

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 Diez¹

Mine Çetinkaya-Rundel²

Christopher D Barr³

(译者)

王世尧⁴

李雪琦⁵

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 国际货币基金数据分析师, 跨越数据银河创始人

⁵ 世界银行数据模型顾问, 跨越数据银河联合创始人

(原著著作权声明)

Copyright © 2019. Fourth Edition.

Updated: November 12th, 2019.

原著 PDF 版本可以通过 openintro.org/os 网站免费下载。在遵守美国 CCI (Creative Common License) 著作许可的前提下，本书源文件可以通过 Github 获取。

目录

1. 走近数据	8
1.1 案例分析：使用颅内支架来预防脑中风	10
1.2 数据基础	13
1.3 抽样原则和策略	22
1.4 试验	31
2. 总结数据	35
2.1 研究数值型数据	37
2.2 研究分类数据	55
2.3 案例分析：疟疾疫苗	64

译者的话

翻译此书的初衷，是帮助广大对计量和统计感兴趣的小伙伴提供一份生动、有趣并且案例丰富的学习教材。我是在美国研究生的统计课堂上第一次接触到这本书，在下载免费电子版阅读之后颇为喜爱，狠狠心在亚马逊上购买了纸质版。而这本纸质版的 OpenIntro Statistics 也确实在之后的学习和研究中多次帮助到我。从学校毕业直到在国际货币基金组织工作，我还是会时不时把它从书架上拿下来，翻阅巩固一些基本定义，以避免工作中的低级分析失误。

我和雪琦都是武汉大学毕业的。最开始，在和武大学生交流的时候，我们经常向大家推荐这本统计学教材。可是，我们很快发现该书没有中文译本，这无疑增加了没有基础的同学入门的难度。幸而我们本科均在外国语言文学学院就读，对翻译略知一二。加上武大校内 Galadata 数据银河社社团成员们的帮助，我们得以在 2022 年 4 月正式开始对本书的翻译。希望能够在秋季开学之前截稿，以帮助即将到来的新一届大学生。

在工作中，我也逐渐发现数据和统计的知识越来越多地帮助到我。例如在回复 Q&A 邮件的时候，传统的方式是每次撰写新的回复。但是我逐渐意识到可以运用数据分析手段对过往 Q&A 类邮件进行统计，并用关键词判定程序来归类，匹配后台的数据库，进而提供相关度极高的答案。于是之后我开始更多使用这种数据筛选 > 匹配 > 自动化撰写回复的思路，果然极大提升了工作效率。此类的案例还有很多，总之数据思维不仅仅可以被用在专业的计量课题上，更可以帮助我们去分析自己所处的情境，发现可以提升效率的地方，并结合计算机工具实现技术和技能的跃迁。

正因如此，我希望把数据的价值更广泛地传播。和几位创始人一起，我们在 2021 年 8 月成立了一个以趣味科普为导向，以微信公众号为平台，以视频和互动推文等新媒体形式为载体的数据公益项目：跨越数据银河。本书的翻译就是我们进行新知科普的一个具体尝试。为了迎合日益增长的移动端阅读需要，我们将翻译稿件分章节不断发布在微信公众号「跨越数据银河」上。在本页下方和每个章节的开头，都可以找到公众号和推文合集的二维码，欢迎大家扫描在手机上打开阅读。每期推文中，我还加入了一些基于自己理解的选择题，希望能够帮助大家加深印象，巩固理解。

我要特别感谢 Galadata 数据银河社的成员们帮忙对翻译初稿进行审阅和批注，尤其是时任社团社长：吕润洋的整体协调和大力支持。此外，我还要感谢对成稿做出卓越贡献的老师和学生们：孙斐，陈睿，官杨颖，江静婧，杨梦迪和邹子晴。正是他们的支持和反馈，让翻译内容得到了极大的补充和完善。



跨越数据银河



系列推文合集

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

- 1.1 案例分析：使用颅内支架来预防脑中风
- 1.2 数据基础
- 1.3 抽样原则和战略
- 1.4 试验

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。



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



跨越数据银河



系列推文合集

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

(G)

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（1）试验组和（2）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	interest_rate	term	grade	state	total_income	homeownership
	贷款额	利率	期数	级别	州	总收入	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	Cahoun	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。

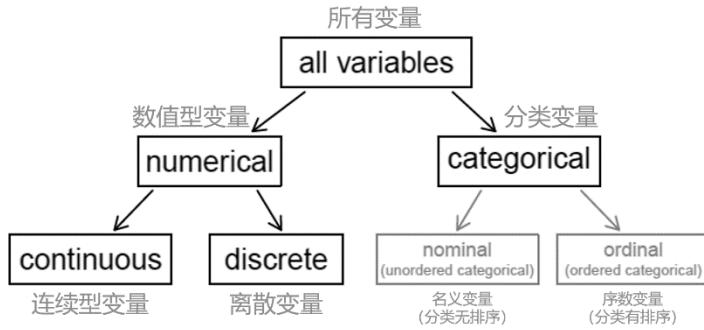


图 1.7：各类型变量细分

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。

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

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。

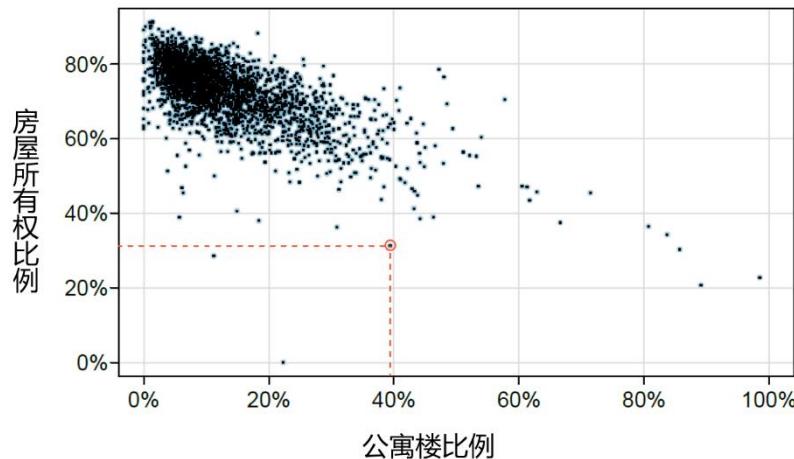


图 1.8：一张散点图：体现了美国各郡「房屋拥有者占总人口比例」和「公寓楼占全部楼房比例」关系的散点图。图中圈出的红点代表了佐治亚州的查塔胡奇郡，该郡的公寓楼占比为 39.4%，房屋所有权比例为 31.3%

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

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 指导练习 1.7 答案：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。

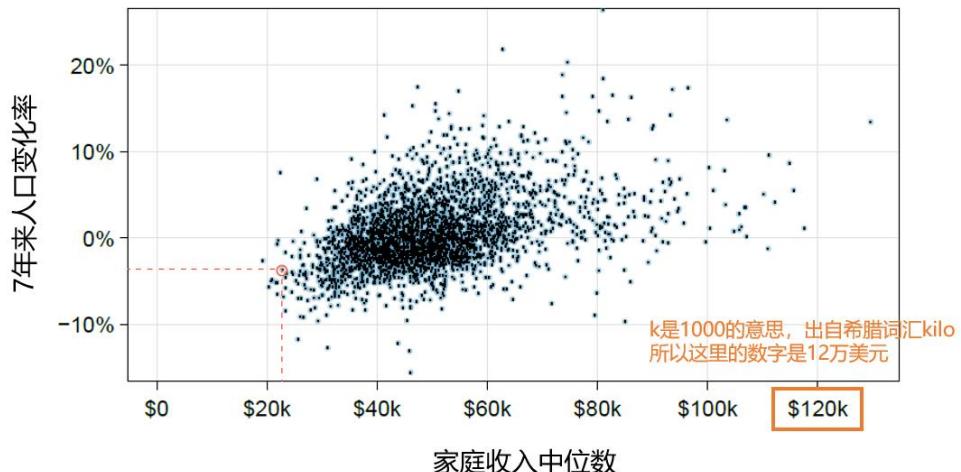


图 1.9：关于「7 年来人口变化率」和「家庭收入中位数」的散点图。肯塔基州的奥斯利郡在图中左下角被标出，它自 2010 年来人口缩减了 3.63%，家庭收入的中位数是 22736 美元

