

统计学B

彭世喆
经济与管理学院
长沙理工大学

What is Statistics?

These resources give a taste of what statisticians, also known as *data scientists*, do in the real world.

▶ 200 years of history through health & wealth

A tour with Hans Rosling

▶ Bringing life to global health statistics

A longer tour led by Hans Rosling

▶ Building a better NBA team through analytics

Ivana Seric, college basketball player becomes a data scientist

▶ Steven Levitt explores the value of car seats

Disclaimers at the end

▶ Using statistics to treat chronic illnesses

MacArthur Fellow Susan Murphy

▶ Using data to better understand agriculture

MacArthur Fellow David Lobell

Why the term "Data Science" is so confusing

The two main *types* of data scientists: Analysis and Building

案例研究： 治疗慢性疲劳综合征

治疗慢性疲劳综合征

目标：评估认知行为疗法对慢性疲劳综合征的疗效（有效性）。

参与者库：由初级保健医生和专科医生转诊至一家专门治疗慢性疲劳综合征的医院诊所的142名患者。

实际参与者：转诊的142名患者中仅有60人参与研究。有些人因不符合诊断标准而被排除，有些人有其他健康问题，还有些人拒绝参与研究。

研究设计

患者被随机分配到治疗组和对照组，每组30名患者：

治疗组：认知行为疗法——协作式、教育式，并侧重于行为方面。向患者展示如何在不加重症状的情况下稳步安全地增加活动量。

对照组：放松疗法——没有给出关于如何增加活动量的建议。而是教授了渐进式肌肉放松、可视化和快速放松技巧。

结果

下表显示了6个月随访时获得良好结果的患者的分布情况。
请注意，有7名患者退出研究：治疗组3名，对照组4名。

		<i>Good outcome</i>		Total
		Yes	No	
<i>Group</i>	Treatment	19	8	27
	Control	5	21	26
	Total	24	29	53

- 治疗组获得良好结果的比例
 $19/27 \approx 0.70 \rightarrow 70\%$
- 对照组获得良好结果的比例
 $5/26 \approx 0.19 \rightarrow 19\%$

理解结果

这些数据是否显示两组之间存在“真实”差异？

- 假设你抛一枚硬币100次。尽管任何一次抛掷硬币正面朝上的概率是50%，但我们可能不会恰好观察到50次正面。这种波动几乎是所有类型数据生成过程的一部分。
- 观察到的两组之间的差异 ($70 - 19 = 51\%$) 可能是真实的，也可能是由自然变异导致的。
- 由于差异相当大，因此更可信的是这种差异是真实的。
- 我们使用统计工具来确定差异是否大到足以拒绝“差异由偶然因素造成”这一观点。

推广结果

这项研究的结果是否可推广到所有慢性疲劳综合征患者？

这些患者具有特定特征，并且自愿参与这项研究，因此他们可能不代表所有慢性疲劳综合征患者。虽然我们不能立即将结果推广到所有患者，但这项初步研究是令人鼓舞的。该方法对具有某些狭窄特征的患者有效，这为该方法至少在一定程度上对其他患者也有效带来了希望。

