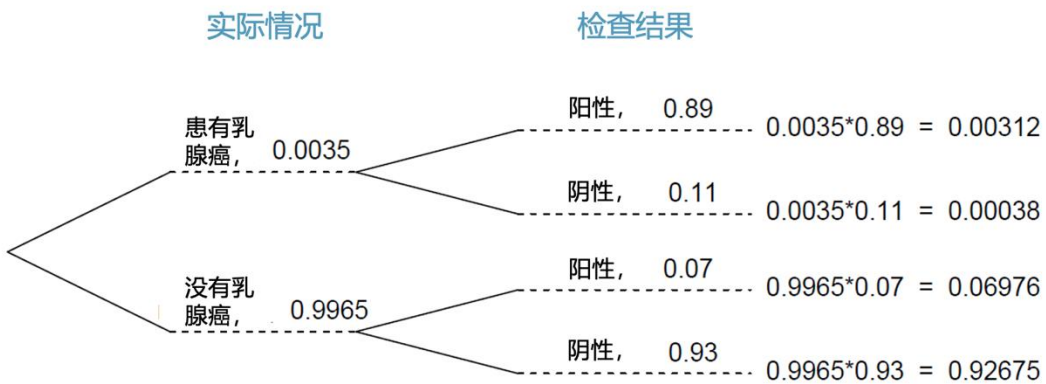
示例 3.42

在加拿大，超过 40 岁的女性每年约有 0.35%会患乳腺癌。人们常用乳房 X 光检测来筛查乳腺癌，但这种检测的结果并不精确。约 11%的乳腺癌患者在乳房 X 光检测后得到**假阴性 false negative** 结果，也就是她患有乳腺癌却没有被检测出来。同样地，约有 7%没有得乳腺癌的女性得到**假阳性 false positive** 的结果，即他们被检测出患有乳腺癌但实际上并没有。在一次实验中，我们随机选取一个超过 40 岁的女士，通过乳房 X 光检测为她筛查乳腺癌，结果为阳性，即检测表明她得了乳腺癌。那么，这位女士确实患有乳腺癌的概率是多少?

答案：利用题干中已知信息，很容易计算出一位乳腺癌患者得到阳性检测结果的概率（1.00-0.11=0.89)。然而，我们需要取反，求出一位女士在检测结果为阳性的条件下确实患乳腺癌的概率。（注意一个隐含的医学术语：检测结果为 *positive* 阳性，往往指检查结果显示患有某种疾病，例如这里通过乳房 X 光检测，判断一个人可能患有癌症)。要计算的概率可以转换成计算公式右侧的两部分：

$$P(\text{得了乳腺癌} \mid \text{X 光阳性}) = \frac{P(\text{得了乳腺癌 \& X 光阳性})}{P(\text{X 光阳性})}$$

然后构建如下树形图：



一位女士患有乳腺癌且检查结果为阳性的概率为：

$$P(\text{患有乳腺癌 \& 检查结果为阳性}) = P(\text{检查结果为阳性} \mid \text{患有乳腺癌})\, P(\text{患有乳腺癌})$$

$$= 0.89 \times 0.0035 = 0.00312$$

一位女士检查结果为阳性的概率可以通过求和得到

$$P(\text{检查结果为阳性}) = P(\text{检查结果为阳性 \& 患有乳腺癌}) + P(\text{检查结果为阳性 \& 没有乳腺癌})$$

$$= P(\text{患有乳腺癌})\, P(\text{检查结果为阳性} \mid \text{患有乳腺癌})$$

$$+ P(\text{没有乳腺癌})\, P(\text{检查结果为阳性} \mid \text{没有乳腺癌})$$

$$= 0.0035 \times 0.89 + 0.9965 \times 0.07 = 0.07288$$

然后可以计算出一位女士在检查结果为阳性的条件下确实患乳腺癌的概率

$$P(\text{患有乳腺癌} \mid \text{检查结果为阳性}) = \frac{P(\text{患有乳腺癌 \& 检查结果为阳性})}{P(\text{检查结果为阳性})}$$

$$= \frac{0.00312}{0.07288} \approx 0.0428$$

也就是说，当一位女士得到乳腺癌阳性检测结果时，她只有约 4.3%的概率确实患病。

通过示例 3.42 就不难理解，在得到一次阳性检查结果后，为什么医生一般不会直接提出治疗，而是会优先安排更多轮次和种类的检查加以确诊。对于罕见的疾病，一次阳性检测结果通常不足以断定真实的患病情况。那么我们再看示例 3.42 中最后一个公式，从树形图可知等号右侧的分子部分可以表示为：

$$P(\text{患有乳腺癌 \& 检查结果为阳性}) = P(\text{检查结果为阳性} \mid \text{患有乳腺癌})\, P(\text{患有乳腺癌})$$

而分母部分，也就是检查结果为阳性的概率，等于不同条件下检查结果为阳性的概率之和：

$$P(\text{检查结果为阳性}) = P(\text{检查结果为阳性 \& 患有乳腺癌}) + P(\text{检查结果为阳性 \& 没有乳腺癌})$$

本例中，树形图上右边的联合概率都可以表示为条件概率和边际概率的乘积：

$$P(\text{检查结果为阳性}) = P(\text{检查结果为阳性 \& 患有乳腺癌}) + P(\text{检查结果为阳性 \& 没有乳腺癌})$$

$$= P(\text{患有乳腺癌})\, P(\text{检查结果为阳性} \mid \text{患有乳腺癌})$$

$$+ P(\text{没有乳腺癌})\, P(\text{检查结果为阳性} \mid \text{没有乳腺癌})$$

这里应用了贝叶斯公式，把原条件概率公式中的分子和分母替换为相应的概率计算表达式。

贝叶斯公式：概率取反

关于变量 1 和变量 2，计算条件概率：

$$P(\text{变量 1 的}A_1\text{ 结果} \mid \text{变量 2 的 B 结果})$$

利用贝叶斯公式可以把这个条件概率表示成如下分数：

$$\frac{P(B \mid A_1)P(A_1)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \cdots + P(B \mid A_k)P(A_k)}$$

这里A_2，A_3，… A_k 代表变量 1 的除 A_1 以外的所有可能结果。

贝叶斯公式实质上是把树形图进行了规律化的总结。上述式子中，分子表示了 A_1 和 B 同时发生的概率，分母表示了B 发生的边际概率。而分母之所以这么长的原因，是因为我们需要把各种情况下 B 发生的概率相加，这一步可以直接在树形图中体现出来。在贝叶斯公式的实际操作中，为了对计算过程进行简化，我们通常会先做一些前期的准备，再套用公式，这样整个公式就不会显得太过复杂了。我们需要的前期准备包括：

(1) 首先计算出变量 1 各种可能结果的边际概率，即 $P(A_1), P(A_2), \ldots, P(A_k)$；

(2)然后计算 B 基于变量 1 各种可能结果下发生的概率，即 $P(B \mid A_1), P(B \mid A_2), \ldots, P(B \mid A_k)$。

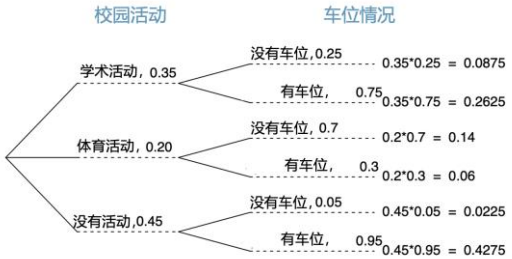
当以上所有概率都计算出来以后，我们把它们逐一带入贝叶斯公式，就可以计算出 $P(\text{变量 1 的}A_1\text{ 结果} \mid \text{变量 2 的 B 结果})$。当变量 1 有很多种可能出现的结果时，画树形图会太过复杂，贝叶斯公式就更为适用。

指导练习 3.43

G

何塞每周四晚上去学校，但有时因为校园活动，学校的停车场会没有空余车位。学校 35%的晚上会有学术活动，20%的晚上会有体育活动，40%的晚上没有活动。在有学术活动时，25%的情况下会没有车位；在有体育活动时，70%的情况下会没有车位；在没有活动时，只有 5%的情况下会没有车位。如果何塞来到学校发现没有车位了，那么当天晚上有体育活动的概率是多少？请用树形图来解答这个问题。[1]

[1] 树形图如下，由图可得：（1）举办体育活动并且没有车位的概率为 0.14；（2）没有车位的概率为 0.0875 + 0.14 + 0.0225 = 0.25。因此，在没有车位的情况下，当天晚上有体育活动的概率为 0.14/0.25 = 56%。



示例 3.44

同样是指导练习 3.43 的问题，我们现在来用贝叶斯公式进行解答。

答案：我们用 A_1 表示学校举办体育活动，用 A_2 表示学校举办学术活动，用 A_3 表示学校没有活动，用 B 表示停车场没有空余车位。那么，根据题目已知信息，我们可以得到：

$$P(A_1) = 0.2 \quad P(A_2) = 0.35 \quad P(A_3) = 0.45$$

E

$$P(B \mid A_1) = 0.2 \quad P(B \mid A_2) = 0.35 \quad P(B \mid A_3) = 0.45$$

根据贝叶斯公式，可以得到在停车场没有车位（B）的条件下，学校举办体育活动（A_1）的概率为：

$$P(A_1 \mid B) = \frac{P(B \mid A_1)P(A_1)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + P(B \mid A_3)P(A_3)}$$

$$= \frac{(0.7)(0.2)}{(0.7)(0.2) + (0.25)(0.35) + (0.05)(0.45)}$$

$$= 0.56$$

因此，在没有车位的情况下，学校举办体育活动的概率为 56%。

G

指导练习 3.45

通过之前的指导练习和示例，证明在没有车位的情况下，学校举办学术活动的概率是 0.35。[1]

G

指导练习 3.46

通过指导练习 3.43 和 3.45，我们知道了在没有车位的情况下，学校举办体育活动和学术活动的概率分别是 0.56 和 0.35，由此计算$P(\text{学校没有活动} \mid \text{停车场没有车位})$。[2]

通过以上几个练习，我们对于在停车场没有车位的前提下，学校举办活动的各种情况有了进一步的认识。实际上，相同的思路奠定了统计学中一个重要篇章—贝叶斯统计—的基础。不过我们在本书中不会对贝叶斯统计进行深入的讲解。

---

[1] $P(A_2 \mid B) = \frac{P(B \mid A_2)P(A_2)}{P(B \mid A_1)P(A_1)+P(B \mid A_2)P(A_2)+P(B \mid A_3)P(A_3)} = \frac{(0.25)(0.35)}{(0.7)(0.2)+(0.25)(0.35)+(0.05)(0.45)} = 0.35$

[2] 因为三种情况都是基于停车场没有车位这个条件，所以计算结果为$1 - 0.56 - 0.35 = 0.09$。

3.3 小样本取样

我们从总体中抽样调查时，通常只能选取一小部分个体。然而，有时当我们的资源和精力允许我们去选择足够大的样本或要研究课题本身的总体数量就很小的时候，我们就能以无放回的方式（无放回抽样指同一个体不能被重复取样）抽取超过总体数量 10%的样本，甚至有时候对全部总体进行分析。抽样比例的多少将直接决定我们即将采取的分析方法，举个简单的例子，如果样本覆盖了全部总体，那所谓抽样其实也就是总体分析了，即枚举遍历总体中的每个观测个体。

示例 3.47

E

老师通常会随机选一个学生回答某个问题。如果每个同学都有同等的机会被选中，且你的班上有 15 名同学，那么老师在下个问题点到你的概率是多少?

答案：如果班上有 15 名同学，且没有人逃课，那么所求概率就是 1/15，约 0.067。

示例 3.48

如果老师准备提问三个问题，且不会在一节课上重复提问同一个人，那么你都没被选到的概率是多少?

E

答案：对于老师提的第一个问题，她提问其他人的概率是 14/15。老师要提问第二个问题时，班上只有 14 个同学还没有被提问过。因此，如果你在第一个问题没被选到，再次不被选到的概率其实只有 13/14。类似地，你在前两个问题中落选后，第三个问题也没有被提问到的概率是 12/13。那么，三题都没有被提问的概率为：

$$\begin{aligned} &P(\text{三次都落选}) \\ &\quad = P(\text{第一个问题落选 \& 第二个问题落选 \& 第三个问题落选}) \\ &\quad = \frac{14}{15} \times \frac{13}{14} \times \frac{12}{13} = \frac{12}{15} = 0.80 \end{aligned}$$

G

指导练习 3.49

示例 3.48 中概率相乘的指导原理是什么？[1]

[1] 我们计算出来的三个概率实际是一个边际概率加两个条件概率，即 $P(\text{第一个问题落选})$，和两个条件概率，即 $P(\text{第二个问题落选} \mid \text{第一个问题落选})$ 和 $P(\text{第三个问题落选} \mid \text{前两个问题都落选})$。使用普适乘法法则，三个概率概率相乘可以得到三个问题都落选的概率。

示例 3.50

在 3.48 的题设下，假设老师提问时不会考虑已经提问过哪位同学，也就是一个同学可能会被多次提问。那么一个同学三次都没有被提问的概率是多少？

答案：三次提问的结果相互独立，那么每次没被提问到的概率都是 14/15。因此，对于相互独立的过程我们可以直接使用乘法法则计算。

E

$$\begin{aligned} &P\big(\text{三次都落选}\big) \\ &\quad = P\big(\text{第一个问题落选 \& 第二个问题落选 \& 第三个问题落选}\big) \\ &\quad = \frac{14}{15} \times \frac{14}{15} \times \frac{14}{15} = 0.813 \end{aligned}$$

所以，相较于不重复提问同一个人，老师可能重复提问时，一个同学三次都没被提问到的概率更高一点，但同时也有可能被提问不止一次。

指导练习 3.51

G

在示例 3.50 的条件下，一个同学被重复提问三次的概率是多少？[1]

指导练习 3.52

在某部门为了反馈员工的辛勤劳动，决定举行福利抽奖活动，假设部门内总共有 30 人，每人把自己的名字写到一个纸条上放在抽奖箱中，然后通过抽奖送出七个奖品。

G

(a) 把写有所有员工名字的纸条放在不透明抽奖箱中，然后抽取 7 次，每次抽出的员工可以领一个奖品。我们规定以无放回地方式抽取奖券，也就是拿出的纸条不会再放回去。那么对于每一个员工来说，中奖的概率是多少？

(b) 如果每次抽出的纸条会放回，那么对每个员工来说，7 次中至少中一次奖的概率又是多少呢？[2]

指导练习 3.53

G

比较指导练习 3.52 中的两个答案，抽样方式对中奖概率有多大的影响？[3]

如果重复指导练习 3.52 中的抽样，但把 30 张纸条换成 300 张纸条，会发现结果很有趣：两种抽样方式得到的中奖概率几乎相等。以无放回地方式抽奖时中奖概率为 0.0233，纸条放回时中奖（这里的中奖指至少中一次奖）概率是 0.0231，当样本大小（该场景下为 7）只占总体（按照上方新作的假设为 300）的很小一部分时（低于 10%），即使以有放回的方式取样，抽样过程也近似相互独立。

---

[1] $P\big(\text{被重复提问三次}\big) = \frac{1}{15} \times \frac{1}{15} \times \frac{1}{15} = 0.00030$

[2] (a) 为了练习普适乘法法则和概率论，我们这次先计算未中奖的概率。以无放回的方式取票，第一次未中奖概率为 29/30，第二次 28/29，…，第七次未获奖的概率是 23/24。未中奖的概率是这些概率的乘积 23/30。那么中奖的概率就是 $1 - 23/30 = 7/30 = 0.233$；(b) 奖券会被放回时，七次抽取的结果相互独立。我们还是先计算未中奖的概率：$(29/30)^7 \approx 0.79$。因此，至少中奖一次的概率是 0.21。

[3] 以无放回地方式抽取奖券时，中奖概率会高出 10%，但同时最多只能赢得一个奖品。

3.4 随机变量

使用**随机变量 random variable** 模拟现实过程非常有用，它能帮助我们利用数学体系和统计原理来更好地理解和预测现实。但首先，我们要先通过案例来感受下随机变量的产生过程。该过程也经常被称为**随机过程 random process**。

示例 3.54

某大学的统计学课会指定两本教材：一本教科书和一本相应的辅导书。该大学书店发现有 20%的学生一本书都不买，55%的学生只买教科书，25%的学生则两本都会买。经长期观察，这些比例在每个学期都相似。如果已知有 100 名新生将要入学，该书店需要两种书各准备多少本?

E

答案：我们按比例做简单的乘法估算，发现约有 20 个学生一本书都不买，对应 0 本。55 个同学只买教科书，共计 55 本。约 25 个同学买两本书，共计 50 本。所以预计书店会需要准备 105 本书，其中教科书 80 本，辅导书 25 本。

指导练习 3.55

G

如果这家书店卖出的书略多于或少于 105 本，你会感到惊讶吗？[1]

示例 3.56

教科书单价 137 元，辅导书单价 33 元。预计书店能收入多少钱?

答案：约 55 个同学只买一本书，收入总计有：

E

$$137 \times 55 = 7535$$

约 25 个同学买两本书，收入总计：

$$(137 + 33) \times 25 = 170 \times 25 = 4250$$

因此，预计书店向这 100 个同学卖书能取得收入：$7535 + 4250 = 11785$ 元。

然而，随机变量存在波动性，所以实际收入数量可能有些不同。

示例 3.57

按照上例的计算过程，书店平均从每个同学能获得多少收入?

E

答案：预计书店的总收入是 11785 元，有 100 个同学。平均从每个同学获得收入：

$$11785/100 = 117.85$$

[1] 书最后可能卖得多一点或少一点，这不足为奇。本书第一章中，我们已经介绍过了实际观测到的统计结果相比概率计算来说难免存在天然偏差。例如，如果我们抛一枚硬币 100 次，正面出现的频率通常不会正好是 50%，但应该非常接近 50%。

3.4.1 期望

如果某随机过程会产生一个数字的统计结果，我们就把该过程产生的这个结果称为**随机变量 random variable**，通常用大写字母 X，Y，Z 表示。例如，一个学生买统计学课本（过程）花费的钱数（数字统计结果）就是一个随机变量，可以用 X 表示。这里需要注意，在上面学生买课本的例子中，我们说一个学生买课本的过程是个随机过程，这是从全体学生角度出发考虑的。它的随机性体现在当我们随机从全体学生中挑选一个学生进行观察时，我们可能随机到以下三种结果之一：

这名学生只购买了一本教科书

这名学生只购买了一本辅导书

这名学生选择两本书都买了

在这里我们要区分两个概念：随机变量和随机变量的一次取值。例如当我们研究 30 岁男性身高这个随机变量的时候，如果我们已经选定了某个个体，例如我名为玉晨的好朋友，那么他的个人身高的统计结果显然是个确定数字，比如说 180 公分。这里「玉晨的身高：180 公分」不是一个随机变量，而是 30 岁男性身高这一随机变量的一次取值。所以随机变量的随机性不体现在其确定的统计结果和取值中，而体现在从总体中随机抽取个体的过程中。

随机变量

随机变量代表了一个会产生数字统计结果的随机过程。从统计学角度出发，我们可以按观测值分行，然后把统计结果记录在一列，这就对应了二维数据集中的变量。

回到书店销售教科书的案例上来，如果我们用 X 代表书店从一位学生处可能取得的收入这一随机变量，而具体取到的值则用对应小写字母 x 加下标来表示。例如，$x_1 = 0$ 的概率是 0.20，$x_2 = 137$ 的概率是 0.55，$x_3 = 170$ 的概率是 0.25。图 3.18 和图 3.19 总结了 X 的分布情况。

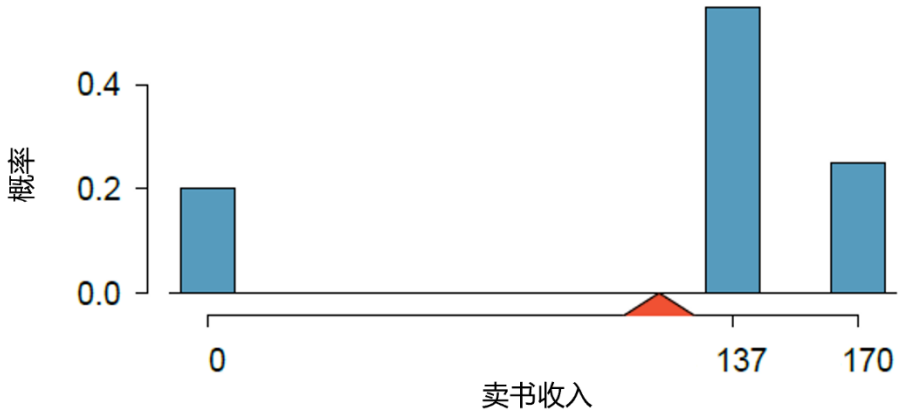


图 3.18：书店向每名学生销售教科书和辅导书收入的概率分布，其中三角形代表了在之前 100 名学生购书示例中书店从每名学生处获得的平均收入。

i	1	2	3	Total
x_i	\$0	\$137	\$170	–
$P(X = x_i)$	0.20	0.55	0.25	1.00

图 3.19：代表书店收入的 X 随机变量的概率分布表。

示例 3.57 中计算出 X 的平均结果为 117.85 元。我们称这个均值为 X 的期望值，用 $E(X)$ 表示。随机变量的期望值是将其可能取到的结果按概率加权后相加得到的：

$$E(X) = 0 \times P(X = 0) + 137 \times P(X = 137) + 170 \times P(X = 170)$$
$$= 0 \times 0.20 + 137 \times 0.55 + 170 \times 0.25 = 117.85$$

离散随机变量的期望值

如果变量 X 取值结果为 x_1，…，x_k，对应概率为 $P(X = x_1)$，…，$P(X = x_k)$，X 的期望就等于每个结果的取值乘以对应概率之后再相加：

$$E(X) = x_1 \times P(X = x_1) + \cdots + x_k \times P(X = x_k)$$
$$= \sum_{i=1}^{k} x_i P(X = x_i)$$

有时候，我们也可以用希腊字母 μ 可以在书写中代替 $E(X)$。

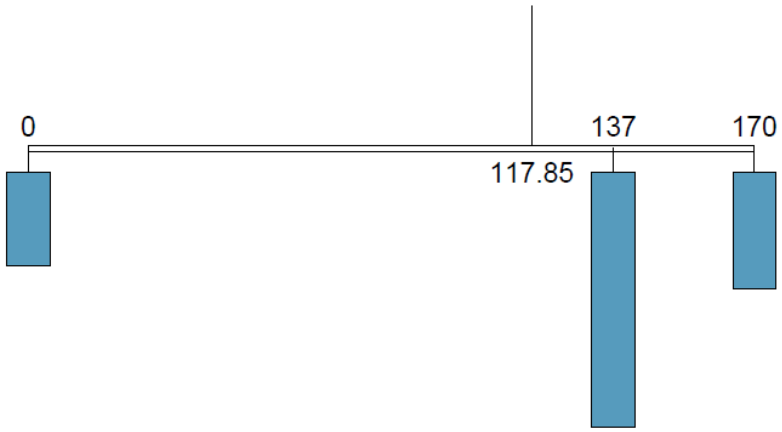


图 3.20：我们通过一个类似杆秤的系统来表示变量 X 的概率分布。上方垂下的绳索正好位于均值的位置，这样的布局保证了整个系统的平衡。

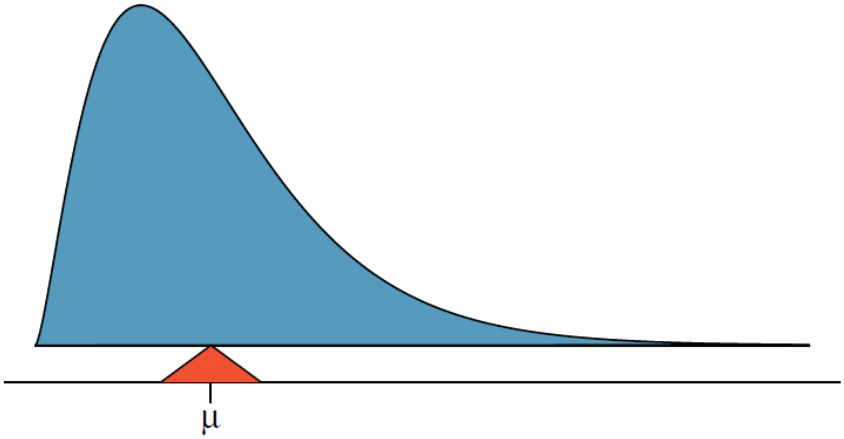


图 3.21：连续变量的分布也可以在其均值处保持平衡。

随机变量的期望值代表着平均的结果。例如，$E(X) = 117.85$ 就是书店期望从每个学生身上收到的平均金额，也可以写作 $\mu = 117.85$。连续随机变量的期望值也是可以计算的（见本章第 5 节）。不过对于连续变量的计算需要用到一些微积分知识。关于微积分的这部分我们在本书中不会提及，希望在之后更高级的数学课程中能够和大家进一步探讨分享。

在物理学中，期望和物体重心概念对应。随机变量的分布就好像物体的质量分布，而均值或者说期望值就对应了物体质量分布的平衡点，也就是重心。在图 3.18 和图 3.20 中可以看出，重心的概念也可以从离散随机变量扩展到连续随机变量。图 3.21 中，我们把一个连续随机变量的概率分布想象成一个物体，那么该变量的数学期望就是用于让整个物体平衡的红色楔子。

3.4.2 随机变量的离散程度

假设你就是之前书店案例中的书店经营者，除了预计的收入值，你还想知道收入是否稳定，也就是探索收入变量的波动性，或者说离散性。

方差和标准差可以用来描述随机变量的离散性。第 2.1.4 节中介绍了计算一个数据集的方差和标准差的方法：我们首先计算每个取值有别于均值的偏差 $(xi - \mu)$，对这些偏差进行平方，然后再取均值，从而求得方差。对于随机变量，我们还是沿用同样的方法。然而不同的是，我们在计算完偏差的平方后，不再取简单平均，而是需要先用它们对应的概率加权后再求和。

相信不少同学可能会记得在初级和中级的数学教学中，学过方差的计算公式。而彼时所学的公式中是对偏差的平方和直接取简单平均，而非加权平均。关于这点我们要做一个小小的展开：即某种程度上说，简单平均其实也是一种加权平均，只不过所有权重都一致。

相信这对于大家来说也不难理解，因为其实取简单平均的过程就可以理解为按照等概率加权的过程。例如我们观察下式，是对 1 到 6 这六个观测值进行简单取平均计算：

$$\frac{1+2+3+4+5+6}{6}$$

而该计算过程还可以写成：

$$1\times\frac{1}{6}+2\times\frac{1}{6}+3\times\frac{1}{6}+4\times\frac{1}{6}+5\times\frac{1}{6}+6\times\frac{1}{6}$$

这样改写后，就可以清晰地看到，1 到 6 六个数字对应了六个观测值，而 $1/6$ 则对应了每个观测值出现的概率，也就是权重。这也是为什么我们在前文说简单平均可以被视为一种特殊的加权平均。不过在实际计算简单算数平均数的时候，我们大可不必按照求期望的模式来书写。因为那样无疑是把原本简单的问题复杂化了。此处的说明只是为了帮助大家理解加权平均运算和求期望的关系，以及求平均运算的本质。

让我们回到方差的计算公式上来：我们先通过减去均值的方式计算偏差，然后对这些偏差平方，然后再对偏差平方进行加权平均，计算出最终的方差。计算出方差后，我们还可以通过计算方差的平方根来得到标准差，就和第 2.1.4 节中做法一样。

普适方差公式

如果变量 X 取值结果为 x_1，…，x_k，对应概率为 $P(X=x_1)$，…，$P(X=x_k)$，X 的期望值，也即是均值$\mu=E(X)$，用$Var(X)$或者 σ^2 表示 X 的方差：

$$\begin{aligned}\sigma^2 &= (x_1-\mu)^2\times P(X=x_1)+\cdots+(x_k-\mu)^2\times P(X=x_k)\\ &= \sum_{j=1}^{k}\left(x_j-\mu\right)^2 P\left(X=x_j\right)\end{aligned}$$

X 的标准差等于方差的平方根，用 σ 表示。

示例 3.58

若 X 指的是一名学生为书店贡献的收入，请计算 X 的期望，方差还有标准差。

答案：可以画表记录每个取值的计算，然后把结果相加：

i	1	2	3	Total
x_i	0	137	170	
$P(X = x_i)$	0.20	0.55	0.25	
$x_i \times P(X = x_i)$	0	75.35	42.50	117.85

E

根据上表，可以得到期望 $\mu = 117.85$，接着我们对上表计算方差：

i	1	2	3	Total
x_i	0	137	170	
$P(X = x_i)$	0.20	0.55	0.25	
$x_i \times P(X = x_i)$	0	75.35	42.50	117.85
$x_i - \mu$	-117.85	19.15	52.15	
$(x_i - \mu)^2$	13888.62	366.72	2719.62	
$(x_i - \mu)^2 \times P(X = x_i)$	2777.7	201.7	679.9	3659.3

可以得到 X 的方差 $\sigma_2 = 3659.3$，标准差 $\sigma = \sqrt{3659.3} = 60.49$。

指导练习 3.59

这家书店还提供一本售价 159 元的化学教科书和一本售价 41 元的增刊。据统计，他们已知大约 25% 的化学系学生只买教科书，60%的学生两样都买。

G

(a) 假定不买教科书的同学也不会买增刊，那么一本书都不买的学生占比多少？

(b) 如果用 Y 表示一个学生在购买化学教科书和增刊时为书店贡献的收入，请写出 Y 的概率分布，即画出一个表格表示每种可能的取值结果及其对应的概率；

(c) 计算平均一个化学系学生为书店贡献的收入；

(d) 计算出 Y 的标准差，用以描述一个化学系学生为书店贡献的收入的离散性。[1]

[1] (a) 100% − 25% − 60% = 15%；(b) 请见下表的前两行；(c) 由下表可知，期望是 159.75；(d) 标准差为 69.28。

i（场景）	1（不买书）	2（仅买书）	3（都买）	Total
y_i	0.00	159.00	200.00	
$P(Y = y_i)$	0.15	0.25	0.60	
$y_i \times P(Y = y_i)$	0.00	39.75	120.00	$E(Y) = 159.75$
$y_i - E(Y)$	-159.75	-0.75	40.25	
$(y_i - E(Y))^2$	25520.06	0.56	1620.06	
$(y_i - E(Y))^2 \times P(Y)$	3828.0	0.1	972.0	$Var(Y) \approx 4800$

3.4.3 随机变量的线性组合

到目前为止，我们已经研究了随机变量的定义，并学习了一个随机变量的期望、方差、标准差等的计算。然而有些场景更适合将多个随机变量组合起来。例如，如果要计算一个人每周在通勤上花费的总时间，把它写作周一到周五的通勤时间之和更容易计算。同样，对股票投资组合的总收益可以通过对其各个部分收益进行加权求和得到。

示例 3.60

老王每周有五天去上班。我们用 X_1 表示他周一的通勤时间，X_2 表示他周二的通勤时间，并以此类推。请用 X_1，…，X_5 列出方程来表示他一周的通勤时间，记作 W。

E

答案：一周的总通勤时间等于五天的通勤时间之和：

$$W = X_1 + X_2 + X_3 + X_4 + X_5$$

把每周通勤时间分成几部分，这样我们就对 W 完成了一个简单的线性建模。

示例 3.61

已知老王平均每天的通勤时间是 18 分钟。请问他每周通勤时间的期望是多少?

E

答案：我们已知平均每天的通勤时间是 18 分钟，也就是期望 $E(X_i) = 18$。要计算五天通勤时间之和的均值，只要把每天通勤时间的均值加起来即可（这里需要用到数学期望的可拆分性质；关于数学期望的性质，我们在本书中不做展开，此处用到的是可以把数学期望 E 符号后括号内用加减相连的内容拆成独立的多个期望相加的性质）：

$$\begin{aligned} E(W) &= E(X_1 + X_2 + X_3 + X_4 + X_5) \\ &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\ &= 18 + 18 + 18 + 18 + 18 = 90\,minutes \end{aligned}$$

通勤总时长的期望等于每天通勤时长的期望之和，这里应用到了数学期望的加法性质。

指导练习 3.62

G

老马准备在现金拍卖会上出售一台电视，同时购买一台烤面包机。如果 X 代表出售电视的收入，用随机变量 Y 代表购买烤面包机的支出，请用含 X 和 Y 的方程表示出老王的净收入或支出。[1]

指导练习 3.63

G

基于示例 3.62，老马出售电视预计收入 175 元，购买烤面包机支出 23 元，所以预计净收入或支出多少元? [2]

[1] 出售电视收入 X 元，购买烤面包机支出 Y 元，所以净收入或支出 $X - Y$ 元。

[2] $E(X - Y) = E(X) - E(Y) = 175 - 23 = 152$，按此计算，老马将收入 152 元。

指导练习 3.64

如果老王一周的通勤时间实际上不等于 90 分钟，或者老马实际收入也不是 152 元，你应该感到惊讶吗？[1]

截至目前我们已经介绍了关于随机变量线性组合的两个重要概念。第一，一个组合式的最终结果有时可以写作式子中各部分之和。第二，一般来说，组合式的期望等于各部分的期望相加。需要注意的是，第二点仅适用于随机变量的线性组合，即对进行加减运算，或者乘以或除以常数。变量和变量之间的相乘关系，或者变量自身的平方计算都属于典型非线性组合方式，它们也都不再适配数学期望运算的拆分性质。

综上，两个随机变量 X 和 Y 的**线性组合 linear combination** 可以写作：

$$aX + bY$$

其中 a 和 b 是已知的某个固定值，称作系数。以老王的通勤时间为例，每个工作日的通勤时间都是一个随机变量，一周共有五个随机变量。每个随机变量的系数都是 1，得到方程：

$$1X_1 + 1X_2 + 1X_3 + 1X_4 + 1X_5$$

对于老马拍卖的净收入或支出，随机变量 X 的系数是+1，Y 的系数则是-1。

要计算随机变量的线性组合的均值，我们可以把每个随机变量的均值代入，相加得到最后结果。再次强调一下，对于随机变量的非线性组合，这一方法并不适用。例如，X 和 Y 是随机变量，要计算两者的非线性组合 $X^2 + Y$，$X \times Y$，X/Y 等，把各部分随机变量的均值代入公式并不能得到组合的均值。

随机变量的线性组合和均值

已知随机变量 X 和 Y，两者的线性组合是 $aX + bY$，a 和 b 是常数。

要计算随机变量的线性组合的均值，把各个随机变量的均值代入公式：

$$a \times E(X) + b \times E(Y)$$

随机变量的均值也是该变量的期望，也就是说 $E(X) = \mu_X$。

[1] 不会惊讶，因为实际观测的结果出现波动很正常。尽管平均下来每天的通勤时间为 18 分钟，也完全有可能某天走得慢一点，或者遇到老奶奶过马路上前乐于助人等等而导致实际结果和预期期望之间出现偏差；对于老马的例子来说，拍卖品的质量和竞买人的兴趣等因素的变化都会使拍卖价格有所不同。

示例 3.65

在美国留学期间，小李通过移动交易软件 Robinhood 购买了卡特彼勒公司（股票代码 CAT）价值 6000 美金的股票，以及埃克森美孚公司（股票代码 XOM）价值 2000 美金的股票。如果 X 代表 CAT 股票下个月股票的价格变化百分比，Y 代表 XOM 下个月股票的价格变化百分比，请用含X 和 Y 的式子表示下个月小李的股票会赚多少钱或损失多少钱。

E

答案：为简单起见，我们假定 X 和 Y 不采用百分位制而是小数形式（例如，如果 CAT 的股票价格增长了 1%，那么 $X = 0.01$；如果损失 1%，则 $X = -0.01$)。那么小李的收益为：

$$\$6000 \times X \ + \ \$2000 \times Y$$

把 X 和 Y 的值代入方程就可以得到这个月小李的股票投资组合的价值变化。结果得到正数代表盈利，负数代表亏损。

指导练习 3.66

G

若卡特彼勒和埃克森美孚的股价该月预计的涨跌幅分别为 2.0%和 0.2%，计算下个月小李股票投资组合的预期变化。[1]

指导练习 3.67

G

在指导练习 3.66 中可知小李下个月的股票预计会有盈利，但若实际发生亏损，你会感到惊讶吗？[2]

3.4.4 随机变量线性组合的波动性

当我们计算出某个随机变量线性组合的预期结果时，同时也要注意，该组合所产生的实际结果是存在波动性的。在指导练习 3.66 中，我们计算了小李股票投资组合的预期净收益或亏损，却没有对这个投资组合净收益或亏损存在的波动性进行定量讨论。例如，虽然我们计算出了每月预期的收益约为 124 元，但每个月实际发生的情况却并不一定如此。而且可以想象，收益情况的波动性往往很大，因为股票市场本身风险就很高。因此，在投资股票时量化这些波动性，或者说风险尤为重要。

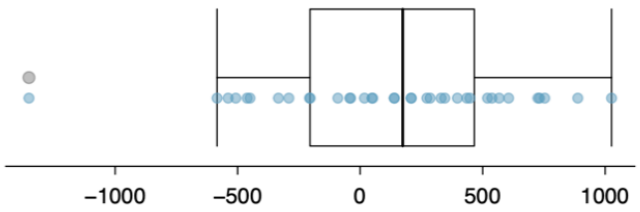


图 3.22：类似小李的投资组合在三年中（36 个月）的月收益情况。该组合中，6000 元用于投资卡特彼勒，2000 元用于投资埃克森美孚。

[1] $E(6000 \times X \ + \ 2000 \times Y \ ) \ = \ 6000 \ \times \ 0.020 \ + \ 2000 \ \times \ 0.002 \ = \ 124$

[2] 不会，即使股票总体呈现涨势，短期内也会有波动，数学期望计算的均值只能代表理想情况的预期，而实际观测结果往往会与之有所出入。

根据所学知识，我们可以用方差和标准差来描述小李每月收益存在的不确定性。图 3.23 展示了两种股票月收益的方差和标准差，并且我们认为两种股票的投资收益近似相互独立。

	均值 ($\bar{x}$)	标准差 (s)	方差 (s^2)
CAT	0.0204	0.0757	0.0057
XOM	0.0025	0.0455	0.0021

图 3.23：卡特皮勒和埃克森美孚两只股票的均值、标准差和方差。这些统计数据是基于历史股票数据得出的，因此采用了样本统计符号。

这里我们要用到概率论中的一个公式来描述小李每月收益的不确定性（关于这个公式的证明过程，我们在本书中不做解释，可能会考虑把它放到之后的课程中再做详细介绍）。随机变量线性组合的方差可以通过将单个随机变量的方差和系数的平方代入方程来计算：

$$Var(aX + bY) = a\text{^}2 \times Var(X) + b\text{^}2 \times Var(Y)$$

需要特别指出的是，这个方程成立的前提是各随机变量相互独立。如果各随机变量不相互独立，方程需要有所调整。如果了解协方差这一概念的小伙伴应该会意识到，线性组合变量的方程需要考虑到两个变量间相互作用的影响，即加入协方差在方程中。而对于独立变量组合来说，协方差为零，因此公式中就将这一部分隐去以避免过多赘述对统计学入门学习带来的不必要困惑。

代入数值可以计算出小李每月投资收益的方差：

$$\begin{aligned} Var(6000 \times X + 2000 \times Y) &= 6000\text{^}2 \times Var(X) + 2000\text{^}2 \times Var(Y) \\ &= 36{,}000{,}000 \times 0.0057 + 4{,}000{,}000 \times 0.0021 \\ &\approx 213{,}600 \end{aligned}$$

标准差等于方差的平方根：$\sqrt{213{,}600} \approx 463$。因此，虽然小李 8000 元的投资组合每月平均能有 124 元收益，但该收益却是非常不稳定的。

随机变量线性组合的波动性

要计算随机变量线性组合的方差，可以将常数平方，并代入各随机变量的方差：

$$Var(aX + bY) = a^2 \times Var(X) + b^2 \times Var(Y)$$

需要注意，该方程的成立条件是各随机变量要相互独立。标准差可通过取方差的平方根得到。

示例 3.68

假定老王每天通勤时间的标准差是 4 分钟。那么他一周总通勤时间的波动性有多大?