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 铁事证据

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。

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以考虑。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。

(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。

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。

总体：全部毕业生

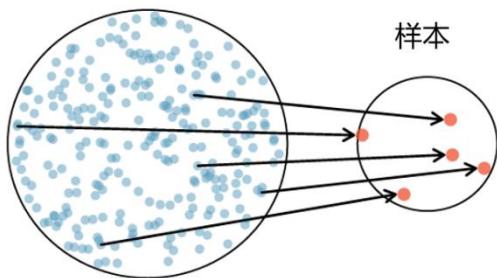


图 1.11：该图中，从总体（全部毕业生）中随机抽出了五位学生组成了一个样本

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。

总体：全部毕业生

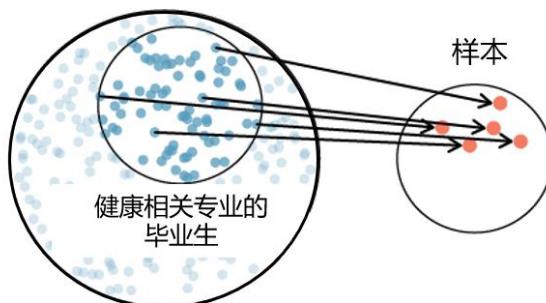


图 1.12：在取样时，一个营养学专业的学生可能会不经意间选择更多健康相关专业的毕业生，从而使得样本中各专业的占比不合理

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**。

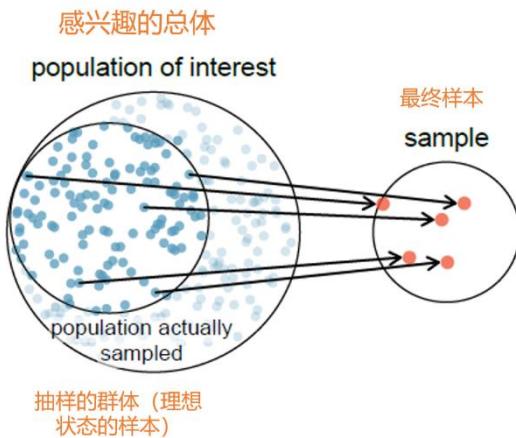


图 1.13：现实中，由于总有人（出于客观原因）无法或者（主观原因）不愿意接受调查，所以真正进入样本的可能只是总体中特定的“愿意提交答案的”个体¹。这个问题几乎不可能被彻底解决

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

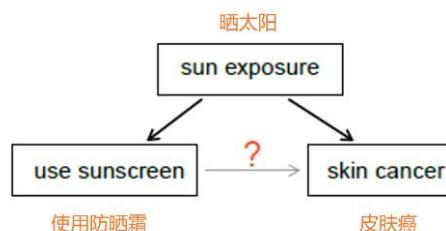
如果不对数据的产生过程进行干预，那么收集上来的数据就被称为**观察性数据 observational data**。例如，之前在章节 1.2 举例时用到的个人贷款数据（loan50 数据集）和美国州郡数据（county 数据集）都属于观察性数据。我们前面讲了，基于试验来进行因果推断是更合理的。对应的，使用观察性数据进行因果推断往往是有风险的，所以不被推荐。因此，观察性数据往往仅被用来去发现相关性，进而做出假设，而这些假设需要通过试验加以验证。

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。

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。

指导练习 1.13

G

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。

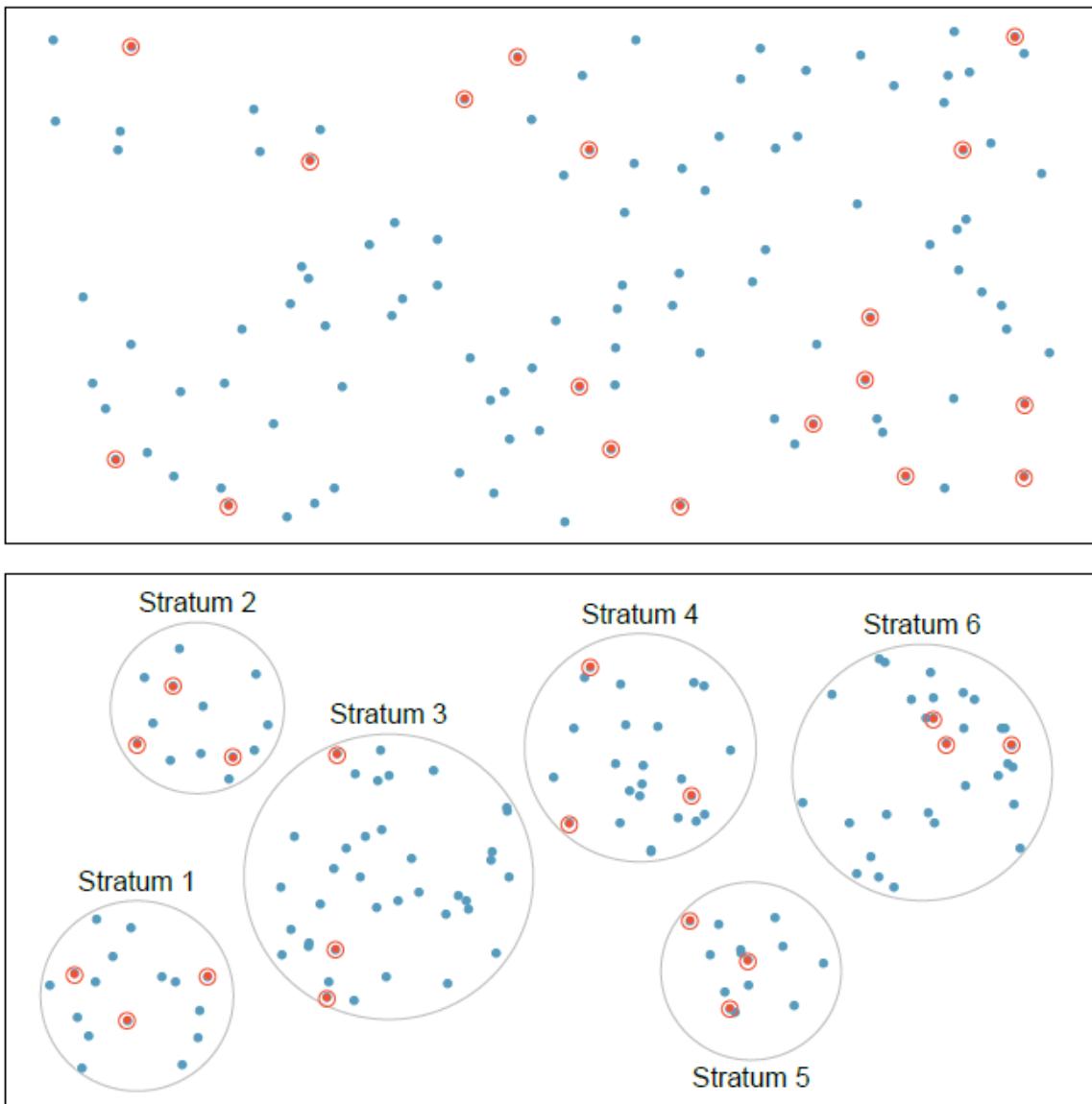


图 1.14：简单随机抽样（上方）和分层抽样（下方）的图解。在上面的图中，展示了使用简单随机抽样方法抽取 18 个个体。在下面的图中，我们可以看到分层抽样的概念设计：Stratum 是「层」的意思，总体首先被不重叠地划分成多个「子总体」，每个子总体也就被称为一个「层」。接着在每层中，我们使用简单随机抽样的方法抽取若干（图中是三个）个体