数据基础

课堂调查

对一门统计学入门课程的学生进行了一项调查。以下是调查中的一些问题，以及响应数据中存储的对应变量：

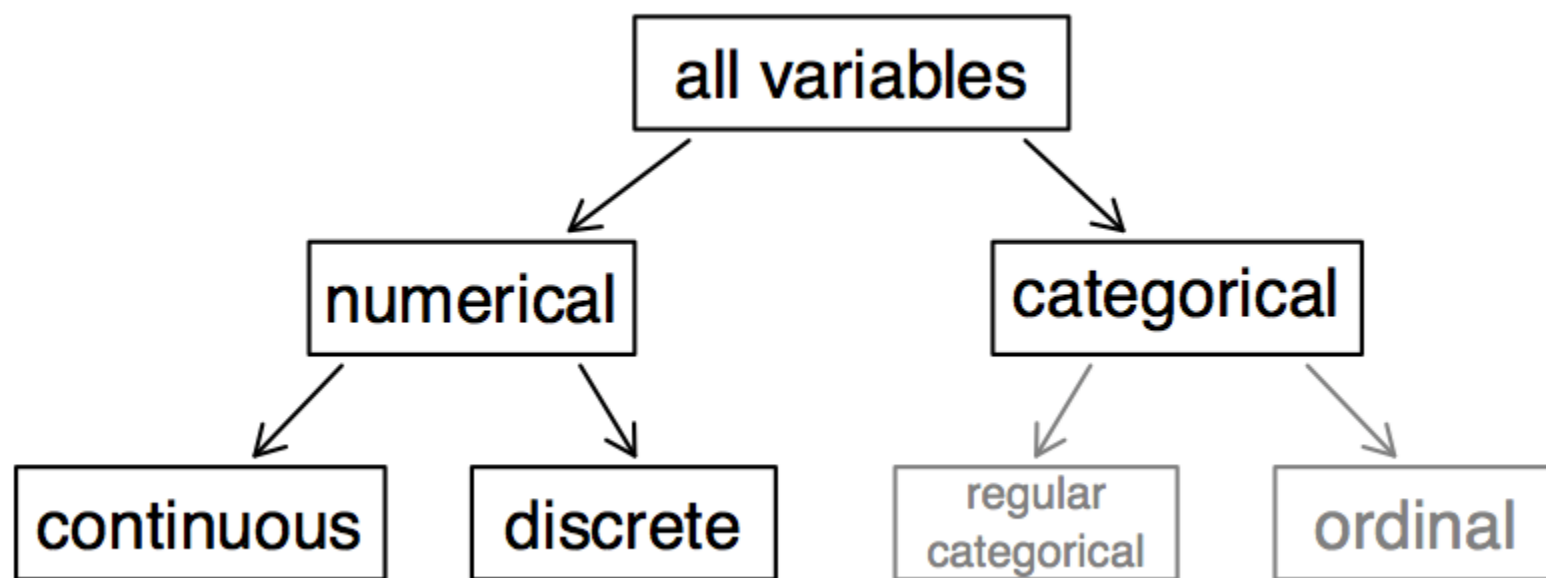
- 性别：你的性别是什么？
- 内向_外向：你是内向还是外向？
- 睡眠：你平均每晚睡多少个小时？
- 就寝时间：你通常几点上床睡觉？
- 国家：你访问过多少个国家？
- 恐惧程度：1-5分，你有多害怕待在这里？

数据矩阵

从统计课学生那里收集的关于多个变量的数据：

<i>variable</i>					
Stu.	gender	intro_extra	...	dread	
1	male	extravert	...	3	← <i>observation</i>
2	female	extravert	...	2	
3	female	introvert	...	4	
4	female	extravert	...	2	
⋮	⋮	⋮	⋮	⋮	
86	male	extravert	...	3	

变量类型



变量类型（续）

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- 性别：分类变量
- 睡眠：数值型、连续型
- 就寝时间：分类变量、有序型
- 国家：数值型、离散型
- 恐惧程度：分类变量、有序型——也可作为数值型使用