答案：我们用以下式子表示老王一周通勤的总时间

E

$$X_1 + X_2 + X_3 + X_4 + X_5$$

其中，每个随机变量代表了一周中某一天的通勤时间，并且系数都是 1，由题目可知，每天通勤时间的方差是 $4^2 = 16$。因此，一周通勤总时间的方差为

$$1^2 \times 16 + 1^2 \times 16 + 1^2 \times 16 + 1^2 \times 16 + 1^2 \times 16 = 5 \times 16 = 80$$

标准差为 $\sqrt{80} = 8.94$，约 9 分钟。

指导练习 3.69

G

示例 3.68 中的计算有一个非常重要的前提：每一天通勤时间相互独立。请问这种假设成立吗？[1]

指导练习 3.70

G

我们再来看指导练习 3.62 中埃琳娜的两次拍卖。假设两次拍卖近似相互独立，电视拍卖价格的标准差为 25 元，烤面包机拍卖价格的标准差为 8 元，请计算埃琳娜净收益的标准差。[2]

---

[1] 不一定。如果老王开车上班，且交通状况每周五会更糟糕，这种假设就不成立。因为在这样的背景下，周五的通勤时间就不再独立于其他日子的通勤时间。例如我们仅来讨论周四和周五的通勤时间，如果这两个时间对应的随机变量相互独立，那么二者时间的大小应该没有任何关系。但是按照上述背景，周五总是更堵也就意味着周五的通勤时间基本上恒大于周四的通勤时间，这样一来二者就不能说是独立的了，因为如果周四的通勤时间变大，周五的通勤时间也只会更大。但如果他步行去上班，通勤时间就不再受到每周交通状况的影响，这样每天走路通勤的时长间可以评估为是相互独立的，也就符合上述方差和标准差计算的前提条件。

[2] 埃琳娜的收益可以写作

$$(1) \times X + (-1) \times Y$$

由题目可以计算出 X 的方差为 625，Y 的方差为 64，然后将系数的平方和方差代入公式可得：

$$(1)^2 \times Var(X) + (-1)^2 \times Var(Y) = 1 \times 625 + 1 \times 64 = 689$$

所以净收益的方差是 689，标准差约为 26.25 元。

3.5 连续分布

在本章中，我们已经讨论过随机变量的取值是离散的情况。接下来在本节中，我们来看连续随机变量的分布特征。首先，我们需要在离散变量和连续变量之间搭建一个桥梁，这就是之前章节中有介绍过的：直方图。直方图统计了某变量取值落在不同区间内的频数（或者概率密度）。譬如对于成年人身高这个变量来说，它本身是一个连续的随机变量，即特征是可以从数轴上连续取值（参考第 1 章第 2.2 小节）。一个人的身高可以是 175cm，也可以是 175.01cm，甚至可以是 175.0001cm，这些都是有意义的取值。而如果我们调查某个样本，例如某区域内随机抽出的若干成年人，然后把统计得到的身高数字划分到几个类别中，例如小于 175cm，175-185cm，大于 185cm，然后绘制直方图。这个时候我们直方图展示的内容其实就具有了离散性质，而这种划分类别的过程就自然产生了一个具备离散特征的分类变量。希望这个自然段能够让你对从连续变量统计出分类变量的流程有个大概的认识。

如果我们按照上述规则对身高分类之后，绘制直方图，得到的就是一个包含三根柱子的直方图。而如果我们在分类的时候分得更精细，那么直方图单根柱子的宽度就会更窄。而假设样本足够大的时候，那么我们就可以无限精细地进行划分。这种情况下，直方图就会变成一张密度图，即不再是柱子与柱子的横向排列，而是一条曲线下方覆盖了一些面积。

示例 3.71

图 3.24 展示了关于美国成年人身高的几个不同的空心直方图。箱子数量的不同会对分析数据产生什么影响？

E

答案：直方图中箱子越多，可以提供的信息越详细。本例中样本非常大，因此很窄的箱子也可以有效地提供信息。但样本容量较小时，我们在直方图中通常不会选用非常窄的箱子，因为箱子越窄，高度越不稳定。

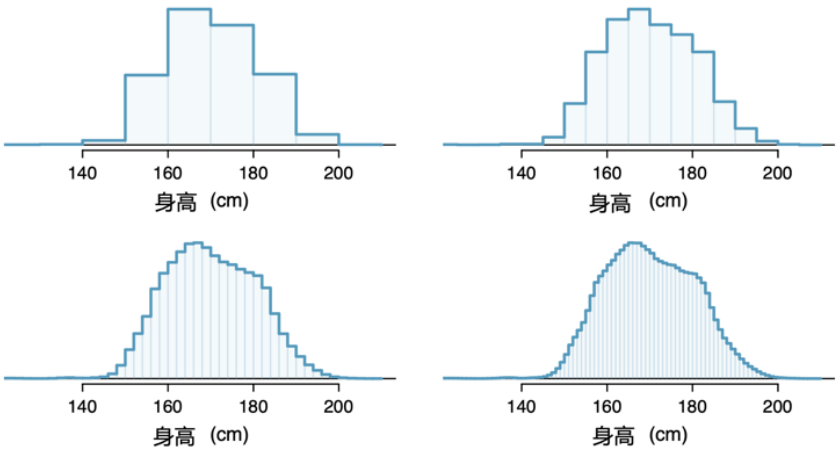


图 3.24：四个空心直方图，显示了美国成年人的身高数据，每个直方图箱子的宽度不同。

示例 3.72

样本中身高在 180cm 到 185cm 的比例是多少?

E 答案:我们可以借助第二张图,把直方图中在 180cm 到 185cm 的箱子高度相加,再除以样本数量。在 180cm 到 185cm 的箱子如图 3.25 所示。这个区域里的两个箱子分别包含 195307 人和 156239 人，除以样本数量 3000000 得到:

$$\frac{195307 + 156239}{3000000} = 0.1172$$

这个分数等于 180 到 185 范围内的箱子面积在直方图上所占比例。

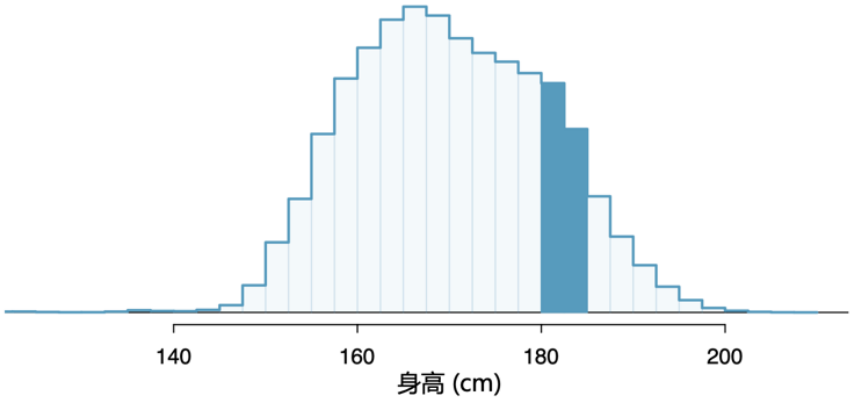


图 3.25：箱子宽度为 2.5cm 时的身高分布直方图，其中阴影部分面积代表了身高在 180cm 和 185cm 之间的成年人。

3.5.1 从直方图到连续分布

在图 3.24 中，左上方空心直方图呈箱型，而右下角的直方图则变得平滑很多。在最后一个图中箱子非常窄，以至于空心直方图趋向于形成一条平滑的曲线。由此可知，身高作为一个可以取到连续数值的随机变量，在样本数量足够的大的时候，描述它的分布用非常窄的箱子连成的平滑曲线非常合适。

这条平滑曲线代表一个**概率密度函数 probability density function** (也可以称为一个**密度状态 density** 或一个**分布 distribution**)，如图 3.26 所示，该曲线叠加在样本的直方图上。密度分布在定义时有一个特殊的性质：密度曲线下的总面积为 1，这也是为了规范化表示密度函数更强调不同区间内分布的比较关系，而非总样本的数量大小。

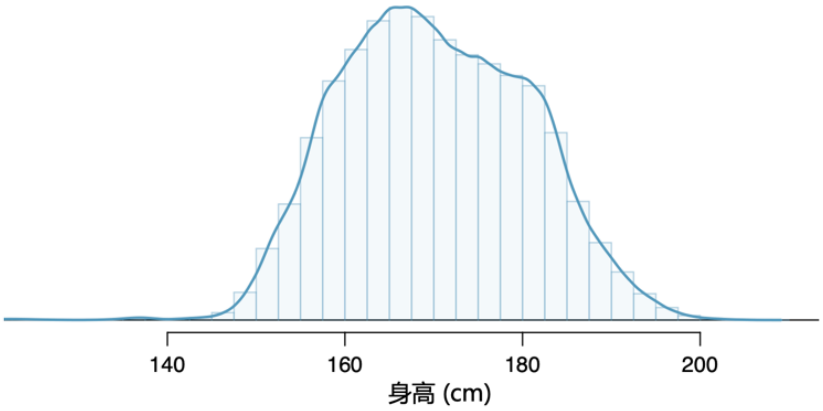


图 3.26：美国成年人身高的连续概率分布。

3.5.2 连续分布的概率

在示例 3.72 中，我们可以把身高在 180cm 到 185cm 之间的人数比例写作一个分数：

$$\frac{\text{身高在 180cm 到 185cm 之间的人数}}{\text{样本总人数}}$$

我们通过直方图在这个区域的面积比例，计算出了身高在 180cm 到 185cm 之间的人数。类似地，我们可以使用曲线下阴影区域的面积来计算概率（借助计算机）：

$$P(\text{身高在 180cm 到 185cm 之间}) = \text{阴影部分面积} = 0.1157$$

因此随机选取一个人身高在 180cm 到 185cm 之间的概率是 0.1157，这个结果与示例 3.72 中依赖较宽区间划分统计出的 0.1172 非常接近。

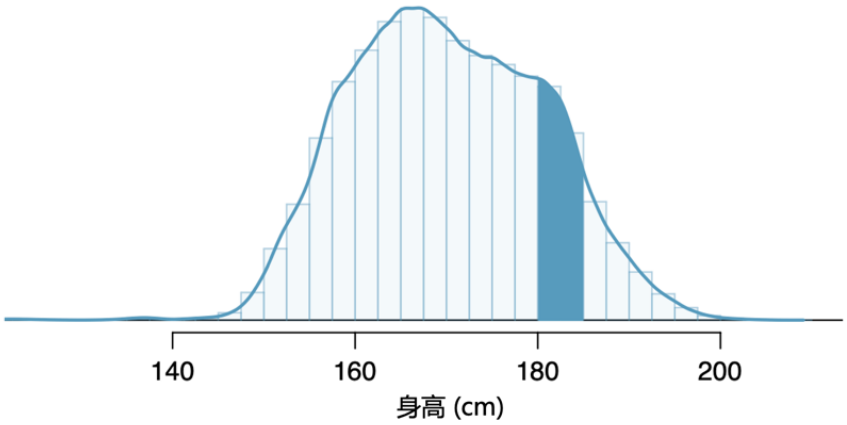


图 3.27：美国成年人身高的密度分布，其中 180cm 到 185 cm 之间的区域用阴影标记了出来。将此图与图 3.25 进行比较。

指导练习 3.73

随机选三个美国成年人，一个人的身高在 180 到 185 之间的概率约为 0.1157。请计算：[1]

(a) 三个人的身高都在 180 到 185 之间的概率是多少?

(b) 没有人身高在 180 到 185 之间的概率是多少?

示例 3.74

随机选取一个人，身高恰好为 180cm 的概率是多少？假设可以精确地测量身高。

答案：概率为 0。一个人身高可能接近 180cm，但不可能恰好等于 180cm。因为我们把连续分布的概率定义为某个区域的面积，而 180cm 和 180cm 之间的面积为 0。

指导练习 3.75

现在假设一个人的身高四舍五入到个位数，随机选取一个人，他的身高有无可能正好是 180cm？[2]

[1] (a) $0.1157 \times 0.1157 \times 0.1157 = 0.0015$; (b) $(1 - 0.1157)^3 = 0.692$。

[2] 有可能。身高在 179.5cm 到 180.5cm 之间的任何人测量出的身高都是 180cm。与示例 3.74 相比，这和现实情况更为相符。

第 4 章

随机变量的分布 Distributions of random variables

本章中，我们来讨论在数据分析或者统计推断中常见的一些数据分布。我们会首先从正态分布讲起。正态分布也是本书其余部分中引用最频繁的一种分布。其余的几种分布在本书的剩余环节中提到的次数没那么多，所以大家也可以把它们当作选读内容来根据自己情况学习。

跨越数据银河

系列推文合集

更多视频，演示文稿，和其他相关资源，请访问：
http://www.openintro.org/os

4.1 正太分布

我们在统计实践中会观察到各种各样的分布状态，其中有一种绝对可以说是最常见的。那就是对称、单峰的钟[1]型曲线。原著中甚至使用了 ubiquitous，也就是「无处不在的」来形容这种曲线在统计学中的普遍性。事实上，也正是因为它太常见了，所以人们会叫他**正态[2]曲线 normal curve** 或者**正态分布 normal distribution**[3]。如图 4.1 所示，例如学生的 SAT 分数（又称美国高考，全称是 Scholastic Aptitude Test，是美国高中生升学必考考试）或者美国成年男子的身高分布都几乎遵守正态分布。

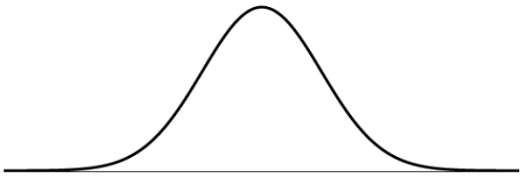


图 4.1：一条正态曲线

关于正态分布的事实

尽管没什么变量是严格地服从「完美的」正态分布，但不得不说，很多变量的分布都非常趋近于正态分布。因此正态分布尽管不完美，但是在众多问题的分析中还是非常好用。我们也会使用它来探索数据，并解决重要的统计学问题。

4.1.1 正态分布模型

正态分布 normal distribution 往往描述的是一个对称、单峰的钟形曲线对应的分布。而现实中，因为具体模型的不同，不同的正态分布曲线看起来也会不尽相同，尽管它们都具备上述的三个特征（对称、单峰和钟形）。更具体点来说不同正态分布曲线间的差别，我们可以通过调整两个关键变量：分布的均值和标准差，来调整曲线的形状。你大约也能想象，通过改变分布的均值，我们可以把整个钟形曲线在形状不变的前提下整体沿数轴向左或者向右移动；而通过改变分布的标准差，我们可以把曲线「拽」得更瘦高，或者「压」得更扁平。

[1] 译者注：钟其实是由英文词汇：Bell 翻译而来，而这个词汇其实更多时候是指铃铛。大家可以通过百度搜索一下「铃铛」一词，可能能够更形象地看到铃铛的形状和正态分布曲线的形状更为贴合。而如果我们搜索「钟」，则很容易搜出来一大堆闹钟，钟表，或者中国古代的青铜钟、编钟一类，混淆对钟形曲线中的「钟」的理解。

[2] 译者注：正态一词是由英文词汇：normal 翻译而来，其本意是正常、普通的、正规的。这条曲线首先非常漂亮对称，所以说它是「正规的」也不会让人意外，而原书中这里的意思主要是，它的存在很普遍、普通，所以统计学家们会觉得见到钟形曲线是一种很「正常的」事情，因而人们就称呼它是「正常曲线」，即 normal distribution。不过由于「正常曲线」一名不够专业，所以我们更多地会采用正态这一个词汇，并且渐渐习惯于把这种分布成为「正态分布」。

[3] 正态分布又被称为高斯分布，因为弗雷德里克·高斯是首个计算出了该分布的数学表达式的学者。

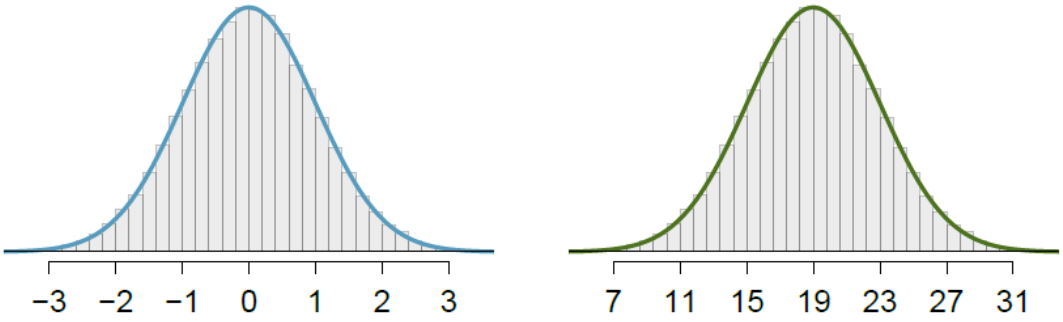


图 4.2：两条曲线都是正态分布。然而它们的区别在于中心和离散程度。

图 4.2 就展示了一个原本均值为 0，标准差为 1 的正态分布和一个调整后均值为 19，标准差为 4 的正态分布。注意在图 4.2 上，右图的横坐标的跨度和左图的不一样（右边是 4，而左边是 1）。图 4.3 中展示了基于同一个坐标轴的两个分布的对比。

$$N(\mu = 0, \sigma = 1) \text{ and } N(\mu = 19, \sigma = 4)$$

因为均值和标准差可以确定一个正态分布，所以它们被称为分布的**参数 parameters**。特别地，我们把均值 $\mu = 0$，标准差 $\sigma = 1$ 的正态分布称为**标准正态分布 standard normal distribution**。

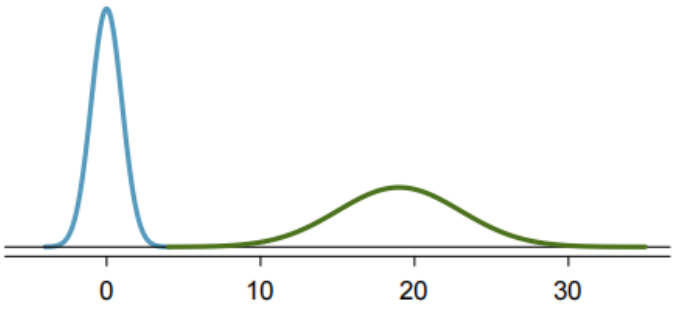


图 4.3：把图 4.2 中的两个正态分布按统一的尺度绘制同到一个数轴上。

指导练习 4.1

写出下列正态分布的简写

(a) 均值是 5，标准差是 3

(b) 均值是-100，标准差是 10

(c) 均值是 2，标准差是 9[1]

[1] (a) $N(\mu = 5, \sigma = 3)$；(b) $N(\mu = -100, \sigma = 10)$；(c) $N(\mu = 2, \sigma = 9)$。

4.1.2 Z 分数与标准化

进行数据分析时，我们经常需要把数据放在一个标准化的尺度上，使得不同变量间的对比更直观且具有说服力。于是我们引入 **Z 分数 Z-score** 对变量进行标准化处理：Z 分数能够真实地反应正态分布中随机变量取得的一个原始取值距离均值的相对标准距离。如果我们把原始取值转换成 Z 分数，那么 Z 分数会以标准差为单位，表示出原始取值到均值的距离：即如果 Z 分数为 1，说明原始取值的位置距离均值相差 1 个标准差。

示例 4.2

图 4.4 展示了某次 SAT 和 ACT 总分的均值和标准差。已知 SAT 和 ACT 分数的分布都近似为正态分布。假设安（人名）在 SAT 考试中得了 1300 分，汤姆（人名）在 ACT 考试中得了 24 分，请问谁的成绩更好?

E

答案：我们以标准差为单位来描述和对比。安的 SAT 成绩比均值高出 1 个标准差：1100 + 200 = 1300。汤姆的 ACT 成绩比均值高出 0.5 个标准差：21 + 0.5 × 6 = 24。如图 4.5，我们可以看出与汤姆相比，安的 SAT 分数与这次 SAT 均值的距离相差更大，也就是说安的分数在同批次中占比更加靠前，所以相对来说安的成绩更好。

	SAT	ACT
Mean	1100	21
SD	200	6

图 4.4：SAT 和 CAT 的均值与标准差。

示例 4.2 使用 Z 分数对变量的原始取值进行标准化处理，这种方法最常用于近似正态分布的实例中，同时也可以用于任何分布。一个随机变量取值的 Z 分数等于其高于或低于均值的标准差数，也就是它本身与均值的差再除以标准差，比均值大的为正值，比均值小的为负值。如果随机变量的一次取值恰巧高于均值一个标准差，则其 Z 分数为 1。如果它比均值低 1.5 个标准差，那么它的 Z 分数是-1.5。如果 x 是分布 $N(\mu,\sigma)$ 的一个取值，我们在数学上将它的 Z 分数定义为：

$$Z = \frac{x - \mu}{\sigma}$$

例如，在示例 4.2 中已知 μ_{SAT} = 1100，σ_{SAT} = 200，$x_{安}$ = 1300，那么安的 Z 分数的计算我们可以把对应的数字带入上面的公式中。需要注意的是，在带入均值和标准差的时候，应当带入随机变量取值对应背后的正态分布的均值以及标准差。我们可以得出如下的计算式：

$$Z_{安} = \frac{x_{安} - \mu_{\mathrm{SAT}}}{\sigma_{SAT}} = \frac{1300 - 1100}{200} = 1$$

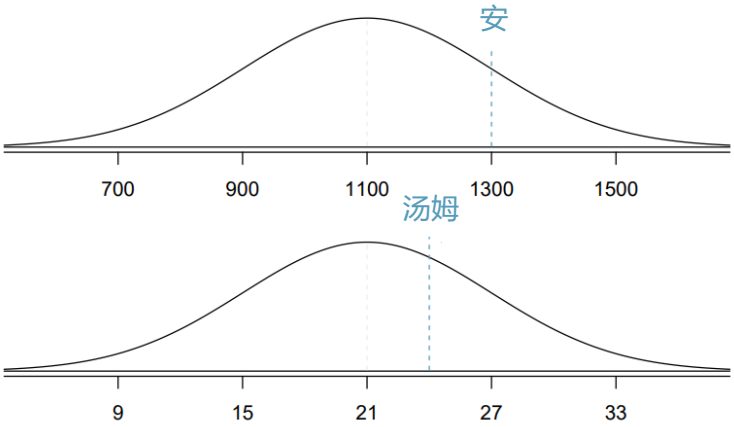


图 4.5：安和汤姆的成绩展示在对应的 SAT 和 CAT 成绩分布上。

Z 分数

一个观测值的 z 分数等于它高于或低于均值的标准差数。如果变量 x 遵循一个均值为 μ、标准差为 σ 的分布，那么 x 的 Z 分数可以根据下式计算：

$$Z = \frac{x - \mu}{\sigma}$$

指导练习 4.3

示例 4.2 中，已知汤姆的分数是 24，ACT 成绩分布均值为 21，标准差是 6，请计算他的 Z 分数。[1]

指导练习 4.4

随机变量 x 遵循一个分布 $N(\mu = 3, \sigma = 2)$，假设取值 x = 5.19：

(a) 计算 x 的 Z 分数；

(b) 使用 Z 分数计算出 x 比均值高或低几个标准差。[2]

指导练习 4.5

一种名叫帚尾袋貂的动物头长遵循正态分布，均值 92.6 毫米，标准差 3.6 毫米。请计算头长分别为 95.4 毫米和 85.8 毫米的帚尾袋貂的 Z 分数。[3]

指导练习 4.6

指导练习 4.5 中两个观测值哪个更不寻常？[4]

---

[1] $Z = \frac{x-\mu}{\sigma} = \frac{24-21}{6} = 0.5$

[2] (a) $Z = \frac{x-\mu}{\sigma} = \frac{5.19-3}{2} = 1.095$；(b) x 比均值高 1.095 个标准差。

[3] 令 $x_1 = 95.4$：$Z_1 = \frac{x_1-\mu}{\sigma} = \frac{95.4-92.6}{3.6} = 0.78$；令 $x_2 = 85.8$：$Z_2 = \frac{x_2-\mu}{\sigma} = \frac{85.8-92.6}{3.6} = -1.89$。

[4] 由于 Z_2 的绝对值大于 Z_1 的绝对值，所以头长 85.5 毫米更不寻常。

4.1.3 计算尾端面积

在统计学中，计算出一个分布的尾端面积是非常有用的。例如，有多少人的 SAT 分数低于安的 1300 分？要计算的结果和安的分数所在的百分位数是一样的，也等于分数低于安的人数所占比例。如图 4.6，我们可以把要计算的结果可视化处理，表示为这个正态分布曲线中的阴影部分，计算出阴影部分的面积也就是尾端面积即可。

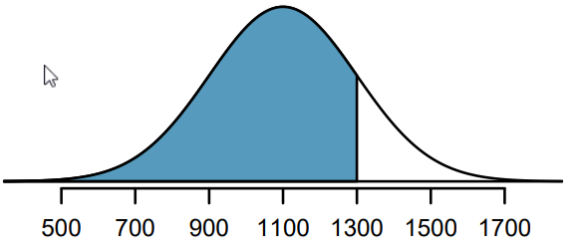


图 4.6：Z 分数左侧的区域相对面积就代表了有百分之多少的考生比安的成绩低。

有许多方法都可以计算出尾端面积，这里介绍其中三种：

(1) 实践中最常用的方法是利用数据分析软件。有一个程序 RStudio（是一个基于 **R** 语言的集成开发环境），可以输入 Z 分数，得到观测值左侧的尾端面积。例如，在程序 RStudio 中输入下列指令，就可以得到图 4.6 中的阴影部分的面积。

```
> pnorm(1)
[1] 0.8413447
```

根据程序计算结果显示，低于 1300 的部分区域对应的考生比例是 84.1%，这些考生的 Z 分数也都低于 1。在使用 **R** 语言计算的时候，我们除了直接提供 Z 分数作为唯一参数外，还可以指定具体的观测值（或者说切割点），均值以及标准差来计算对应的比例：

```
> pnorm(1300, mean = 1100, sd = 200)
[1] 0.8413447
```

除了 **R** 语言外，也有许多其他语言/软件可供选择，包括 Python 和 SAS 等，甚至电子制表程序如 Excel，Google Sheets 等等也支持这种运算。

(2) 统计学课堂上常使用图形计算器，如 TI 或卡西欧计算器。不过使用这些计算器操作有些复杂，不太容易描述。你可以在 OpenIntro 视频库中找到使用这些计算器计算正态分布尾端面积的教程：

www.openintro.org/videos

(3) 还有一种计算尾端面积的方法是使用概率表。但只是偶尔会在课堂上使用，在实践中很少会选择它。附录 C.1 中有概率表，并附有使用指南。

在本节中，我们总是需要先计算出 Z 分数，再解决正态分布问题。因为我们将在第 5 章开始遇到类似的统计分析问题,且很多时候实质上都要计算 Z 分数。

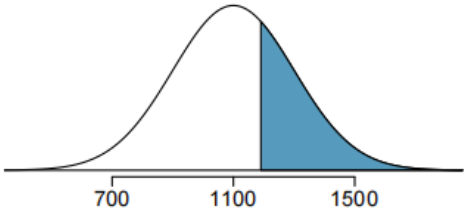
4.1.4 正态分布示例

前两节中的 SAT 分数近似于正态分布模型，$N(\mu = 1100, \sigma = 200)$，本节中把它作为已知条件。

示例 4.7

随机选取一个考生小明，我们并不了解他的水平。他在 SAT 考试中至少得 1190 分的概率是多少？

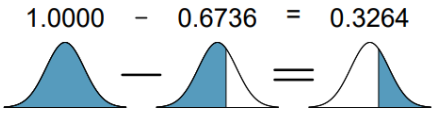
答案：画出正态分布曲线图，并做标注，画图不要求精准。我们要计算他得分高于 1190 分的概率，所以把右侧画为阴影部分。



该图的横轴显示了均值以及在均值上下 2 个标准差处的值。要计算曲线下阴影部分的面积，最简单的方法是使用 1190 分作为截断值求得 Z 分数。当 μ = 1100，σ = 200，截断值 x = 1190 时，Z 分数为：

$$Z = \frac{x - \mu}{\sigma} = \frac{1190 - 1100}{200} = 0.45$$

使用统计软件（或其他简便方法），我们可以计算出 Z = 0.45 左侧区域的面积等于 0.6736。要计算 Z = 0.45 右侧区域面积，我们可以用 1 减去左侧区域的面积。



最后可以知道他在 SAT 考试中至少得 1190 分的概率是 0.3264。

先画图再求 z 分数

对于任意一个正态分布的情况，要解决问题的首要步骤一定是画出正态分布曲线图并标出对应的阴影部分。该图可以帮助我们估计相应的概率值。画出图形后，再确定具体值的 Z 分数。

指导练习 4.8

如果小明在 SAT 考试中至少得 1190 分的概率是 0.3264，那么他得分少于 1190 的概率是多少？画出对应的正态分布曲线图，标出正确的阴影部分。[1]

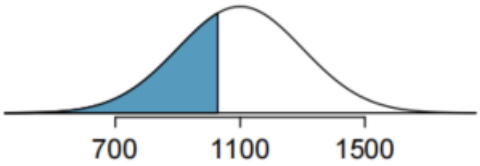
[1] 在示例 4.7 中可知他得分少于 1190 的概率是 0.6736，图略。

示例 4.9

小红在 SAT 考试中得了 1030 分，她的百分位数是多少?

答案：首先，需要画图。小红的百分位数等于得分低于 1030 分的考生的比例。得分低于 1030，在图中显示为 1030 左侧区域。

E



确定均值 μ = 1100，标准差 σ = 200，截断值，尾部区域截断值 x = 1030 时，可以计算出 Z 分数:

$$Z = \frac{x - \mu}{\sigma} = \frac{1030 - 1100}{200} = -0.35$$

使用统计软件可得尾部面积是 0.3632，也就是小红排在第 36 个百分位。

G

指导练习 4.10

使用示例 4.9 的结果，计算本次 SAT 中得分比小红高的考生的比例。注意画正态分布曲线图。[1]

寻找右侧的面积

很多统计软件/语言都会在计算 Z 分数的时候返回左侧的面积。而如果你想要知道截断值右侧的面积是多少，可以使用正态分布曲线下方面积为 1 这个定义，用 1 减去左侧面积就得到右侧的面积啦，是不是很方便呢?

G

指导练习 4.11

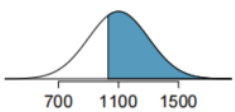
小欧在 SAT 考试中得到 1500 分，画出正态分布曲线图并回答问题:

(a) 他的百分位数是多少?

(b) 得分比小欧高的考生占比是多少? [2]

据统计，美国成年男性的身高近似符合正态分布，均值是 70.0 英寸，标准差是 3.3 英寸。现在有一个 100 位男性的样本。

[1] 如果小红得分比 36%的考生都高，那么有 64%的考生得分高于小红。



[2] 图略。(a) $Z = \frac{1500-1100}{200} = 2 \rightarrow 0.9772$; (b) 1 − 0.9772 = 0.0228。

指导练习 4.12

已知，小李身高 5 英尺 7 英寸，小拜身高 6 英尺 4 英寸（1 英尺等于 12 英寸），且他们都是美国成年男性。

(a) 小李身高的百分位数是多少?

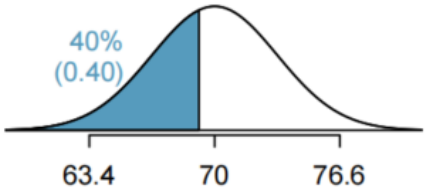
(b) 小拜身高的百分位数是多少? [1]

前面几个问题都是已知某一特定观测值，确定它的百分位数（下尾）或上尾。如果已知百分位数，如何确定观测值呢?

示例 4.13

小陈的身高在 100 人中的百分位数是 40，那么他身高是多少?

答案：还是和之前一样，首先我们先画图。



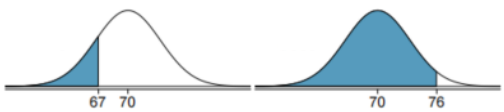
本例中，40 的百分位数对应了左尾的概率等于 0.40，即图中的阴影部分。我们需要根据这个百分位数计算出观测值。首先，我们需要计算出第 0.40 对应的 Z 分数。使用统计软件得到，对应的 Z 分数约等于-0.25。

已知 $Z_{小陈}$ = -0.25，样本参数 μ = 70，σ = 3.3。接着用 x 指代要计算的小陈的身高，带入 Z 分数的取值，我们可以列出如下方程：

$$-0.25 = Z = \frac{x-\mu}{\sigma} = \frac{x-70}{3.3}$$

解方程可得：x = 69.2 英寸，再进行进一步换算可得小陈的身高约为 5 英尺 9 英寸。

[1] 首先把身高单位换算成英寸：67 英寸和 76 英寸，如图

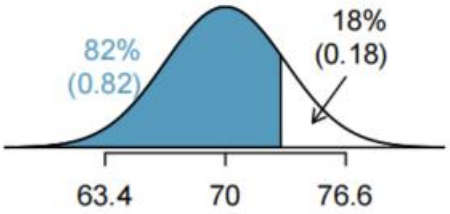


(a) $Z_{小李} = \frac{67-70}{3.3} = -0.91 \rightarrow 0.1814$；(b) $Z_{小拜} = \frac{76-70}{3.3} = 1.82 \rightarrow 0.9656$。

示例 4.14

如果一个在美国的成年男性的身高百分位数是 82，他的身高是多少?

答案：还是需要先画图：



然后，我们需要确定百分位数 82 对应的 Z 分数，用统计软件计算求得 Z = 0.92。最后，利用已知的均值 μ，标准差 σ，0.92 的 Z 分数以及 Z 分数的计算公式，可以列出方程：

$$0.92 = Z = \frac{x-\mu}{\sigma} = \frac{x-70}{3.3}$$

计算得到 73.04 英寸或大约 6 英尺 1 英寸，是第 82 个百分位的美国成年男性的身高。

指导练习 4.15

已知一次 SAT 考试的所有考生分数服从一个正态分布 $N(1100, 200)$。

(a) 第 95 个百分位数对应的分数是多少?

(b) 第 97.5 个百分位数对应的分数是多少? [1]

指导练习 4.16

已知某地成年男性身高遵循一个正态分布 $N(70, 3.3)$，单位在这里依然是英寸。

(a) 随机抽选一名成年男性，他的身高至少是 6 尺 2 寸（74 英寸）的概率是多少?

(b) 当地一名成年男性的身高低于 5 尺 9 寸（69 英寸）的概率是多少? [2]

[1] 参考答案：(a) 1429；(b) 1492。

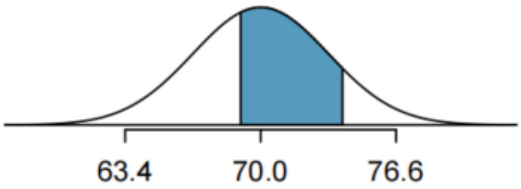
[2] 参考答案：(a) Z = 1.21 对应概率 0.8869，用 1 减去这个数字得到 0.1131；(b) 同理，得到 0.3821。

示例 4.17

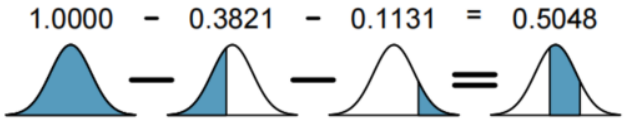
随机抽选一名成年男性，他的身高在 5 尺 9 寸到 6 尺 2 寸之间的概率是多少?

答案：首先进行单位换算，得出 69 英寸和 74 英寸。接着我们画图来辅助，可以从下图中看出，蓝色部分的区域面积就是我们感兴趣的区间概率。

E



然后，在指导练习 4.16 中我们已知这左侧尾端面积和右侧尾端面积的分别是 0.3821 和 0.1131：



随机抽选一名成年男性，他的身高在 5 尺 9 寸到 6 尺 2 寸之间的概率是 0.5048。

G

指导练习 4.18

一次 SAT 考试分数遵循正态分布 $N(1100, 200)$。那么得分在 1100 和 1400 之间的考生占比多少? [1]

G

指导练习 4.19

已知某地成年男性身高遵循一个正态分布 $N(70.0, 3.3)$，单位是英寸。那么请问身高在 5 尺 5 寸到 5 尺 7 寸之间的男性占比多少? [2]

4.1.5 68-95-99.7 法则

我们在实践中总结出了一个非常有用的法则，适用于正态分布中落在均值的 1、2 和 3 个标准差内的概率。这个法则在实践中非常有用，特别是适用于在没有计算器或 Z 分数表的情况下需要进行快速估计的时候。

[1] 一定要先画图！利用已知条件确定两个数值对应的 Z 分数，然后可以确定分数低于 1100 和高于 1400 的考生占比：$Z_{1100} = 0.00 \rightarrow 0.5000$，$Z_{1400} = 1.5 \rightarrow 0.0668$。所以答案是 $1.0000 - 0.5000 - 0.0668 = 0.4332$。

[2] 5 尺 5 寸等于 65 英寸，对应 Z 分数 $Z = -1.52$；5 尺 7 寸等于 67 英寸，对应 Z 分数 $Z = -0.91$；同样参照上面的指导练习，可以计算出左侧尾端和右侧尾端的面积分别是 0.0643 和 0.8186。进而我们可以计算出 $1.000 - 0.0643 - 0.8186 = 0.1171$：即身高在 5 尺 5 寸到 5 尺 7 寸之间的男性占比 11.71%。

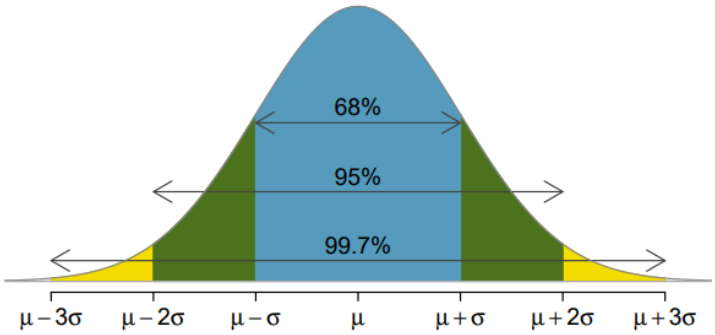


图 4.7：落在均值左右 1、2 和 3 个标准差的概率。

指导练习 4.20

使用软件、计算器或概率表来确认在正态分布中，偏离均值 1 个标准差之间的面积约 68%、2 个标准差之间的面积约 95%、3 个标准差之间的面积约 99.7%。例如，首先找到 Z =-1 和 Z = 1 之间的区域，它的面积应该约为 0.68。同样，在 Z = -2 和 Z = 2 之间的面积大约是 0.95。[1]

一个随机变量的某次随机取值是有可能与均值相差 4、5 甚至更多个标准差的。然而，如果数据接近正态分布，即大量数据都环绕在均值附近时，那么出现这种情况是非常罕见的。对于正态分布来说，偏离均值超过 4 个标准差的双尾区间概率约为 15000 分之一。对于 5 个和 6 个标准差，分别约为 200 万分之一和 5 亿万分之一。因此对于一个服从正态分布的随机变量，如果我们随机取到了一个与均值相差太远（比如有 4 到 5 个标准差）的值，那么我们可以怀疑：可能是正态分布的均值假设不准确，亦或是抽样过程并不是完全随机的。

指导练习 4.21

已知 SAT 分数非常接近正态分布，均值 μ = 1100，标准差 σ = 200：

(a) 分数在 700 到 1500 的考生数量比例是多少?