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。

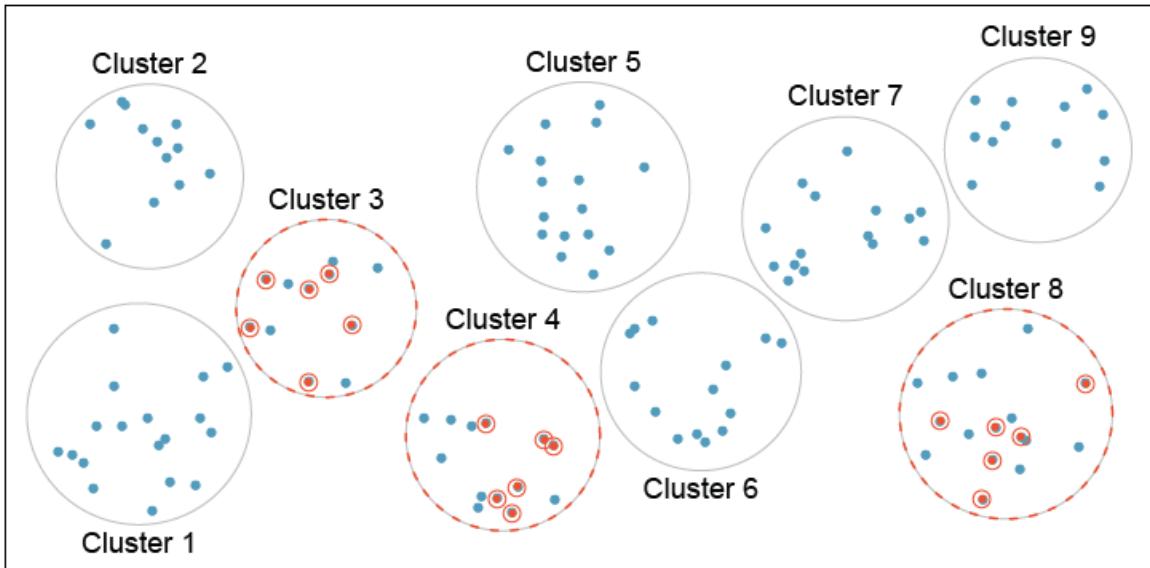
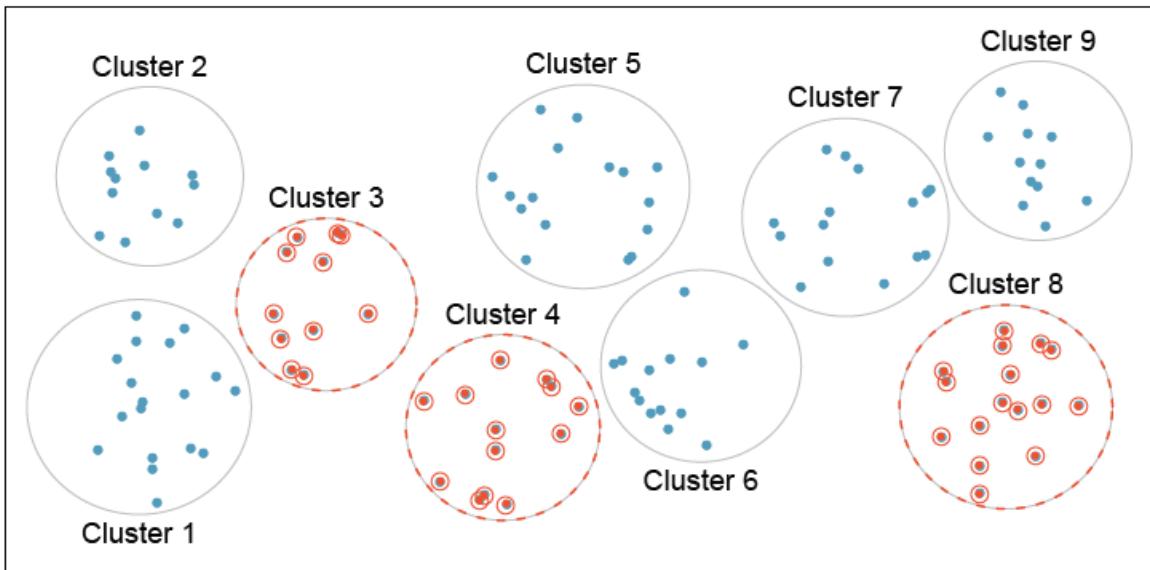


图 1.15：整群抽样（上方）和多阶段抽样（下方）的图解。在上面的图中，展示了整群抽样的概念：首先把所有数据划分成 9 个群集，然后从中取出 3 个群集，并抽取每个群集中所有的个体组成样本。在下面的图中，我们可以看到多阶段抽样的示例：同样是首先划分成 9 个群集，同样是抽取其中 3 个，唯一不同点在于多阶段抽样中我们对每个群集只抽取其中部分个体（而不是全部）加入样本。

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。

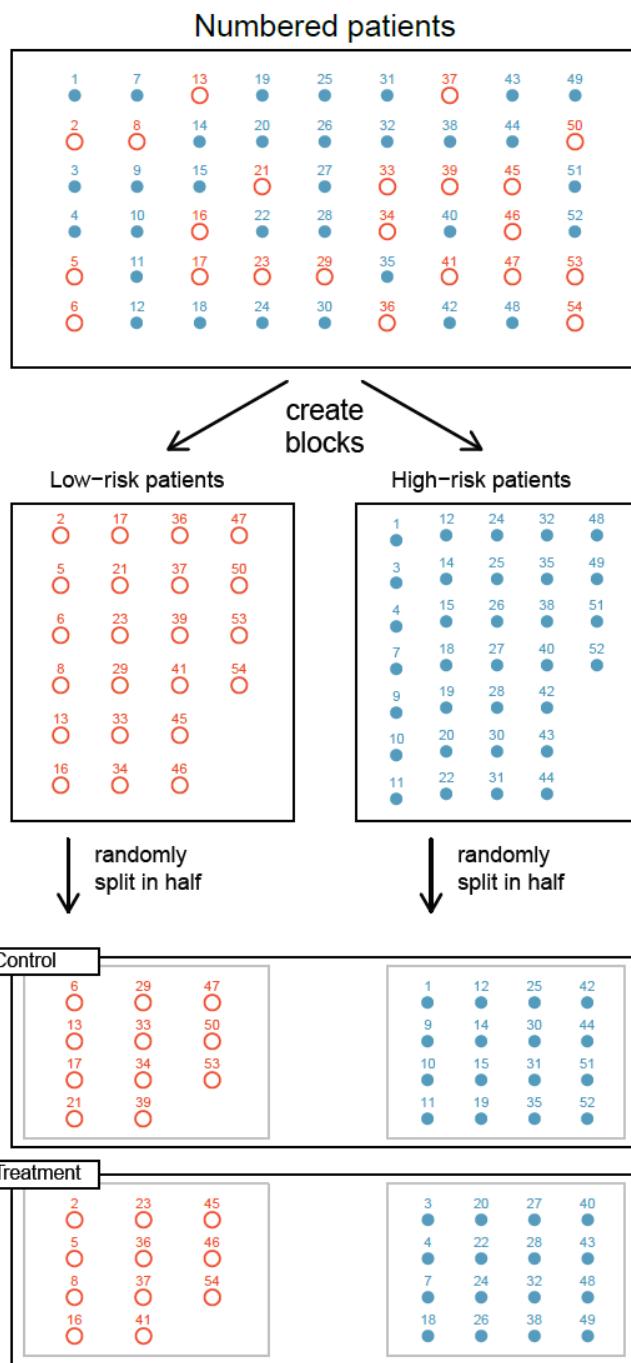


图 1.16：使用风险变量进行分区：所有患者首先被分成低风险区和高风险区，之后在每个区内再随机分成试验组和对照组。这样可以保证试验组和对照组中的两种风险患者相对比例是相同的。

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。以人作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