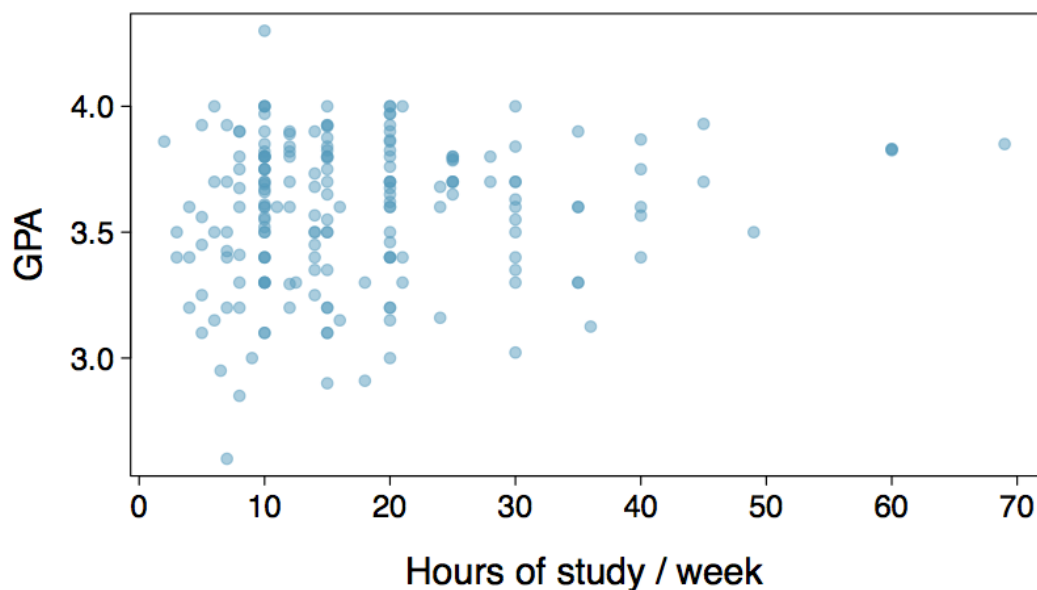
练习

电话号码的地区代码属于什么类型的变量？

- (a) 数值型、连续型
- (b) 数值型、离散型
- (c) 分类变量
- (d) 分类变量、有序型

变量之间的关系

每周学习时长与学生的GPA之间是否存在关系？

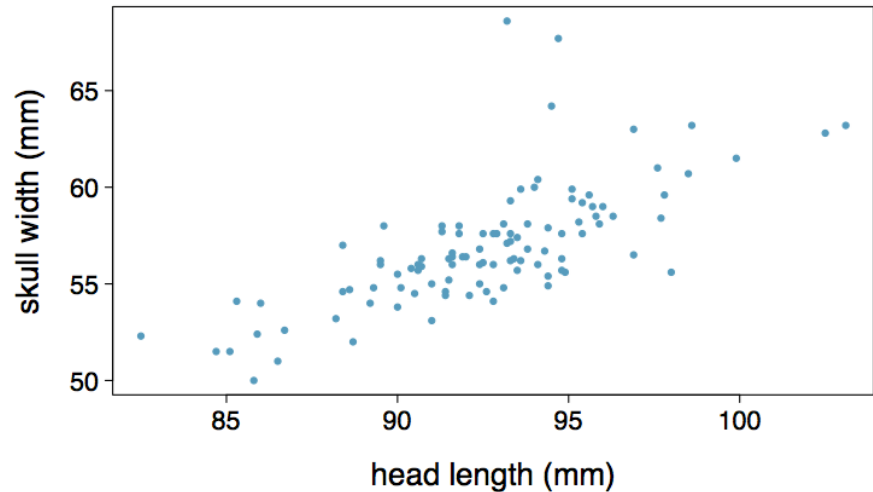


你能发现任何数据点中存在异常情况吗？

有一名学生的GPA超过4.0，这很可能是个数据错误。

练习

根据右侧的散点图，关于负鼠的头长与颅长，以下哪项陈述是正确的？



- (a) 头长与颅宽之间没有关系，即这两个变量是独立的。
- (b) 头长与颅宽呈正相关。
- (c) 颅宽与头长呈负相关。
- (d) 头越长导致颅骨越宽。
- (e) 颅骨越宽导致头部越长。

关联vs.独立

- 当两个变量之间表现出某种联系时，它们被称为关联变量。
 - 关联变量也可称为因变量，反之亦然。
- 如果两个变量没有关联，即两者之间不存在明显联系，则称它们为独立的。