(b) 分数在 1100 到 1500 的考生数量比例是多少？[2]

[1] 首先画出正态分布曲线图。借助软件可以得到，1 个标准差之间的面积为 0.6827，2 个标准差之间的面积为 0.9545，3 个标准差之间的面积为 0.9973。

[2] (a) 700 比均值低两个标准差，1500 比均值高两个标准差，所以约有 95%的考生分数在 700 到 1500 之间；(b) 已知约有 95%的考生分数在 700-1500，且考生分数以均值 1100 为界分成两等份，所以 95%/2=47.5%，也就是约有 47.5%的考生分数在 1100 到 1500 的分数区间内。

4.2 几何分布

抛一枚硬币，预计需要几次才会出现正面朝上的结果？掷一个骰子，预计经过几次才会第一次得到点数 1？想回答这两个问题我们可以借助几何分布。我们首先使用伯努利分布的相关知识，定义好什么是单次试验。例如抛一次硬币或掷一次骰子。然后将这些试验与第 3 章中的概率工具相结合来构建几何分布。

4.2.1 伯努利分布

我们在实践中总结出了一个非常有用的法则，适用于正态分布中落在均值的 1、2 和 3 个标准差内的概率。这个法则在实践中非常有用，特别是适用于在没有计算器或 Z 分数表的情况下需要进行快速估计的时候。

为帮助大家更形象地理解伯努利分布，我们以一个捉鱼实例来看。一个鱼塘中的鱼到了成熟时节，身长达到 20cm 即合格，并可以被高价收购；未达到 20cm 的鱼则只能卖到较低的价格。

假设一个鱼塘经统计发现 70%的鱼成熟后身长达到了 20cm。在这里，每捉一次鱼都可以视为一次**试验 trial**。如果随机捉到的一条鱼身长超过 20cm，我们会将其标记为**成功 success**；否则，我们将这次试验标记为**失败 failure**。 因为 70%的鱼身长会达到 20cm，我们将成功的概率表示为 $p = 0.7$。而失败的概率为 $q = 1 - p$，本例中 $q = 1 - 0.7 = 0.3$。

当一次独立的试验只有两种可能的结果即成功或失败时，我们把这中试验的结果称为一个**伯努利随机变量 Bernoulli random variable**。我们可以把身长达到 20cm 的鱼标记为“成功”，把其他的鱼标记为“失败”。同时，我们也可以很容易地更换这些标签。我们要构建的数学框架并不取决于结果的标签。无论哪个结果被贴上成功的标签，哪个结果被贴上失败的标签，只要我们在研究中自始至终坚持一个标准就可以了。

伯努利随机变量通常用 1 表示成功，0 表示失败。这样标记不仅方便记录数据，在数学上也很方便。假设我们进行十次试验，记录并观察结果：

1 1 1 0 1 0 0 1 1 0

那么**样本比例 sample proportion** 就是这些观测值的样本均值，用 p 表示为：

$$p = \frac{\#成功次数}{\#总试验次数} = \frac{1+1+1+0+1+0+0+1+1+0}{10} = 0.6$$

然后，对伯努利随机变量的数学探究可以进一步扩展。因为 0 和 1 是数值结果，我们可以定义

伯努利随机变量的均值和标准差。(见指导练习 4.15 和 4.16)

伯努利随机变量

如果 X 是一个随机变量，取值是 1 时为成功，概率为 p；取值是 0 时为失败，概率为 $1-p$。那么 X 就是一个伯努利随机变量，其均值和标准差如下：

$$\mu = p \qquad \sigma = \sqrt{p(1-p)}$$

4.2.2 几何分布

几何分布 geometric distribution 的定义是：在一系列伯努利试验中，试验 n 次才得到第一次成功的概率分布。它也常常被用于描述计算首次观察到成功的结果需要经过多少次试验。我们先来看一个例子：

示例 4.22

假设一个鱼塘的老板需要捕捞出一条身长达到 20cm 的鱼，以便展示给采购方作为样品。沿用之前假设，已知随机捉一条鱼身长达到 20cm 的概率是 0.7，我们进行多次有放回随机抽样，那么第一条鱼身长达到 20cm，即成功的概率是多少？ 第一条不满足条件的前提下，第二个条鱼才达到 20cm 的概率又是多少？ 依次类推的第三条呢？假设经过 $n-1$ 次实验后得到成功的结果，也就是说捕到第 n 条鱼时，我们获得首次成功，即找到一条身长达到 20cm 的鱼，其概率又是多少，你能用一个和 n 相关的式子表示吗？（例如，捕到第五条鱼时首次获得成功，那么我们说 n =5。）

(E) 答案：捕一次就成功的概率，等于第一条鱼的身长达到 20cm 的概率：0.7；而捕两次才观察到首次成功的概率是：

$$\begin{aligned} &P(\text{第二个鱼身长达到 20cm}) \\ &= P(\text{第一条鱼身长小于 20cm，第二条鱼身长达到 20cm}) \\ &= 0.3 \times 0.7 = 0.21 \end{aligned}$$

同样，第三次才获得成功的概率是 $0.3 \times 0.3 \times 0.7 = 0.063$

如果捕第 n 条鱼时获得第一次成功，那么意味着前面经过了 $n-1$ 次失败，最后有 1 次成功。这对应了概率 $(0.3)^{n-1}(0.7)$，我们也可以吧失败的概率用 1 减去成功的概率表示，即：$(1-0.7)^{n-1}(0.7)$。

示例 4.22 阐述了**几何分布 geometric distribution**。它描述了**独立同分布 independent and identically distributed (iid)** 的伯努利随机变量取得首次成功时的试验次数。其中，独立性指试验中的个体不会相互影响，相同性指每次试验获得成功的概率相同。

示例 4.22 的几何分布如图 4.8 所示。一般来说，几何分布的概率呈**指数式 exponentially** 特

征快速下降。

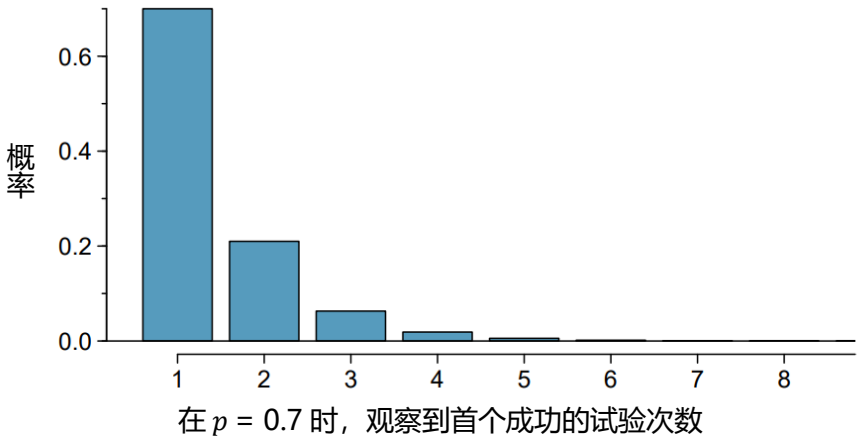


图 4.8：成功概率为 0.7 的几何分布。

通过例子，结合上图，我们不难得出几何分布的一些特点：

(1) 几何分布的变量是进行伯努利试验的直到成功的次数，例如投掷一枚硬币 n 次直到首次观察到正面朝上，投骰子 n 次直到首次观察到数字 1，随机从群体中有放回抽访 n 个人直到首次观察到一个吸烟者等等……

(2) 几何分布的变量是离散的，取值范围是从 1 到正无穷；同时由于其变量的离散性，我们绘制分布图的时候往往可以用一些列柱子而不是曲线来表示

(3) 几何分布呈现较明显的单峰右偏特征；

(4) 几何分布中所有柱子加起来概率为 1，这点我们可以从上图有一个直觉上的认识，而关于它的证明则需要依赖严密的数学推导，涉及到对一个无穷项数列的求和，即无穷几何级数的知识。在这本旨在面向更多非数理背景的教材中我们不做过多展开证明。

考虑到篇幅限制，我们在本文中不会推导第一次成功所需的平均（预期）试验数的公式或该分布的标准偏差或方差，只提供它们的通用公式。

几何分布

如果一次试验的成功概率为 p，失败概率为 $1-p$，则第 n 次试验获得首次成功的概率为：

$$(1-p)^{n-1}p$$

相应的均值，方差和标准差分别是：

$$\mu = \frac{1}{p} \qquad \sigma^2 = \frac{1-p}{p^2} \qquad \sigma = \sqrt{\frac{1-p}{p^2}}$$

这里我们使用符号 μ 表示均值，因为均值和数学期望概念上是高度一致的。

在几何分布中，平均需要 $1/p$ 次试验才能获得成功。这个数值符合我们的直觉。如果成功的概率很高（例如 0.8），那么我们直觉上也会觉得不需要为成功等待很长时间，带入数字就是平均需要 1/0.8 = 1.25 次试验即能获得成功 ；如果成功的概率很低（例如 0.1），那么直觉告诉我们，在观察到成功情形之前，很可能要经过多次试验，其期望是：1/0.1 = 10 次。

指导练习 4.23

G 沿用前例，假设从鱼塘中捕一条达到 20cm 的鱼的概率是 0.7。如果从鱼塘中捕鱼，直到捕到第一条身长达到 20cm 的鱼为止，预计需要捕多少次？[1]

示例 4.24

三次内我们能捕到身长达到 20cm 的鱼的概率是多少?

答案：三次内成功的概率等于第一次成功（$n = 1$）的概率、第二次成功（$n = 2$）的概率与第三次获得成功（$n = 3$）的概率之和。因为我们从总体中随机取样，$n = 1$，$n = 2$，$n = 3$ 是三个相互独立的结果。所以计算每种情况的概率，并相加：

E

$$
\begin{aligned}
&P(n = 1, 2, or\ 3) \\
&= P(n = 1) + P(n = 2) + P(n = 3) \\
&= (0.3)^{1-1}(0.7) + (0.3)^{2-1}(0.7) + (0.3)^{3-1}(0.7) \\
&= 0.973
\end{aligned}
$$

三次内我们能捕到身长达到 20cm 的鱼的概率是 0.973。

指导练习 4.25

G 尝试寻找一个更好的方法来解答例示例 4.24，得到同样的结果则表明你的方法正确。[2]

示例 4.26

假设一个公司老板发现 88%的员工会有早上吃早点的习惯。如果他随机从员工名单里挑选记录，直到记录到第一个吃早点的员工为止，那么预计他需要记录下多少个名字？对应的标准差是多少?

E 答案：在这个例子中，因为当老板观察到员工吃早点的时候就会停下，所以成功对应的事件是挑选出的员工吃早点，概率 $p = 0.88$。预期记录次数为 $1/p = 1/0.88 = 1.14$，标准差为$\sqrt{(1-p)/p^2} = 0.39$。

最后提醒大家，吃早点有助于促进新陈代谢，提高身体能量，改善心情和认知能力，所以还是建议大家再忙也要记得吃早饭，照顾好自己!

---

[1] 预计捕 $1/0.7 \approx 1.43$ 次后，会得到第一条身长达到 20cm 的鱼。

[2] 首先计算三次内没有成功的概率：$P(3\text{ 次试验均未成功}) = 0.3 \times 0.3 \times 0.3 = 0.027$。然后，用 1 减去这个计算出来的概率就可以得出前三次内至少有一次成功的概率了：$1 - P(3\text{ 次试验均未成功}) = 1 - 0.027 = 0.9$。

指导练习 4.27

G

沿用示例 4.26 的结果，$\mu = 1.14$ 和 $\sigma = 0.39$，是否可以使用正态分布模型来计算在 3 次内结束试验的概率？[1]

在几何分布中，独立性假设对准确描述试验非常重要。从数学上来说，为描述第 n 次试验的成功概率，我们必须对独立过程使用乘法法则。因此，试验不相互独立时，很难用几何分布来研究。

[1] 不可以。几何分布是向右呈指数下降的，不能用单峰对称的正态分布来描述。

4.3 二项分布

二项分布用于描述在一定次数的试验中成功的次数。这与几何分布不同，几何分布描述了在观察到成功之前我们必须等待的试验次数。

4.3.1 二项分布

还是以捉鱼为例，已知一个鱼塘里有 70%的鱼身长达到 20cm，我们把达到 20cm 的鱼称为「达标」，即对应伯努利试验中的成功概念；而未达到 20cm 的鱼称为「未达标」，即对应伯努利试验中失败的概念。

示例 4.28

假设某采购方来考察鱼塘，计划随机捕捞上来四条鱼。请计算其中三条鱼身长达到 20cm，一条鱼身长没有达到 20cm 的概率是多少？方便起见，我们把这四条鱼标记为 A、B、C、D。

答案：只有 A 身长没有达到 20cm 的概率：

$$
\begin{aligned}
&P(A = \text{未达标}, B = \text{达标}, C = \text{达标}, D = \text{达标}) \\
&= P(A = \text{未达标}) \times P(B = \text{达标}) \times P(C = \text{达标}) \times P(D = \text{达标}) \\
&= (0.3)(0.7)(0.7)(0.7) \\
&= (0.7)^3(0.3)^1 \\
&= 0.103
\end{aligned}
$$

此外，还有其他三种可能：只有 B 身长没有达标、只有 C 身长没有达标、只有 D 身长没有达标。在每一种情况下，概率都是 $(0.7)^3(0.3)^1$。这四种情况涵盖了（穷尽了）这四条鱼中只有一条鱼身长没有达标的所有可能，因此总概率为 $4 \times (0.7)^3(0.3)^1 = 0.412$。

指导练习 4.29

请用计算验证只有 B 身长没有达到 20cm 的概率为 $(0.7)^3(0.3)^1$。[1]

示例 4.28 中概述的情况是二项式分布的一个实例。二项式分布描述了在 n 次独立的伯努利试验中恰好有 k 次成功的概率，其中每次伯努利试验成功的概率为 p（在示例 4.28 中，$n = 4$，$k = 3$，$p = 0.7$）。接下来我们想要更一般地确定与二项式分布相关的概率，也就是说，我们想要一个公式，使用上面提到的参数 n，k 和 p 来表达概率。为此，我们重新审视示例 4.28 的每个部分。

[1] $P(A = \text{达标}, B = \text{未达标}, C = \text{达标}, D = \text{达标}) = (0.7)(0.3)(0.7)(0.7) = (0.7)^3(0.3)^1$。

有四个鱼身长可能没有达到 20cm，这四种情况中的每一种都有相同的概率。因此，我们可以确定最终概率为：

$$[\#\text{成功情景数}] \times \mathrm{P}(\text{单次试验成功})$$

结合上面捕鱼的例子，这个等式的第一个组成部分，是在 $n = 4$ 次试验中出现 $k = 3$ 次成功的方式数量。第二个组成部分是四种情况中任何一种（可能性相等）的概率。

在 n 次试验中有 k 次成功和 $n - k$ 次失败的情况下，考虑 $\mathrm{P}(\text{单次试验成功})$ 的一般情况。在任何这样的情况下，我们应用独立事件的乘法法则：

$$p^k(1-p)^{n-k}$$

这就是我们得出的 $\mathrm{P}(\text{单次试验成功})$ 的一般公式。其次，我们引入一个公式，表达在 n 次试验中出现 k 次成功的方式数量，即出现 k 次成功和 $n - k$ 失败：

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

其中，数量 $\binom{n}{k}$ 读作 ***n* 选 *k***[1]，感叹号符号（例如 $k!$）表示**阶乘 factorial** 表达式。

$$0! = 1$$
$$1! = 1$$
$$2! = 2 \times 1 = 2$$
$$3! = 3 \times 2 \times 1 = 6$$
$$4! = 4 \times 3 \times 2 \times 1 = 24$$
$$\ldots$$
$$n! = n \times (n-1) \times \ldots \times 3 \times 2 \times 1$$

使用该公式，我们可以计算在 $n = 4$ 次试验中出现 $k = 3$ 次成功的方式数量：

$$\binom{4}{3} = \frac{4!}{3!\,(4-3)!} = \frac{4!}{3!\,1!} = \frac{4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1)(1)} = 4$$

这个结果与我们在示例 4.28 中仔细思考每种可能情况所得到的结果完全相同。用 n 选 k 代替成功情景的数量，用 $p^k(1-p)^{n-k}$ 代替单个情况的，可以得到一般二项式公式。

[1] n 选 k 还有其他的一些写法，例如 C_n^k 还有 $C(n,k)$。

二项分布

假设单次试验成功的概率是 p。则在 n 次独立试验中恰好观察到 k 次成功的概率为：

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

相应的均值，方差和标准差分别是：

$$\mu = np \qquad \sigma^2 = np(1-p) \qquad \sigma = \sqrt{np(1-p)}$$

是否符合二项式分布，检查如下四个条件

(1) 试验间相互独立；

(2) 试验的总数 n 是固定的；

(3) 每个试验的结果可以且仅能被简单归纳为两种情况：成功和失败；

(4) 试验出现成功情形的概率 p 对于各试验来说是统一的，这也意味着失败的概率对各试验来是也是一致的。

示例 4.30

已知一个鱼塘中的鱼有 70%身长达到 20cm。请计算：随机选取 8 条鱼中有 3 条鱼身长没有达到 20cm，即 8 条鱼中有五条身长达到 20cm 的概率是多少?

E

答案：我们想应用二项式分布，首先可以参照上方检查一下各个条件。试验次数是固定的（$n = 8$）符合条件 2，每个试验结果可以分为成功或失败符合条件 3， 因为样本是随机抽取的，所以试验是独立的符合条件 1，并且每个试验的成功概率是相同的符合条件 4。所以确定该分布符合二项式分布。

我们设定的结果是，在 $n = 8$ 次试验中有 $k = 5$ 次成功。回想一下，成功是鱼的身长达到 20cm，成功的概率为 $p = 0.7$。因此，随机选取 8 条鱼中有 3 条鱼身长没有达到 20cm，即 8 条鱼中有五条身长达到 20cm 的概率是：

$$\binom{8}{5} (0.7)^5 (1-0.7)^{8-5} = \frac{8!}{5!\,(8-5)!} (0.7)^5 (1-0.7)^{8-5} = \frac{8!}{5!\,3!} (0.7)^5 (0.3)^3$$

处理阶乘部分：

$$\frac{8!}{5!\,3!} = \frac{8\times7\times6\times5\times4\times3\times2\times1}{(5\times4\times3\times2\times1)(3\times2\times1)} = \frac{8\times7\times6}{3\times2\times1} = 56$$

然后计算$(0.7)^5(0.3)^3 \approx 0.00454$，所以最终概率约为 $56 \times 0.00454 \approx 0.254$。

计算二项式概率

使用二项式模型的第一步是检查模型是否适用，第二步是确定 n，p 和 k。最后，使用软件或公式来确定概率，然后解释结果。

如果你必须手动计算，通常有用的方法是在二项式系数的分子和分母中尽可能地约分。

指导练习 4.31

如果我们从前面讨论过的鱼塘中随机捕捞 40 条鱼，预计有多少条身长达标？身长达标的鱼的数量的标准差是多少？[1]

指导练习 4.32

随机选取一个吸烟者，他一生中患严重肺病的概率约为 0.3。如果你有 4 个吸烟的朋友，是否满足二项式分布的条件？[2]

指导练习 4.33

假设这四个朋友彼此不认识，我们可以把他们当作从人群中随机抽取的样本。这时应用二项式模型合适吗？计算一下情况的概率：

(a) 他们都不会患上严重的肺病；

(b) 其中一个人会患上严重的肺病；

(c) 最多有一个人会患上严重的肺病。[3]

指导练习 4.34

你的四个吸烟朋友中至少有两个会患上严重肺病的概率是多少？[4]

---

[1] 我们要计算期望（均值）和标准差，两者都可以直接通过公式计算出来：$\mu = np = 40 \times 0.7 = 28$，$\sigma = np(1-p) = \sqrt{40 \times 0.7 \times 0.3} = 2.9$。因为已知大约 95%的观测值落在均值的 2 个标准差之内（具体计算方法见第 2.1.4 节），预计在我们的样本中，可以观察到至少 22 条但少于 34 条鱼身长达到 20cm。

[2] 如果朋友们彼此认识，那么独立性假设很可能不满足。例如，熟人可能有相似的吸烟习惯，或者那些朋友可能会约定一起戒烟。

[3] 为了确定二项式模型是否适用，我们必须验证条件。
(1) 由于我们假设可以将朋友视为随机样本，因此它们是相互独立的。
(2) 试验次数是固定的 $n = 4$。
(3) 每次实验结果都是成功或失败。
(4) 每次试验的成功概率是相同的：这点显然满足，且如果我们说「成功」是某人得了肺病，那成功的概率 $p = 0.3$。

接着使用二项式公式计算(a)和(b)部分：
$P(0) = \binom{4}{0}(0.3)^0(0.7)^4 = 1 \times 1 \times 0.7^4 = 0.2401$（注：$0! = 1$），$P(0) = \binom{4}{1}(0.3)^1(0.7)^4 = 0.4116$
(c)部分为(a)和(b)部分之和：$P(0) + P(1) = 0.2401 + 0.4116 = 0.6517$。
也就是说，有大约 65%的可能性，你的四个吸烟朋友中至多有一个会患上严重的肺病。

[4] 在指导练习 4.33 中可知，至多有一个朋友患肺病的概率是0.6517，而 $P(\text{至多有一个朋友患肺病})$ 与 $P(\text{至少有两个会患上严重肺病})$ 互补，所以 $P(\text{至少有两个会患上严重肺病}) = 1 - P(\text{至多有一个朋友患肺病}) = 1 - 0.6517 = 0.3483$

指导练习 4.35

假设你有 7 个吸烟的朋友，且他们可以作为抽烟者的随机样本。

(a) 预计有多少人会患上严重的肺病，也就是均值是多少?

(b) 你的 7 个朋友中最多有 2 个患上严重肺病的概率是多少? [1]

接下来我们考虑在某些特定情形下，二项式概率中的第一项 n 选 k 的一些特殊情况:

指导练习 4.36

为什么对于任意数字 n，$\binom{n}{0} = 1$ 和 $\binom{n}{n} = 1$ 都成立? [2]

指导练习 4.37

在 n 次试验中，出现 1 次成功和 $n-1$ 次失败有多少种可能? 在 n 次试验中，出现 $n-1$ 次成功和 1 次失败有多少中可能? [3]

---

[1] (a) $\mu = 0.3 \times 7 = 2.1$；(b) $P(0, 1, or\ 2\ 患严重肺病) = P(k=0) + P(k=1) + P(k=2) = 0.6471$。

[2] 用文字表达出这两个式子：第一个可以写成：n 次实验中出现 0 次成功和 n 次失败的方式数量是多少？；第二个可以写成：n 次实验中出现 n 次成功和 0 次失败的方式数量是多少？这样就可以直观地看出对于任意数字 n，这两个式子都成立。

[3] 出现 1 次成功和 $n-1$ 次失败：我们可以在 n 次试验中任意一次获得成功，所以有 n 情况会出现一次成功和 $n-1$ 次失败；出现 $n-1$ 次成功和 1 次失败也同理。在数学上，我们通过验证下面两个公式表达这两个结果（此处不再展开验证）：

$$\binom{n}{1} = n \qquad \binom{n}{n-1} = 1$$

4.3.2 二项分布的正态近似估计法

当样本大小 n 很大时，尤其是当我们要考虑一系列观察值时，二项式公式很麻烦。在某些情况下，我们可以使用正态分布作为一种更容易便捷的方法来估计二项式概率。

示例 4.38

假设已知大约 15%的美国人吸烟。某当地政府认为他们的社区吸烟率较低，并随机选取了 400 个人进行了调查。调查发现，400 名受访者中只有 42 人吸烟。 如果社区中吸烟者实际占比（上帝视角下的总体参数）为 15%，那么在 400 人的样本中观察到 42 名或更少吸烟者的概率是多少?

答案：首先采用二项式分布解答本题，即把调查每名受访者吸烟与否当作一个独立的伯努利试验，这样的试验进行了 400 次从而构成了一个二项分布。首先要验证本例中的分布符合二项分布的四个条件，我们把它留作课后作业，在此不再赘述。

E

题干中的问题相当于，在 $n = 400$，$p = 0.15$ 的样本中，观察到 $k = 0$，1，2，...，42 的概率是多少？ 我们可以计算这 43 种不同的概率，并将它们相加，得出答案：

$$\begin{aligned} &P(k = 0 \ or \ k = 1 \ or \ \dots \ or \ k = 42) \\ &= P(k = 0) + P(k = 1) + \cdots + P(k = 42) \\ &= 0.0054 \end{aligned}$$

如果社区中吸烟者的真实比例为 $p = 0.15$，那么在 $n = 400$ 的样本中观察到 42 名或更少吸烟者的概率为 0.0054。

示例 4.38 中要算 43 个概率.....这着实有些过于繁琐。 一般来说，如果存在更快速简便且依然准确的替代方法，我们应该避免这样的计算。回想一下，「观察到 42 名或更少吸烟者」和之前正态分布章节中例如「成绩比某某分还低」的场景有些类似。而在正态分布中，使用 Z 分数和表格计算一个区间概率要容易得多。那么，用正态分布的方法计算二项式分布合理吗？答案是合理的，只要二项分布满足特定条件，就可以使用正态分布的方法计算。

示例 4.39

已知一个二项分布，成功概率 $p = 0.10$。图 4.9 中有四个空心直方图，分别描述了样本大小 $n = 10$、30、100 还有 300 的二项分布。样本大小变大时，分布的形状会发生什么变化？最后一个空心直方图类似于什么分布?

E

答案：观察四个直方图可以知道，样本容量变大时，分布从明显的块状和右倾形状逐渐变得圆滑对称，而最后一个空心直方图已经近似于正态分布。

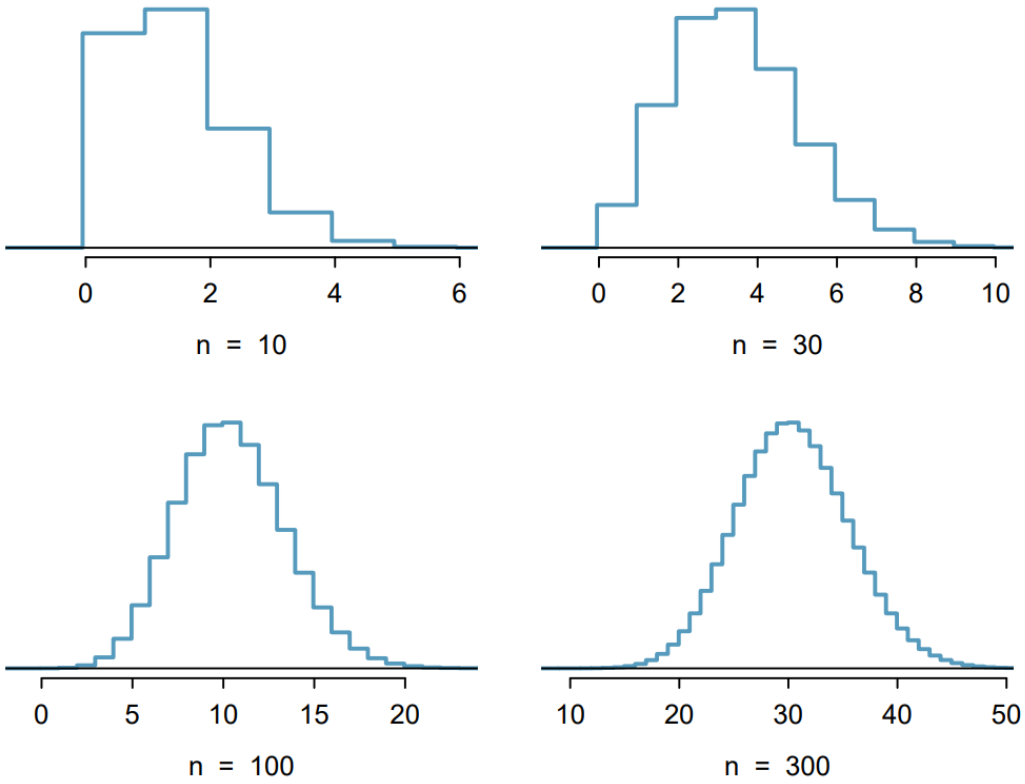


图 4.9：当 p = 0.10 时不同样本量的二项式空心直方图模型
四幅图的样本量分别为 n =10、30、100 和 300。

二项分布的正态近似模拟

当样本大小 n 足够大，使得 np 和 $n(1-p)$ 都至少等于 10 的时候，成功概率为 p 的二项分布近似为正态分布。近似的正态分布和二项分布的均值与标准差相等：

$$\mu = np \qquad \sigma^2 = np(1-p) \qquad \sigma = \sqrt{np(1-p)}$$

当计算二项分布中试验成功的概率分布时，可以使用正态分布近似。例如，我们可以将正态分布应用于示例 4.38 的情境中。

示例 4.40

如果吸烟者的真实比例是 $p = 0.15$，我们如何使用正态分布来估计在 400 个样本中观察到 42 个或更少吸烟者的概率?

E

答案:我们已经在示例 4.38 中验证过本案例适合使用二项分布模型。我们还需要验证 np 和 $n(1-p)$ 都至少为 10:

$$np = 400 \times 0.15 = 60 \qquad n(1-p) = 400 \times 0.85 = 340$$

验证过条件成立后，我们可以使用正态分布代替二项分布，并用二项分布的均值和标准差进行计算:

$$\mu = np = 60 \qquad \sigma = np(1-p) = 7.14$$

我们想用这个模型计算观察到 42 个或更少吸烟者的概率。

指导练习 4.41

G

使用正态分布 $N(\mu = 60, \sigma = 7.14)$ 估计观察到 42 名或更少吸烟者的概率。注意，这里的答案应该约等于示例 4.38 中得到的结果：0.0054。[1]

4.3.3 二项分布的正态近似法在小区间内不适用

二项分布正态近似方法对于小区间的概率估计内不太适用。当研究的区间相对分布全部范围较小时，即使样本量满足 np 和 $n(1-p)$ 都至少等于 10 这个条件，用二项分布的近似正态分布来计算概率，效果并不好。

例如，假设我们想计算当 $p = 0.15$ 时，在 400 人中观察到 49、50 或 51 名吸烟者的概率。面对如此大的样本，我们可能会尝试应用近似的正态分布方法。同时我们也使用二项分布的计算式来算出 49 到 51 的区间成功概率（即三次连续的独立试验结果均为成功的概率），并与正态分布近似法得到的结果进行比较。我们会发现用二项分布和近似的正态分布得到的两种结果明显不同:

$$\text{二项分布结果：}0.0649 \qquad \text{正态分布结果：}0.0421$$

从图 4.10 中我们可以看出产生偏差的原因。图中红线框出的面积对应二项分布的概率，阴影部分对应正态分布的面积。需要注意的是，正态分布下的区域宽度比二项分布的区间两侧都窄了 0.5 个单位。

[1] 首先计算 Z 分数：$Z = \frac{42-60}{7.14} = -2.52$，对应的左尾面积为 0.0059，所以计算出观察到 42 名或更少吸烟者的概率是 0.0059。

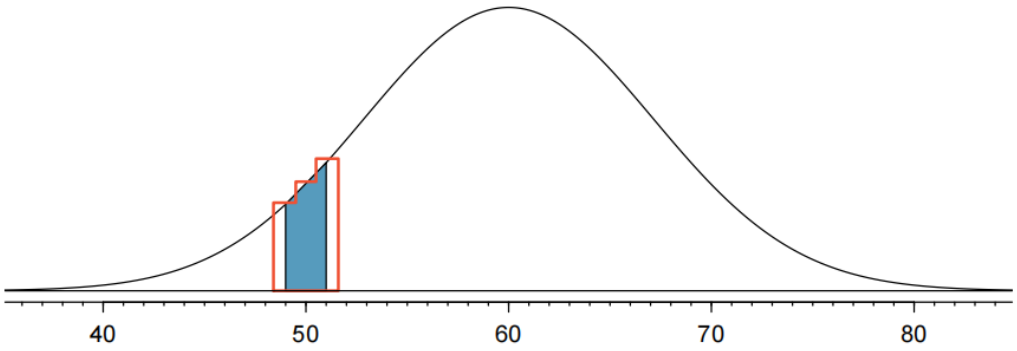


图 4.10：描述上述案例的一条正态曲线，其中阴影面积对应了 49 到 51 之前的区间概率，而红色框线代表了用二项式计算出的相应概率。

改进二项分布的正态近似方法

通过修改临界值值，二项式法和正态分布近似法算出的区间面积不同的问题可以得到改善。具体来说，对于正态分布阴影区域下限的截断值应减小 0.5，上限的截断值应增加 0.5。

这一技巧在考量一个由若干观测值构成的较窄的区间时很好用。在上述例子中，经过修正后的正态分布估计值为 0.0633，与实际值 0.0649 更接近。尽管在计算尾部面积时也可以应用这种校正方法，但由于计算的尾部区间通常很宽，修改的效果并不明显。比如可以想象，0.05 和 0.04 之间就差了前者的百分之 20，而 0.60 和 0.59 则只差了前者的百分之 1.67。

4.4 负二项分布

虽然名字中有「二项」一词，但**负二项分布 negative binomial distribution** 和几何分布从产生原理上相似度更高。几何分布描述了在第 n 次试验时观察到了首次成功的概率。负二项分布则更为一般化，它描述了在第 n 次试验时观察了第 k 次成功的概率。简单来说，之前的几何分布章节我们举过投骰子的例子，即投一个六面均匀的骰子，计算第 n 次时观察到第一次出现数字 1 的概率。这里随着 n 的不同：

$$
\begin{aligned}
&\text{第 1 次就观察到首个 1 的概率为 } 0.166\\
&\text{第 2 次才观察到首个 1 的概率为 } 0.139\\
&\text{第 3 次才观察到首个 1 的概率为 } 0.116\\
&\text{第 4 次才观察到首个 1 的概率为 } 0.096\\
&\text{第 5 次才观察到首个 1 的概率为 } 0.080\\
&\cdots\cdots
\end{aligned}
$$

这里的 1、2、3、4、5 等次数就是随机变量的取值，而 0.166、0.139、0.116、0.096、0.080 等就是取值所对应的场景出现概率。现在，我们来到负二项分布，讨论的变量取值依然是次数 1、2、3、4、5 等等，但在计算概率时不再计算首次观察到 1 的概率，而是计算最后一次出现了第 3 次数字 1 的概率。这里的「最后一次出现第 3 次[1]数字 1」就和上文「第 n 次（n 对应了最后一次）出现第 k 次（$k=3$）成功（成功 = 观察到 1）」对应。可以想象，当总共投掷 1 次和 2 次的时候，不可能在最后一次出现第 3 次数字 1（因为本身就只投了两次），所以对应概率都是 0。而总共投 3 次的时候，在最后一次出现第 3 次数字 1 的概率为 $1/6 \times 1/6 \times 1/6 = 1/216 = 0.005$。我们还可以计算第 4 次，第 5 次，将会依次得到 0.012 和 0.019 等等。这样我们就可以获得一个基于投骰子的 n 次试验中，在最后一次观察到第 3 次（$k = 3$）的负二项分布。可能读完这个例子后，你依然会觉得负二项分布有点抽象，我们不妨再通过几个其他的示例场景来展开，进一步感受负二项分布的概念。

示例 4.42

一个足球运动员每天需要成功踢出四个 35 码的射门后，才可以结束训练回家。假定每一次射门成功的概率为 p。如果 p 很小，例如接近 0.1，难预计他要尝试多少次才能取得四次成功？

E

答案：我们实际上是在等待第四次成功（$k = 4$）。如果成功的概率 p 很小，那么最终尝试的次数 n 可能会很大。结合本例子，成功概率小就意味着运动员脚法不够好，那么想要挑战 4 次成功就意味着需要更多次尝试。换句话说，最后的总尝试次数 n 取值较小的概率应该是很低的。

[1] 译者注：这里特别强调出现「第」三次数字 1 而不是出现了 3 次数字 1 是因为：如果投掷 5 次，前三次都是数字 1 则满足「出现了 3 次」但却不满足「出现了第 3 次」

判断一个分布是否符合负二项分布，我们需要验证四个条件。其中，前三个条件在二项分布中很常见。

是否符合负二项分布，检查如下四个条件

(1) 试验间相互独立；

(2) 每个试验的结果可以且仅能被简单归纳为两种情况：成功和失败；

(3) 试验出现成功情形的概率 p 对于各试验来说是统一的，这也意味着失败的概率对各试验来是也是一致的；

(4) 最后一次试验必须是成功的。

指导练习 4.43

G 假设这名足球运动员训练时非常勤奋，技术大幅提升，现在他成功踢出 35 码射门的概率 $p = 0.8$。请问，他在第 4 次成功之前，需要尝试多少次？[1]

示例 4.44

在昨天的训练中，这个运动员只试了 6 次就踢进了他的第 4 个球。写出每种可能的踢球顺序。

E 答案：他试了 6 次才第 4 次成功，且最后一脚一定是成功的。剩下 3 次成功的踢球和 2 次不成功的踢球 (我们称之为失败) 构成了前 5 次尝试。那么前 5 次尝试有 10 种可能的顺序，如图 4.11 所示。所以，如果他在第 6 次尝试（$n = 6$）时获得了第 4 次成功（$k = 4$），他的成功和失败顺序必定是这 10 个可能序列中的一个。