- 2.1 研究数值型数据
- 2.2 研究分类数据
- 2.3 案例分析：疟疾疫苗

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。



更多视频，演示文稿，和其他相关资源，请访问：

<http://www.openintro.org/os>



跨越数据银河



系列推文合集

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。

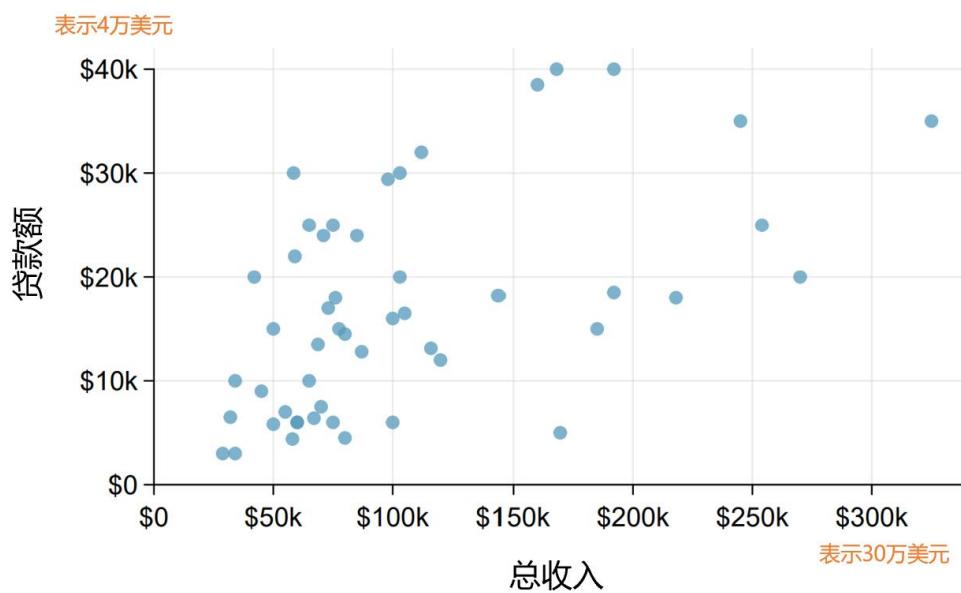


图 2.1：loan50 数据集中对比「总收入」和「贷款额」的散点图

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。

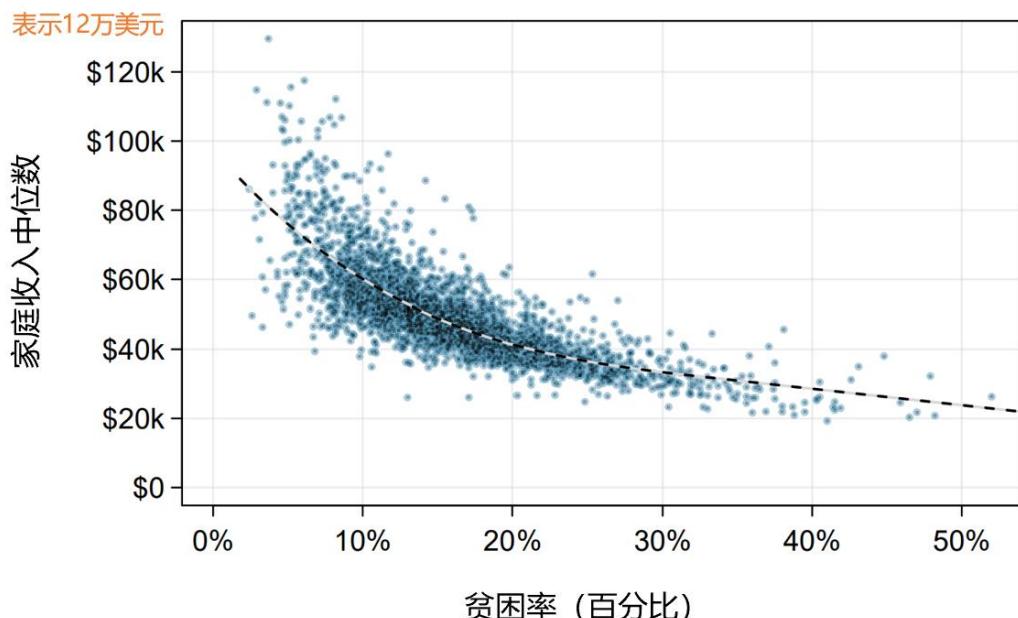


图 2.2：关于 county 数据集各郡「家庭收入中位数」和「贫困率」的散点图。我们已经找到了和数据拟合的统计模型，并在图上用一条虚线标识了出来

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： \cap ）？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。

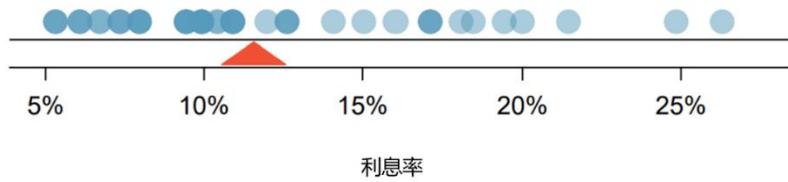


图 2.3：loan50 数据集中「利息率」变量的点图，该分布的均值在图上以红色三角标识

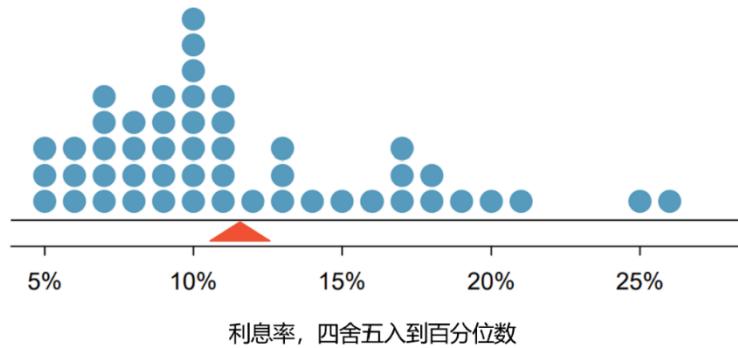


图 2.4：loan50 数据集中「利息率」变量的堆积点图，该图中利息率被四舍五入到最近的百分位数，分布均值同样用红色三角标出

[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。

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

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：

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

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

指导练习 2.4

(G) 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

指导练习 2.5

(G) 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 $x/y/z/\dots$ 。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 。

示例 2.6

(E) 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：

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：

E 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

E 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。

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

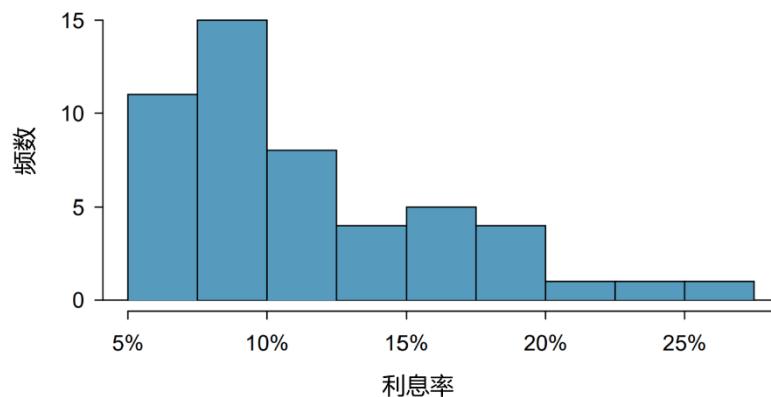


图 2.6：「利息率」数据分布直方图，可以看出数据明显呈现右偏趋势

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**。

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 mode⁴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，分别叫做：单峰 unimodal、双峰 bimodal 和多峰 multimodal。

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容易看出偏度的。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。

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。

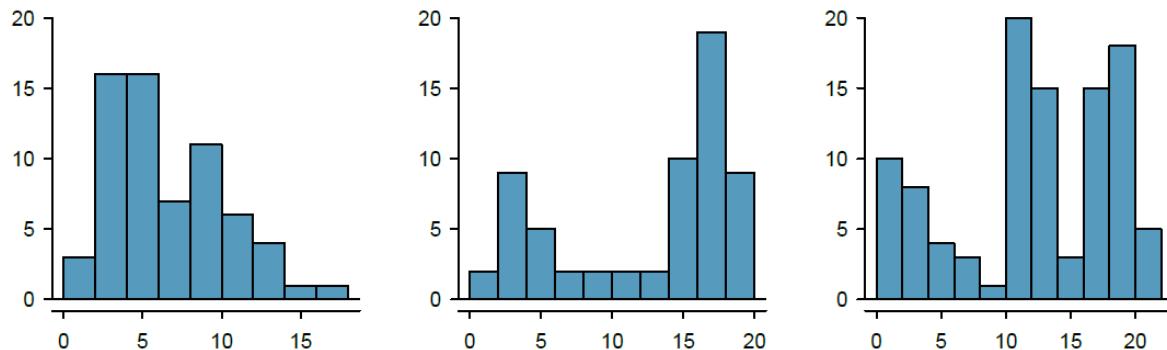


图 2.7：针对「明显突起」的山峰个数，从左到右分别对应了：单峰分布、双峰分布和多峰分布。
对于最左边的这张图，因为我们只考虑「明显突起」的山峰个数，而非任何山峰都考虑，所以它只算单峰分布