观察性研究与抽样策略

总体与样本

PHYS ED | AUGUST 29, 2012, 12:01 AM | 21 Comments

Finding Your Ideal Running Form

By GRETCHEN REYNOLDS



David De Lossy/Getty Images

研究问题：人们能否仅仅通过跑步，自行成为更好、更高效的跑者？

关注总体：所有人

<http://well.blogs.nytimes.com/2012/08/29/finding-your-ideal-running-form>

样本：一组最近加入跑步团体的成年女性

可以推广结果的总体：成年女性，前提是数据为随机抽样所得

轶事证据与早期吸烟研究

- 反吸烟研究始于20世纪30年代和40年代，当时吸烟变得日益流行。虽然一些吸烟者似乎对香烟烟雾敏感，但其他人则完全不受影响。
- 反吸烟研究面临基于轶事证据的阻力，例如“我叔叔每天抽三包烟，身体却非常健康”，这些证据基于有限的样本量，可能无法代表总体情况。
- 当时得出的结论是：“吸烟是一种复杂的人类行为，其本质难以研究，且因人类个体差异而混淆不清。”
- 随着时间的推移，研究人员得以检查更大的病例（吸烟者）样本，表明吸烟对健康有负面影响这一趋势变得更加清晰。

普查

- 难道不能把所有个体都包括进去，即对整个总体进行“抽样”吗？
 - 这被称为普查。
- 进行普查存在一些问题：
 - 完成普查可能很困难：总有一些个体难以定位或难以测量。而这些难以找到的人可能具有某些特征，使其与总体中的其他人区分开来。
 - 总体很少是静止不动的。即使你能进行普查，总体也在不断变化，因此永远不可能获得完美的测量结果。
 - 进行普查可能比抽样更为复杂。

Illegal Immigrants Reluctant To Fill Out Census Form

by PETER O'DOWD

March 31, 2010 4:00 AM

 from **KJZZ**



Listen to the Story 

Morning Edition

3 min 48 sec

+ [Playlist](#)

↓ [Download](#)

There is an effort underway to make sure Hispanics are accurately counted in the 2010 Census. Phoenix has some of the country's "hardest-to-count" districts. Some Latinos, especially illegal residents, fear that participating in the count will expose them to immigration raids or government harassment.

从探索性分析到推断

- 抽样是自然的。
- 想想你做菜时的情况——你会品尝（检查）一小部分成品，以求对整道菜有所了解。
- 当你尝了一勺汤并判断它不够咸时，这就是探索性分析。
- 如果你概括并得出结论，认为整锅汤都需要加盐，那么这就是推断。
- 要使你的推断有效，你尝的那勺汤（样本）必须代表整锅汤（总体）。
 - 如果这勺汤只来自汤的表面，而盐沉在锅底，那么你尝到的很可能不代表整锅汤。
 - 如果你在品尝之前充分搅拌汤，尝到的那勺就更可能代表整锅汤。

抽样偏差

- 无应答：如果随机抽样的人中只有一小部分选择回复调查，那么样本可能不再代表总体。
- 自愿应答：当样本由自愿答复的人组成（因为他们对问题有强烈看法），也会导致样本不能代表总体。

Quick vote

Do you get paid sick days at your job?

- ☐ Yes ☐ No
- ☐ What job?

[VOTE](#) or view results

Quick vote

Do you get paid sick days at your job?