指导练习 4.45

G 图 4.11 中的每个序列正好有 2 次失败和 4 次成功，且最后一次尝试总是成功。如果成功的概率 $p = 0.8$，求第一个序列发生的概率。[2]

[1] 从直觉来讲，他至少要尝试四次，最多可能不会超过六或七次，因为现在每脚球都很可能成功。

[2] 第一个序列：$0.2 \times 0.2 \times 0.8 \times 0.8 \times 0.8 \times 0.8 = 0.01$。

	尝试次数					
	1	2	3	4	5	6
1	F	F	$\overset{1}{S}$	$\overset{2}{S}$	$\overset{3}{S}$	$\overset{4}{S}$
2	F	$\overset{1}{S}$	F	$\overset{2}{S}$	$\overset{3}{S}$	$\overset{4}{S}$
3	F	$\overset{1}{S}$	$\overset{2}{S}$	F	$\overset{3}{S}$	$\overset{4}{S}$
4	F	$\overset{1}{S}$	$\overset{2}{S}$	$\overset{3}{S}$	F	$\overset{4}{S}$
5	$\overset{1}{S}$	F	F	$\overset{2}{S}$	$\overset{3}{S}$	$\overset{4}{S}$
6	$\overset{1}{S}$	F	$\overset{2}{S}$	F	$\overset{3}{S}$	$\overset{4}{S}$
7	$\overset{1}{S}$	F	$\overset{2}{S}$	$\overset{3}{S}$	F	$\overset{4}{S}$
8	$\overset{1}{S}$	$\overset{2}{S}$	F	F	$\overset{3}{S}$	$\overset{4}{S}$
9	$\overset{1}{S}$	$\overset{2}{S}$	F	$\overset{3}{S}$	F	$\overset{4}{S}$
10	$\overset{1}{S}$	$\overset{2}{S}$	$\overset{3}{S}$	F	F	$\overset{4}{S}$

图 4.11：第 6 次踢出第 4 次成功的 10 种情况枚举。

如果该运动员踢一个 35 码的射门成功的概率 $p = 0.8$，那么他要尝试 6 次后恰巧获得第 4 次成功的概率是多少？我们可以把它写作：

$$\begin{aligned} &P\big(\text{他试了 6 次才获得 4 次成功}\big) \\ &\quad = P\big(\text{他在前 5 次尝试中获得 3 次成功，且第 6 次尝试一定成功}\big) \\ &\quad = P\big(\text{第 1 个序列 or 第 2 个序列 or ... or 第 10 个序列}\big) \end{aligned}$$

这里的序列来自图 4.11。我们可以把 $P\big(\text{第 1 个序列 or 第 2 个序列 or ... or 第 10 个序列}\big)$ 写作 10 个相互独立的可能序列之和：

$$\begin{aligned} &P\big(\text{第 1 个序列 or 第 2 个序列 or ... or 第 10 个序列}\big) \\ &\quad = P\big(\text{第 1 个序列}\big) + P\big(\text{第 2 个序列}\big) + \cdots ... + P\big(\text{第 10 个序列}\big) \end{aligned}$$

在指导练习 4.45 中，我们计算得出第一个序列的概率是 0.0164，其他每个序列具有相同的概率。因为 10 个序列中的每一个都有相同的概率，所以总概率是任意序列的概率的 10 倍。计算负二项式概率的方法类似于第 4.3 节中解决二项式问题的方法，我们也是把概率分为两部分：

$$\begin{aligned} &P\big(\text{第 6 次尝试时获得第 4 次成功}\big) \\ &\quad = \big[\text{\# 可能序列数}\big] \times P\big(\text{单个序列}\big) \end{aligned}$$

接着我们去分别计算出各个部分的结果，然后我们相乘得到最终结果。

我们首先确定后半部分，也就是单个序列出现的概率。这里的计算可以投机取巧一些，因为各种序列出现的概率应该相等，我们不妨考虑一种特殊情况：先得到所有失败的结果，然后再出现成功的结果。

$$\begin{aligned} &P(\text{单个序列}) \\ &\quad = P(n-k \text{ 次失败，接着 } k \text{ 次成功}) \\ &\quad = (1-p)^{n-k}p^k \end{aligned}$$

然后，我们还需要确定一般情况下的序列数。在前面的练习中，对于第 6 次尝试时获得第 4 次成功，我们确定了 10 个可能的序列。这些序列都是先把最后一个观察结果确定为成功，然后对剩下的 3 次成功和 6 次失败的结果进行排列。换句话说，在 $n-1$ 次试验中出现 $k-1$ 次成功，有多少种排列组合方式？可以用二项式 n 选 k 公式计算，不过需要把 n 和 k 换成 $n-1$ 和 $k-1$。

$$\binom{n-1}{k-1} = \frac{(n-1)!}{(k-1)!\,((n-1)-(k-1))!} = \frac{(n-1)!}{(k-1)!\,((n-k))!}$$

这就是在 $n-1$ 次试验中出现 $k-1$ 次成功和 $n-k$ 次失败的组合数。如果对这个公式里的阶乘符号（感叹号）不熟悉，请查阅第 150 页。

负二项分布

负二项分布描述了在第 n 次试验时获得第 k 次成功的概率，其中每次试验相互独立：

$$P(\text{第 } n \text{ 次试验时恰巧出现第 } k \text{ 次成功}) = \binom{n-1}{k-1}(1-p)^{n-k}p^k$$

其中 p 代表了每次独立试验获得成功的概率。

示例 4.46

使用负二项分布的公式，计算那名足球运动员在第 6 次尝试时实现第 4 次成功射门的概率，注意正确结果应该是 0.164（成功概率按照 0.8 计算）。

E

答案：单次成功的概率 $p = 0.8$，成功次数 $k = 4$，该场景下必要的尝试次数 $n = 6$：

$$\binom{n-1}{k-1}(1-p)^{n-k}p^k = \frac{5!}{3!\,2!}(0.8)^4(0.2)^2 = 10 \times 0.0164 = 0.164$$

指导练习 4.47

G

要满足负二项分布，该足球运动员每次踢球的尝试都应该是相互独立的。你认为，判定他每一次踢球尝试相互独立，是合理的吗？[1]

指导练习 4.48

G

[1] 答案不唯一。我们不能肯定地判断，它们是否相互独立。不过许多运动类数据统计显示，每次踢球几乎相互独立。

假设该运动员每一次踢球尝试相互独立。他在 5 次尝试中有 4 次成功射门的概率是多少？[1]

二项分布与负二项分布

二项分布和负二项分布都是描述随机试验结果的离散概率分布，但它们所描述的问题略有不同。

二项分布描述在一定数量的独立重复试验中，成功的次数的概率分布，而且每次试验成功的概率相同。负二项分布描述的是在一系列独立重复试验中，需要进行多少次试验才能达到指定数量的成功次数。在对比这两种分布的时候，需要注意如下几点：

(1) 不同点：从随机变量可以取到的值的角度：二项分布的次数变量取值最大不能超过预先确定的试验总次数，并且可以取 0（因为可能会出现所有试验都得不出成功结果）；而负二项分布的次数变量取值往往从 1 开始（因为显然进行 0 次试验是不可能观察到 1 次或者多次成功的），而且理论上可以取到正无穷大；

(2) 相同点：这两种分布的随机变量都是离散的，取值只能是整数，如果把分布绘制出来，比较推荐的方法都是用一系列柱子表示；

(3) 相同点：尽管两种分布涉及的概率计算方式不同，但是所有可能取到的随机变量取值对应的概率之和也是 1。

指导练习 4.49

在 70%的日子里，医院至少会收治一名心脏病患者。30%的日子里，没有心脏病患者入院。判断下面的情况符合二项分布还是负二项分布，并计算概率。

(a) 医院在本周刚好有三天接收到心脏病患者的概率是多少？

(b) 医院在周四接收到本周第二个心脏病患者的概率是多少？

下个月的第五天接收到第一个病人的概率是多少？[2]

[1] 如果他第 4 次射门是在 5 次尝试内完成的，那么他需要 4 次或 5 次尝试，即 $n = 4$或 $n = 5$。我们之前已知成功概率 $p = 0.8$，然后可以使用负二项分布计算 $n = 4$ 和 $n = 5$ 次的概率，并相加：

$$P(n = 4 \text{ OR } n = 5) = P(n = 4) + P(n = 5)$$
$$= \binom{4-1}{4-1}0.8^4 + \binom{5-1}{4-1}(0.8)^4(1-0.8) = 1 \times 0.41 + 4 \times 0.082 = 0.41 + 0.33 = 0.74$$

当然这道题还有另一种解法：就是假设无论如何运动员都踢了 5 脚，然后计算踢 5 脚进 4 个加上踢 5 脚进 5 个的两种概率，会发现答案与上面提供的解法一致。之所以要加上踢 5 脚进 5 个是因为这种情况也应该算入「在 5 脚内获得 4 个进球」。

[2] 每一部分中都有 $p = 0.7$。(a) 天数固定，可判断出它符合二项分布。其中 $k = 3$，$n = 7$，计算得到 0.097；(b) 要在最后一天接收到最后一个病人，即出现最后一次成功，可判断出它符合负二项分布。其中参数 $k = 2$，$n = 4$，计算得到 0.132；(c) 本例符合负二项分布，$k = 1$，$n = 5$，计算得到 0.006。注意在负二项分布中，当$k = 1$ 时，该分布也属于几何分布。

4.5 泊松分布

示例 4.50

纽约市大约有 800 万人。根据历史记录，预计每天会有约为 4.4 人因急性心肌梗死（AMI），或者说是心脏病发作而住院？除了这个数字，我们还想知道它大致遵循什么分布。如果记录某一年内每天因心脏病发作而住院的人数，那么这个数量形成的直方图会呈现什么形状？

E

答案：图 4.12 中的直方图描述了纽约市随机抽取的 365 天内因心脏病发作而住院的人数分布。样本均值（4.38）与历史平均值 4.4 相近。样本标准差约为 2，直方图表明约 70%的数据落在 2.4 和 6.4 之间。分布的形状是单峰的，向右倾斜。

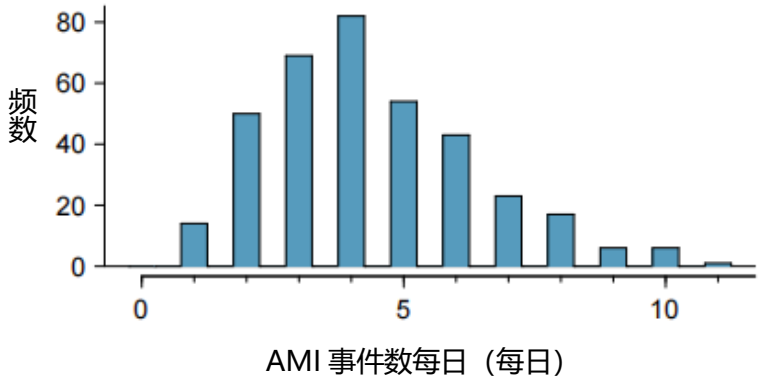


图 4.12：纽约中随机 365 天 AMI 事件数的直方图。

估计单位时间内在大群体中的「事件数量」时，泊松分布很有用。例如，可以考虑以下事件：

- 心脏病发作
- 结婚
- 被闪电击中

泊松分布有助于我们描述一天内这类事件在一个固定的群体中发生的数量，当然需要注意群体中的个体相互独立。泊松分布也可以用于其他时间单位，如一小时或一周。图 4.12 中的直方图近似于速率为 4.4 的泊松分布。泊松分布的**速率 rate** 是单位时间内在基本固定的群体中某个事件发生的平均次数。在示例 4.50 中，时间单位是一天，群体是所有纽约市居民，历史速率是 4.4。泊松分布中的参数是速率，即我们预期观察到的事件数量，通常用 λ（希腊字母 lambda，读作兰姆达）或 μ 表示。使用速率的概念，我们可以描述在单位时间内恰好观察到 k 个事件的概率。

泊松分布

假设我们正在观察事件，并且观察到的事件数遵循一个比率为 λ 的泊松分布。那么

$$P(\text{观察 } k \text{ 个事件}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

其中 k 可以取值 0，1，2，依此类推；$k!$ 表示 k 的阶乘，如第 150 页所述。字母 e 对应的是自然对数的底：$e \approx 2.718$。该分布的均值和标准差分别为 λ 和 $\sqrt{\lambda}$。

根据目前掌握的信息，我们可以对泊松分布也做一个大致的总结：

(1) 泊松分布的随机变量是单位时间区间，例如在最开始的例子中，我们把每天当做一个单位时间区间，统计计算每天内的一些事件发生的数字；
(2) 泊松分布的随机变量也是离散的，我们是从一个单位时间跳到下一个单位时间，比如 1 月 1 日跳到 1 月 2 日；
(3) 与之前介绍的正态分布，几何分布，二项分布和负二项分布都不同的是，泊松分布统计展现的不是概率，而是事件发生的次数。概率是一个从 0 到 1 的连续值，而次数是一个大于等于零的离散整数值，这点也可以从图 4.12 直观地看到；
(4) 尽管泊松分布展现的是次数，但由于其离散特性，它和几何分布，二项分布还有负二项分布一样，常常可以用柱状图来绘制展示。

关于泊松分布更详细的条件和内容，我们在后面的课程中再做介绍。不过现在我们提供了一些简单的准则，可用于初步判断泊松分布是否适用。当事件数量很大，或者事件发生在很大的群体中，且事件都相互独立，那么随机变量很可能遵循泊松分布。

此外，即使事件实际上不是相互独立的，只要我们在不同时间使用不同速率，泊松分布有时仍然成立。例如，人们更倾向于选择周六和周日举办婚礼，使得举办时间不再相互独立。但我们可以在建模时，使周末的速率高于工作日的速率。针对第二个变量为泊松分布设置不同速率的方法，是**广义线性模型 generalized linear models** 中一些更高级方法的基础。在第 8 章和第 9 章，我们将会讨论线性模型的基础。

第 5 章

统计推断 Foundations for Inference

统计推断的主要目的是了解和量化使用参数估计的不确定性。虽然根据具体场景不同，我们使用的公式还有涉及到的流程细节都会发生变化，但统计推断的理论基础在整个统计学中是基本一致的。

我们从一个熟悉的话题开始：使用样本比例估计总体比例。然后，我们会建立所谓的置信区间。置信区间往往是一个取值范围，该范围内将很可能包含真实的总体值。最后，我们引入假设检验框架，该框架允许我们用一种统计学中比较正规的流程来评估关于总体的某个主张。例如某项调查是否提供了足够有力证据，以判断某位候选人是否得到了大多数投票人的支持。

跨越数据银河

系列推文合集

更多视频，演示文稿，和其他相关资源，请访问：
http://www.openintro.org/os

5.1 点估计和样本的波动性

皮尤研究中心 Pew Research Center 这样的调查公司经常通过民意调查来了解人们对政治事件、科学知识、品牌形象等方面的理解和观念。这类调查的最终目的是用抽样调查得到的「样本结果」估计其背后「更大范围的总体」的认知与观点。

5.1.1 点估计和误差

由于现实中很难通过收集总体中所有个体的信息来计算获取总体的情况，一般都会先计算样本信息，再用计算得到的值作为实际总体的估计值。假设有一项民调显示美国总统的支持率是 45%。在这一语义下，45%从统计术语上可以称作**点估计 point estimate**。关于这一术语的理解可以从两方面入手：首先它是「一次」调查所得出的「单个」结论，几何中与「一次」和「单个」最对应的概念就是「点」；其次这项民意调查的目的肯定是为了用有限的调查到的人群的反映来「预测」全体美国公民对总统的支持率，所以计算 45%这个数字本身就带有「估计」的意味。两下结合，就是「点估计」。

依然是在这个民调的背景下，用 45%想要去预测的总体实际支持率术语上被称为**总体参数 parameter**[1]。当总体参数是一个比例时（一般来说就是所有个体中有百分之多少符合某条件），我们往往用小写英文字母 p 表示总体比例，用上面带了「帽子」的 $\hat{p}$（读作：$p-hat$，中文有时也可以读作 p 帽）表示通过样本得到的点估计值。点估计值和总体比例实际值之间的数值差距叫做估计中的**误差 error**，包括两类：抽样误差和偏差。举例来说，假设有一位全知全能的神，他用自己的超能力在 1 秒内和所有美国公民进行了一次对话，并记录了他们对现任总体支持与否的评价，然后又在 0.5 秒内计算出了实际上有 37.5%的人表示了支持。那么这里的 37.5%就是感兴趣参数，之前调查得出的 45%就是点估计，7.5%就是误差。

抽样误差 sampling error，也称为样本的不确定度。从不同样本得到的不同估计值往往有大有小，抽样误差描述的就是这种不同估计值的变动程度。比如，从一个样本得到的估计值可能比实际值高了 1%，从另一个样本得到的估计值可能比实际值低了 3%。很多统计知识都旨在理解并量化抽样误差。在量化抽样误差的过程中，我们往往会讨论到**样本大小 sample size**，它是非常重要的概念，一般用小写 n 表示。

[1] 之所以称这种估计数据叫点估计，是因为我们需要先从样本计算出的一个统计值（例如受调查人群中支持某某某做美国总统人口比例），而这个统计值从几何意义上来说就像是一个点。之所以把总体的未知实际值（例如所有美国公民中支持某某某做总统的人口比例）称作总体参数，是因为该术语能够体现我们进行统计学研究的目的和总体参数是变量的本质。

偏差 bias 则是一种系统性倾向，它是由抽样过程本身决定的，和单次抽样没有关系。例如，如果我们在统计鱼塘里面鱼身长度的时候，采用网捕法，那么比网眼小的鱼就会漏网而出。这样无论怎么捕捞取样，最后统计结果都会偏大。从字面意思来说，偏差这一词包括了「偏见」和「误差」的含义，是用带有误导性的方式收集数据造成的误差。再举一个例子：设想当我们调查学生们是否支持在校内新建体育馆时，如果问题的表述是：「你愿意支持学校投资新建体育馆吗？」，那这个问题本身就容易诱导人们去进行正面回答，即做出支持的答案。这样最终得到的很可能是不准确的结论。我们需要高质量地收集数据以尽可能地减少偏差，在第一章中我们已经对此进行过探讨，之后的章节也会对其进行进一步的介绍。

5.1.2 理解用样本计算点估计的波动性

假设美国支持推广太阳能的成年人占全部成年人的真实比例是：$p = 0.88$。如果我们让随机 1000 名美国成年人投票表决是否支持推广，得到的样本中投支持票的比例估计值大概率不会等于实际值 88%，但这一样本比例估计值会有多接近实际值呢？换句话说，我们想知道的是，*当实际总体比例是 0.88 时，样本比例估计值 $\hat{p}$ 将会是怎样的*？如果我们进行多次抽样，那么 $\hat{p}$ 有没有什么规律或者特征？让我们做个模拟试验来获得结论！由于已经知道了推广太阳能的实际支持率是 0.88，我们可以模拟当我们询问随机 1000 个美国成年人时得到的可能回复。以下是我们如何设计这个模拟试验的：

(1) 2018 年，美国约有 2.5 亿的成年人。取 2.5 亿张白纸条，在 88%中写上「支持」，在剩余 12%中写上「不支持」。

(2) 把所有纸条混在一起，取出其中的 1000 张，当作是抽取的 1000 个美国成年人的样本。

(3) 计算样本中写有「支持」的纸条的比例。

有人愿意进行这一模拟吗？可能不会有人。用 2.5 亿张纸条来进行模拟听起来就像是天方夜谭，但我们可用计算机写代码进行模拟。图 5.1 中是相关代码的写法。在这次模拟中，样本得出的点估计 $\hat{p}_1 = 0.894$。我们知道总体比例 $p = 0.88$，所以估计值的误差是 0.894 - 0.88 = +0.014。

```
# 1. Create a set of 250 million entries, where 88% of them are "support"
#    and 12% are "not".
pop_size <- 250000000
possible_entries <- c(rep("support", 0.88 * pop_size), rep("not", 0.12 * pop_size))

# 2. Sample 1000 entries without replacement.
sampled_entries <- sample(possible_entries, size = 1000)

# 3. Compute p-hat: count the number that are "support", then divide by
#    the sample size.
sum(sampled_entries == "support") / 1000
```

图 5.1：这是统计软件 **R** 模拟单一样本估计值 $\hat{p}$ 的代码。以#开头的行是代码注释，用来描述代码的作用。如果你想进一步学习，我们在 openintro.org/stat/labs 提供了更多的资料。

一次模拟不足以获得关于估计值分布的普遍结论，所以我们应该多进行几次模拟。在第二次模拟中，我们得到 $\hat{p}_2$ = 0.885，其误差是+0.005。第三次模拟得到 $\hat{p}_3$ = 0.878，其误差是-0.002。第四次模拟得到 $\hat{p}_4$ = 0.859，其误差是-0.021。在计算机的帮助下，我们进行了 10,000 次模拟，并绘制了从这 10,000 次模拟得到的结果构成的直方图，如图 5.2 所示。这些样本比例估计值的分布叫做**抽样分布 sampling distribution**。这一分布的概率的特点如下：

中心 Center 该分布的中心是 $\bar{x}_{\hat{p}}$ = 0.880,这个值正是总体的参数（回顾上一章节，参数即我们感兴趣的统计量的总体实际值）。请注意，模拟试验模仿了对总体的简单随机抽样，这是一种很直接的抽样策略，有助于避免抽样偏差，关于简单随机抽样可以本书参考第 1.3.5 小节。

散布 Spread 该分布的标准差 $s_{\hat{p}}$ = 0.010。当我们讨论一个抽样分布，或者说点估计的波动性特征时，我们一般用**标准误 standard error** 这个统计术语，而不是「标准差」一词。符号 $SE_{\hat{p}}$ 用来表示样本比例的标准误。

形状 Shape 这一分布是对称的，呈钟形，它近似于正态分布。

这些发现令人兴奋！当总体比例 p = 0.88，样本大小 n = 1000 时，样本比例估计值 $\hat{p}$ 对总体真实比例参数的估计非常准确。我们也观察到一个很有趣的现象：抽样分布的直方图类似正态分布。

抽样分布无法被观察到，但我们应该记着它们

在现实应用中，我们没法实际观测到抽样分布，因为我们从来都不太可能真正去花费大量成本反复抽样。因此，抽样分布也往往只是作为一个理论假设存在。然而，把点估计看做来自于这种假设分布的（随机变量的）一次取值是很有用的。了解抽样分布有助于我们更好地描述和理解观察到的点估计。

示例 5.1

E 如果我们用一个更小的样本，例如把样本大小从 n = 1000 修改为 n = 50，你觉得的 $\hat{p}$ 标准误会比 n = 1000 时的更大还是更小?

答案：直觉上来讲，数据似乎越多越好。这在理论上也是正确的！当 p = 0.88 时，n = 50 时得到的误差会比 n = 1000 时得到的误差更大。

示例 5.1 强调了我们会在之后多次见到的一条重要性质：样本大小越大，得到的点估计就越准确。

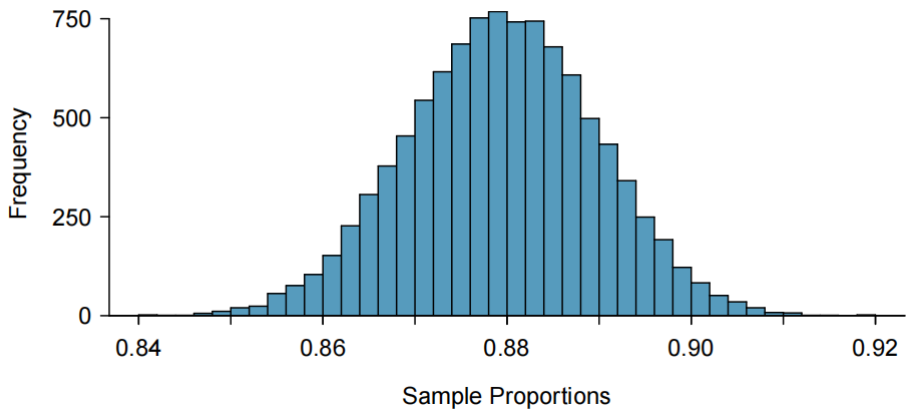


图 5.2：由 1000 个样本比例累计而成的柱状图。其中样本大小 n = 1000，每个样本都从实际比例为 0.88 的总体中抽取。

5.1.3 中心极限定理

图 5.2 中的分布看上去非常像正态分布。这种结果并非个例，而是具有普遍性。我们将这种具有一般性的规律称为**中心极限定理 Central Limit Theorem**。

中心极限定理和成功-失败条件

如果观测结果相互独立，且样本大小足够大，通过样本得到的比例估计值 $\hat{p}$ 的分布将趋向正态分布，或者说近似服从正态分布；对应正态分布的平均值和标准误如下。

$$u_{\hat{p}} = p \qquad SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

我们将右侧的计算结果称为「标准误」而非标准差的原因是：我们在讨论的已经不再是简单统计计算，而是用样本的估计量进行统计推断。我们在意的不再是数据分布离散情况（标准差的定义），而是对总体进行「估计」时产生的误差大小，用以评估统计推断的可靠程度。标准误某种意义上来说也是一个标准差，但它是样本点估计值的标准差。如果脱离了用样本估计总体的研究背景，那么标准误概念也就会失去相应的意义。

大家在不少统计软件中都能看到 std.err.的字样，这也是在提醒我们这个数据的产生不仅仅是为了算出离散，而是为了帮助我们进行统计推断。因为标准误本质也是标准差，所以标准误越小，样本统计量的分布越集中，对总体就越有代表性。一般来说，要想使中心极限定理成立，样本大小 n 需要同时满足 $np \geq 10$ 和 $n(1-p) \geq 10$。这个关于样本大小的前提条件被称为**成功-失败条件 success-failure condition**。

中心极限定理极其重要，是很多统计学理论的基石。当你运用中心极限定理时，请注意上述的两个条件：观测结果之间必须互不影响，样本大小 n 需要同时满足 $np \geq 10$ 和 $n(1-p) \geq 10$ 同时成立。后一个条件其实也可以理解为，进入成功组的样本数和进入失败组的样本数都至少有 10 个。

示例 5.2

我们之前用模拟数据，估计了在 p = 88，n = 1000 的情况下，$\hat{p}$ 的平均数和标准误。接下来我们要证明中心极限定理在这一情况下适用，即样本比例分布近似于正态分布。

E

答案：**独立性** 每个样本估计值 $\hat{p}$ 的样本大小都为 1000，并且这些样本都是通过独立抽取得到的。往往简单随机抽样得到的观测值互相也是独立的。

成功失败条件 我们可以通过应用成功失败条件和上述提及的计算公式检验样本大小是否足够大。

$$np = 1000 \times 0.88 = 880 \geq 10 \qquad n(1-p) = 1000 \times (1-0.88) = 120 \geq 10$$

独立性和成功失败条件都成立，所以对 $\hat{p}$ 来说中心极限定理适用，进而我们可以说用正态分布对 $\hat{p}$ 的分布建模是合理的。

如何证明样本观测值是独立的

如果试验对象是被随机分配到试验组的，那他可以被视为是独立的。

如果观测对象来自于简单随机抽样，那他们就是独立的。

如果样本来自于随机过程，比如，生产线上的偶然失误，那检测独立性就变得很难了。在这种情况下，尽你所能去判断。

判定中心极限定理时，有时我们还会加上一个额外的条件：从总体抽取的样本大小不能超过总体的 10%。当样本大小超过了总体大小的 10%时，相比其他更先进的方法[1]，本章介绍的均值和误差的计算公式可能会稍稍高估样本误。这一般都不会是个问题。而且在少数会造成问题的场景下，我们往往会选择非常保守的统计方法。因为选择的统计方法本身已经足够保守了，10%的额外检验相对也就没那么必要了。

[1] 比如，我们可能会采用**有限总体校正**：如果样本大小为 n，总体大小为 N，那么我们会将标准误公式乘以 $\sqrt{\frac{N-n}{N-1}}$ 来获得一个对实际标准误更小更精准的估计。当 $n < 0.1 \times N$ 时，校正系数会相对较小。

示例 5.3

当 p = 0.88，n = 1000 时，根据中心极限定理计算 $\hat{p}$ 分布的平均数和标准误的理论值。

E

答案：$\hat{p}$ 的平均值就是总体比例：$\hat{p}$ = 0.88

$\hat{p}$ 的标准误由下述公式计算得出：

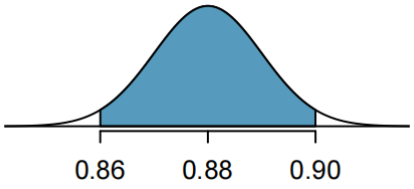
$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.88(1-0.88)}{1000}} = 0.010$$

示例 5.4

估计样本比例落在总体比例参数值 p 左右 0.02（2%）的范围中的频率。基于示例 5.2 和 5.3，我们知道这个分布近乎正态分布，即服从$N(u_{\hat{p}} = 0.88, SE_{\hat{p}} = 0.010)$。

答案：经过 4.1 节中大量的练习，我们应该会对这个正态分布的例子很熟悉，也应该能理解 $\hat{p}$ 的值会处于 0.86 和 0.90 之间。

E



当 $u_{\hat{p}} = 0.88, SE_{\hat{p}} = 0.010$ 时，我们可以计算出左右两个边界点的 Z 分数。

$$Z_{0.86} = \frac{0.86 - 0.88}{0.010} = -2 \qquad Z_{0.90} = \frac{0.90 - 0.88}{0.010} = 2$$

我们可以用统计软件，图形计算器或表格计算位于分布两侧的尾部的面积。单个尾部的面积是 0.0228，他们的总面积就是 2 x 0.0228=0.0456，剩余的阴影部分的面积则为 0.9544。图 5.2 中，大约 95.44%的样本分布都落在总体比例 0.88 的±0.02 范围区间内。

G

指导练习 5.5

在示例 5.1 中，我们讨论了小样本得到的估计不太可靠，请尝试解释这是为什么。提示，你可以参考使用公式 $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$。[1]

[1] 由于样本大小 n 在分母上，样本越大，整体值越小。也就是说，较大的样本大小对应较小的标准误。

5.1.4 将中心极限定理应用于真实世界

除非花费量财力精力调查总体中的每个个体，否则我们不可能知道实际的总体比例。我们之前设定p = 0.88 是由于皮尤研究中心调查发现 1000 个美国成年人中赞成推广太阳能的人数占比为 $\hat{p}$ = 0.887。研究人员们可能会好奇，民调得到的样本比例会不会近似服从正态分布？我们可以通过检验中心极限定理的条件来解答这个问题。

独立性 Independence：民调是一个对美国成年人的简单随机采样，这意味着观测值之间是互相独立的。

成功失败条件 Success-failure conditions：为了检验这个条件，我们需要总体比例 p 来确认 np 和 $n(1-p)$ 是否都大于 10。然而我们并不知道 p 的值（这正是民调机构采取抽样调查的原因）。在这种情况下，我们经常用 $\hat{p}$ 作为评估成功-失败条件的次佳方法。样本比例 $\hat{p}$ 在这个检验中发挥着 p 的合理替代品的作用。两个计算结果都远超最低要求 10。

$$n\hat{p} = 1000 \times 0.887 = 887 \qquad n(1-\hat{p}) = 1000 \times (1-0.878) = 113$$

当计算样本比例的标准误时，这种用 $\hat{p}$ 代替 p 的**近似替换 substitution approximation** 也很有用。这种替换技巧有时也被称为「插入式原则」。通过下方公式可以看到，在这一案例中，和我们之前用总体参数 0.88 时得到的计算结果相比，使用点估计值 0.887 得出的 $SE_{\hat{p}}$ 的变化非常微小（小数点后三位的值都没有改变）。也是因为同样的原因，即使从不同样本中观察到的比例均值略有不同，我们使用点估计值带入公式得出的标准误也会是趋于稳定的。

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.887(1-0.887)}{1000}} = 0.010$$

5.1.5 关于中心极限定理的更多细节

至此，我们已经将中心极限定理应用到了多个例子中。

当观测值独立，样本大小足够大时，$\hat{p}$ 的分布近似于参数如下的正态分布。

$$u_{\hat{p}} = p \qquad SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

当 $np \geq 10$ 且 $n(1-p) \geq 10$ 时，样本大小被视为足够大。

在这一小节中，我们会进一步探索成功失败条件的满足情况，寻求对中心极限定理更深层次的理解。

一个有趣的问题是，当 $np \geq 10$ 且 $n(1-p) \geq 10$ 时会发生什么？正如我们在第 5.1.2 小节做的，我们能模拟从不同大小中抽样。实际比例 $p = 0.25$。下面是一个容量为 10 的样本：

$$否，否，是，是，否，否，否，否，否，否$$

在这次抽样中，我们观测到样本中「是」的回答占比为 $\hat{p} = 2/10 = 0.2$。我们可以对这个比例进行多次模拟，以理解 $N = 10, p = 0.25$ 时 $\hat{p}$ 的抽样分布。图 5.3 画出了这个分布，并附上了与其平均数和方差相同的正态分布。可以发现，从直观视觉上来说，这二者之间还是有很多区别的。

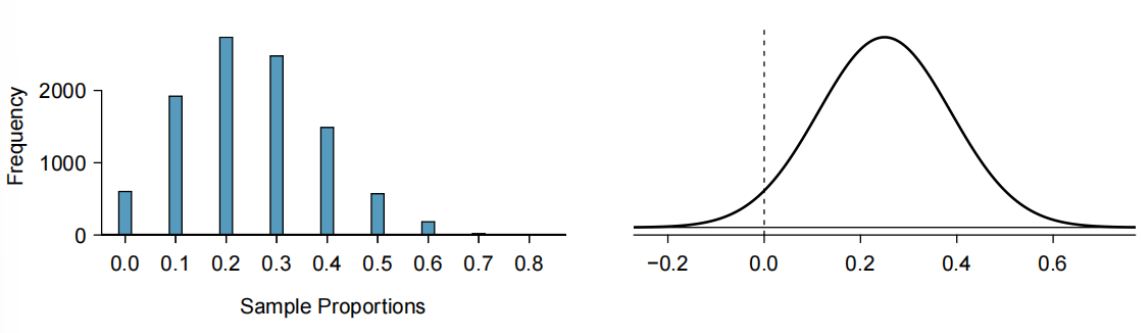


图 5.3：左图展示：当样本大小 $n = 10$，总体比例 $p = 0.25$ 时，通过反复模拟记录得到 $\hat{p}$ 的分布。右图展示：平均数为 0.25，标准差为 0.137 的正态分布。

	Unimodal?	Smooth?	Symmetric?
Normal: $N(0.25, 0.14)$	Yes	Yes	Yes
$n = 10$, $p = 0.25$	Yes	*No*	*No*