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。

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。

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：

$$x_1 - \bar{x} = 10.90 - 11.57 = -0.67$$

$$x_2 - \bar{x} = 9.92 - 11.57 = -1.65$$

$$x_3 - \bar{x} = 26.30 - 11.57 = 14.73$$

⋮

$$x_{50} - \bar{x} = 6.08 - 11.57 = -5.49$$

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, \dots, 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。

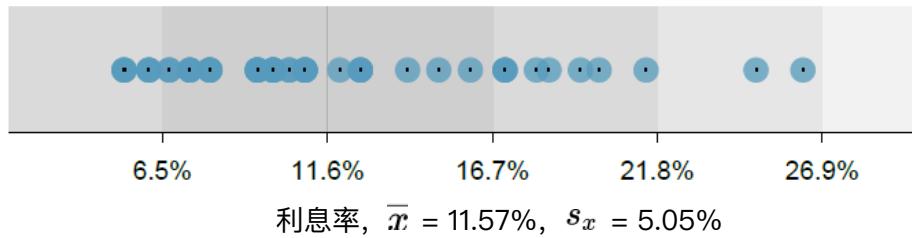


图 2.8：对于利息率这个变量，50 笔贷款中，34 笔的利息率都落在（离均值的）一个标准差以内，48 笔都落在（离均值的）两个标准差以内。通常情况下，70% 的观测值都落在一个标准差以内，95% 都在两个标准差以内，但也并非所有情况都是这样

¹ 译者注：至于到底为什么用 $n - 1$ 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。

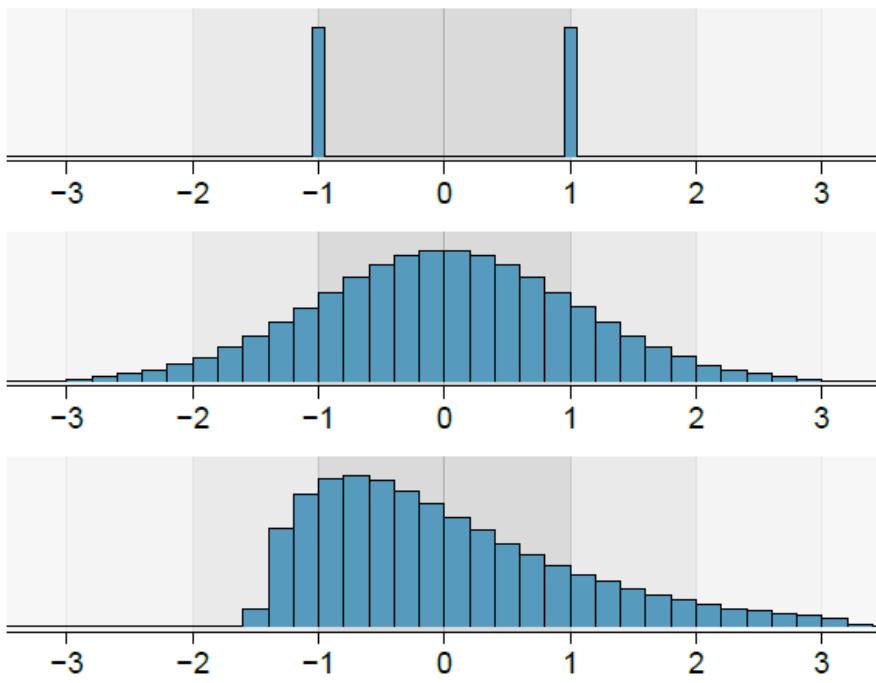


图 2.9：三种非常不同的分布，它们的均值和标准差完全相等，均值 $\mu = 0$ ，标准差 $\sigma = 1$ 。

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 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。

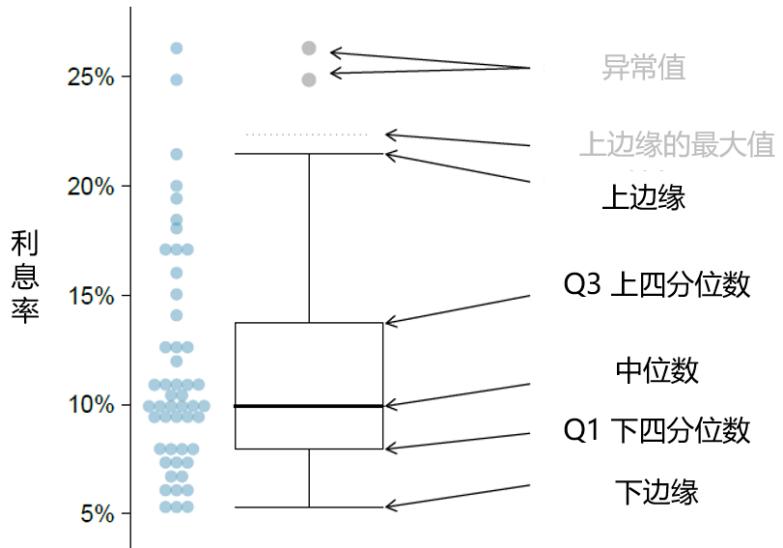


图 2.10：「利息率」变量的竖直点图和箱形图（带标签）

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 Q_1 和中位数之间？又有多少落在中位数和 Q_3 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）。图中箱的高度，它的计算公式是：

¹ 分别各是 25%。

G

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(1)Q1, (2)Q3 和(3)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：

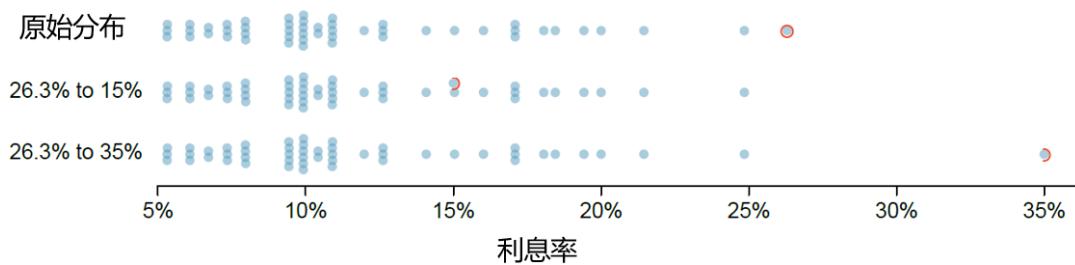


图 2.11：原始「利息率数据」和修改后的数据的对比点图

情形	稳健		不稳健		
	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，谁更容易受到极端值的影响？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。

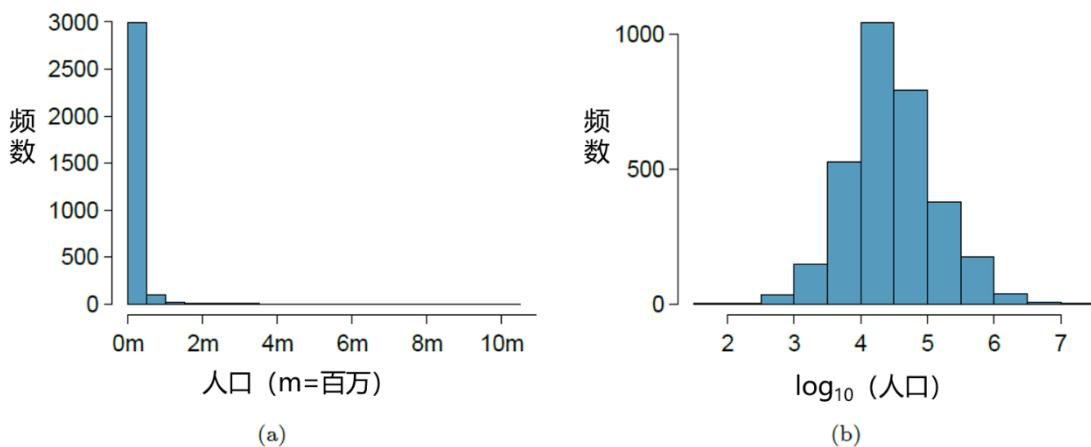


图 2.13：(a) 基于美国各郡人口的直方图；(b) 基于各郡人口取（以 10 为底的）对数之后的直方图。

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是很常见的方法。进行数据转换的目的包括：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。