[Read Related Articles](#)

Yes		63%	20056
No		21%	6816
What job?		15%	4885

Total votes: 31757
This is not a scientific poll

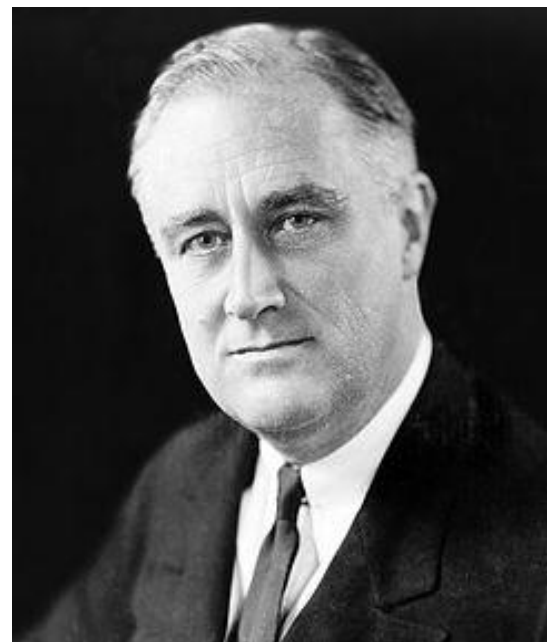
- 便利样本：容易被接触到的个体更有可能被包含在样本中。

抽样偏差实例： 兰登 vs. 富兰克林·德拉诺·罗斯福

一个因有偏样本而产生误导结果的历史实例

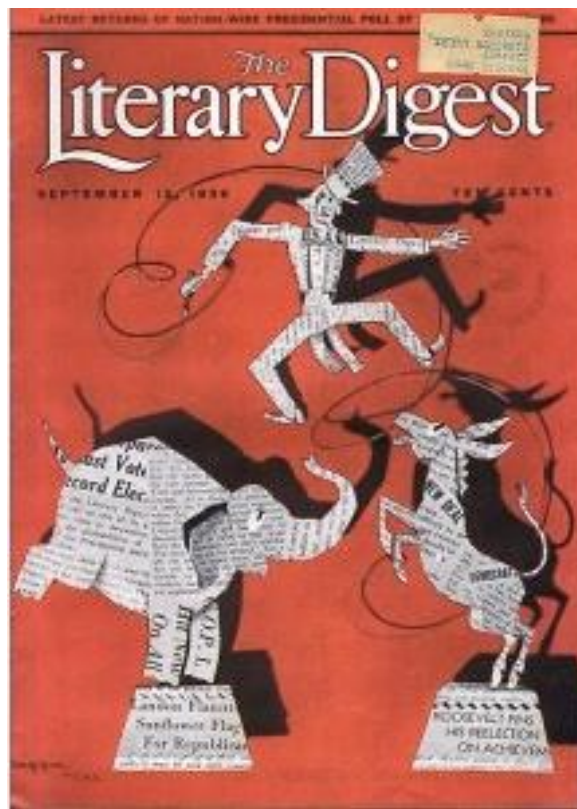


1936年，兰登寻求共和党总统候选人提名，与竞选连任的罗斯福对决。



《文学文摘》民意调查

- 《文学文摘》对约一千万美国人进行了民意调查，收到了约二百四十万人的回复。
- 调查显示兰登似乎会成为压倒性赢家，而罗斯福只能获得43%的选票。
- 选举结果：罗斯福获胜，获得了62%的选票。
- 该杂志因这次民意调查彻底信誉扫地，不久即停刊。



《文学文摘》民意调查—— 出了什么问题？

- 该杂志调查了
 - 其自己的读者，
 - 已登记的汽车车主，以及
 - 已登记的电话用户。
- 这些群体的收入远远高于当时全国平均水平（请记住，这是大萧条时期），因此这份选民名单中的选民比当时真正有代表性的投票者更可能支持共和党——也就是说，该样本并不代表当时的美国人口。

大样本固然可取，但是...