注意，当 $n = 10$，$p = 0.25$ 时，成功失败条件不成立。

$$np = 10 \times 0.25 = 2.5 \qquad n(1-p) = 10 \times 0.75 = 7.5$$

这种简单抽样分布并不能说明成功失败条件是完美准则，但它确实正确地判断出正态分布可能不合适。

我们可以再做几次模拟，如图 5.4 和图 5.5。可以看出以下这些趋势：

(1) np 或 $n(1-p)$ 较小时，分布较为离散。

(2) 当 np 或 $n(1-p)$ 小于 10 时，分布显著偏斜。

(3) np 和 $n(1-p)$ 越大，分布越趋向于正态。这有些难以从图中有着较大样本大小的分布看出，因为方差也变小得多。

(4) 当 np 和 $n(1-p)$ 都很大时，分布几乎不呈离散，且看上去更像正态分布。

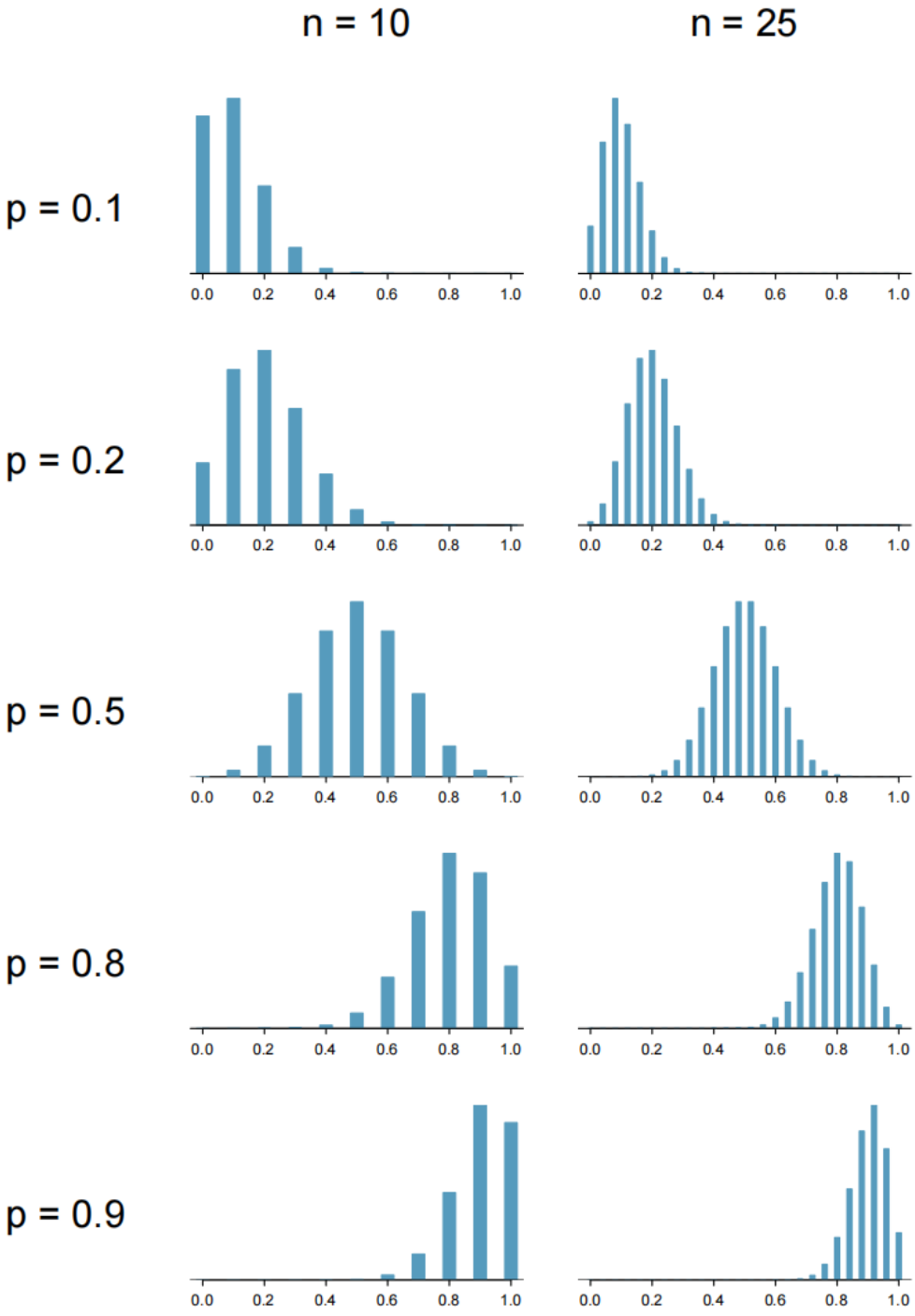


图 5.4：不同 p 和 n 之下的抽样分布

横向：$p = 0.10,\ p = 0.20,\ p = 0.50,\ p = 0.80,\ p = 0.90$

纵向：$n = 10,\ n = 25$

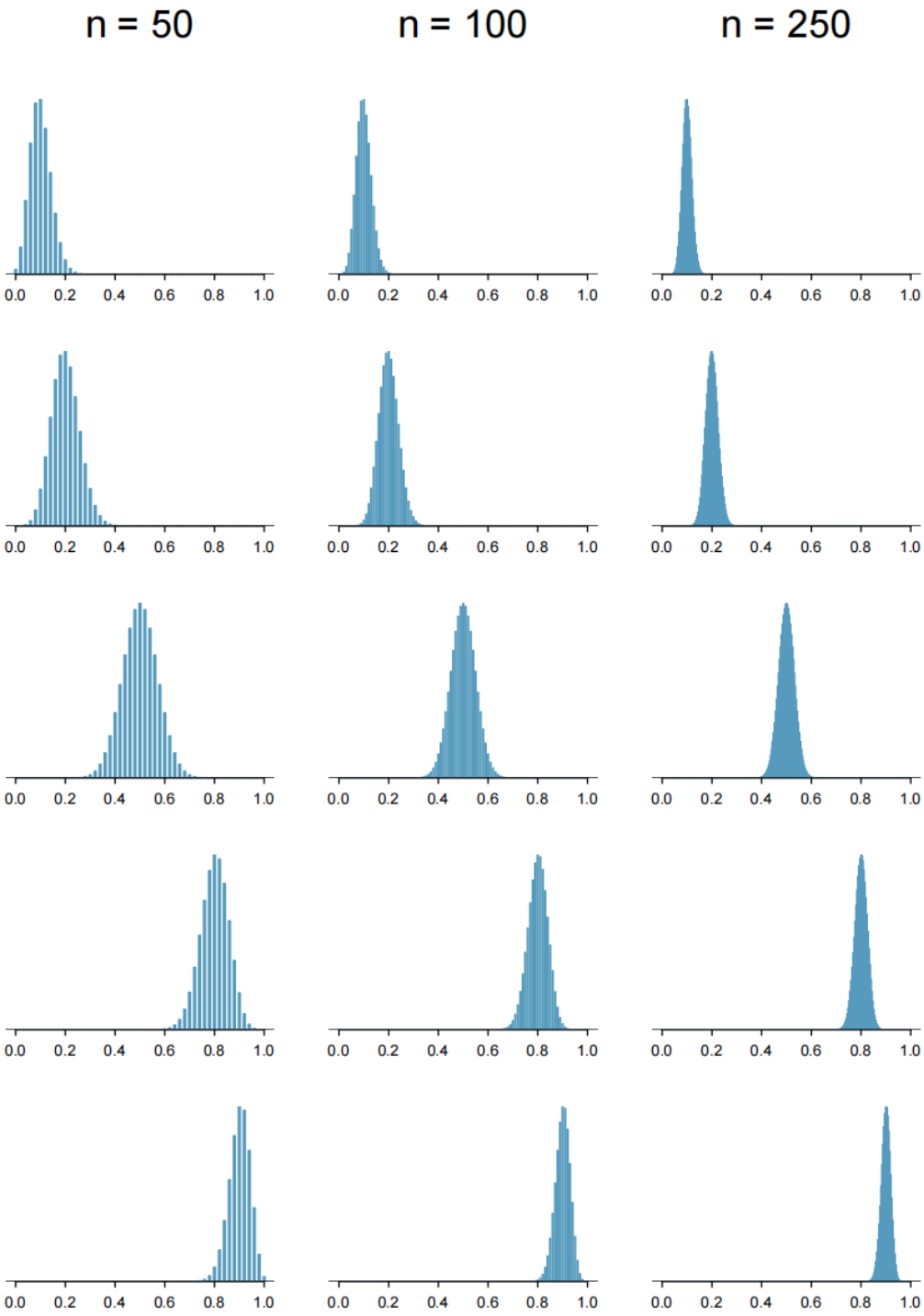


图 5.5：不同 p 和 n 之下的抽样分布

横向：$p = 0.10,\ p = 0.20,\ p = 0.50,\ p = 0.80,\ p = 0.90$

纵向：$n = 50,\ n = 100,\ n = 250$

到目前为止，我们只关注了分布的偏斜和离散程度，还没有考虑分布的平均数和标准差是如何变化的。带着以下三个思路，我们一起回顾一下这些图：

(1) 分布的中心总是总体比例 p。抽样分布 $\hat{p}$ 总是以总体参数 p 为中心，这意味着当数据是从总体中独立抽样获得时，样本比例是**无偏 unbiased** 的。

(2) 对于特定的总体比例 p,随着样本大小 n 变大，多次抽样的样本均值，或者说点估计值分布的波动性减小。这可能和你的直觉一致：一个基于较大样本大小的估计会更准确。

(3) 对于特定的样本大小，样本比例点估计的波动性在 $p = 0.5$ 时达到最大。由于图和图的差异比较小，所以这一点可能需要仔细看才能观察到 --->

这一点实际反映了比例 p 在标准误公式 $SE = \sqrt{\frac{p(1-p)}{n}}$ 的作用。即 $p = 0.5$ 时标准误最大。

$\hat{p}$ 的分布不会百分百完全服从正态分布，因为 $\hat{p}$ 实际上总是离散取值的(x/n)。换句话说，在显示场景下，由于样本大小不可能是无限大，那么从样本中统计的比例就不可能百分百覆盖数轴上从 0 到 1 区间的所有点，而只能是算出一个点之后跳动到下一个点。尽管如此，我们依然可以说根据中心极限定理，在特定条件下样本比例均值的抽样分布是「近似」服从正态分布。在这本书中，我们会用标准成功失败条件，即衡量 np 和 $n(1-p)$ 是否都大于 10，来作为判断一个点估计背后的分布是否「近似」服从正态分布的依据。

5.1.6 将这一框架推广至其他统计量

用样本统计量估计总体参数是很常见的策略。我们可以将它应用到除了比例以外的统计量。比如，如果我们想估计某个学校的研究生平均薪酬，我们可以先对研究生们随机抽样并调查。在这个例子中，我们会用一个样本均值 $\bar{x}$ 来估计所有研究生的总体真实均值 μ。又比如，如果我们想估计两个网站的产品价格区别，我们可能对两个网站上售卖的产品进行随机抽样，得到它们各自的价格，并计算两个网站产品价格的均值差。这种策略肯定会让我们通过一个点估计，了解实际差异。尽管这章节围绕着比例进行讨论，这本书中还会将这一方法运用到其他问题上。虽然细节上有一些不同，但原理和总体思路是一样的。

5.2 单比例点估计的置信区间

样本比例 $\hat{p}$ 为总体参数比例 p 提供了一个合理的数值估计。然而，仅仅使用样本比例均值来代表总体参数并不完美。在上节中我们介绍到了，抽样估计的过程往往是伴随着误差的，而如果我们有资源进行多次抽样，还可以通过统计计算多次抽样误差的波动性，数值上可以用标准误衡量。因此，在用点估计来评估总体参数的时候，更好的做法是除估计值外再提供一个范围，用于表示真实总体值可能落在的区间，这样会让估计结论更合理和有说服力。

5.2.1 用置信区间「捕捉」总体参数

只使用一个点估计，就像在一个浑浊的湖里用长矛捕鱼。我们可以在看到鱼影的地方投掷鱼叉，但很可能投不中；而如果我们在那个地方撒网，就有很大的机会抓到鱼。**置信区间 confidence interval** 就像捕鱼时撒下去的网，它代表了一个很可能包括总体参数的范围。

指导练习 5.6

G 如果我们希望能够非常确定地在一个区间内囊括总体比例，我们应该使用更大的区间还是更小的区间？[1]

5.2.2 构建 95%的置信区间

根据前面的学习我们得知，通过使用样本计算出的比例 $\hat{p}$ 来估计总体比例最为合理，所以围绕这个点估计建立一个置信区间是有意义的。而我们究竟需要建立多大的置信区间？标准误这一统计量则直接为我们提供了指导。

标准误代表点估计值的波动性。虽然现实中我们往往只进行一次抽样，但是可以想象假设资源无限的情况下我们可以进行无数轮抽样，每次抽样得到的均值（即点估计值）会组成一组数，这组数的标准差就是标准误。根据标准误能反映波动性的特点，结合当中心极限定理的条件得到满足时，点估计值近似服从正态分布的信息。我们就可以构建置信区间了。

[1] 如果我们想更加确定我们将捕到鱼，我们可以使用更大的网。同样地，如果我们想更加确定我们包括的参数，我们就使用更大的置信区间。

在正态分布中，95%的数据都在距离平均值的 1.96 个标准差之内。利用这一原则，我们可以构建一个置信区间，从样本比例，也就是点估计值中，延伸出 1.96 个标准误，从而使我们有 95%的**信心 confident** 认为总体的真实比例会落在该区间内：

$$\text{点估计值} \pm 1.96 \times \text{标准误}$$

$$\hat{p} \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

但「95%的信心」又是什么意思？假设我们选取了许多样本，并在每个样本中都建立一个 95%的置信区间。记得在第 5.1.2 小节中，我们使用一种叫做模拟的技术手段不断抽样，并对这些样本进行研究分析。

简单回顾一下这个场景：假设美国民众对于扩大太阳能这种新能源使用的支持率的真实值为 0.88，即为感兴趣的参数实际值为 0.88。我们使用电脑模拟试验的方式，将 2.5 亿美国成年人的支持/不支持结果写在纸片上。按照 0.88 的真实值，2.5 亿张纸片中有 88%的纸片，也就是 2.2 亿张纸片上写着「支持」，其余 3000 万张纸片上写着「不支持」。让机器从这 2.5 亿张在线纸片中随机抽取 1000 张，并统计在抽取出的纸片样本集合中，有多少比例的纸片写着「支持」。我们将每抽取 1000 张纸片的一轮过程称为一次模拟。每一次模拟试验都会对应产生一个随机样本。

图 5.6 展示了按照这种模拟方法构建的 25 个样本，并在其中对每个样本的点估计值建立置信区间的过程。可以看到在下图示例中，有 24 个置信区间包含我们最初假设的总体比例 p = 0.88，1 个区间不包含。不包含的样本使用红色标注了出来。

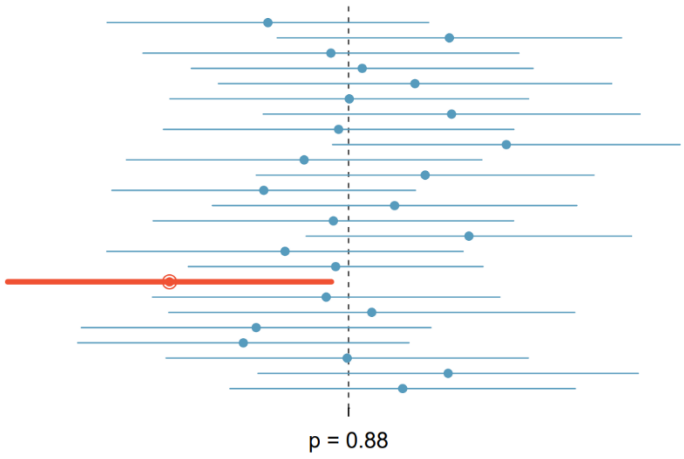


图 5.6：按照第 5.1.2 小节的模拟方法，随机抽取 25 个样本对应的 25 个点估计值和置信区间。这些区间是相对于总体比例 p = 0.88 显示的，其中只有一个不包括总体比例，已在图中用红色加粗进行特别标注。

示例 5.7

在图 5.6 中，有一个区间不包含 $p = 0.88$。这是否意味着我们在构建模拟时使用的真实总体比例不可能是 $p = 0.88$?

E

答案：就像一些观察结果自然而然地出现在超过平均值 1.96 个标准差以上的地方一样，一些点估计值也会超过 1.96 个标准误。置信区间只是提供一个合理的数值范围。虽然我们可以说，根据数据，其它数值是不可靠的，但这并不意味着它们是不可能的。

根据点估计值对总体参数构建 95%的置信区间

当一个点估计值的分布符合中心极限定理的要求，并因此近似服从正态分布时，我们可以构建一个 95%的置信区间，如下所示：

$$\hat{p} \pm 1.96 \times SE$$

示例 5.8

在第 5.1 节中，我们了解到皮尤研究中心的一项民意调查。在随机抽样的 1000 名美国成年人中，有 88.7%支持对新能源太阳能的扩大使用。请基于这个点估计值，计算并解释一个用于估计总体比例的 95%置信区间。

E

答案：根据上节的内容，我们此前已经确认，$\hat{p}$ 遵循正态分布，其标准误为 $SE_{\hat{p}} = 0.010$。为了计算 95%的置信区间，将点估计值 $\hat{p} = 0.887$ 和标准误代入 95%的置信区间公式。

$$\hat{p} \pm 1.96 \times SE_{\hat{p}} \rightarrow 0.887 \pm 1.96 \times 0.010 \rightarrow (0.8674, 0.9066)$$

因此我们有 95%的信心认为，支持扩大太阳能作用的美国成年人的实际比例在 86.7%和 90.7%之间。(在报告置信区间时，通常会四舍五入到千分位或者万分位，这样如果用百分数形式表示，就是小数点后保留一位或两位)。

5.2.3 改变置信区间

假设我们想要考虑置信度高于 95%的情形，例如置信度为 99%的置信区间。回想一下捕鱼的比喻：如果我们想更有把握捕到鱼，我们应该使用更大的网。相应的，为了达到 99%置信度，我们也必须拓宽 95%的置信区间。而如果我们想要一个更低的置信水平，如达到 90%的置信度，我们可以使用比 95%稍窄的置信区间。

95%置信区间结构为如何构建不同置信度的区间提供了指导。对于遵循正态分布的点估计值，一般的 95%置信区间是：

$$\hat{p} \pm 1.96 \times SE$$

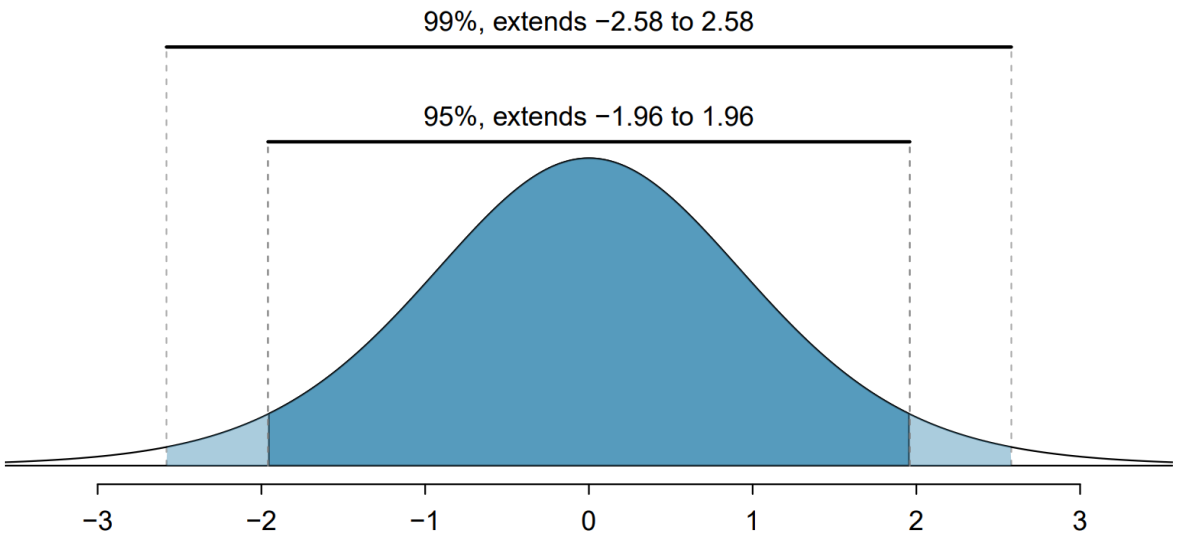
该置信区间由三部分组成：点估计值、「1.96」和标准误。选择 1.96 倍的标准误，是因为按照正态分布对应的概率密度函数，正负 1.96 的 Z 分数的面积占总体分布面积的 95%。这样，样本的点估计值应该有 95%的时间会落在从实际值出发的 1.96 倍的标准误的范围内。或者说，如果取无数次样本，对其计算点估计，并从点估计出发构建 1.96 倍标准误的范围，构建范围的 95%会包括总体参数的实际值，就如图 5.6 展示的那样。倍数 1.96 的选择是由 0.95 对应的 Z 分数得来，对应了 95%的置信度。

指导练习 5.9

G 如果 X 是一个服从正态分布的随机变量，那么 X 的一次随机取值在其均值的 2.58 个标准差之内的概率是多少？[1]

指导练习 5.9 强调，服从正态分布的随机变量的取值，99%的情况下会落在平均值的 2.58 个标准差之内。为了构建一个 99%的置信区间，可以将 95%置信区间公式中的 1.96 改为 2.58。也就是说，99%置信区间的公式是：

$$\hat{p} \pm 2.58 \times SE$$



*图注：99%对应了从-2.58 到 2.58；95%对应了从-1.96 到 1.96

图 5.7：当 z^* 变大时，$-z^*$ 和 z^* 之间的区域随之增加。如果需要定义 99%的置信区间，我们就需要选择了一个恰当的 Z 分数，使得该正态分布的 99%的面积都处在$-z^*$ 和 z^* 之间，这相当于要留出 0.5%的面积在左尾，0.5%的面积在右尾：通过借助标准正态分布的 Z 分数表格，或者一些统计工具，可以计算出符合 99%面积，或者说 99%置信度的 Z 分数为：$z^* = 2.58$。

[1] 这相当于询问 Z 分数有多少可能会大于-2.58 但小于 2.58，图片请见图 5.7。为了确定这个概率，我们可以用统计软件、计算器或表格来查询正态分布的-2.58 和 2.58 概率值：0.0049 和 0.9951。因此，有 0.9951 - 0.0049 ≈ 0.99 的可能，确定一个未观测到的正态随机变量 X 将在 μ 的 2.58 个标准差内。

这种使用 Z 分数来计算置信区间的方法，在点估计值近似服从正态分布时是合适的。对于其它某些点估计值，正态模型并不适合用于描述其分布；在这样的情况下，我们将使用更有代表性的其它分布以及对应的统计分数来构建置信区间。

使用任何置信度的置信区间

如果一个点估计值近似服从标准误为 SE 的正态模型，那么，当 z^* 对应了所选置信度的 Z 分数时，对某点估计值构建（用于估计总体参数的）：

$$\hat{p} \pm z^* \times SE$$

图 5.7 展示了如何根据置信度来确定 z^*。其代表的原理是，我们需要选择合适的 z^*，以便在标准正态分布 $N(0,1)$ 中，$-z^*$ 和 z^* 之间的区域面积占曲线下总面积比例与置信度相对应。

误差范围

在置信区间的概念中，$z^* \times SE$ 术语上被称为**误差范围 margin of error**。

示例 5.10

使用示例 5.8 中的数据，为支持扩大使用太阳能的美国成年人的比例估计量构建一个 90%的置信区间。注意在上一节中，我们已经完成了正态分布条件的二步检验。

E

答案：我们首先找到对应的 z^*，使 90%的标准正态分布面积落在 $-z^*$ 和 z^* 之间。我们可以使用图形计算器、统计软件或概率表格，寻找 5%的右尾（其它 5%在左尾），或者数值 0.90 对应的 Z 分数：z^* = 1.65。据此，可以计算出 90%的置信区间为：

$$\hat{p} \pm 1.6449 \times SE_{\hat{p}} \rightarrow 0.887 \pm 1.65 \times 0.010 \rightarrow (0.8705, 0.9034)$$

也就是说，我们有 90%的信心认为，支持在 2018 年扩大太阳能发电的美国成年人占所有成年人的比例在 87.1%到 90.3%之间。

单样本比例的置信区间计算流程

对于单样本比例的置信区间构建来说，可以分成以下四个步骤：

准备 Prepare：确定样本估计量 $\hat{p}$ 和样本大小 n，并确定你希望使用的置信度。

检查 Check：验证条件以确保 $\hat{p}$ 近似服从正态分布。对于单样本比例置信区间，用 $\hat{p}$ 代替 p 来检查成功-失败的检验。

计算 Calculate：如果检验通过，用 $\hat{p}$ 计算标准误，根据所选的置信度找到 z^*，并按照本节中介绍的公式构建置信区间。

总结 Conclude：在问题的背景下解释置信区间。

5.2.4 更多案例研究

2014 年 10 月 23 日的纽约，一名当时在几内亚治疗埃博拉患者的医生因轻微发烧前往医院，随后被诊断出患有埃博拉。此后不久，NBC 纽约四台/《华尔街日报》/马里斯特民意 Marist Poll 调查发现，82%的纽约人赞成 「对任何接触过埃博拉患者的人进行为期 21 天的强制隔离」。这项民意调查包括在 2014 年 10 月 26 日和 28 日之间，1042 名纽约成年人的答复。

示例 5.11

在这种情况下，点估计值是什么？用正态分布来模拟这个点估计值是否合理?

E

答案：基于 n = 1042 大小的样本，点估计值为 $\hat{p}$ = 0.82。为了检查 $\hat{p}$ 是否可以用正态分布进行合理建模，我们进行了独立性检验（调查基于简单随机抽样和成功-失败检验（$1042 \times \hat{p} \approx 854$ 和 $1042 \times (1-\hat{p}) \approx 188$，都很容易大于 10)。在满足这些条件的情况下，我们确信，$\hat{p}$ 的抽样分布可以使用正态分布来合理地建模。

示例 5.12

从民意调查中估计 $\hat{p}$ = 0.82 的标准误。

E

答案：我们使用 $p \approx \hat{p}$ = 0.82 的替代近似值来计算标准误。

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{0.82(1-0.82)}{1042}} = 0.012$$

示例 5.13

构建 p 的 95%置信区间，即支持对接触过埃博拉患者的人进行隔离的纽约成年人的比例。

E

答案：使用示例 5.12 中的标准误 SE = 0.012，点估计值 0.82，z^* = 1.96，置信度为 95%，置信区间为：

$$\text{点估计值} \pm z^* \times \text{标准误} \rightarrow 0.82 \pm 1.96 \times 0.012 \rightarrow (0.796, 0.844)$$

因此我们有 95%的信心认为，在 2014 年 10 月，支持对任何接触过埃博拉患者的人进行隔离的纽约成年人的比例在 0.796 和 0.844 之间。

指导练习 5.14

G

请回答以下两个关于示例 5.13 的置信区间的问题：

(a) 在这种情况下，95%的信心是什么意思?

(b) 你认为这个置信区间对如今纽约人的意见是否仍然有效? [1]

[1] (a) 如果我们选取许多这样的样本，并为每个样本计算出 95%的置信区间，那么这些区间中将有大约 95%的纽约成年人支持对任何接触过埃博拉患者的人进行隔离。(b) 不一定。该民意调查是在重要公共安全议题出现的时候进行的。现在距离那时已经有些时间，人们可能已经改变了观点。如果我们想得到目前的比例，我们需要进行一次新的民意调查。

指导练习 5.15

G

在上文所述的皮尤研究中心关于太阳能的民意调查中，研究者也询问了针对其他能源类型选择意向，在 1000 名受访者中，84.8%的人支持扩大使用风力涡轮机：

（a）使用正态分布来模拟支持扩大风力涡轮机的美国成年人的比例是否合理?

（b）为支持扩大使用风力涡轮机发电的美国人建立一个 99%的置信区间。[1]

我们也可以为其它参数构建置信区间，比如说总体平均数。在这些情况下，置信区间的计算方法与单样本比例的计算方法类似：一个点估计值加上/减去一些误差范围。我们将在后面深入探讨这些细节。

5.2.5 解释置信区间

在上述每个例子中，我们都把置信区间放在数据背景下描述，同时也使用统计学语言进行解释：

太阳能 Solar：我们有 90%的信心认为，在 2018 年，87.1%至 90.4%的美国成年人支持扩大太阳能发电。

埃博拉病毒 Ebola：我们有 95%的信心认为，在 2014 年 10 月，支持对任何接触过埃博拉患者的人进行隔离的纽约成年人的比例在 0.796 和 0.844 之间。

风力涡轮机 Wind Turbine：我们有 99%的信心认为，在 2018 年，81.9%至 87.7%的美国成年人支持扩大使用风力涡轮机。

仔细阅读观察这些术语，你有什么感悟吗?

首先，因为我们做统计研究的目的往往是为了估计总体，所以一定要注意这些陈述总是关于总体参数的：在能源民意调查中考虑「所有」美国成年人，在隔离民意调查中考虑「所有」纽约成年人。

其次，我们也要避免另一个常见错误：把置信区间的置信度和概率等同起来。置信区间中的置信度并不代表「可能性」。所以我们绝不能在划定 95%置信区间范围后，告诉大家「有 95%的概率总体参数落在这里面」。

---

[1] （a）调查是随机抽样，计数都≥10（$1000 \times 0.848 = 848$，$1000 \times 0.152 = 152$），所以满足独立性和成功-失败检验，而且 $\hat{p} = 0.848$，可以使用正态分布建模。（b）指导练习 5.15 证实了$\hat{p}$近似服从正态分布，因此，我们可以使用置信区间公式：点估计值 $\pm\ z^* \times$ 标准误。此时，点估计值 $\hat{p} = 0.848$。对于一个 99%的置信区间而言，$z^* = 2.58$。计算标准误：$SE_{\hat{p}} = \sqrt{\frac{0.848(1-0.848)}{1042}} = 0.0114$。最后，我们计算区间为 $0.848 \pm 2.58 \times 0.014 = (0.8186,\ 0.8774)$。同样记得，在每次计算完成都要为区间提供解释：我们有 99%的信心认为，支持在 2018 年扩大使用风力涡轮机的美国成年人的比例在 81.9%和 87.7%之间。

置信度只是量化了参数在给定区间内的信心程度，是一个带有一定主观表述色彩，为了增强表述信心的指标，而绝非一个客观概率。尽管也有人提出把置信度当做概率解释也有一定意义，但在本书涉及的经典统计范畴中，需要明确这种表述是不正确的。我们可以做一个简单的思维试验：假设某真实参数为 0.5，而通过统计得出的估计值和对应的置信区间范围分为为 0.7 和 0.6 到 0.8，那么实际情况就是总体参数 100%不在置信区间中，而不能说有 95%的概率总体参数在置信区间中。

关于置信区间的另一个重要考虑是，它们只涉及总体参数。置信区间无法说明数据中单个观测值的具体情况或者由单个观测值计算出的点估计值的情况，置信区间只能提供总体参数的合理范围。

最后，请记住，我们讨论的方法只适用于抽样误差，而不适用于偏差（关于这两项定义可以参考第 5.1.1 小节）。如果一个数据集的收集方式倾向于系统性低估（或高估）总体参数，我们上述讨论的内容将无法解决这个问题。我们所考虑的例子是依靠谨慎的数据收集程序，帮助我们防止偏差，这也是数据科学家们的常见做法。

指导练习 5.16

G

太阳能调查的 90%置信区间是从 87.1%到 90.4%。如果我们再次进行调查，能说我们有 90%的信心，新调查的比例也将在 87.1%和 90.4%之间吗？[1]

[1] 不能。置信区间只提供该参数的合理范围，而非关于未来的点估计值。

5.3 关于正确率的假设检验

下面的问题来自于汉斯-罗斯林 Hans Rosling、安娜-罗斯林-罗恩隆德 Anna Rosling Roonnlund，和奥拉-罗斯林 Ola Rosling 所写的一本书，名为《事实》。

今天世界上有多少 1 岁的儿童已经至少接种了一种疫苗：

a. 20%

b. 50%

c. 80%

写下你的答案（或你的猜测），正确答案见脚注。[1]

5.3.1 假设检验的知识框架

我们想知道人们对世界卫生和世界发展情况的了解程度。换个说法，假设让所有人都去做上面这道选择题，我们想知道以下两个判断哪个更符合实际：

H_0：无论学识，世界卫生对于人们来说都太陌生了，所以所有人的回答只是基于随机的猜测。

H_A：掌握一定知识的人们（体现在学历上）相比于低学历人群会有不同的表现。比如说，拥有的知识可以帮助他们比随机猜测回答得更好，或者也有可能，他们在学习中形成了一些错误的刻板印象，导致他们还不如随机猜测回答的正确率高。

这些相斥的想法在统计学中被称为**假设 hypothesis**。我们称 H_0 为原假设，H_A 为备择假设。当有像 H_0 中的下标 0 时，数据科学家将其读作 「nought」，中文听起来像「闹特」，H_0 也就读作「H-nought」。

原假设和备择假设

原假设 null hypothesis（H_0）通常代表一种怀疑的观点或一种要挑战的主张。

备择假设 alternative hypothesis（H_A）代表了正在考虑的另一种主张，通常由可能的参数值范围表示。作为数据科学家，我们的工作就是扮演一个怀疑者的角色：在我们相信备择假设之前，我们需要看到强有力的支持证据。否则，我们一般会说「还不能拒绝」原假设。

[1] 正确答案是 c。世界上 80%的 1 岁儿童已经接种了某种疾病的疫苗。

原假设往往代表一种怀疑的立场或「无差异」的观点。在我们的第一个例子中，我们将考虑关于开篇提到的婴儿接种疫苗的问题上，拥有特定知识的人群是否与随机猜测的人表现有所不同。

备择假设一般代表一种新的或更强的观点。在这个例子中在关于婴儿接种疫苗的问题上，判断高学历人群是否会回答得更好这点自然很有趣。如果我们了解到这些拥有更多知识的人的表现甚至比随机猜测的那批人还要差，那也会非常有趣，因为这表明有很多与世卫组织相关的不正确的信息在广为流传，并影响人们的判断。

假设检验框架是一个非常通用的工具，我们经常会不假思索地使用它。请思考，如果一个人提出了一个有点不可思议的主张，我们最初会持怀疑态度。然而，如果有足够的证据支持这一说法，我们就会抛开怀疑，拒绝原假设，从而支持这种主张。这种做法在美国法院系统中也可以找到。

指导练习 5.17

G

当被告被美国法院审判时，存在两种可能的说法：无辜或有罪。若放入假设检验的框架里，哪个是原假设，哪个是备择假设？[1]

陪审员应分析是否有足够令人信服的证据表明被告有罪。但请注意，即使没有充分的证据让陪审员做出有罪推定的时候，在话语表述上陪审员也只会说「无法做出被告有罪的推定」，这并不意味着他们就要百分百相信被告是无辜的。把这个逻辑套用过来，假设检验的情况也是如此：即使未能拒绝原假设，我们通常也不会因此接受原假设为真。所以没能找到备择假设的有力证据并不等同于就要直接接受原假设，这才是严谨的统计推断。

在考虑开篇处关于婴儿接种疫苗的问题时，原假设代表着我们的实验对象--受过大学教育的成年人，我们需要分析他们的回答是否和未受过大学教育的人有着相似的准确率。在这种情况下，我们假设受过大学教育的受访者选对正确答案的比例为 p，如果知识无法帮助人在这个问题上更好判断，那么大家应该都是蒙出来的答案，正确率 p 的值也就应当约为 33.3% (三个选项，蒙对的概率)。那么 p = 33.3%就是原假设。对应的备择假设也就是，受过大学教育的群体中，正确回答率不等于33.3%。我们用数学符号来表达上面的两个假设也可以很清晰：

$$H_0: \ p = 0.333$$
$$H_A: \ p \neq 0.333$$

在这个假设检验中，我们想对参数 p 做一个统计学定义。我们所比较的参数值原假设下的取值被称为**原取值 null value**，在本例中是 0.333。在书面标记上，常见的原取值标记做法是用同一个字母加上下标「0」来表示。也就是说，我们可以把原取值记作 $p_0 = 0.333$（读作「p-nought 等于

[1] 陪审团应分析证据是否足够令人信服，以至于可以充分对该人的罪行产生合理的怀疑；在这种情况下，陪审团拒绝无罪推定（原假设），并得出被告有罪的结论（备择假设）。

0.333」)。

示例 5.18

看看这个原假设：受过大学教育的人群，在做这道题目的时候，正好有 33.3%的答对。这听起来似乎是不可能的。 那么如果我们觉得一个原假设听起来不太可能，我们能否干脆不依赖统计推断，草率地拒绝它?

答案：不能，注意这是一本统计学教材！我们的判断需要有统计推断依据，而不能「凭感觉」。虽然我们可能不相信这个比例正好是 33.3%，但假设检验框架要求我们必须要有强有力的证据才能拒绝原假设，这样才能进一步得出一些更有趣的结论。

毕竟，即使我们不相信这个比例正好是 33.3%，这种直观上的怀疑也没有真正告诉我们什么有用的东西。也就是说，我们仍然会停留在原来的问题上：有特定知识的人群对婴儿疫苗问题的回答会更准确还是更不准确？就算和别人分享关于这个问题的观点时，也只能说一句「是靠感觉得出的结论」。这显然没有说服力，也无法推动学科的发展和进步。因此，如果不依赖数据有力地回答，而只是凭感觉去拒绝或者接受猜测，这样做是是毫无意义的。

指导练习 5.19

现实世界中假设检验的另一个例子是评估一种新药在治疗某种特定疾病方面的效用是否比现有药物更好。在这种情况下，我们应该用什么来表示原假设和备择假设？[1]

5.3.2 假设检验的置信区间

我们将使用罗斯林回答（rosling_response）数据集来评估假设检验，分析受过大学教育的成年人在关于婴儿疫苗接种的问题上的正确率是否与 33.3%不同。rosling_response 数据集总结了 50 名受过大学教育的成年人的答案。在这 50 名成年人中，有 24%的受访者回答正确，即选择 80%的 1 岁儿童已经至少接种了一种疾病的疫苗这一问题。

我们之前的讨论还一直停留在理论层面。然而，现在我们有了数据之后，我们可能会问自己：数据是否提供了强有力的证据，能够说明受过大学教育的成年人正确回答这个问题的比例与 33.3%不同？我们在第 5.1 节中了解到，不同的样本之间存在误差，我们的样本概率 $\hat{p}$ 不太可能完全等于 p，但我们又想通过 $\hat{p}$ 对 p 做出一个结论。这时我们难免会担心：这个 33.3%的 24%的偏差仅仅是偶然一次抽样造成的（即仅存在于该数据集中），还是数据已经揭露了强有力的真相，表明受过大学教育真实情况与 33.3%不同?