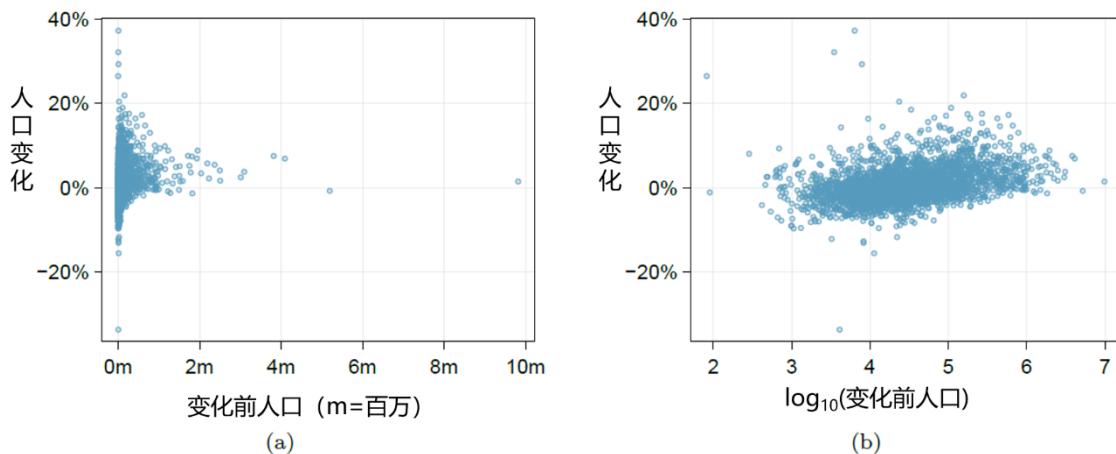


图 2.14：(a)基于人口变化百分比和变化前人口数量之间关系的散点图；(b)基于人口变化百分比和取对数后的变化前人口之间关系的散点图

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发现吗？

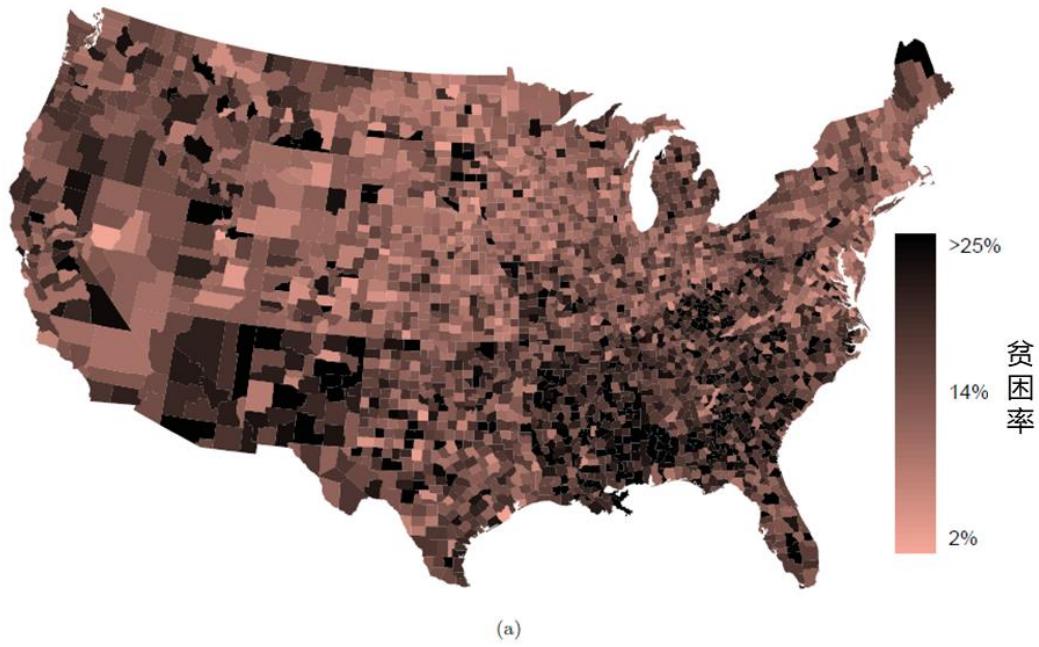
E 答案：可以看出，贫困率在某几个地方显著地高。具体点儿说，在很靠南的位置，例如亚利桑那州和新墨西哥州的一些郡县的颜色明显比别的地方要深。此外，还有其他部分区域，例如密西西比州和肯塔基州的贫困率也较高。

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。

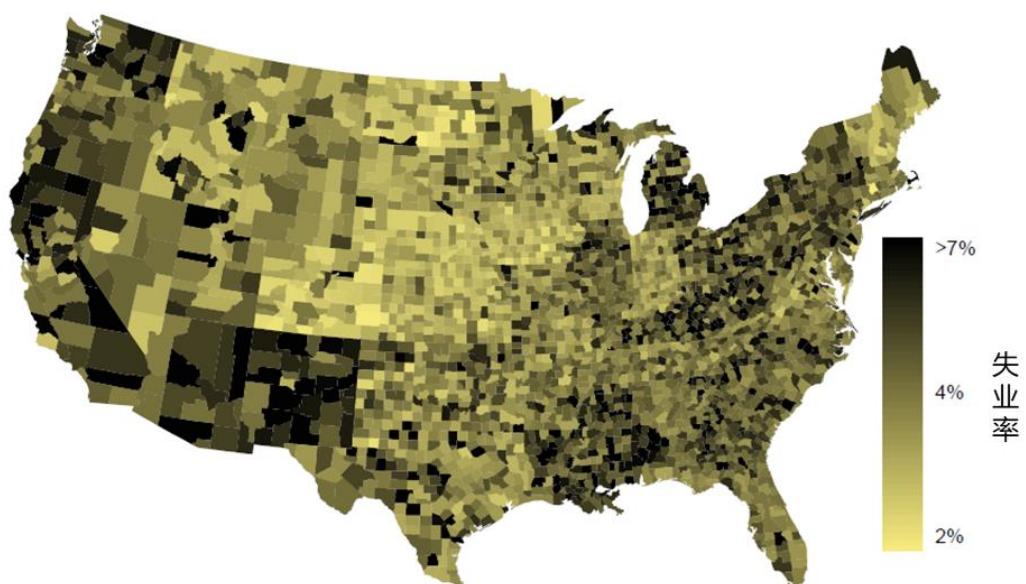
指导练习 2.22

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)