[1] 在这种情况下，原假设（H_0）是指两者没有差异，即药物的效果相同。备择假设（H_A）是指新药的药效与现有药物不同，即它的药效可能更好或更差。这样设计会符合我们之前提到的：把有趣的发现和推断放在备择假设中，而把值得怀疑的基础判断放在原假设中。

在第 5.2 节中，我们学习了如何用置信区间来量化我们估计的不确定性，这种方法同样可以用于假设检验。

示例 5.20

使用样本数据构建 p 的置信区间合理吗？如果合理，请构建一个 95%的置信区间。

答案：首先保证 $\hat{p}$ 满足近似正态的条件：数据来自简单的随机样本（满足独立性），$n\hat{p} = 12$ 和 $n(1-\hat{p}) = 38$ 都至少是 10，即为通过成功-失败检验。

为了构建置信区间，我们需要计算点估计值 $\hat{p}$ = 0.24 的 95%置信度的 **Z 分数临界值 critical value**（根据前面的内容，可以知道应为 1.96），以及 $\hat{p}$ 的标准差（$SE_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n} = 0.060$）。有了这些，就可以构建 p 的置信区间：

E

$$\hat{p} \pm z^* \times SE_{\hat{p}}$$

$$0.24 \pm 1.96 \times 0.060$$

$$(0.122, 0.358)$$

我们有 95%的把握，在所有受过大学教育的成年人中，正确回答这个关于婴儿疫苗接种的问题的人数比例在 12.2%到 35.8%之间。

因为假设检验中的原假设的值是 p_0 = 0.333，属于置信区间的可信值范围，所以我们不能说原假设的值是在观测到的样本数据背景下完全取不到的，或者说，是不能说原假设的值是完全不合理的。[1]也就是说，数据没有提供足够的证据让我们拒绝「受过大学教育的成年人的表现与随机猜测相同」的说法。根据已有的数据统计和推断信息，我们不能拒绝原假设，即不能拒绝 H_0。

示例 5.21

解释一下为什么我们不能得出如下结论：受过大学教育的成年人在婴儿疫苗接种问题上也是仅凭随机猜测的。

E

答案：虽然我们未能拒绝 H_0，但这并不一定意味着原效假设是真的。也许是否受过大学教育确实会影响题目回答的正确率，但我们无法在相对较小的 50 个样本以及未经严密设计的统计观测中得出结论。

双重否定句有时可用于统计学中

在许多统计学解释中，我们使用双重否定词。例如，我们可能会说，原假设并非不可信，或者我们未能拒绝原假设。双重否定句可以用来表示虽然我们没有拒绝一个立场，但我们也没有说它是正确的。

---

[1] 这种方法是否足够精确还存在一定争议。正如我们将在几页之后所看到的，在对比例进行假设检验的情况下，标准误的计算方法往往和这里提到的会略有不同。

指导练习 5.22

让我们看看和开篇问题相似的的第二个问题：

今天世界上有 20 亿 0-15 岁的儿童，根据联合国的数据，2100 年将有多少儿童？

a. 40 亿

b. 30 亿

c. 20 亿

G

提出适当的假设，来评估受过大学教育的成年人在这个问题上是否回答得比随机猜测要好。另外，在查看脚注中的答案之前，看看你是否能猜出正确答案！[1]

指导练习 5.23

G

这次我们抽取了 228 个受过大学教育的成年人作为样本来回答指导练习 5.22 里的问题，其中 34 人（14.9%）选择了正确答案：20 亿。我们能否用正态分布建立样本比例模型，并构建一个置信区间吗？[2]

示例 5.24

计算「受过大学教育的成年人中正确回答上述问题的比例」的 95%的置信区间，分析指导练习 5.22 中的假设。

答案：为了计算标准误，我们将再次使用 $\hat{p}$ 来代替 p 进行计算。

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.149(1-0.149)}{228}} = 0.024$$

E

在指导练习 5.23 中，我们发现可以用正态分布来模拟 p，这就保证了 95%的置信区间可以被准确地构建为：

$$\hat{p} \pm z^* \times SE \rightarrow 0.149 \pm 1.96 \times 0.024 \rightarrow (0.103, 0.195)$$

因为原假设的值是 $p_0 = 0.333$，它不在置信区间内，所以按照统计推断的置信原理，我们可以据此判断 0.333 的这个假设是不可靠的，我们选择拒绝原假设。也就是说，数据提供了有统计学意义的证据，这表明在有大学学历的成年人中，答对上述问题的人数比例与 33.3%不同。因为整个 95%的置信区间都低于 0.333，我们可以得出结论，受过大学教育的成年人在这个问题上的主观表现比随机猜测的要差一些。

---

[1] 适当的假设是：
H_0：得到正确答案的比例与随机猜测相同。1/3，或 $p = 0.333$。
H_A：得到正确答案的比例与随机猜测不同，$p \neq 0.333$。
该问题的正确答案是 20 亿。虽然世界人口预计会增加，但平均年龄也会上升。也就是说，大部分的人口增长将发生在老年人群体中，这意味着预计未来世界大部分地区的人们将活得更长。

[2] 我们还是首先检查两个正态分布要求的条件，它们都得到了满足，所以对 $\hat{p}$ 使用正态分布构建模型是合理的：独立性 >> 由于数据来自于一个简单随机抽样，所以观测结果是独立的。成功-失败检验 >> 我们用 $\hat{p}$ 代替 p 来分析：$n\hat{p} = 34$, $n(1-\hat{p}) = 194$。两者都大于 10，所以成功-失败条件得到满足。

关于上个示例中现象出现的原因，很有可能是因为如果没有大学知识背景，不经思考的随机猜测还有一定可能会让人们选择 20 亿这个正确答案。而在接受高等教育的过程中，人们接触到了例如「世界人口不断膨胀」,「新生儿死亡率越来越低」之类的信息，导致人们更倾向于相信未来 0-15 岁年龄区间内儿童的数量会比现在更多。带着这种倾向，自然也就更难选对正确答案。

一个奇妙的想法是，我们当前使用的是 95%的置信区间。如果我们使用的是 99%的置信度呢？或者甚至是 99.9%的置信度？如果使用不同的置信度，就有可能得出不同的结论。因此，当我们根据置信区间做出结论时，我们也应该确保清楚地知道我们选择的是什么置信水平。

事实上，受过大学教育的人们的主观表现比客观猜测要差并非偶然：在世界健康领域问题上，无论受过大学教育与否，只要是让人们主观去判断，往往都比纯随机猜测的正确率要低。总的来说，人们对世界发展的看法往往容易比现实更悲观。这个话题在本章开头介绍的书籍《事实》中讨论得更为详细。

5.3.3 决策误差

我们刚刚介绍的假设检验并不是万能的：我们也会在基于数据的统计假设检验中做出错误的决定。正如在法院系统中，无辜的人有时会被错误地定罪，而真正的罪犯有时会逍遥法外。与之相比，统计假设检验的一个关键区别是，我们可以利用必要的工具来分析概率，量化我们在结论中犯错误的频率，从而让推断结果尽可能向真相靠拢。

回顾一下前文，我们介绍了两个相互矛盾的假设：原假设和备择假设。在假设检验中，我们会选择拒绝一个假设而倾向于另一个假设。但我们也有可能选错了，这就对应了四种可能的情况，在图 5.8 中进行了总结。

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	okay	**Type 1 Error**
	H_A true	**Type 2 Error**	okay

上方注释：　　不拒绝 H_0　　拒绝 H_0 转而选择 H_A

图 5.8：假设检验可能出现的四种情形

第一类错误 Type 1 Error（又称 I 类错误）是指在 H_0 实际为真时拒绝了原假设。**第二类错误 Type 2 Error**（又称 II 类错误）是在备择假设实际为真时未能拒绝原假设。

指导练习 5.25

G 在美国法庭上，被告要么无罪（H_0），要么有罪（H_A）。在这种情况下，第一类错误代表什么？第二类错误代表什么？图 5.8 可以参考。[1]

示例 5.26

我们如何能减少美国法院犯第一类错误的概率？这对犯第二类错误的概率会有什么影响？

E 答案：为了降低犯第一类错误的概率，我们可以将我们的定罪的标准提高，只要案件有一点点不确定性，就不予定罪。这样宽松的治理毋庸置疑可以减少被错误定罪的人。然而需要注意的是，宽松的治理也会让罪犯更容易逃脱法律的制裁，即会使真正有罪的人更难被定罪，也就意味着更多的第二类错误。

指导练习 5.27

G 我们该如何减少美国法院的第二类错误率？降低犯第二类错误的概率会对犯第一类错误的概率有什么影响？[2]

指导练习 5.25 到指导练习 5.27 提供了很好的例子：通常情况下，如果我们减少犯一种类型的错误的频率，那么另一种类型的错误率会上升。假设检验是围绕着拒绝或不拒绝原假设而进行的。也就是说，除非我们有强有力的证据，否则我们不会拒绝 H_0。但这里所谓的强有力的证据究竟意味着什么？通常来说，对于实际为真的原假设，我们不希望错误地拒绝 H_0 的概率超过 5%。即对应的**显著性水平 significance level** 为 0.05。（显著性水平往往是人为的共识达成的，比如在多数社会科学学科中，大多默认为 0.05。但某些自然学科中，有时会要求更为严格，定为 0.001。）也就是说，如果原假设为真，显著性水平表示我们错误地拒绝 H_0 的频率。我们通常用 α（希腊字母 a）来表示显著性水平。往往记作：$\alpha = 0.05$. 我们将在第 5.3.5 小节中讨论正确使用不同显著性水平的具体内容。

如果我们使用 95%的置信区间来评估一个假设检验，而原假设恰好为真，那么只要点估计值与人口参数的距离相差大于 1.96 个标准差，那就我们就会犯错。（考虑到原假设实际是真的，正确的判断应该是不拒绝原假设，但由于此次抽样结果显示出我们应该拒绝原假设，我们在这里就出现的判断失误）。这种出现错误的比例大概是 5%（两边尾部各有 2.5%）。同样地，使用 99%的置信区间来评估一个假设，相当于$\alpha = 0.01$ 的显著性水平。置信区间可以帮助我们去决定是否应该拒绝原假设。然而，置信区间的方法并不是万能的。在一些章节中，我们将遇到无法构建置信区间的情况。例如，如果我们想评估一个假设，且该假设里面多个部分的比例都是均等的，我们将无法构建和比较很多置信区间。接下来，我们将介绍一种叫做 p值 的统计量，这将使我们能够更好地理解所谓证

[1] 如果法院出现了第一类错误，这意味着真相就是被告是无辜的（H_0 为真），但被错误地定罪。请注意，只有在我们拒绝了原假设的情况下，才可能出现第一类错误。第二类错误意味着法院未能拒绝 H_0（即未能将此人定罪），而她实际上是有罪的（H_A 为真）。请注意，只有在我们未能拒绝原假设的情况下，才可能出现第二类错误。

[2] 为了降低第二类错误率，我们要给更多的违法者定罪。，因此我们可以将定罪的标准降低。比如，把定罪标准从「排除一切合理怀疑」到「排除一些合理怀疑」。然而降低定罪标准也会导致更多的冤案发生，即第一类错误发生的可能性增加。

据的强度，并让我们能够在后面的章节中处理更复杂的数据。

5.3.4 使用 p 值进行更正式的假设检验

p 值是一种量化拒绝原假设和支持备选假设的证据强度的方法。统计学假设检验通常是使用 p 值，而不是单纯根据置信区间来做决定。

p 值

之前提到的 **p 值** **p-value** 就是当原假设为真时，出现比得到的样本观察结果更极端的结果出现的概率。我们通常使用数据统计量来总结，在本节中该统计量为样本比例，以帮助计算 p 值和评估假设检验。

示例 5.28

皮尤研究公司 Pew Research 随机抽样询问了 1000 名美国成年人是否支持增加使用煤炭投入来生产能源。他们构造假设检验，试图去分析大多数美国成年人是否支持或拒绝增加煤炭的投入。

E 答案：然而实际结果是，两者都不占多数：一半的美国人支持，另一半反对扩大使用煤炭来生产能源。备择假设是，有大多数人支持或反对（尽管我们不知道是哪一个，但是会有大多数人有一种倾向！）扩大煤炭的使用。如果 p 代表支持的比例，那么我们可以把假设写成：

$$H_0: p = 0.5 \qquad H_A: p \neq 0.5$$

本例中，原假设是 $p_0 = 0.5$

检查成功-失败检验和在假设检验中计算 $SE_{\hat{p}}$ 的值

当使用 p 值评估假设检验时，我们检查 $\hat{p}$ 是否满足独立性以及能否通过成功-失败检验，但在之后我们使用原取值 p_0 构建标准误，而不是像之前那样使用点估计值带入公式。

关于带入原取值的解释：在有 p 值的假设检验中，我们假设原假设为真，这与我们计算置信区间时的心态不同。这就是为什么我们在检查条件时使用 p_0 而不是 $\hat{p}$。

当我们确定原假设下的抽样分布时，它有一个特殊的专有名字：原假设分布。原假设的分布构建与抽样均值 p 的分布构建逻辑一直，但它是围绕原取值 p_0 而建立的。也就是说，原假设分布是建立在原假设为真的情况下的一个近似正态分布模型。而 p 值就是在这样一个分布中，也就是假设原假设为真的前提下，观察到的 $\hat{p}$ 或更极端的 $\hat{p}$ 的概率。我们一般会利用原假设分布找到与我们的点估计 $\hat{p}$ 相对应的尾部区域，进而找到 p 值。

示例 5.29

皮尤研究的样本显示，37%的美国成年人支持增加煤炭的投入。我们现在想知道，这里的 37%是否表示与原假设的 50%存在显著差异？如果原假设为真的话，$\hat{p}$ 的抽样分布会是什么样子？

答案：**独立性** 该民意调查是基于简单的随机抽样，所以独立性得到了满足。

成功-失败检验 根据投票的样本大小 n=1000，成功-失败检验的条件得到满足，因为：

$$np \overset{H_0}{=} 1000 \times 0.5 = 500 \qquad n(1-p) \overset{H_0}{=} 1000 \times (1-0.5) = 500$$

E 二者均大于 10，请注意，成功-失败检验的条件是用原取值来检查的，即要用 p_0 = 0.5 带入；这是与置信区间检验法的第一个明显差异。

如果原假设为真，抽样分布表明，n=1000 个观测值的样本比例将服从正态分布。接下来，我们可以计算标准差，在计算中我们将再次使用原取值 p_0 = 0.5。

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \overset{H_0}{=} \sqrt{\frac{0.5 \times (1-0.5)}{1000}} = 0.016$$

这是与置信区间的另一个区别：由于抽样分布是在原比例下确定的，所以在计算比例的时候使用了原取值 p_0，而不是 $\hat{p}$。最终，如果原假设为真，那么样本比例应该遵循正态分布，平均值为 0.5，标准差为 0.016。这个分布如图 5.9 所示。

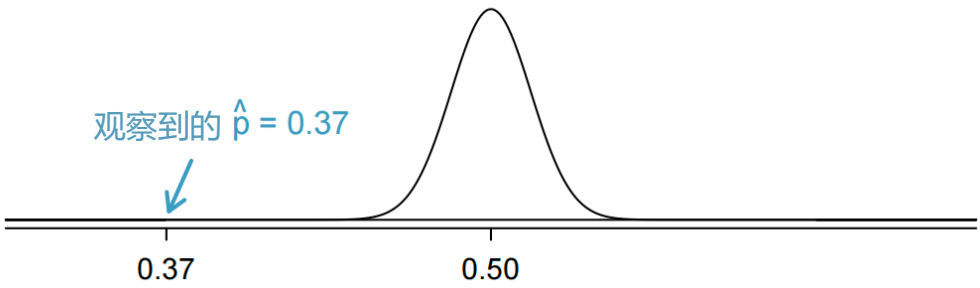


图 5.9：如果原假设为真，这个正态分布描述了 $\hat{p}$ 的分布。

当使用 p 值法评估比例假设检验时，我们将稍微修改我们检查成功-失败条件的方式，并计算单一比例情况下的标准差。这些改动并不会特别大，但要注意我们使用原取值 p_0 的方式。

示例 5.30

如果原假设为真，确定在原假设分布下找到 $\hat{p}$ 至少与尾部距离为 0.37 的概率，该分布是具有均值 $\mu = 0.5$ 和标准误 $SE = 0.016$ 的正态分布。

答案：这是一个正态分布的概率问题，其中 $x = 0.37$。首先，我们绘制一个简单的图形来表示这种情况，类似于图 5.9。由于 $\hat{p}$ 的尾部距离很远，因此尾部区域将非常小。我们计划从使用均值 0.5 和标准误 0.016 计算 Z 分数开始计算：

E

$$Z = \frac{0.37 - 0.5}{0.016} = -8.125$$

我们可以使用软件找到尾部区域：2.2×10^{-16}（0.00000000000000022）。如果查询附录 C.1 中的正态分布概率表，我们会发现 Z = -8.125 超出了范围，因此我们将使用最小的区域：0.0002。在图 5.10 中显示的超过 0.63 的上尾部分的潜在 $\hat{p}$ 也代表了至少与观察值 0.37 一样极端的观测值。考虑到在假设环境下更极端的数值，我们将下尾部翻倍以得到 p 值的估计：4.4×10^{-16}（或者如果使用查表法，为 0.0004）。

这里的 p 值就代表了如果原假设为真的话，我们只有非常小的概率在一次随机抽样中观察到如此极端的情况。

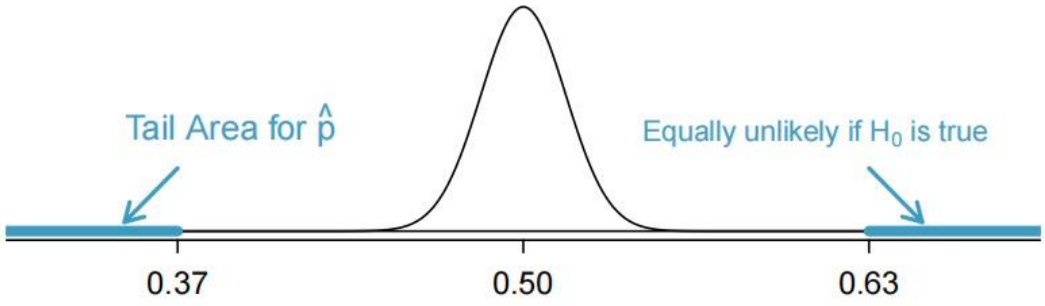


图 5.9：如果原假设为真，则通过一次随机抽样，观察到大于 0.63 的值和观察到小于 0.37 的值的概率是相同的，即两者同样不太可能发生。

示例 5.31

我们应该如何评估 p 值为 4.4×10^{-16} 的假设？在评估中请使用 $\alpha = 0.05$ 的显著性水平。

答案：如果原假设为真，观察到 $\hat{p}$ 与 0.5 之间的极端偏差的概率非常小。这意味着以下两种情况之一必须成立：

E

（1）原假设为真，我们只是碰巧观察到如此极端的情况，这种情况仅在每 23 万亿次观测中才会发生一次（1 万亿 = 1 百万 × 10 亿）；

（2）备择假设为真，这将让我们观察到的「一次随机抽样的样本比例远离 0.5」更容易发生。

第一种情况几乎不可能发生，而第二种情况似乎更加合理。正式地说，当我们评估假设检验时，我们将 p 值与显著性水平进行比较，而在这种情况下显著性水平是 $\alpha = 0.05$。由于 p 值小于 α，因此我们拒绝原假设。也就是说，数据提供了强有力的证据反对 H_0。数据表明差异的方向：大多数美国人不支持扩大使用燃煤能源。

将 p 值与 α 进行比较以评估原假设 H_0

当 p 值小于显著性水平 α 时，选择拒绝 H_0。我们会得出结论：数据提供了支持备择假设的有力证据。 当 p 值大于 α 时，则选择不拒绝 H_0，因为我们没有足够的证据来拒绝原假设。在任何情况下，基于数据去做结论都是很重要的。

G

指导练习 5.32

大多数美国人支持还是反对核武器裁减？请建立一个可用于评估此问题的假设。[1]

示例 5.33

在 2013 年 3 月，对 1028 名美国成年人的简单随机抽样结果显示，56%的人支持核武器裁军。这是否在 5%的显著性水平上提供了令人信服的证据，表明大多数美国人支持核武器裁军?

答案：首先，检查一些条件：

独立性 该调查是对美国成年人进行的简单随机抽样，这意味着观察结果是独立的。

成功-失败条件 在单比例假设检验中，这个条件是依据原假设比例来确定的，在这种情况下，原假设比例为 p_0 = 0.5：$np_0 = n(1-p_0) = 1028 \times 0.5 = 514 \geq 10$。

得到了这些条件后，我们可以使用正态模型来计算 $\hat{p}$。SE

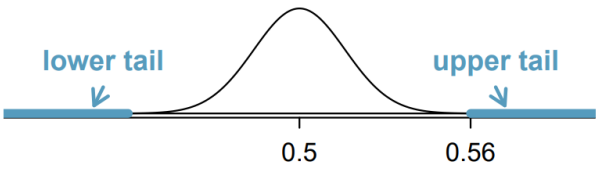
$$SE_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{1028}} = 0.0156$$

E

接下来，可以计算标准误。在这里再次使用原假设值 p_0，因为这是一个单比例假设检验。

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.56 - 0.50}{0.0156} = 3.85$$

一般来说，画出原假设分布和尾部区域对计算 p 值很有帮助。



上尾部面积约为 0.0001，我们将这个尾部面积加倍，得到 p 值为 0.0002。由于 p 值小于 0.05，我们拒绝 H_0。这项调查提供了有力的证据，表明在 2013 年 3 月，大多数美国人支持核减军备。

[1] 我们希望了解大多数美国人是否支持或反对核武器削减，或者最终是否存在差异。如果 p 是支持核武器削减的美国人的比例，则 H_0：$p = 0.50$，H_A：$p \neq 0.50$。

单一比例的假设检验

有四个步骤来完成检验：

准备：确定参数，列出假设，确定显著性水平，并确定 $\hat{p}$ 和 n。

检查：验证条件，确保在 H_0 下 $\hat{p}$ 接近正常。对于单比例假设检验，使用原假设值来检查成功-失败的条件。

计算：如果条件成立，计算标准误，同样使用p_0，计算 Z 分数，并确定 p 值。

结论：通过比较 p 值和 α 来评估假设检验，并结合问题给出结论。

5.3.5 选择置信区间

在许多情况下，为测试选择一个显著性水平是很重要的，传统的水平是 $\alpha = 0.05$。然而，根据应用情况调整显著性水平可能会有帮助。我们可以选择一个小于或大于 0.05 的水平，这取决于从检验中得出的任何结论的后果。

如果犯第一类错误是危险的或者代价特别高,我们应该选择一个较小的显著性水平(例如 0.01)。在这种情况下，我们关于拒绝原假设要很谨慎，因为显著性水平很小，也就意味着在拒绝 H_0 之前，我们需要非常有力的证据，或者说观察到非常非常极端的情况来支持我们放弃原假设转而选择 H_A。

如果第二类错误相对更危险或者代价比第一类错误高得多，那么我们可能会选择一个较高的显著性水平（例如 0.10）。在这种情况下，我们需要的谨慎是由于容忍度的提高，我们也会相对来说更容易犯第一类错误，也就是在实际上备择假设为真时，依然没能拒绝 H_0。

此外，如果相对于第二类错误的代价而言，收集数据的成本较小，那么选择较高的显著性水平同时伴随收集更多的数据也可能是一个好策略。在这种策略下，通过更多大的样本或者多轮次的分析比对，可以减少第二类错误而不影响第一类错误率。当然，收集额外的数据通常是有成本的，因此在决定采用这种策略前，还需要进行充分的成本效益分析。

示例 5.34

一家汽车制造商正在考虑转换到一种新的、质量更高的设备，用于制造汽车门铰链。他们认为，如果这台新机器生产的铰链有缺陷的概率小于 0.2%，从长远来看他们将节省资金。然而，如果铰链的缺陷超过了 0.2%，他们将无法从新设备中获得足够的投资回报，会亏钱。在这种假设检验中，是否有充分的理由修改默认的 0.05 的显著性水平？

E

答案：原假设是铰链有缺陷的概率为 0.2%，而备择假设是铰链的缺陷率与 0.2%不同。根据上下文场景我们发现，该制造商并没有刻意强调「如果新铰链缺陷率恰好是 0.2%但未被采用」带来的负面影响（第一类错误），也没有强调如果新铰链缺陷率不是 0.2%但采用后造成的重大隐患（第二类错误）。这让我们可以暂时认为本次假设检验的结果对应的 I 和 II 类错误[1]对公司影响都不太大。这样一来，选择 0.05 的显著性水平应该是合理的。

示例 5.35

同一家汽车制造商正在考虑为一个安全相关的部件（这次不是门铰链了）更换一个稍贵的供应商。如果相关安全部件的耐用性被证明显著优于当前的供应商，他们就会更换制造商。在该场景下，我们是否有充足的理由去修改显著性水平？

E

答案：首先进行假设定义梳理：原假设是供应商的零件具有相同的可靠性，即不更换新制造商。

因为涉及到安全问题，即使只有强度一般的理由来证明安全性的提升，可能这种提升也是非常必要的。也就是说对于新制造商选择的更换条件应该松一些。汽车公司在管理决策时可能也应该热衷于转向略微昂贵的安全部件制造商，这些都支持我们在预设期望时需要更容易拒绝「不更换」对应的原假设 H_0。体现在统计参数上就是稍大一些的显著性水平会更适用，例如 $\alpha = 0.10$。

指导练习 5.36

G

某机器内部的某零件更换起来非常昂贵。然而，即使该零件损坏了，机器通常也能正常工作，所以只有在我们非常确定零件已经严重损坏的情况下，才会对零件进行更换。请用简单的语言为此次测试确定合适的假设，并建议选择一个适当的显著性水平。[2]

为什么选择 0.05 作为默认显著性水平

$\alpha = 0.05$ 的显著性水平最常见。但这是为什么呢？如果你在认真的思考这个问题并提出质疑，说明你正在以更加批判的眼光来阅读本书，为你鼓掌！为了解释这个问题，我们设计了一个 5 分钟的线上小课堂，用以帮助澄清为什么选择 0.05 作为默认水平：

www.openintro.org/why05

---

[1] 译者注：这是个很好的从直觉上体验理解 I 和 II 类错误概念的例子。一般来说我会喜欢把 I 类错误理解成：「执法太严」，而 II 类错误理解成：「执法太松」。套用这里就是 I 类错误相当于如果执法太严，必须完全按照 0.2%的缺陷率走，看是否有什么危害？而 II 类错误相当于如果执法太松，明明缺陷率不是 0.2%却被我们按照 0.2%来推进决策的背景下，可能带来的潜在风险和隐患。

[2] 这里的原假设是零件没有损坏，而备择假设是零件已损坏。如果我们没有足够的证据来拒绝原假设，我们就不会更换零件。根据描述，如果零件损坏但未修复（原假设为假，备择假设为真）的问题并不大，而更换零件的成本很高。因此，在更换零件之前，我们需要找到对原假设有非常有力的证据。可以选择一个较小的显著性水平，例如 $\alpha = 0.01$。

5.3.6 统计显著性与实际显著性

当样本量变大时，点估计变得更加精确，即观测值与原假设值之间的任何实际差异都更容易被识别出来。如果我们采用足够大的样本，把非常小的差异也检测出来，这有时候会导致那些没有实际价值的差异也被纳入统计结果之中。在这种情况下，我们仍然会说差异在统计上是显著的，但在实际上却不显著。例如，通过线上调查可能会发现，在电影评论网站上投放更多广告在统计学上显著地增加了某电视节目的收视率 0.001%，但这 0.001%的增长可能没有任何实际价值。

在进行研究时，数据科学家通常需要规划研究的规模。数据科学家可以首先咨询专家或查阅科学文献，了解与原假设取值的最小有意义差异是什么，例如电视节目收视率增长 1%有无实际意义，增长10%有无实际意义。数据科学家还会收集其他信息，例如真实比例 p 的一个非常粗略的估计值，以便大致估计标准误。建立在这些信息的基础上，数据科学家可以不用永远追求大样本，而是灵活选择一个「够用」的样本，以便一旦存在实际意义上的差异，样本就能捕捉到它。虽然我们可能还是会倾向于使用较大的样本，但是在考虑成本或潜在风险（如医学研究中志愿者可能的健康影响）时，了解一个「够用」的样本有多大显然是有好处的。

5.3.7 单尾假设检验（专题）

目前为止，我们只考虑了所谓的**双尾假设检验 two-sided hypothesis tests**，即我们关心的是检测 p 是否高于或低于某些原取值 p_0。现在，我们介绍名为**单尾假设检验 one-sided hypothesis test** 的第二种类型假设检验。对于单尾假设检验，我们的假设会采取以下形式之一：

1. 只有在检测总体参数是否「小于」某个值 p_0 时才有价值。在这种情况下，备择假设可以被写成该式：$p < p_0$，原取值为 p_0。
2. 只有在检测总体参数是否「大于」某个值 p_0 时才有价值。在这种情况下，备择假设可以被写成该式：$p > p_0$，原取值为 p_0。

虽然我们调整了备择假设的形式，但在单尾假设检验的情况下，我们继续使用等号来书写原取值。在整个假设检验过程中，评估单尾假设检验与双尾假设检验只有一个区别：如何计算 p 值。在单尾假设检验中，我们将 p 值计算为仅*在备择假设对应的那一侧尾部区域*。例如，在单尾假设检验的情况下，如果 $p = 0$ 是原假设，$p < 0$ 就是备择假设，那么计算 p 值时只需要计算 0 左侧那部分的阴影面积。这就是为什么单尾假设检验有时很吸引研究者：如果我们不需要把尾部面积翻倍来得到 p 值，p 值就会变小，那么在备择假设对应的一侧上想确定一个有趣的发现时，所需要的证据等级也会下降。然而，这种「更容易发现一侧的显著性」的代价是：任何相反方向潜在的显著性发现都势必会被忽略，这也是为什么我们在选择单尾检验前要慎重，因为它并不总意味着「阳光和彩虹」。

示例 5.37

在第 1.1 节的例子中，医生们对于「颅内支架是否能帮助潜在的中风高风险患者」感兴趣。研究人员相信使用支架会有帮助。不幸的是，数据显示情况正好相反，即使用支架的病人情况更糟。结合这个例子和你对单双尾检验的理解，尝试解释为什么采用双尾假设检验是非常重要的?

E

答案：现有研究表明，使用心血管支架对心脏病患者是有帮助。因此，在本研究之前，研究人员很可能会倾向于猜测：使用颅内支架也会对病情有好处。在这样的背景下，选择单尾假设检验肯定是很诱人的；但如果这样做，就会限制研究人员识别支架对病人的潜在伤害。而试验结果也恰恰表明，使用颅内支架反而导致了更高的发病率。

示例 5.37 强调，使用单尾假设检验是有风险的：它会让研究者容易忽视「支持相反结论的数据」。我们重温一下本节最开始的罗斯林问题，彼时也可能犯类似的错误；如果我们有一个先入为主的观念，认为受过大学教育的人的成绩不会比随机猜测的人的差，进而使用了单尾假设检验，我们就会错过真正有趣的发现：即许多受过教育的人对全球公共卫生有不正确的认识。

什么时候使用单尾假设检验是合适的？其实答案是「只有很少情形适用」。如果你发现自己曾经考虑使用单尾假设检验，请在决定前认真思考以下问题：

> 如果数据恰好与我的备择假设走向明显相反，我或其他人会得出什么结论？

如果你或其他人发现：对数据做出「与单尾假设检验相反方向的结论」有任何价值，那么实际上应该选择使用双尾假设检验。有些情形下结论与价值考虑可能是很微妙的，正因如此所以更要谨慎行事。在本书的其余部分，我们将只使用双尾假设检验。

示例 5.38

为什么我们不能简单地进行单尾假设检验，按照数据的方向（比如统计均值小于零，备择假设就选择小于零的方向）去做?

答案：我们一直在建立一个谨慎的框架，控制第一类错误，也就是假设检验中的显著性水平 α。为了计算简便，我们使用 $\alpha = 0.05$。

E

想象一下，我们在看到数据后可以选择单尾假设检验。会出现什么问题呢?

如果 $\hat{p}$ 小于原取值，那么 $p < p_0$ 的单尾假设检验将意味着在原分布中，少于尾部 5%的任何观测都会让我们拒绝 H_0。而如果 $\hat{p}$ 大于原取值，那么 $p > p_0$ 的单尾假设检验将意味着在原始分布中，多于尾部 5%的任何观测都会让我们拒绝 H_0。

那么，如果 H_0 为真，则在观测中我们有 10%的观测情况都会犯下 I 类错误，在原假设为真的时候拒绝它。这导致我们实际上是按照 α =0.10 来进行假设检验的，而不再是 0.05。也就是说，如果没有充分理由，而仅仅根据样本数据来采用单尾假设，实际是破坏了我们努力开发和长期使用的假设检验框架。

第 6 章

基于分类数据的推断

Inference for categorical data

本章中，我们将把第 5 章里讨论的方法和思想用于对分类数据的统计推断之中。我们首先来重温一下之前提到的「单样本比例」的概念，并考虑使用正态分布来描述和计算「使用估计值去推断总体参数」过程中的不确定性。紧接着，我们把这些想法用于分析双样本间的比例差，并依旧在正态分布假设的基础上来进行统计推断。在章节的后半部分，我们会把统计推断技巧应用在列联表中，直接探究如何对分类变量的某个分布状态（以多个比例值形式呈现）进行估计检验。而在列联表相关的复杂问题里，我们可能要摆脱正态分布转而采用新的分布模型。不过无论哪种分布模型，假设检验的核心思路都是不变的：即通过 Z 分数和 p 值来判断是否可以接受或者不得不拒绝原假设。

跨越数据银河

系列推文合集

更多视频，演示文稿，和其他相关资源，请访问：
http://www.openintro.org/os

6.1 单样本比例的推断

我们在第 5 章接触到了单样本比例的推断方法，探讨了点估计、置信区间和假设检验。在本节中，我们将对这些主题做一个回顾，同时也将介绍：在对一个简单的比例进行估计的统计场景下，如何选择合适的样本大小来收集数据。

6.1.1 确认样本比例接近正态分布

当样本观测值相互独立，并且样本大小足够大时，样本比例 $\hat{p}$（即符合某特定条件的个体数除以总个体数）的分布接近正态分布。这是我们统计推断的前提。

比例 $\hat{p}$ 的抽样分布

根据中心极限定理（见第 5 章第 5.1.3 小节）在以下条件满足的前提下，从具有真实比例为 p 的总体中抽取大小为 n 的样本，其样本比例 $\hat{p}$ 的抽样分布接近于正态分布：

1. 样本的观测值是独立的，例如，使用简单随机抽样（见第 1 章第 1.3.3 小节）。
2. 需要在样本中至少观察到 10 次成功和 10 次失败，即符合特定条件的个体数和不符合条件的个体数都大于 10，总样本数量也自然地应大于 20。用数学语言来描述就是 $np \geq 10$ 和 $n(1-p) \geq 10$。这项条件对于单比例估计至关重要，被称为**成功-失败检验 success-failure condition**。

当以上这些条件得满足的时候，根据样本计算出的 $\hat{p}$ 就将近似服从一个均值为 p，标准误为 $SE = \sqrt{\frac{p(1-p)}{n}}$ 的正态分布。

通常情况下，我们不知道真正的比例 p，所以我们用一些观测到的数值来确认条件并估计标准误。对于置信区间的计算也是一样，例如样本比例 $\hat{p}$ 作为一个观测到的数值，往往被用来进行成功-失败检验，也被用于计算标准误，这个计算出的标准误又会进一步被用于计算置信区间。对于本书中已经介绍到的假设检验场景，我们又通常用原假设取值来代替总体参数 p。

6.1.2 单样本比例估计的置信区间

我们在用 $\hat{p}$ 估计 p 的时候，是不能够直接把「样本观察到的估计值」与那个「无从得知的总体实际值」画等号的。这也正是统计推断的严谨之处。因此我们再次引入置信区间的概念，它为我们对参数 p 的估计提供了一个合理的数值范围，也就是我们有一定的信心说最终总体参数 p 应该会落在这个范围内。当 $\hat{p}$ 满足正态分布时，p 的置信区间为：

$$\hat{p} \pm Z^* \times SE$$

示例 6.1

我们对小额贷款 APP 使用者进行了一次简单的随机抽样调查，以更好地了解他们对监管和成本的偏好。在抽选的 826 名使用者中，70%的人支持针对小额贷款机构加强监管。在这种情况下，我们相当于已知一次抽样，样本大小为 826，样本估计的结果是 $\hat{p}$ = 0.70。根据这些信息，我们能继续假设其服从正态分布进而进行统计推断吗?

E

答案：由于数据是随机抽样的，因此观察结果互相独立，同时能够代表我们感兴趣的群体。其次我们还必须确认成功-失败检验的条件被满足，而由于 p 的实际取值未知，所以在下面的计算中我们用 $\hat{p}$ 代替 p。

支持: $np \approx 826 \times 0.70 = 578$　　不支持: $n(1-p) \approx 826 \times (1-0.70) = 248$

由于这两个值都大于 10，我们可以用正态分布来分析 $\hat{p}$。

指导练习 6.2

G

计算 $\hat{p}$ = 0.7 的标准误。因为实际总体 p 是未知的，所以参照本节开头所述，我们用 $\hat{p}$ 来代替 p。[1]

示例 6.3

对 p（即支持加强监管的用户比例）构建一个 95%的置信区间。

答案：使用点估计值 0.7， 95%的置信区间的 Z 分数 1.96，和来自指导练习 6.2 中的标准误 SE = 0.016，得出的置信区间为:

E

$$\text{点估计} \pm Z^* \times SE \rightarrow 0.70 \pm 1.96 \times 0.016 \rightarrow (0.669, 0.731)$$

我们有 95%的把握相信，支持加强监管的用户比例的真实值在 0.669（66.9%）到 0.731（73.1%）之间。

[1] $SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{0.70(1-0.70)}{826}} = 0.016$

单样本比例置信区间的过程

对于单样本比例置信区间的构建来说，可以分成以下四个步骤：

准备 Prepare：确定样本估计量 $\hat{p}$ 和样本大小 n，并确定你希望使用的置信度。

检查 Check：验证条件以确保 $\hat{p}$ 近似服从正态分布。对于单样本比例置信区间的抽样估计，用 $\hat{p}$ 代替 p 来检查成功-失败条件。

计算 Calculate：如果检验通过，用 $\hat{p}$ 计算标准误，根据所选的置信度找到 Z^*，并按照本节中介绍的公式构建置信区间。

总结 Conclude：在相关背景下解释置信区间的实践含义。

该重点框其实在第 5.2 节中就已经出现过，但因为这个过程太重要了所以在这里再次展示一遍以方便诸君加深记忆。

6.1.3 单样本比例的假设检验

一种对小额贷款 APP 的规范化管理办法是：强制他们在提供贷款前必须审查借款人的信用状况，并根据借款人的财务状况评估其偿还能力。这样虽然能让贷款流程更加合规，但无疑会让用户借钱变得麻烦，同时也会要求借款人披露更多的隐私信息。在这个背景下，我们希望探究借款人是否支持这项管理办法。

指导练习 6.4

请提出合适的统计假设，以评估多数借款人对该类型的监管是否支持。[1]

为了在比例的假设检验中使用正态分布框架，涉及的比例估计量必须满足独立性和成功-失败检验的条件。在假设检验中，成功-失败检验是用原假设下的比例来计算判断的，这里就对应了 0.5：我们验证 np_0 和 $n(1-p_0)$ 都必须大于 10，其中 p_0 代表原假设下的 p 的取值，英文对应的词汇是 null value。

指导练习 6.5

在随机抽取的 826 名借款人中，51%的人说他们会支持这样的规定。那么在这里使用正态分布进行假设检验，并建立 $\hat{p} = 0.51$ 的正态分布模型是否合理？[2]

[1] $H_0: p = 0.50$; $H_A: p \neq 0.50$。

[2] 合理：独立性成立，因为投票是基于随机抽样的。成功-失败检验的条件也是成立的，这一点通过空值 $(p_0 = 0.5)$ 来检查：$H_0: np_0 = 826 \times 0.5 = 413, n(1-p_0) = 826 \times 0.5 = 413$。因为两个条件都满足，所以可以使用正态分布模型。

示例 6.6

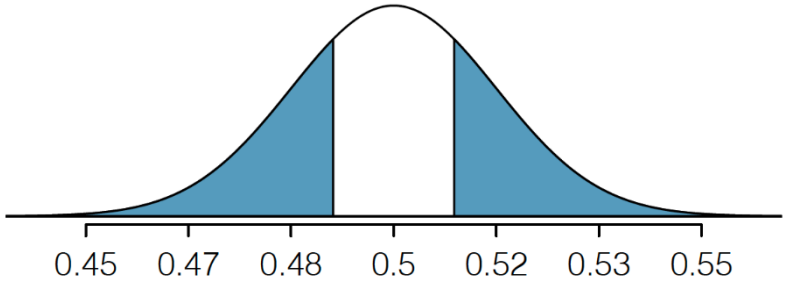
使用指导练习 6.4 和 6.5 中的假设和数据，评估该民意调查是否提供了令人信服的证据，证明大多数借款人支持新的管理办法。

答案：在建立假设和确定通过成功-失败检验后，可以开始进行统计计算。在单比例假设检验的背景下，我们使用原假设 p_0 来计算标准误。

$$SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{0.70(1-0.70)}{826}} = 0.016$$

下面是一张正态模型的图片，p 值由阴影区域表示。

E



基于正态分布模型，我们需要使用点估计的 Z 分数来计算上图阴影区面积，因而要先计算 Z 分数：

$$Z = \frac{\text{点估计} - \text{原假设取值}}{SE} = \frac{0.51 - 0.50}{0.017} = 0.59$$

借助统计软件或者正态分布表，单尾区域为 0.2776，两个尾部区域共同代表的 p 值为 0.5552。因为 p 值大于 0.05，我们没有充分的理由拒绝原假设 H_0。因而该民意调查没有提供令人信服的证据，证明大多数小额贷款 APP 的借款人支持新的管理办法。我们观察到的 51%很可能只是由于单次统计偏差导致的，而不足以作为显著有效的证据。

单样本比例假设检验的过程

有四个步骤来完成检验：

准备 Prepare：确定参数，列出假设，确定显著性水平，并确定 $\hat{p}$ 和 n。

检查 Check：验证条件，确保在 H_0 下 $\hat{p}$ 服从正太分布。对于单样本比例的假设检验，使用原假设取值来检查成功-失败条件。

计算 Calculate：如果正态分布的验证通过，则进一步计算标准误，同样注意在计算中需要使用p_0。然后计算 Z 分数，并确定 p 值。

结论 Conclude：通过比较 p 值和 α 来评估假设检验，并结合场景给出结论。

有关其他单样本比例假设检验的例子请见第 5.3 节。

6.1.4 当成功-失败检验中一个或多个条件没有得到满足时

我们已经花了很多时间来讨论使用正态分布分析 $\hat{p}$ 的条件。那么当成功-失败检验无法通过的话会发生什么？当不满足独立性条件时又会怎样呢？在这两种情况下，置信区间计算和假设检验的一般思路是相同的，但置信区间的构建方法发生了变化。当无法满足成功-失败检验的条件时，我们可以围绕原假设取值 p_0 来建立分布，而不再基于 $\hat{p}$ 建立分布模型。用 p_0 进行模拟的概念类似于第 2.3 节介绍的疟疾案例研究中使用的思路。我们在本书中不再展开，但大家可以前往下述的线上网页参考学习这种解决方法：

www.openintro.org/r?go=stat_sim_prop_ht

对于不满足成功-失败条件时的置信区间搭建，我们可以使用**克罗普尔-皮尔森区间 Clopper-Pearson interval**。具体的细节超出了本书的范围。不过有许多网上资源涉及到了这一主题，大家也可以利用例如 ChatGPT 这样的语言模型工具来辅助搜索和学习。

独立性条件实质上是一个更细的要求。当它没有得到满足时，我们需要了解未能满足的背后原因和解决方法。例如，如果我们分析一个聚类样本时，就无法假设样本中个体的抽取符合独立随机的特征。而这时也会有专门相应的统计方法供我们使用，但这超出了我们本书讨论的范围。与聚类抽样类似，方便抽样（见本书 1.3.3 小节）也会因为不能满足独立性条件而产生固有偏差。

本书会更聚焦已满足独立性和成功-失败检验的统计问题，这类问题也被称为**良好约束的统计问题 well-constrained statistical problems**。也请大家在学习时谨记，本书中的内容只是浩如烟海的统计知识的冰山一角，主要用于数据和统计理论的启蒙。而在现实中统计学家需要去研究更加复杂的数据，也会需要借助更多更专业的统计理论。

6.1.5 样本大小的选择

在收集数据时，我们需要选择一个适合的样本大小。通常情况下，我们会建议选择一个足够大的样本，大到让误差范围[1]足够小，从而使得抽样过程有相应研究价值。因为如果样本选的太小，尽管数据收集的成本会大幅降低，但是误差范围可能会过大导致结论本身失去意义。

例如我们通过问卷调查研究某提案的支持率，发出的问卷数就代表了样本大小。而假如因为发放问卷太少而导致误差范围在正负 40%……而如果调查得到的支持率是 60%，那么这个 40%误差就会让我们的结论变得很尴尬：我们可以有 95%的信心认为支持率在 20%到 100%的区间内。这无疑像是一句废话！但如果我们能找到一个样本大小 n，使在 95%的置信区间内的误差在正负 4%以内，对应有 95%的信心认为实际支持率在 56%到 60%之间，这样的结果就显得更加合理和有价值。注意，上一句话中提到的 4%只是为了说明而随口举例的值。

[1] 也就是我们在置信区间内对点估计值进行加减的部分。

示例 6.7

一家大学的报纸正在进行一项调查，以确定有多少学生支持每年增加 200 美元的费用来造新的足球场。在 95%的显著性水平下，需要多大的样本才能保证误差率小于 0.04?

答案：根据上文讲过的内容，样本比率的单边误差范围是

$$Z^* \times \sqrt{\frac{p(1-p)}{n}}$$

我们的目标是找到最小的样本大小 n，使上式小于 0.04。在 95%的置信水平下，Z 分数是 1.96。

$$1.96 \times \sqrt{\frac{p(1-p)}{n}} < 0.04$$

E

方程中有两个未知数：p 和 n。如果我们从之前的调查中的到 p，那我们就可以求出 n。事实证明，当 p 为 0.5 时，误差率最大，所以如果在不清楚调查得到的估计值的时候，我们通常使用这个最坏的情况。

$$1.96 \times \sqrt{\frac{0.5(1-0.5)}{n}} < 0.04$$

$$1.96^2 \times \frac{0.5(1-0.5)}{n} < 0.04^2$$

$$1.96^2 \times \frac{0.5(1-0.5)}{0.04^2} < n$$

$$600.25 < n$$

通过计算得出，我们需要超过 600.25 个参与者，这意味着我们需要 601 个参与者或更多，以确保样本比例在 95%置信水平下样本比例的单边误差范围在 0.04 以内。

指导练习 6.8

G

一位经理即将监督她的工厂大规模生产一种新型号的轮胎，她想估计这些轮胎中有多大比例会被质检部门拒绝。质检小组对工厂过去生产的三个型号的轮胎进行了检测，第一个型号的轮胎有 1.7%不合格，第二个型号有 6.2%不合格，第三个型号有 1.3%不合格。关于新轮胎，经理希望检测足够多的样本，以尽可能准确估计新轮胎型号的故障率，使其在 90%置信水平下的误差小于 1%。请参考上例，通过列方程的方式分别把 1.7%，6.2%和 1.3%作为 p 带入式中，并分别确定要达到目标置信水平和误差要求的最小样本 n 的大小。[1]

[1] 90%的置信区间的 Z 分数是 1.6449，同时比率估计的值是 0.017。我们将数值带入单边误差范围的公式：

$$1.6449 \times \sqrt{\frac{0.017(1-0.017)}{n}} < 0.01 \quad \rightarrow \quad \frac{0.017(1-0.017)}{n} < \left(\frac{0.01}{1.6449}\right)^2 \quad \rightarrow \quad 452.15 < n$$

这道题对于样本大小的计算我们使用四舍五入，所以第一个轮胎模型表明 453 个轮胎就足够了。类似的计算可以用 0.062 和 0.013 的 p 值来完成。使用这些比例的最小样本大小分别为 1574 和 348 个轮胎。

示例 6.9

在完成指导练习 6.8 后，我们发现满足要求的最小样本数差别很大：分别是 453，1574 和 348。如果要选择一个样本大小，你建议使用以上三者中的哪一个？什么会影响你的选择?

(E) 答案：我们可以研究哪个旧型号与新型号轮胎最接近，然后选择相应的样本大小。例如过去三种型号轮胎分别是纵向花纹，横向花纹以及交叉花纹轮胎，而新型号轮胎是纵向花纹轮胎，那么我们就可以参照使用 453 作为最终样本大小选择。或者，我们可以考虑上题所述的三个样本估计值(1.7%，6.2%和 1.3%) 产生过程用的样本大小，如果有两个估计值是基于小样本计算得出的，而另一个是基于更大样本计算得出的，我们可以考虑选择对应过去较大样本的那个样本大小。具体来说，假设在统计出 1.7%，6.2%和 1.3%时分别抽取了 100，100 还有 450 个轮胎，那么我们就可以选择 348 作为新一轮抽样的样本大小。此外，一定还有其他一些合理的方法，可以根据实际情况进行判断分析。

请注意，最后选择的样本数量应当也能通过成功-失败检验。例如，如果我们对 $n = 1584$ 个轮胎进行抽样，发现故障率为 0.05%，那么这时 $np = 1584 \times 0.05\% = 0.79$，这样情况下直接使用正态分布的就不合理了，我们需要用更高级的统计方法来建造置信区间。

指导练习 6.10

(G) 假设我们想持续跟踪小额贷款公司新监管规定的民意支持情况，我们可以每个月进行一次新的调查。但进行这样频繁的大样本民意调查是很昂贵的，所以我们需要想办法降低成本。而降低成本的途径之一就是使用小样本，但这样无疑又会增加误差。经过研究分析，我们最终决定，只要每次调查的误差范围在 5%以内就可以接受。已知在原始样本中有 70%的人支持新监管规定，我们应该选择多大的样本，才能保证不超过 0.05 的单边误差范围和 95%的置信水平？ [1]

---

[1] 我们完成与之前相同的计算，只是现在 p 是 0.7 而不是 0.5。

$$1.96 \times \sqrt{\frac{p(1-p)}{n}} \approx 1.96 \times \sqrt{\frac{0.70(1-0.70)}{n}} \leq 0.05 \quad \rightarrow \quad n \geq 322.7$$

323 个或更多的样本大小将是合理的。(提醒：计算样本大小时一定要四舍五入!) 鉴于我们计划长期跟踪这项投票，我们可能还想定期重复这些计算，以确保我们在推荐样本量时考虑周到，避免参考了错误的基准导致样本大小估计不准。

6.2 双样本比例差

接下来，我们将扩展第 6.1 节的方法，将置信区间和假设检验应用于两个总体比例间差异的估计：从统计写法上我们可以把这种差异记作 $p_1 - p_2$，而对于该表达式的估计值，你或许已经能猜到它的写法了：$\hat{p}_1 - \hat{p}_2$。这里也稍微温习一下，估计值指的是通过一组抽样后根据样本计算出的统计值。在本节中，我们将延续使用与单样本比例估计相同的步骤：即首先验证感兴趣的总体参数是否可以用正态分布来建模，接着计算标准误，并最终进行统计推断和假设检验。

6.2.1 双样本的比例差的分布

与 $\hat{p}$ 一样，如果需要做出「双样本比例之差 $\hat{p}_1 - \hat{p}_2$ 也近似服从正态分布」的判断，根据中心极限定理（见第 5 章第 5.1.3 小节），样本必须满足以下条件：首先是一个广义的独立性条件，其核心是两个样本的个体在跨样本级别上也需要相互独立；其次是两个样本都要通过成功-失败检验。

双样本比例差 $\hat{p}_1 - \hat{p}_2$ 服从正态分布的条件

当以下条件满足时，$\hat{p}_1 - \hat{p}_2$ 可以用正态分布来建模分析：

- 独立性条件的延伸：因为我们要计算两个比例估计值，所以势必会涉及两个样本。那么每个样本个体的观测值不仅需要在该样本内部相互独立，在样本间的尺度上也需要是相互独立的。一般来说，如果数据来自两个独立的随机抽样过程，或者数据来自于一个经过严密设计的随机试验，就可以满足这一点。
- 成功-失败检验：同样地，该检验需要分别对两组都成立。

当以上两个条件都被满足时，我们可以用下式计算两个总体比例差的标准误：

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

其中 p_1 和 p_2 代表总体比例，n_1 和 n_2 代表样本容量。

6.2.2 双样本比例差的置信区间

对于双样本情形来说，置信区间的计算依然可以使用通用公式，即点估计加减 Z 分数乘以标准误。如下，其中 $\hat{p}_1 - \hat{p}_2$作为点估计带入，而标准误 SE 则用上一个小节中介绍的公式替换：

$$点估计值 \pm z^* \times SE \quad \rightarrow \quad \hat{p}_1 - \hat{p}_2 \pm z^* \times \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

在介绍单样本比例的置信区间和假设检验时，我们引入了一套「准备、检查、计算和结论」的四步走方法。对于双样本比例差的置信区间计算和假设检验，我们也可以套用同样的四步走方法。尽管整体逻辑依然是内在统一，但是在细节处理上还是会略有不同。在本小节的学习中，大家也可以时时和之前的四步走方法对照比较，加深对该逻辑框架的认识。

示例 6.11

我们来探讨一个关于心脏病人发病和抢救的试验案例。该案例中的病人在心脏病发作后都接受乐心肺复苏急救（CPR），紧接着被送往医院。在医院接受下一步治疗时，病人被随机分为两组：试验组的病人接受抗凝血药物的治疗，而对照组的病人则不使用抗凝血药物。我们重点关注这两组的患者是否能挺过最危险的 24 小时，统计结果见图 6.1。请先判断我们是否可以应用正态分布模型。

答案：我们首先检查独立性：因为这是一个随机试验，所以所有病人都相互独立，因而该条件满足。接下来，我们分别进行每组的成功-失败检验。在各个试验组对应的统计结果（11、14、39 和 26）中，我们都至少有 10 个成功案例和 10 个失败案例，因而此条件也满足。

E

在这两个条件都满足的情况下，样本比例的差异统计量就可以说是近似服从正态分布，即如果我们取足够多组样本计算比例差，最终应该会观察到比例差的估计值们呈一个钟形的曲线分布。

	存活	死亡	总计
对照组	11	39	50
实验组	14	26	40
总计	25	65	90

图 6.1：心肺复急救苏试验结果。试验组患者使用了抗凝血药，对照组患者则没有。

示例 6.12

创建并解释两组心脏病人存活率之差的置信区间，要求的置信度为 90%。

答案：$\hat{p}_t$ 表示试验组的存活率，$\hat{p}_c$ 表示对照组的存活率。

$$\hat{p}_t - \hat{p}_c = \frac{14}{40} - \frac{11}{50} = 0.35 - 0.22 = 0.13$$

我们使用本小节开头提供的标准误公式。与单样本比例的情况一样，由于总体参数无从得知，对于公式中涉及 p_c 和 p_t，我们使用对应的估计值 $\hat{p}_t$ 和 $\hat{p}_c$ 来带入计算。

E

$$SE \approx \sqrt{\frac{0.35(1-0.35)}{40} + \frac{0.22(1-0.22)}{50}} = 0.095$$

对于 90%的置信区间，我们确定 Z 分数值为 1.6449。

$$\text{点估计} \pm z^* \times SE \quad \rightarrow \quad 0.13 \pm 1.6449 \times 0.095 \quad \rightarrow \quad (-0.026, 0.286)$$

我们有 90%的信心认为两组间差异值在-2.6%到+28.6%之间。也就是说我们有 90%的信心认为，对于突发心脏病并接受了心肺复苏术的病人来说，抗凝血剂的治疗对病人生存率的影响在-2.6%到+28.6%之间。因为 0%被包含在区间内，即试验组和对照组生存率差异为零也是有可能的。因此，我们并没有足够的证据来说明抗凝血剂对接受心肺复苏术后入院的心脏病患者有显著帮助 (或危害)。

指导练习 6.13

我们进行了一项为期 5 年的试验，以评估鱼油对减少心血管疾病发作（本例中特指心脏病）的有效性。每位受试者被随机分到试验组或者对照组中，他们的心脏病发作结果如下图所示：

G

	心脏病发	没有病发	总计
鱼油	145	12788	12933
安慰剂	200	12738	12938

请为鱼油对心脏病发作的影响（即试验组和对照组间的差异值）建立一个 95%的置信区间。同时，请在本题目的研究背景下解释这个置信区间的含义。[1]

---

[1] 因为患者是随机抽取的，所以研究对象在两个实验组内部和之间都是独立的。同时两组都分别满足成功-失败检验的条件，因为所有计数都至少为 10。因此，样本比例的差异可以视为近似服从正态分布。
计算样本比例（$\hat{p}_{鱼油} = 0.0112, \hat{p}_{安慰剂} = 0.0155$），差值点估计（$0.0112 - 0.0155 = -0.0043$）以及标准误（$SE = \sqrt{\frac{0.0112 \times 0.9888}{12933} + \frac{0.0155 \times 0.9845}{12938}} = 0.00145$）。接下来，将这些值代入置信区间的一般公式。对于 95%置信区间，我们确定 Z 分数值为 1.96：

$$-0.0043 \pm 1.96 \times 0.00145 \quad \rightarrow \quad (-0.0071, -0.0015)$$

我们有 95%的信心认为，对于那些与被测试者背景大体一致的群体来说，鱼油可以在 5 年期间内将心脏病发病率降低 0.15 到 0.71 个百分点(从大约为 1.55%的基线)。因为整个置信区间均小于 0，这些数据为「鱼油补充剂可以减少像研究中那样的患者的心脏病的发作」提供了强有力的证据。

6.2.3 双样本比例差的假设检验

乳房 X 光是一种用于检查乳腺癌的技术手段，该技术也是我们在接下来的这个例子中要探讨的话题。继上个小节讨论过比例差的置信区间计算之后，我们将学习当原假设 H_0 为 $p_1 - p_2 = 0$（也就是 $p_1 = p_2$）时的双样本比例差假设检验。

一项为期 30 年的研究对近 9 万名女性参与者进行了跟进调查。首先在为期 5 年的试验期中，每位女性被随机分配到两组中的一组：试验组中，参与者通过拍摄乳房 X 光片以筛查乳腺癌；而在对照组中，参与者只接受常规乳腺癌检查。在接下来的 25 年的数据收集期里，我们没有对这两组进行任何干预。然后我们根据整个 30 年间搜集到的数据，分析因乳腺癌导致的死亡情况。图 6.2 总结了该研究的结果。

如果乳房 X 光检查比常规检查更有效，那么我们将应该观察到对照组死于乳腺癌的人数与试验组不相同。另一方面，如果乳房 X 光检查相比常规检查手段无显著效果，我们就应该观察到试验组和对照组死于乳腺癌的人数基本一致。

	是否因乳腺癌而死亡	
	是	否
乳腺X光检查组	500	44,425
对照组	505	44,405

图 6.2：乳腺癌研究的结果呈现。

指导练习 6.14

这是一个试验性质的研究还是观察性质的研究？[1]

指导练习 6.15

提出假设：试验组和对照组因为乳腺癌死亡的人数是否有差异。[2]

原假设的提出其实很简单，即两个比例差为零，或者说试验组和对照组因为乳腺癌死亡的人数没有显著差异。在示例 6.16 中，我们将检验满足正态分布的各项条件来分析研究结果。具体检验过程与置信区间的计算非常相似。不过，当原假设是 $p_1 - p_2 = 0$ 时，我们可以用一个特殊的比例，即 **混合比例 pooled proportion**，来进行成功-失败检验。检验依然要对两个样本 (试验组及对照组) 进行，但是不再使用各自的 $\hat{p}$，而是使用混合比例，具体请参考下个示例。

$$\hat{p}_{混合} = \frac{\#\ 整个研究中死于乳腺癌的病例数}{\#\ 整个研究中的总病例数} = \frac{500 + 505}{500 + 44{,}425 + 505 + 44{,}405} = 0.0112$$

[1] 这是一个试验。患者被随机分配接受乳房 X 光检查或标准乳腺癌检查。我们基于这项研究将得出因果关系的结论。

[2] H_0: 接受乳房 X 光检查（试验组）的患者乳腺癌死亡率与对照组患者的乳腺癌死亡率相同，$p_{X光} - p_{对照} = 0$。
H_A: 接受乳房 X 光检查（试验组）的患者乳腺癌死亡率与对照组患者的乳腺癌死亡率不同，$p_{X光} - p_{对照} \neq 0$。

这个混合比例是对整个研究中乳腺癌死亡率的估计，也是在 $p_{\text{x 光}} = p_{\text{对照}}$ 的原假设为真的情况下，我们对 $p_{\text{x 光}}$ 和 $p_{\text{对照}}$ 比例的最佳估计。在计算标准误时，我们也将使用这个混合比例。

示例 6.16

在这项研究中，使用正态分布是否合理？

答案：因为病人是随机抽取的，他们可以被视为独立的，无论是组内还是组间。我们还必须进行每组的成功-失败检验。在原假设下，试验组和对照组的比例是相等的，所以我们可以用我们对这些值的最佳估计来进行成功-失败检验，也就是上个指导练习中提到的双样本的混合比例，$\hat{p}_{\text{混合}} = 0.0112$。

E

$\hat{p}_{\text{混合}} \times n_{\text{x 光}} = 0.0112 \times 44{,}925 = 503$　　$(1 - \hat{p}_{\text{混合}}) \times n_{\text{x 光}} = 0.9888 \times 44{,}925 = 44{,}422$

$\hat{p}_{\text{混合}} \times n_{\text{对照}} = 0.0112 \times 44{,}910 = 503$　　$(1 - \hat{p}_{\text{混合}}) \times n_{\text{对照}} = 0.9888 \times 44{,}910 = 44{,}407$

成功-失败检验的条件可以被满足，因为所有数值都至少是 10。满足了以上两个条件，我们可以大胆使用正态分布来进行后续的置信区间计算及假设检验推断。

当 H_0：$p_1 - p_2 = 0$ 时需要使用混合比例

当原假设中两个比例相等时，使用混合比例 $\hat{p}_{\text{混合}}$ 来进行成功-失败检验，并估计标准误：

$$\hat{p}_{\text{混合}} = \frac{\#\text{「成功」的案例个数}}{\#\text{总案例数}} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

这里的 $\hat{p}_1 n_1$ 代表了 1 号样本中的成功案例数量，因为：

$$\hat{p}_1 = \frac{\#\text{1 号样本中「成功」的案例个数}}{n_1}$$

同样地，$\hat{p}_2 n_2$ 代表了 2 号样本中的成功案例数量。

在示例 6.16 中，混合比例被用来进行成功-失败检验。在下一个例子中，我们用混合比例进行标准误的计算。

示例 6.17

计算两组中乳腺癌死亡率差异的点估计值，并使用混合比例 $\hat{p}_{混合} = 0.0112$ 来计算标准误。

答案：乳腺癌死亡率差异的点估计是：

$$\hat{p}_{\text{x 光}} - \hat{p}_{对照} = \frac{500}{500 + 44,425} - \frac{505}{500 + 44,405}$$

$$= 0.01113 - 0.01125$$

$$= -0.00012$$

乳房 X 光检查组的乳腺癌死亡率比对照组低 0.012%。

接下来，用混合比例 $\hat{p}_{混合}$ 来计算标准误。

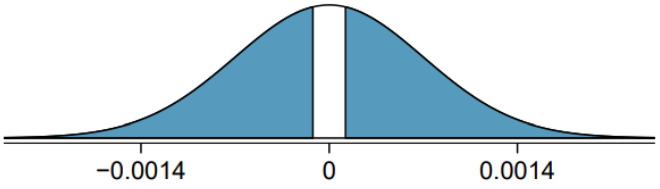
$$SE = \sqrt{\frac{\hat{p}_{混合}(1 - \hat{p}_{混合})}{n_{\text{x 光}}} + \frac{\hat{p}_{混合}(1 - \hat{p}_{混合})}{n_{对照}}} = 0.00070$$

示例 6.18

使用双样本比例差的点估计值 $\hat{p}_{\text{x 光}} - \hat{p}_{对照} = -0.00012$ 和标准误 $SE = 0.00070$，计算假设检验的 p 值，请列出计算过程并写出最终的结论。

答案：就像之前的练习一样，我们首先计算出一个测试统计量，然后画出图像。

$$Z = \frac{\text{点估计} - \text{原假设值}}{SE} = \frac{-0.00012 - 0}{0.00070} = -0.17$$



下尾部区域是 0.4325，它代表着比我们观测到的-0.17 的比例差更小的差异出现的概率是 43.25%。而如果只看数值 0.17，我们还要考虑另一侧的尾端分布，也就是说我们需要将其加倍，得到 p 值为 0.8650。即在试验中会观测到比 0.17 数值更极端的情况的概率为 86.5%。因为这个 p 值大于 0.05，说明我们这次试验观察到的情形并不够极端。因此我们不能拒绝原假设。也就是说，试验组和对照组乳腺癌死亡率的微小差异很可能只是一次偶然事件，而无法说明哪种检查效果更好。通过这个试验，我们没有观察到乳房 X 光检查更有助于减少乳腺癌的死亡率。

既然无法论证 X 光检查更有效，那么通过上面几个示例，我们能否得出「X 光检查没有好处甚至可能存在潜在危害」的结论呢？显然草率得出这样的结论也是不负责任的。其实在现实研究中，尤其是在临床医疗领域的统计研究中，想要得出一个结论远比想象中复杂的多。尽管我们已经进行了试验，也通过较为严密的数据知识进行了统计推断，但是还有很多值得我们考虑的因素。我们在此列举一些与本小节讨论的乳房 X 光检查研究相关的要点，并由此希望引发大家对医学研究的深层思考：

- 根据统计推断结果，我们不能拒绝原假设，也就是说我们没有足够的证据来得出「X 光检查会导致乳腺癌患者的死亡率显著降低或升高」的结论；
- 注意，尽管我们明确「不能拒绝」原假设，但不意味我们「要接受」原假设。这是又一次对统计推断措辞严密性的拓展延伸；
- 在乳腺癌与 X 光的试验中，我们无法得出一个 X 光会导致死亡率降低或者升高的结论，这感觉上有点像是试验白做了。但事实上，在统计探究的过程中，「无法拒绝原假设」往往也是一条很重要的结论，它会指导我们修正试验设计，并鼓励我们继续探究真理。这也是为什么我们不能草率接受原假设，而还必须提及置信度以及分析所有潜在可能；
- 根据数据，如果 X 光检查对乳腺癌预防 (从统计学显著角度) 真有一定的帮助或危害效果，这次试验结果也说明其影响从实际角度其实并不是很大；
- 此外，从现实角度出发，我们必须考虑的一个关键问题是：X 光检查是比其他常规筛查更昂贵还是更便宜？在医疗方案的选择上，一项方案如果又贵又没有特别明显的效果，那么毫无疑问推广这种方案的意义并不大；
- 该研究的作者还发现，X 光检查会导致乳腺癌的过度诊断。有些癌细胞本身是良性的，而有些乳腺癌患者不幸还同时罹患其他更加严重的隐疾。无论哪种情况，X 光检查尽管可以检出乳腺癌，但可能会导致本不需要接受治疗的病人接受了价格不菲的治疗，或者导致本应该优先治疗其他疾病的病人耽误了宝贵的治疗时间在医治乳腺癌上。要知道任何医疗手段都是有成本的，这些成本可能包括高昂的治疗费用，治疗手段带来的副作用，以及因为进行某项治疗而耽误了其他治疗的时间成本。无论如何，过度诊断本身可能并不是一件好事，它有可能会对病人造成不必要的身体或精神伤害。

希望读至此处的读者还能跟上我们思维的脚步，以上这些思考均突出了对医疗护理和治疗进行改进的复杂性。医学专家和委员会常常需要将上述因素纳入考虑范围，并在现有证据的基础上谨慎提出建议。

6.2.4 关于双样本比例之差的假设检验的拓展延伸（特别话题）

当我们进行双比例假设检验时，通常 H_0（原假设）是 $p_1 - p_2 = 0$。在一些比较少见的情况下，我们会想检查 p_1 和 p_2 的非零差异。例如，也许我们关心的是 $p_1 - p_2 = 0.1$ 的原假设。在这样的情况下，我们通常使用 $\hat{p}_1$ 和 $\hat{p}_2$ 来进行成功-失败检验，并计算标准误。

指导练习 6.19

G 一家四轴无人机公司正在考虑一个新的旋翼制造商。这家新制造商将会更贵，但他们声称他们的高质量叶片更可靠，反映在数据上就是他们生产的叶片，相比竞争者来说，会有多 3%的叶片通过质检。请为该检测设置适当的假设。[1]

图 6.3：一架四轴飞行器。

图片来源 David J http://flic.kr/p/oiWLNu

已获得授权取得 CC-BY 2.0 证书

[1] H_0: 高质量叶片通过质检的频率比标准质量的叶片高 3%，$p_{高质量} - p_{标准} = 0.03$。
H_A: 高质量叶片会比标准质量的叶片多通过一定数量的质检（不同于 3%），$p_{高质量} - p_{标准} \neq 0.03$。

示例 6.20

延续指导练习 6.19，无人机公司的质量控制工程师分别从两个供应商处随机收集了 1000 个叶片的样本，然后进行质量检测。她发现有现有供应商提供的叶片中，有 899 个通过了质量检测，而新的供应商提供的叶片中，958 个通过了同样的质量检测。使用这些数据，在 5%的显著性水平下，请评估指导练习 6.19 中提出的假设。

答案：首先，我们需要检验双总体比例之差是否满足正态分布的各项条件。第一，我们假设叶片都是独立的。这里其实更多是为了本书推进的需要，因为或许也会有人为操作或者批量抽样导致不能满足随机条件的情况出现。但在这里，我们姑且假设抽样的独立随机性。成功-失败检验显然也适用于每个样本。因此，我们可以进行下一步，计算出样本的比例差为 0.958 − 0.899 = 0.059，并且可以说这个估计量近似地服从正态分布。

(E) 标准误是用两个样本各自的比例来计算的，因为原假设不再是比例差为零，所以我们在这里不能再使用混合比例。

$$SE = \sqrt{\frac{0.958(1-0.958)}{1000} + \frac{0.899(1-0.899)}{1000}} = 0.095$$

接下来，我们计算检验统计量，并利用它来寻找 P 值，如图 6.4 所示。

$$Z = \frac{\text{点估计} - \text{原假设值}}{SE} = \frac{0.059 - 0.03}{0.0114} = 2.54$$

把这个测试统计量的 Z 分数值放在标准正态分布上，我们确定右尾部区域为 0.006。我们将其加倍，得到 p 值为 0.012。也就是说，出现比样本调查出情况更极端情况的概率仅为 1.2%。因为 0.012 小于我们的阈值 0.05，我们拒绝原假设。因此，我们有 95%的信心说，在统计学上有显著的证据表明，新供应商提供的刀片通过质检的比率比目前使用的刀片高 3%以上。

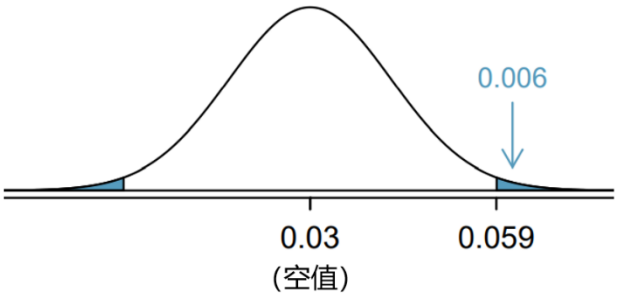


图 6.4：若原假设为真的分布情况，染色区域表示了 p 值。

6.2.5 假设检验的标准误公式（特别话题）

本小节涵盖了更多的理论课题，对双比例差的标准误公式的起源有了更深的讨论。归根结底，我们在本章和第 7 章中遇到的所有标准误公式都可以从第 3.4 节的概率原理中推导出来。

双样本比例差的标准误公式可以解构为单样本比例的标准误公式。回顾一下单一样本比例的标准误 $\hat{p}_1$ 和 $\hat{p}_2$ 是

$$SE_{\hat{p}_1} = \sqrt{\frac{p_1(1-p_1)}{n_1}} \qquad SE_{\hat{p}_2} = \sqrt{\frac{p_2(1-p_2)}{n_2}}$$

双样本比例差的标准误可以解构为各个样本比例的标准误:

$$SE_{(\hat{p}_1-\hat{p}_2)} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

这种特殊关系也遵循我们之前在概率论篇章中讨论的知识框架：即两个相互独立的随机变量之差的标准差应该等于两个随机变量标准差之和再开方。

指导练习 6.21

G

先修前提：本书第 3.4 节。我们可以用不同的方式改写上述方程式。

$$SE_{(\hat{p}_1-\hat{p}_2)}^2 = SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2$$

用两个随机变量之和的公式变形来解释这个公式的来源。[1]

[1] 标准误的平方代表估计值的方差。如果 X 和 Y 是两个随机变量，其方差分别为 σ_x^2 和 σ_y^2，那么 $X-Y$ 的方差是 $\sigma_x^2 + \sigma_y^2$。同样，$\hat{p}_1 - \hat{p}_2$ 相对应的方差是 $\sigma_{\hat{p}_1}^2 + \sigma_{\hat{p}_2}^2$。因为 $\sigma_{\hat{p}_1}^2$ 和 $\sigma_{\hat{p}_2}^2$ 只是 $SE_{\hat{p}_1}^2$ 和 $SE_{\hat{p}_2}^2$ 的另一种写法，所以 $\hat{p}_1 - \hat{p}_2$ 的方差可以对应写成 $SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2$。

6.3 用卡方检验评估拟合度

在这一章节中，我们将会学习一种方法，用于对分组数据进行假设检验以及模型评估。这一方法通常用于两种情况：

- 给出一个可分为几组的样本数据，确定该样本是否代表它对应的总体；
- 评估数据是否满足于一个特定的分布，如正态分布或几何分布。

这些情况都可以用同一个统计检验来解决：卡方检验。上面的这两种情形描述可能会显得有些抽象，请大家耐心往下读，并结合下面的实例来理解卡方检验的作用。

首先我们来举第一个案例。我们研究从美国某郡中随机抽取的 275 名陪审员的样本的数据。在进入案例前做一点背景介绍。美国拥有陪审员制度：即每次司法案件都会有一定有选举权的公民组成陪审团，陪审团会参与决定是否对嫌疑犯进行起诉，或者嫌疑犯是否有罪。有一部非常经典的电影《十二怒汉》，可以作为了解美国陪审团制度的不错的影视资料。从片名中也可以看出，影片中的陪审团由 12 个公民组成。同时在美国，一个郡可以近似地与中国行政级别中市进行类比，当然二者还是有很多区别，这里只是为了让大家有一个近似的可以直观想象的概念。

中国一个城市中总会有大大小小各种案件发生，美国也一样，因此在一个郡中担任过陪审员的人数无疑也不是一个小数字。这些陪审员就构成了我们本次研究的总体。而我们研究的真正课题是，在选择这些陪审员的时候，是否会有种族歧视的现象发生，比如白人会更容易进入陪审团；又或者某人种群体尽管在郡里占很大比重，却很少人进入到陪审团中。

为了进行这项研究，沿用之前的假设检验思路，我们就要提出原假设和备择假设，接着从所有陪审员群体中抽取一个样本，并利用样本计算出一些估计量，最后进行检验。但这里，我们感兴趣的不再是一个比例，也不再是两个比例差。假设我们把人们的种族分成：白人，黑人，墨西哥裔和其他，我们将会同时对这四种人的比例感兴趣。我们可以选择依次去检验这四种人的比例，但是这样也只能就某特定种族得出结论，而无助于我们对「整体人群占比」进行判断。这也是为什么我们在这里舍弃均值（或者说之前的单样本比例值）和均值差（或者说之前的双样本比例差），而将会去构建一个全新的估计量：卡方系数，即卡方值。

介绍完这些背景，相信大家应该对于我们要提取的数据和要解决的问题有了一点点感觉。如图 6.5 所示。我们想确定这些陪审员在族裔分布上是否能代表样本来源的总体，也就是小镇上所有符合陪审员条件的有投票权的选民。

种族	白种人	黑种人	西班牙裔	其它	总计
陪审团代表人数	205	26	25	19	275
注册的投票人数	0.72	0.07	0.12	0.09	1.00

图 6.5：一个城市陪审团和人口中的种族分布。

如果陪审员真的是从登记的选民中随机抽取的，那么本身占比较大的族裔群体应该也会选出较多的陪审员。即便具体比例可能会和族群比例有一定的差异，但差异也不应该太大，否则就说明该陪审团不能够代表郡人口，也就是种族歧视某种程度上讲确实存在。

以上是一个使用卡方检验评估「样本分组能否反映总体构成」的案例。而关于开头处提到的两种情形中的第二种，即用卡方检验「评估分布的拟合度」，我们将会在最后一个小节中详细展开。届时我们将讨论标普 500 指数在过去 25 年中的每日收益率的波动，进而研究：股票当日的表现是否会受到前日表现的影响。在这个问题中，我们希望同时检查样本的多个乃至所有分组的情况，而不是简单地计算样本整体的均值或两个样本的均值差（单样本或者双样本的比例差），这时卡方检验就是不错的选择。

6.3.1 建立一维表的检验统计

示例 6.22

已知在上面提到的郡中，有 72%白人人口，有 7%黑人人口。那么假设需要有 275 人担任陪审员，而且陪审员的选择完全随机的话，我们应该预期 275 人中有多少是白人？有多少人是黑人？

E

答案：由于在总体人口中，有大约 72%的人口是白人，因此我们预计大约 72%的陪审员是白人：0.72 × 275 = 198。同样，在总体人口中，我们预计大约 7%的陪审员是黑人，这将对应大约 0.07 × 275 = 19.25 个黑人陪审员。

指导练习 6.23

G

在同个郡中有 12%的人口是西班牙裔，9%的人口是其他种族。在随机 275 名陪审员中，我们预计有多少人是西班牙裔或来自其他种族？答案见图 6.6。（该指导练习答案不在脚注，直接见下文）

虽然可能会存在抽样差异，但如果陪审团的选择过程是无偏见的，那么我们会认为，不同种族人口在样本中的占比应与在整体人口中的占比非常相近。我们把这些推论转化成统计假设的语言：

H_0 原假设：陪审员的选择过程满足随机抽样，即在选择陪审员的过程中不存在种族偏见，观察到的统计反映了自然抽样的变化。

H_A 备择假设：陪审员的选择不是随机的，也就是说，在挑选陪审员时存在种族偏见。

但我们的假设需要通过检验来证明。我们需要检验是否样本与总体的差异足够大，是否能够提供令人信服的证据证明陪审团的构成不是随机样本。实际和预期的陪审团族裔构成请见图 6.6。

种族	白种人	黑种人	西班牙裔	其它	总计
实际数据	205	26	25	19	275
预期数据	198	19.25	33	24.75	275

图 6.6：陪审员的实际和预期构成。

为了进行假设检验，我们统计了实际样本与预期样本的数值差异。而两组数据间异常大的差异就能够作为我们放弃原假设，更倾向于选择备择假设的强有力证据。也就是说，我们观察到的样本比例与全郡比例不同的情形，不仅仅来源于单次抽样的样本误差，而是可能抽样本身就不够随机。

6.3.2 卡方值

在以前的假设检验中，我们通过如下的公式计算**检验统计量 test statistics**。

$$\frac{\text{点估计} - \text{原假设值}}{\text{点估计的标准误}}$$

这种计算基于的逻辑如下：首先在原假设为真的前提下，我们计算点估计值和预期值之间的差值；其次，使用点估计值的标准误将该差值进行标准化处理。在卡房检验的应用场景中，我们将把这种思路应用在各个组别的计数统计上，并在多个组的检验统计量的基础上计算出卡房统计量。例如，对白人群体就要首先计算：

$$Z_1 = \frac{\text{观察到的白人数量} - \text{原假设的白人数量}}{\text{观察到的白人数量的标准差}}$$

在上式中，点估计的标准误就是基于原假设值的白人数量的平方根，即上方表格中「预期数据」一行对应的白人数量[1]。通过公式和图 6.6 中的数字，我们可以得出：

$$Z_1 = \frac{205 - 198}{\sqrt{198}} = 0.5$$

[1] 利用前几章学到的一些规则，我们可能认为标准误是 $np(1-p)$，其中 n 是样本量，p 是人口中的比例。如果我们只看一个统计结果数，这将是正确的。然而，我们正在计算许多组统计量，并考虑对他们进行融合处理。这时，直接计算数值平方根会是更好的方法，因为人口的比例将在我们对多组统计量进行加总的时候自然地被考虑进去。

这个公式和我们之前接触过的思路很相似，都是首先计算一个差值，然后将其标准化。我们需要对黑人、西班牙裔和其他人种都进行这样的计算：

黑人

$$Z_2 = \frac{26 - 19.25}{\sqrt{19.25}} = 1.54$$

西班牙裔

$$Z_3 = \frac{25 - 33}{\sqrt{33}} = -1.39$$

其他人种

$$Z_4 = \frac{19 - 24.75}{\sqrt{24.75}} = -1.16$$

我们想用一个检验统计来确定这四个标准化差异是否都和零相距甚远，或者说从统计学意义上显著不等于零。那也就意味着，Z_1, Z_2, Z_3 和 Z_4 必须以某种方式结合起来作为一个整体，然后我们只需要对这个整体统计量进行判断。既如此，它们之间「结合」的方法就是我们必须要考虑的事情，一种想法是通过取绝对值：

$$|Z_1| + |Z_2| + |Z_3| + |Z_4| = 4.58$$

这种方法虽然可以把负数转换成正数然后相加，从而让整体差值不会因为正负数而抵消。然而，更常见的是取平方和。

$$Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 = 5.89$$

取平方和的运算有两点好处：

(1) 取完平方之后所有数都是正值，一样可以避免差异正负方向不同导致的抵消情况出现；

(2) 原本不大的差值（小于 1）在平方后会更小，而原本就比较大的差值（例如-2.5）在取完平方后会变得更大，也就是给距离数字零更远的值以更大权重。

卡方值 X^2 的平方和公式可以转化为如下的公式：

$$\mathrm{X}^2 = \frac{(\text{组 1 观测值} - \text{组 1 原假设值})^2}{\text{组 1 原假设数目}} + \cdots + \frac{(\text{组 4 观测值} - \text{组 4 原假设值})^2}{\text{组 4 原假设值}}$$

最后 X^2 概括了观察数值与原假设数值的偏差。在第 6.3.4 小节中，我们将看到，如果原假设为真，那么多次抽样后分组统计计算出的 X^2 奖遵循一个新的分布，即卡方分布。利用这个分布，我们将能够计算出 p 值来进行和之前章节类似的假设检验。

6.3.3 卡方分布和查找区域

卡方分布 chi-square distribution 有时被用来描述那些数值总是正的且右偏的数据。我们回顾一下之前介绍的正态分布：为了描述一个正态分布我们往往需要两个参数，即均值和标准差。而有了这两个具体参数，我们往往也可以绘制出精确的正态分布形状，并且掌握它的确切特征。在卡方分布，我们只需要一个参数，叫做**自由度 degree of freedom (df)**，它将影响分布的形状、中心和扩散程度。自由度也是整个统计学中很重要的概念，这里将会是大家和它的第一次接触。

指导练习 6.24

图 6.7 显示了三个卡方分布。

(a) 当自由度较大时，分布的中点如何变化?