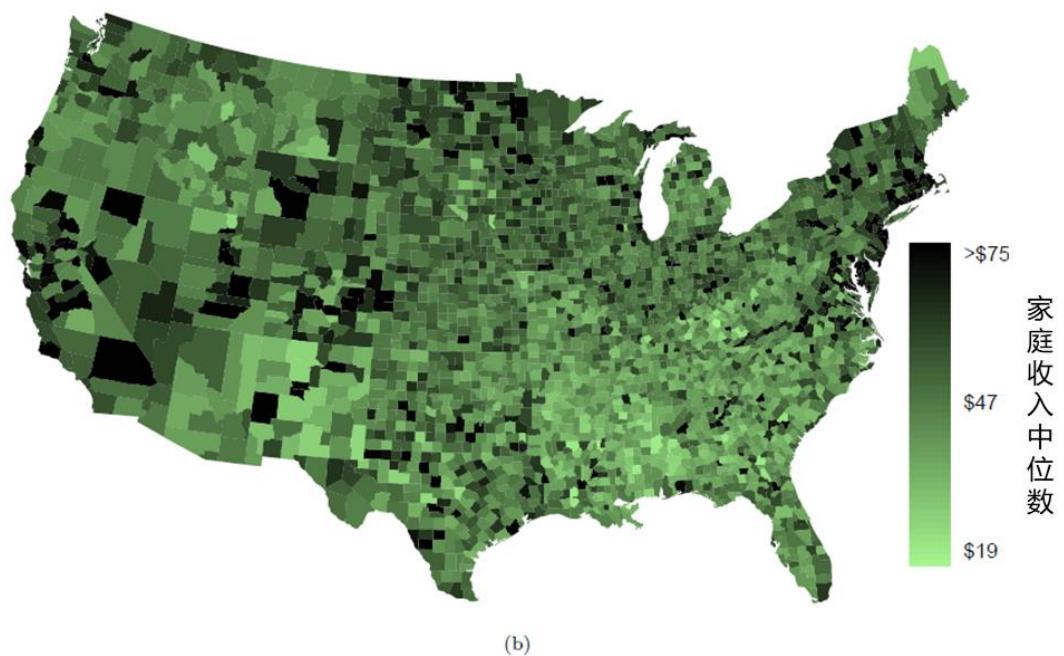
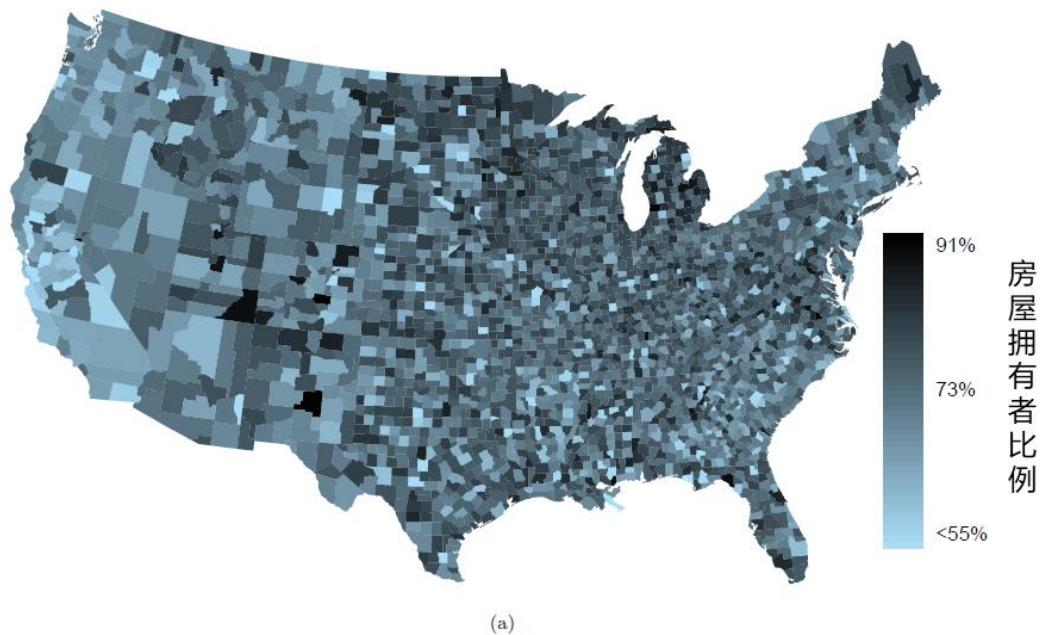


图 2.16: (a)房屋拥有者比例 (百分比) (b)家庭收入中位数 (\$1000s)