(b) 变化程度（扩散）是怎样的?

(c) 图片的形状如何变化? [1]

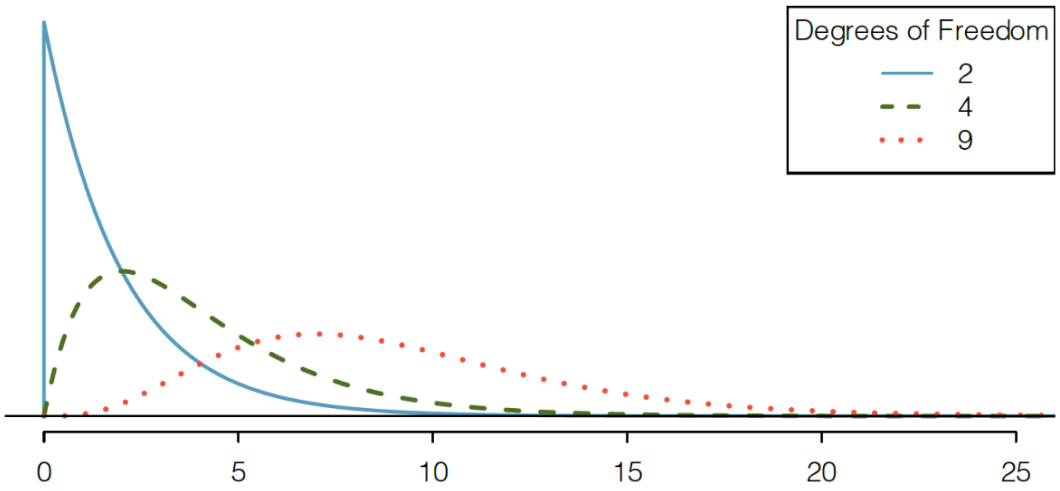


图 6.7：三个自由度不同的卡方分布。

图 6.7 和指导练习 6.24 展示了随着自由度的增加，卡方分布的三个一般性质：分布将变得更加对称，中心将向右移动，分布也会更加离散增加。我们研究卡方分布时，一样会对计算 p 值非常感兴趣，这就意味着我们还是要确定尾部区域的面积。和正态分布章节介绍的一致，最常见的方法是使用图形计算器或使用对照表。对于下面的示例和指导练习，请使用你个人擅长的方法来得出你的答案。对于想使用卡方检测值对照表的朋友，请前往**附录 C.3** 查看概要和详细表格。当然，你也可以选择使用统计软件来进行计算求解，这也会是一种更快捷和准确的方法。

[1] (a) 自由度变大时，中心凸起的面积也会相应变大。如果仔细观察，我们可以看到，每个分布的均值（最高凸起）都等于该分布的自由度。(b) 变化程度随着自由度的增加而增加。(c) 分布在 $df = 2$ 时有很明显的右偏，然后在较大的自由度 $df = 4$ 时分布变得更加对称。如果我们研究更大自由度的分布，我们会看到这种趋势将继续下去。

示例 6.25

图 6.8(a)（所有本页涉及的图均见下页）显示了一个自由度为 3 的卡方分布，其尾部的阴影开始于 6.25。求尾部阴影部分的面积。

E

答案：使用统计软件或图形计算器，我们可以发现，自由度为 3，以 6.25 分割的卡方分布的尾部面积是 0.1001。也就是说，图 6.8(a)中阴影部分的面积为 0.1。

示例 6.26

图 6.8(b)显示了自由度为 2 的卡方分布的尾部阴影。我们以 4.3 为边界，求尾部阴影部分的面积。

E

答案：通过使用统计软件，可以发现图 6.8(b)中阴影部分的尾部面积为 0.1165。而如果使用对照表，根据不同的表格精度，我们有可能只能把尾部面积的范围确定在 0.1 和 0.2 之间。

示例 6.27

图 6.8(c)展示了一个自由度为 5 的卡方分布，同时图上标注出了一定部分的阴影，其对应的分割边界为 5.1。求尾部阴影部分的面积。

E

答案：通过使用统计软件，不难得出尾部阴影部分的面积为 0.4038。如果使用本书后的对照表，我们会发现尾部面积是一个大于 0.3 的值，但无法进一步精确。

指导练习 6.28

G

图 6.8(d)展示了一个自由度为 7 的卡方分布，阴影区域的边界是 11.7。求尾部阴影的面积。[1]

指导练习 6.29

G

图 6.8(e)展示了一个自由度为 4 的卡方分布，阴影区域的边界是 10。求尾部阴影的面积。[2]

指导练习 6.30

G

图 6.8(f)显示了一个自由度为 3 的卡方分布，阴影区域的边界是 9.21。求尾部阴影的面积。[3]

[1] 尾部面积的精确值为 0.1109。通过对照表我们会得到面积在 0.1 和 0.2 之间。
[2] 尾部面积的精确值为 0.0404。通过对照表我们会得出面积在 0.02 和 0.05 之间。
[3] 尾部面积的精确值为 0.0266。通过对照表我们会得出面积在 0.02 和 0.05 之间。

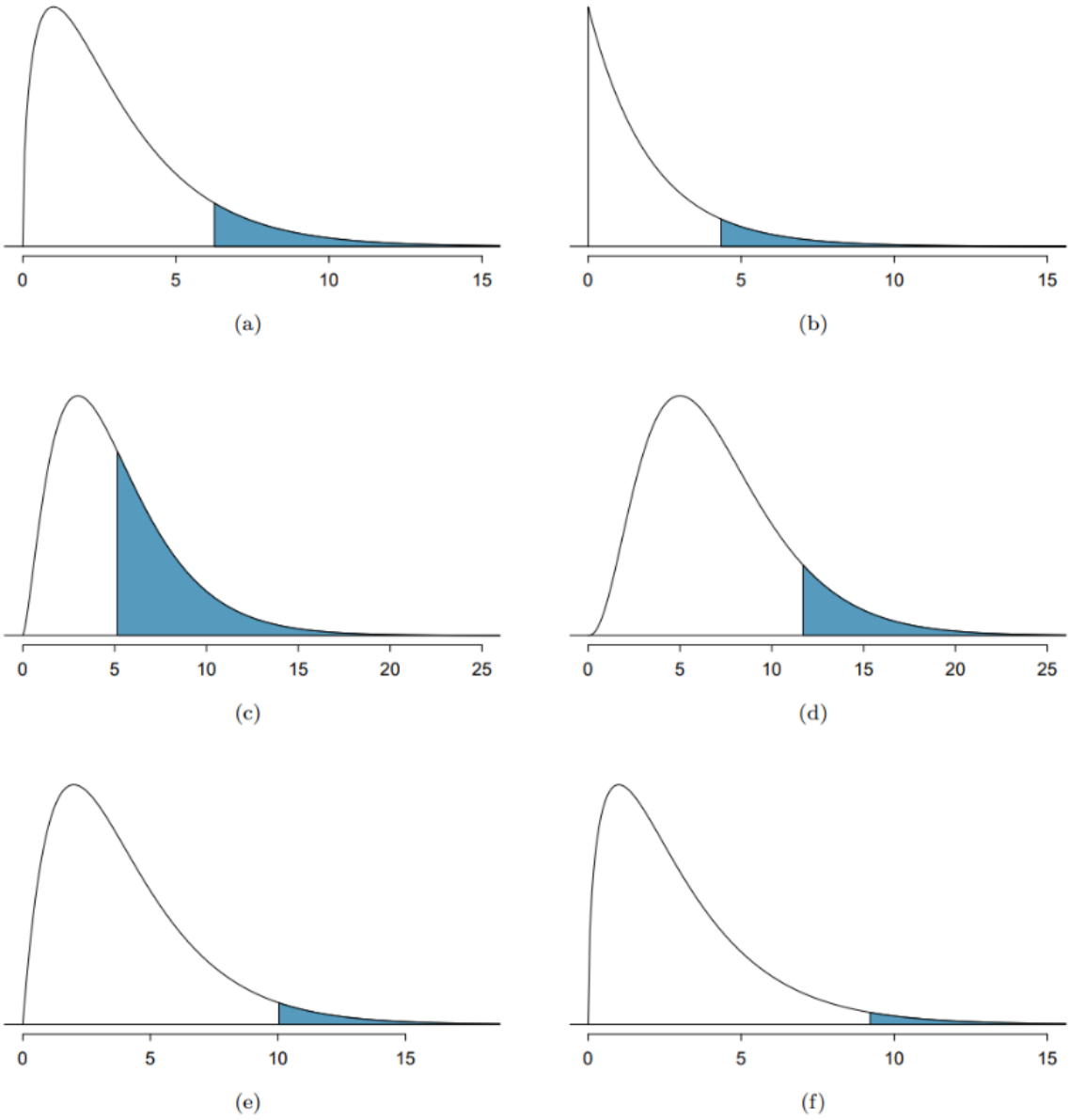


图 6.8：(a) 自由度为 3 的卡方分布，阴影位于 6.25 以上的部分；(b) 自由度为 2，阴影边界为 4.3；(c) 自由度为 5，阴影边界为 5.1；(d) 自由度为 7，阴影边界为 11.7；(e) 自由度为 4，阴影边界为 10；(f) 自由度为 3，阴影边界为 9.21。

6.3.4 寻找卡方分布的 p 值

在第 6.3.2 小节中，我们在评估陪审员选择中是否存在种族偏见时，确定了一个新的检验统计量：卡方值 X^2。当时原假设是假设陪审员是随机选择的，备择假设是在选择过程中存在种族偏见。尽管我们已知，一个较极端的 X^2 值将表明有强有力的证据支持备择假设，然而，我们无法量化在原假设为真的前提下出现如此极端取值的几率是多少。这就是为什么我们需要引入卡方分布。在该例中，如果原假设为真，那么 X^2 将遵循自由度为 3 的卡方分布。在特定条件下，统计量 X^2 遵循自由度为 $k-1$ 的卡方分布，其中 k 是分组的数量。

示例 6.31

在陪审员的例子中，种族有多少个分组？X^2 所使用的卡方分布应该对应怎样的自由度？

E

答案：在陪审员的例子中，有 $k = 4$ 类：白人、黑人、西班牙裔和其他。根据上述规则，如果原假设为真，那么检验统计量 X^2 应该遵循自由度为 $k - 1 = 3$ 的卡方分布。

就像我们在之前正态分布的案例中无一例外都检验了样本量条件一样，我们也必须确认卡方分布的样本量条件。在陪审员的例子中，原假设值为 198、19.25、33 和 24.75，这些样本量都大于 5，所以我们可以将卡方模型应用于检验统计量 $X^2 = 5.89$。

示例 6.32

如果原设为真，检验统计量 $X^2 = 5.89$ 将与自由度为 3 的卡方分布密切相关。请利用卡方分布和卡方值来确定 p 值。

E

答案：图 6.9 展示了对应的卡方分布和 p 值。 不难看出卡方值（对应了图中的阴影左边界）越大越意味着有足够的证据拒绝原假设。我们将尾部区域涂上阴影来代表 p 值的大小。通过使用统计软件（或**附录 C.3** 的表格），我们可以确定该区域的面积为 0.1171。一般来说，在得出这么大的 p 值的时候，我们不会拒绝原假设。换句话说，这些数据并没有提供令人信服的证据，证明在选择陪审员时存在种族偏见。

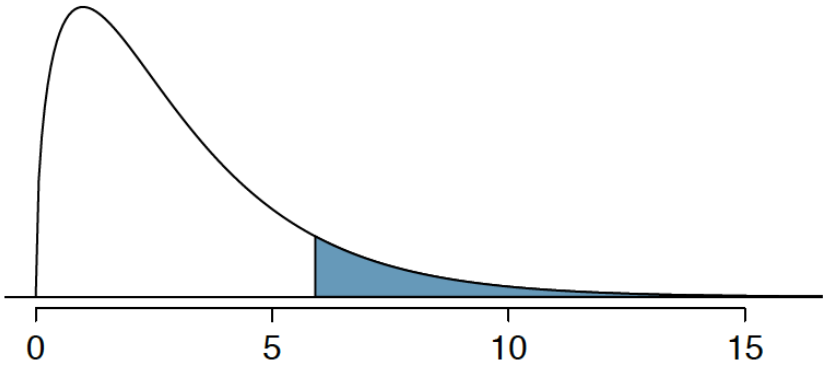


图 6.9：阴影部分面积表示了自由度为 3 的卡方分布和对应 p 值。

一维表格数据的卡方检验

假设我们要检验：是否有令人信服的证据表明，在 k 个组别中观察到的一组计数 $O_1, O_2, ..., O_k$ 与在原假设下得出的一组 k 个期望值有显著差别。我们将基于原假设的预期统计称为 $E_1, E_2, ..., E_k$。如果每个期望值至少为 5，并且原假设为真，那么下面的检验统计量就遵循自由度为 $k - 1$ 的卡方分布。

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k}$$

这个检验统计量，即卡方值的 p 值是通过观察这个卡方分布的尾部区域面积找到的。我们考虑尾部是因为 X^2 的值越大，就说明观测值和期望值相差越大，进而越能提供证据反对原假设。

卡方检验的条件

在进行卡方检验之前，必须检查两个条件。

独立性 Independence：每个为表格提供计数的观测值必须独立于表格中的所有其他观测值；

样本大小/分布 Sample size / distribution：每个特定组别场景必须至少有 5 个观测值。

如果我们不检查以上两个条件，就可能导致检验结果存在偏误。

注意：当我们处理一个只分成两组的表格时，应该选择其中一组并使用第 6.1 节介绍的单比例方法进行假设检验的计算。

6.3.5 评估分布的拟合度

我们在本节的最开始就提出了，可以将卡方检验框架应用于评估数据是否服从特定分布，或者说评估某个统计分布模型是否适用于某个数据集。

标准普尔 500 指数，简称 S&P500 或标普 500，是一个反应美国股票市场行情的重要指数。自 1957 年创建起，它一直被视作一个广泛而具有代表性的股票市场基准。往往我们只需要观察标准普尔指数的变动，就可以整体把握股票市场的波动情况。例如在 2020 年 3 月，受疫情和全球油价大幅波动的影响，在一个月内标普 500 指数就触发了 4 次熔断[1]，标志着美国股票市场短期危机的开始。标普 500 指数涵盖了 500 只代表性的普通股，这些股票的总市值约占整个市场的 80%。现在我们通过分析 10 年间标普 500 的数据，来试着判断股票市场当天的表现是否会受其昨天表现的影响。

[1] 美国标普 500 的熔断机制指的是当市场在短时间内出现大幅度波动时，交易所为了防止恐慌扩大和保护投资者利益而暂停或限制交易的措施。熔断制度在美国股市有明确的规定，分为三个级别的熔断：7%、13% 和 20%，而触发的基准指数就是标普 500。

该问题乍一听会觉得有点复杂，但我们可以用卡方检验的框架来研究。我们根据当天指数的涨跌情况，可以把每一天标记为上涨（Up）或下跌（D）。如下方表格所示，第一天标普 500 指数涨了 2.52%，那么我们把当天标记为 Up，紧接着的一天指数下跌 1.46%，我们把这天标记为 D。然后我们根据第一行列出的指数变化，可以得到其下两行的信息，分别是记入标记结果和每次指数上涨日距离上一次上涨日的间隔天数。对第三行的理解是，例如 0.51%和 2.52%这两次上涨间隔了两天，所以记录 2，而表格最右侧的 1.71%和上次上涨 1.10%间隔了 4 天，所以记录 4。

指数变化	2.52	-1.46	0.51	-4.07	3.36	1.10	-5.46	-1.03	-2.99	1.71
结果	Up	D	Up	D	Up	Up	D	D	D	Up
间隔天数	1	-	2	-	2	1	-	-	-	4

如果日与日间的股票波动确实是相互独立的，那么从理论上讲，每次上涨距离上一次上涨的间隔天数（以下简称间隔天数）应该遵循一个几何分布。因为几何分布描述了在 k 次伯努利试验中首次观察到第一个成功的概率的分布。即几何分布的横轴是试验次数，纵轴是概率（或给定总试验数时的频数）。我们讨论的标普 500 指数变动的情形中，若假设股票市场日与日波动相互独立，那么每天出现上涨或者下跌就像是一次伯努利试验，而间隔天数就对应了横轴的试验次数，10 年间每种间隔天数出现的频数则对应了几何分布纵轴的频率/频数。

于是我们统计了所有间隔天数（几何分布的横轴）对应的出现次数：

间隔天数	1	2	3	4	5	6	7+	总计
观察到次数	717	369	155	69	28	14	10	1362

图 6.10：S&P500 指数在 10 年间出现的间隔天数。

对于 S&P500 股票指数上涨日的间隔天数来说，如果股票活动在每一天都是独立的，而且单日出现的概率是恒定的，那么我们就会期望这个间隔天数遵循一个几何分布。把该情景套用到假设检验的框架中，我们可以得到：

H_0 原假设：股票市场在某一天的上涨或下跌是独立于所有其他交易日的。我们研究的是两个上涨日之间的间隔天数。在该假设下，间隔天数应该遵循一个几何分布。

H_A 备择假设：股票市场每一天的上涨或下跌并非独立于其他交易日。由于在原假设下，两个上涨日之间的间隔天数将遵循几何分布，因此我们要观测到的分布和期望几何分布间的明显偏差来支持选择备择假设。

该研究结果显然对股票交易者们有意义：如果过去的交易信息对于判断今天会发生什么是有用的，那么这种信息就可能会帮助交易者们做判断。

在图 6.11 和图 6.12 中，我们展示了所观察到间隔天数的数据统计，以及它与期望情形下几何分布的数据的对比。在进行期望几何分布的计算中，我们需要知道「指数单日上涨」这一伯努利试验事件发生的概率，而在我们所观察的 10 年的时间区间中，通过统计可知有 54.5%的天数 S&P500 指数是上涨的，所以我们也就把概率设置为 54.5%来进行模型推演计算。

由于卡方检验要求每组无论观测值还是预期值都至少有 5 个样本，所以我们把所有间隔天数至少为 7 天的情形放在一起作为一组，以确保达到使用卡方检验的基本要求。图 6.11 中展示了实际每组统计数据和几何分布模型的期望计数的直观数字比较，并在下方的标题中讨论了计算期望值的方法的方法。关于期望值的计算，大家可以回顾第 4 章中关于几何分布的计算公式。在此我们也简单描述一下期望值的统计方法：首先，通过几何分布的模型计算出该组对应的概率大小，然后把这个概率当做原假设比例去乘以总计数，以得到期望计数。

间隔天数	1	2	3	4	5	6	7+	总计
观察值	717	369	155	69	28	14	10	1362
几何模型预期	743	338	154	70	32	14	12	1362

图 6.11：间隔天数的分布。基于几何模型的期望值显示在第二行。对于每个期望值，我们根据几何模型确定间隔天数为 D 的概率 $\left(P(D) = (1-0.545)^{(D-1)}(0.545)\right)$，然后乘以观察值的总数 1362。例如，根据几何分布模型，两次上涨之间间隔 3 天的情形约占 $0.455^2 \times 0.545 = 11.28\%$ 的比例，这相当于 $0.1128 \times 1362 = 154$ 。

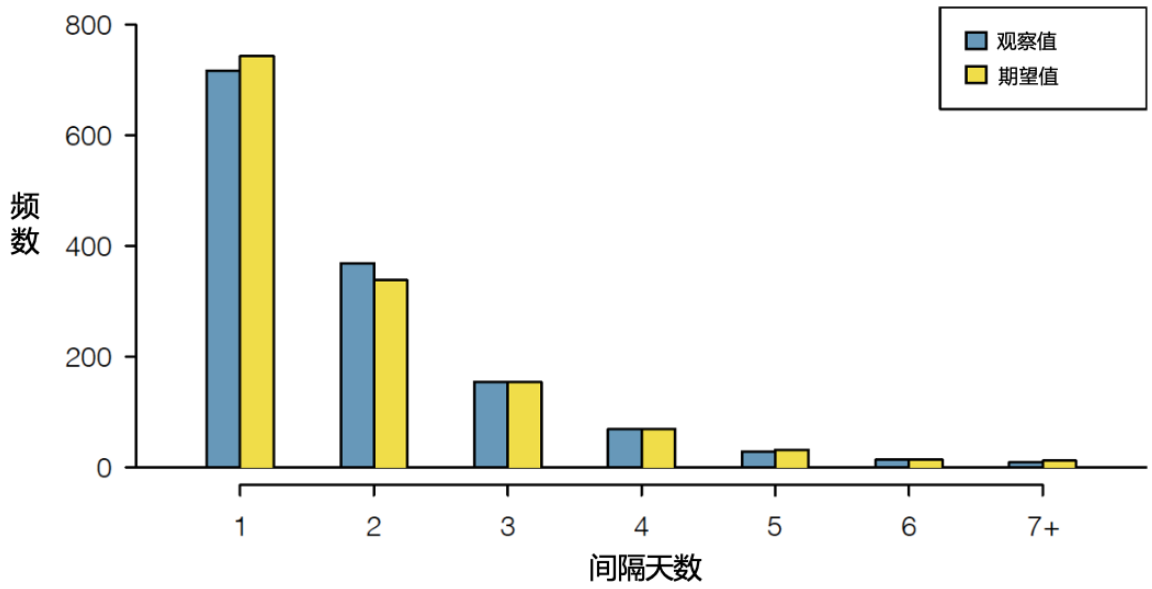


图 6.12：每个间隔天数的观察值和期望值的并列柱形图。

示例 6.33

从图 6.12 中你是否能观察到任何不寻常的偏差？你能仅仅通过观察来判断它们是偶然还是代表着统计学显著差异的吗？

答案：观察到的统计值与几何分布期望值之间的差异并不明显。也就是说，仅从图中我们无法判断这些偏差是否是由于偶然因素造成的。因此仅从图中观察，我们还没能提供令人信服的证据来拒绝原假设。我们还需要对图 6.11 中的统计进行卡方检验。

指导练习 6.34

请根据图 6.11 提供的各组观测数据 ($O_1 = 717$, $O_2 = 369$, ...) 和几何分布假设下计算的期望数据 ($E_1 = 743, E_2 = 338, ...$)，请计算卡方检验的统计量：$X^2$。[1]

指导练习 6.35

因为各组的观测值和期望值都至少为 5，所以我们可以用卡方分布。那么自由度应该选的多少？[2]

示例 6.36

如果原假设为真，即各分组间隔天数的统计值服从几何分布，那么卡方检验统计量 $X^2 = 4.61$ 将遵循 $df = 6$ 的卡方分布。请利用这些信息计算 p 值。

答案：图 6.13 显示了自由度为 6 的卡方分布，4.61 对应的边界值和 p 值对应的尾部阴影面积。通过使用统计软件，我们可以发现 p 值为 0.5951。因此，我们没有足够的证据来拒绝原假设。也就是说，我们不能拒绝标普 500 指数每天的上涨和下跌是相互独立的，昨天的股指表现并不影响第二天。

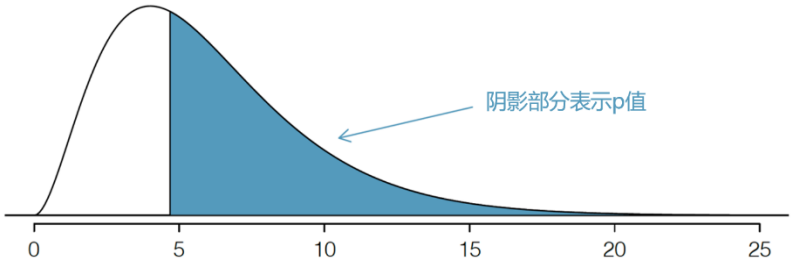


图 6.13：自由度为 6 的卡方分布，阴影部分代表股票分析的 p 值。

示例 6.37

在指导练习 6.36 中，我们没有拒绝股指表现日与日之间相互独立的原假设。请试着解释为什么这一点如此重要？

答案：如果市场连续几天都在下跌，那么我们的潜意识里可能会认为市场即将迎来一个上涨日，这种判断会影响我们的投资决策。然而，我们还没有发现强有力的证据表明市场有这样的属性。如果市场每天的表现其实是独立的，我们最好还是不要基于昨天的涨跌妄加推测第二天的股指涨跌。

[1] $X^2 = \frac{(717-743)^2}{743} + \frac{(369-338)^2}{338} + \cdots + \frac{(10-12)^2}{12} = 4.61$

[2] 由于总共有 $k = 7$ 个类别，所以我们用 $df = k - 1 = 6$。

6.4 检验二维表中的独立性

当我们购买二手产品时（例如汽车、电脑、教科书等等），我们有时会认为这些产品的卖家会直言不讳地说明这些产品的潜在问题，但实际情况却并不一定是这样。我们来介绍这么一个语言心理学的案例：研究人员在一项研究中招募了 219 名参与者，这些参与者的任务是努力售出一些已知存在硬件问题的二手 iPad。参与者被要求尽可能多地从销售这些有问题的二手 iPad 上获得利润。除了 10 美元的固定奖励之外，参与者被许诺可以分配到个人销售利润的 5%。而这项研究中，研究人员还将偷偷扮演消费者，然后会在购买前向参与的销售人员提一个问题，而研究的真实目的就是看哪些类型的提问会让卖家更容易坦诚 iPad 的硬件问题。

由于卖家并不知道研究人员会偷偷扮演消费者，所以他们的行为很大程度会是个人心理和外部诱因共同作用的结果。而在研究人员去购买 iPad 的时候，按剧本会讲出类似以「好吧，所以你的 iPad 已经用了 2 年了..... 」开头的台词，然后用下面三个问题中的一个结束：

- 普通问题：能再告诉我一些关于它性能状态的信息吗?
- 积极暗示：所以它没啥毛病，是吧?
- 消极暗示：能不能给我说说它有啥毛病?

在这个情境中，我们会关注面对买家不同的问题，卖家是否会选择告诉买家 iPad 硬件存在问题的事实。图 6.14 展示了最终结果。而根据研究最后的统计显示，被以消极暗示的方式提问，即被问到「能不能给我说说它有啥毛病」时，卖家更容易向买家披露 iPad 硬件问题的情况。从统计学角度，我们不禁要问：这样的统计结果仅仅是次偶然，还是提问方式的不同确实会影响个人行为?

	普通问题	积极暗示	消极暗示	总计
披露问题	2	23	36	61
隐藏问题	71	50	37	158
总计	73	73	73	219

图 6.14：关于 iPad 研究的统计二维表摘要。

一维表与二维表的区别

一维表描述了单个变量中每种可能结果的数据。二维表描述的则是两个变量交叉统计的结果。对于一个二维表，我们通常想知道展示在表中的两个变量是否有任何关系？换句话说，看它们是否是相关的?

iPad 试验中，我们想要研究是否有统计学上显著的证据表明，卖方披露 iPad 的硬件问题和买方提问方式有关。换言之，我们的目标是建立假设，完成这两个变量间的统计学检验。

6.4.1 二维表中的期望值计算

与一维表一样，我们需要计算二维表中每个单元格的期望值作为基准，好进行统计数字的比较以及检验统计量的构建。与之前一样，我们需要明确该场景下的原假设。这里的原假设显然是：二维表的两个变量相互独立，也就是「卖方是否披露硬件毛病」和「买方沟通时的提问方式」二者不相关。

示例 6.38

(E) 从实验中，我们可以计算出所有卖家中披露 iPad 硬件问题的比例为 61/219 = 0.2785。如果卖方选择披露与否与买家选择提问方式无关，那么无论哪种提问方式，都应该对应 27.85%的卖家会坦诚披露 iPad 硬件毛病。按照这个逻辑推导，请试着计算出在普通提问方式一组的 73 人中，若原假设为真，预计有多少人会坦诚 iPad 已存在的硬件问题?

答案：按照原假设逻辑，在这一组的 73 人中，应该会有 0.2785 × 73 = 20.33 个卖家会坦诚披露问题。显然，我们实际观察到的情况比这要少，尽管目前还不清楚这是否是偶然因素导致的，亦或是不同提问方式真的会影响卖家的选择。

指导练习 6.39

(G) 如果原假设为真，即提问方式不影响卖家的披露选择，那么总有大约 27.85%的卖家会坦诚披露 iPad 的硬件问题。在此前提下，你能估计上面提到的 73 人中有多少卖家会选择隐瞒问题吗？[1]

我们可以用类似示例 6.38 和指导练习 6.39 中的策略，针对二维表的每组进行原假设期望值计算。由于之前我们按照提问方式把卖家分成了人数相同的三部分，所以该二维表的期望值的计算也会变得相对容易，因为每列的期望值数据应该是相同的。如此一来，我们就构建出了图 6.15，该图与图 6.14 结构完全一致，只是原假设下预期的人数被加在了括号里。注意，为了统计学计算需要，我们在此一定程度上忽略了数字的现实意义，否则统计的人数是不可能出现小数的。另一种对小数的解释是这里是数学期望值，仅代表了统计学意义的理论数字，而不代表现实中要有 20.33 个人最后选择披露硬件问题。

	普通问题	积极暗示	消极暗示	总计
披露问题	2 (20.33)	23 (20.33)	36 (20.33)	61
隐藏问题	71 (52.67)	50 (52.67)	37 (52.67)	158
总计	73	73	73	219

图 6.15: 观测计数和期望计数。

[1] 我们预计人数为(1 - 0.2785) × 73 = 52.67.。这个结果和指导练习 6.38 的结果一样是一个分数。

希望通过上面的示例和指导练习，你能对二维表的期望值计算有一定的感觉。一般来说，二维表的期望统计数字可以用行总计、列总计和表总计来计算。例如，如果按提问方式划分成的三组之间相互独立，那么每一列的行分布都应该满足 27.85%的比例。这里的 27.85%又是通过行总计和表总计计算得出的：

$$0.2785 \times (\text{第 1 列列总计：}73) = 20.33$$
$$0.2785 \times (\text{第 2 列列总计：}73) = 20.33$$
$$0.2785 \times (\text{第 3 列列总计：}73) = 20.33$$

如果我们进一步剖析一下前图中蓝色数字 20.33 的计算方法，它还可以被写成下面的式子。相比较上面的式子，式中的 0.2785 被「行总计除以表总计」的逻辑给替换掉了：

$$(\text{第 1 行行总计/表总计：}61/219) \times (\text{第 1 列列总计}) = 20.33$$
$$(\text{第 1 行行总计/表总计：}61/219) \times (\text{第 2 列列总计}) = 20.33$$
$$(\text{第 1 行行总计/表总计：}61/219) \times (\text{第 3 列列总计}) = 20.33$$

我们可以由此总结出一个通用的公式，当我们想要研究二维表中某单元格的期望值的时候，可以考虑用这个下方的通用公式来套用计算。

计算二维表格中的单元格期望值

为了确定第 i 行和第 j 列的预期值，计算公式如下：

$$\text{期望值}_{(\text{第 } i \text{ 行,第 } j \text{ 列})} = \frac{(\text{第 } i \text{ 行总计}) \times (\text{第 } j \text{ 列总计})}{\text{列总计}}$$

6.4.2 二维表的卡方检验

二维表的卡方值的求法与一维表的求法相同。对于每个单元格，我们先使用如下的方法计算单个统计量：

$$\text{一般公式：}\frac{(\text{观测值}-\text{期望值})^2}{\text{期望值}}$$

$$\text{第 1 行，第 1 列：}\frac{(2-20.33)^2}{20.33}=16.53$$

$$\text{第 1 行，第 2 列：}\frac{(23-20.33)^2}{20.33}=0.35$$

$$\cdots\cdots$$

$$\text{第 2 行，第 3 列：}\frac{(37-52.67)^2}{52.67}=4.66$$

将每个单元格的计算值相加，就可以得到整张表格的卡方值 X^2：

$$X^2 = 16.53 + 0.35 + \cdots + 4.66 = 40.13$$

与之前介绍一维表卡方检验时相同，该统计量也遵循卡方分布。但是，对于二维表来说，自由度的计算方法有些不同。在一维表中，我们只需要把分组数减 1 就可以得到自由度。对应到二维表中，直觉告诉我们应该把表格中的单元格数目统计出来减 1 得到自由度。但这样做并不正确。因为二维表是由两个变量构成的，每个变量在行或者列维度上其实都有自己的分组。所以如果真正严格仿照一维表案例去构建自由度，应该遵循如下公式：

$$df=(\text{列数}-1)\times(\text{行数}-1)$$

应用到我们的 iPad 例子中，自由度参数为：

$$df=(2-1)\times(3-1)=2$$

如果原假设为真（即提问方式对试验中的卖家没有影响），那么检验统计量 $X^2 = 40.13$ 就应服从合自由度为 2 的卡方分布。利用这些信息，我们可以计算出检验的 p 值，然后用可视化的思路把其对应的阴影面积在分布上标注出来，如图 6.16 所示。

计算二维表的自由度

当对二维表应用卡方检验时，我们使用：

$$df=(\mathrm{R}-1)\times(\mathrm{C}-1)$$

其中 R 是表格中的行数，C 是列数。

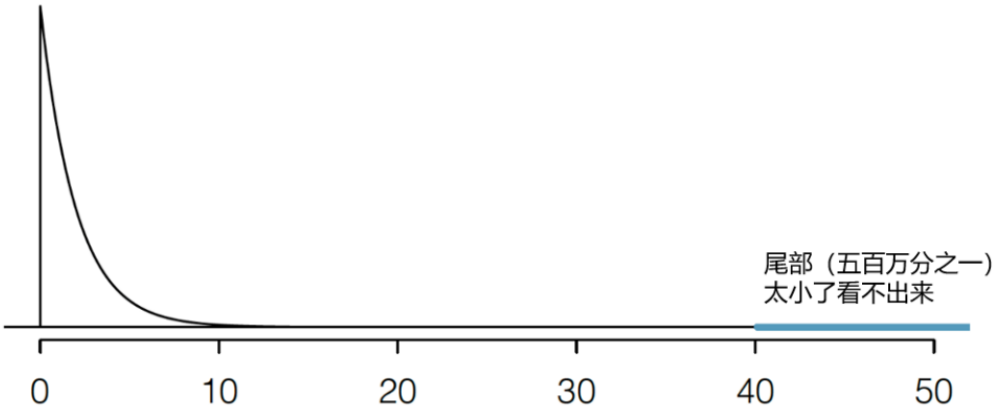


图 6.16：当 $df = 2$ 时，$X^2 = 40.13$ 的 p 值的可视化展示。

示例 6.40

找出 p 值，并针对以下问题得出结论：不同的提问方式会影响卖家披露电子产品问题的选择吗？

E

答案：通过计算，我们可以计算出自由度为 2 的卡方分布的 $X^2 = 40.13$，尾部面积为 0.000000002。(如果使用附录 C.3 中的表格，我们会发现 p 值小于 0.001)。在 5%的置信水平下，我们是把 p 值和 0.05 作比较。因为 p 值显著低于 0.05，所以我们拒绝原假设。也就是说，我们得出的结论是：数据提供了令人信服的证据，证明所问的问题确实影响了卖家披露 iPad 硬件问题的选择。

示例 6.41

图 6.17 总结了一个试验的统计结果，该试验评估了正在接受二甲双胍治疗的 10-17 岁的 2 型糖尿病患者的三种疗法。所考虑的三种治疗方法分别是：（1）继续使用二甲双胍治疗，（2）二甲双胍联合罗格列酮治疗，或（3）生活方式干预疗法。疗法的结论有两种：即要么血糖没有得到有效控制（失败），要么得到了有效控制（成功）。那么该试验研究的合理统计学假设是什么？

E

答案：合理的原假设是：三种治疗方法的效果没有差别。而备择假设是：三种治疗方法之间的效果有差异，例如，也许二甲双胍联合罗格列酮治疗比生活方式干预疗法效果更好。

	失败	成功	总计
生活方式干预	109	125	234
二甲双胍治疗	120	112	232
联合疗法	90	143	233
总计	319	380	699

图 6.17：2 型糖尿病研究的结果。

指导练习 6.42

接下来我们就使用二维表的卡方检验来检验指导练习 6.41 中的假设。首先，请计算表格 6 个单元格中每格对应的期望值。[1]

指导练习 6.43

接着，让我们计算图 6.17 中数据的卡方值量。[2]

指导练习 6.44

由于我们有三行两列，我们的自由度为：$df = (3-1) \times (2-1) = 2$。利用 $X^2 = 8.16$，$df = 2$，来评估在 0.05 的显著性水平下，是否有足够证据拒绝原假设。[3]

---

[1] 第 1 列/第 1 行的期望值通过如下方式计算：第一行行总计(234)×第一列列总计(319)/表总计(699)，即 $\frac{234 \times 319}{699} = 106.8$；第 1 列第 2 行的计算采用相同的方法：$\frac{234 \times 380}{699} = 127.2$。同理，我们可以计算出第 2 行的期望值：105.9 和 126.1，还有第三行的期望值：106.3 和 126.7。

[2] 对于每一个单元格，我们采用 $\frac{(\text{观察值} - \text{预期值})^2}{\text{预期值}}$ 的公式来计算单格的统计量。例如，第 1 行和第 1 列的结果为：$\frac{(109-106.8)^2}{106.8} = 0.05$。将每个单元格的结果相加的最终结果即为卡方值：$X^2 = 0.05 + \cdots + 2.11 = 8.16$。

[3] 通过软件辅助计算，我们可以确定 p 值为 0.017。也就是说，我们拒绝原假设，因为 p 值小于 0.05，我们得出结论，在治疗 2 型糖尿病的血糖控制方面，至少有一种治疗方法比其他方法更有效或更少。