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： $3858/10000 = 38.58\%$ ）。

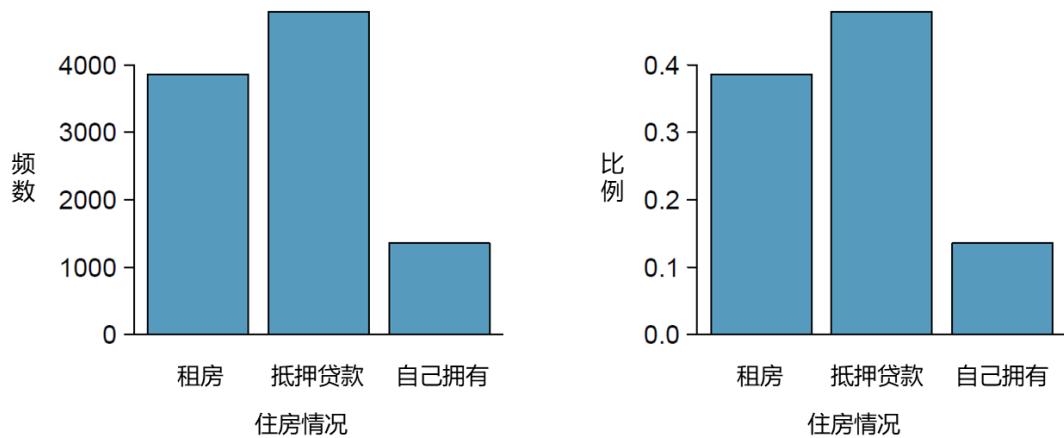


图 2.19：两张「住房情况」变量的柱形图，左侧展示的是频数，右侧展示的是每组比例

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

- (G) (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

- (G) (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。

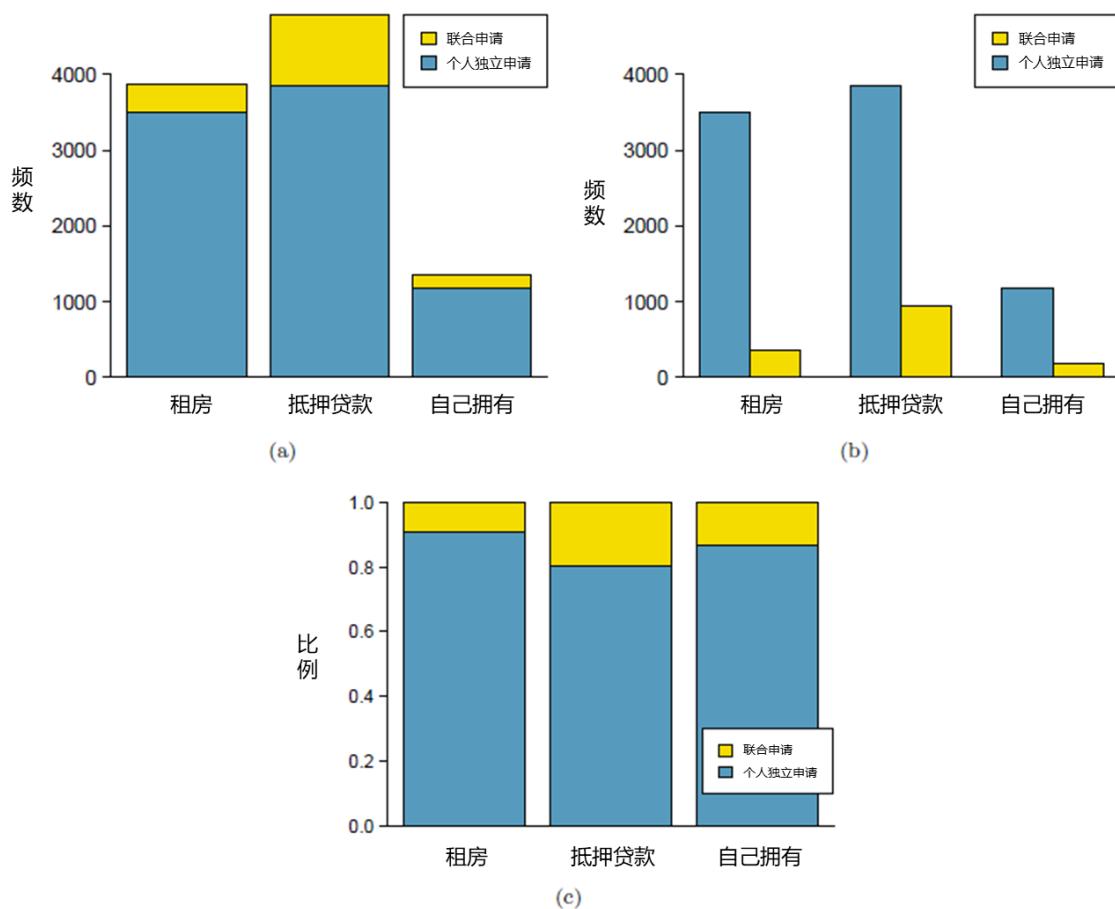


图 2.23: (a)「住房类型」的堆积柱形图，每个柱子按照「申请类型」分成两类；(b)并列柱形图；
(c)百分比堆积柱形图

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。

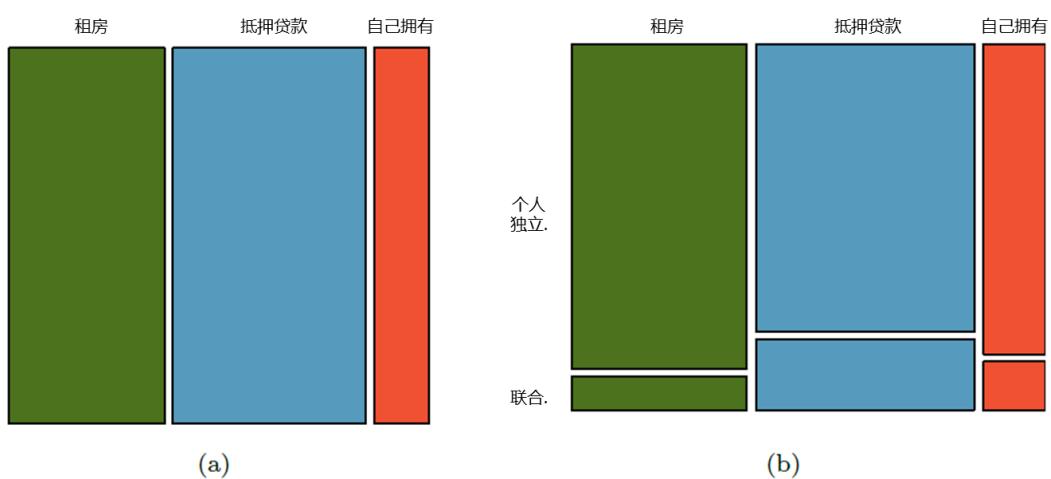


图 2.24：(a)单变量「住房类型」马赛克图；(b)双变量马赛克图

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。

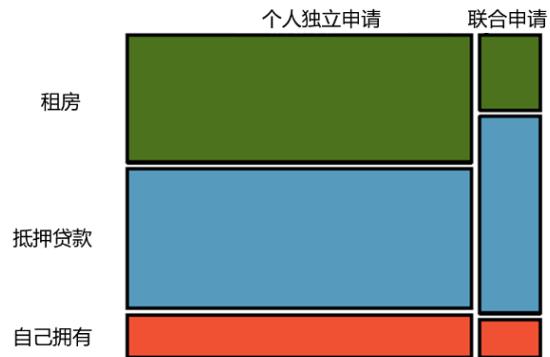


图 2.25：首先按「申请类型」分成两列，然后按照「住房类型」分成三行的马赛克图

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。

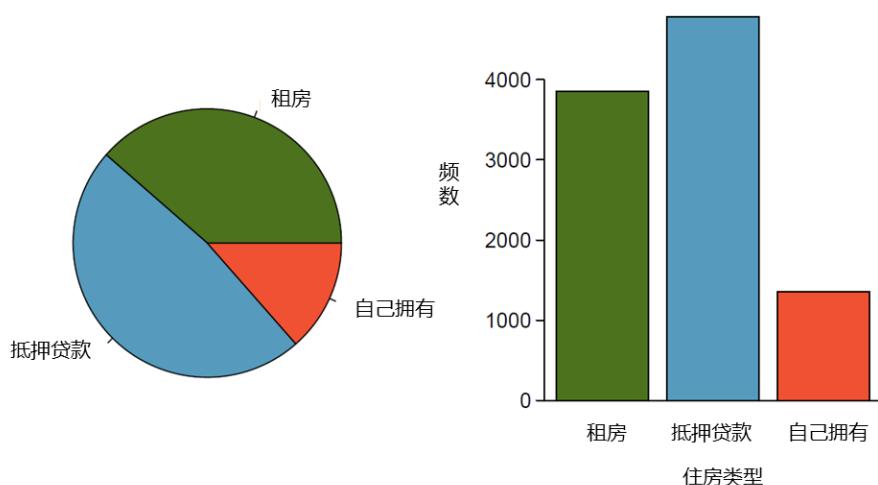


图 2.26：「住房类型」的饼状图和柱形图

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

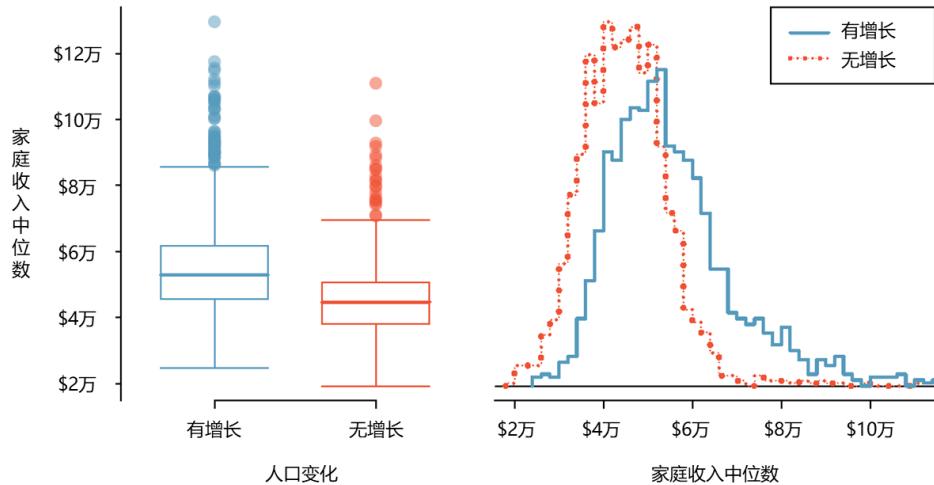


图 2.28：「家庭收入中位数」的并排箱型图（左）和空心直方图（右），每张图中的郡县被按照人口增长与否分成两部分。

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\$4.5万，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

E 某位老师把教室里的学生分成了两组：坐在左边的为一组，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

G 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。

接种	结果		
	感染	未感染	行总计
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

G 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。如果随便便一分组，都能观察到类似 64.3% 这么大的差值，那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。

接种 (模拟)		结果		
		感染	未感染	行总计
	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）。

示例 2.35

E

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：可以大约估算出，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。

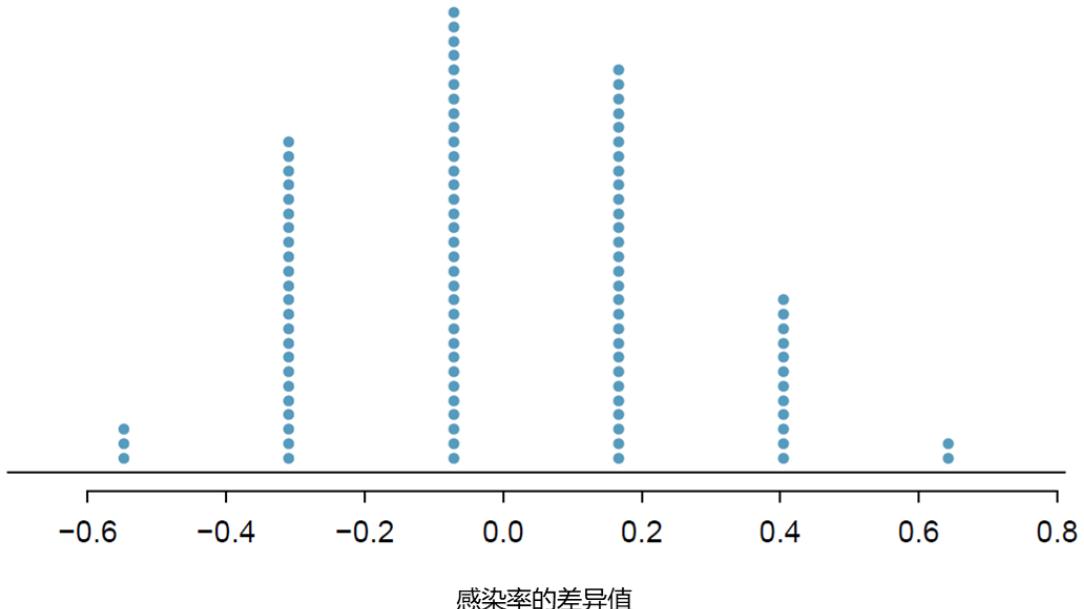


图 2.31：假设原假设 H_0 为真前提下，进行 100 次随机模拟过程的组间感染率差值的堆积点图，这些差值的产生是不受疫苗接种与否的影响的。图上只有右侧两个点的差值不小于 64.3%，即真实研究中得到的差值

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：(1) 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。(2) 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。