- 《文学文摘》的选举民意调查基于二百四十万的样本量，样本量非常庞大；但由于样本存在偏差，未能产生准确的预测。
- 回到汤的比喻：如果汤没有搅匀，汤勺再大也没用，尝起来的味道依然不对。如果汤搅拌均匀了，一把小勺子就足以品尝。

练习题

一个学区正在考虑是否不再允许高中生在学校停车，因为最近发生了两起学生重伤事故。作为第一步，他们通过邮件对家长进行调查，询问他们是否反对这一政策变化。发出的6000份调查问卷中，收回了1200份。在这已完成的1200份调查中，960份同意政策变动，240份不同意。下列哪些陈述是正确的？

- I. 部分邮件可能从未送达至家长手中
- II. 该学区得到了家长们的强烈支持，可以推进政策审批
- III. 可能大多数高中生家长都不同意这项政策变动
- IV. 调查结果不太可能存在偏差，因为所有家长都收到了调查邮件。

(a) 仅I (b) I和II (c) I和III (d) III和IV (e) 仅IV

观察性研究

- 研究者以不直接干扰数据产生的方式收集数据。
- 观察性研究的结果通常可用于建立解释变量与响应变量之间的关联。
- 相关性不等于因果关系。例如，使用防晒霜与皮肤癌。
- 当研究人员想要探究因果关系的可能性时，他们会进行实验。

获取良好的样本

- 几乎所有统计方法都基于隐含随机性的概念。
- 如果观测数据不是以随机框架从总体收集而来，这些统计方法——即估计值及其相关误差——就不可靠。
- 最常用的随机抽样技术有简单随机抽样、分层抽样和整群抽样。

前瞻性研究 vs. 回顾性研究

前瞻性研究通过识别个体并随着事件发展收集信息。

- 例如：护士健康研究自1976年开始招募注册护士，并通过问卷调查收集她们的数据。

回顾性研究则在事件发生后收集数据。

- 例如：研究人员在病历中回顾过去发生的事件。

获取良好的样本

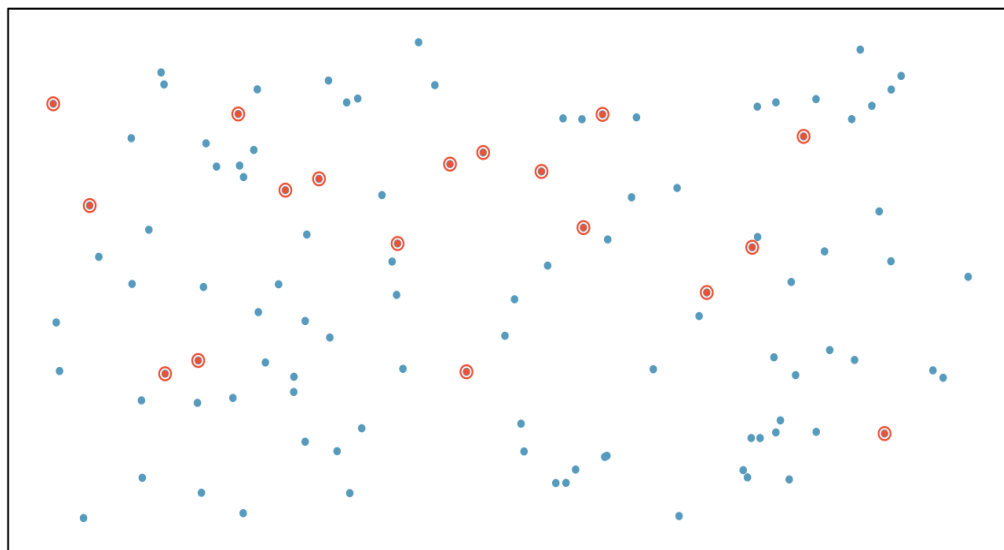
- 几乎所有统计方法都基于隐含随机性的概念。
- 如果观测数据不是以随机框架从总体收集而来，这些统计方法——即估计值及其相关误差——就不可靠。
- 最常用的随机抽样技术有简单随机抽样、分层抽样和整群抽样。

简单随机样本

从总体中随机抽取样本，所选个体之间不存在隐含的关联。

如果无应答率较高，则需谨慎对待。

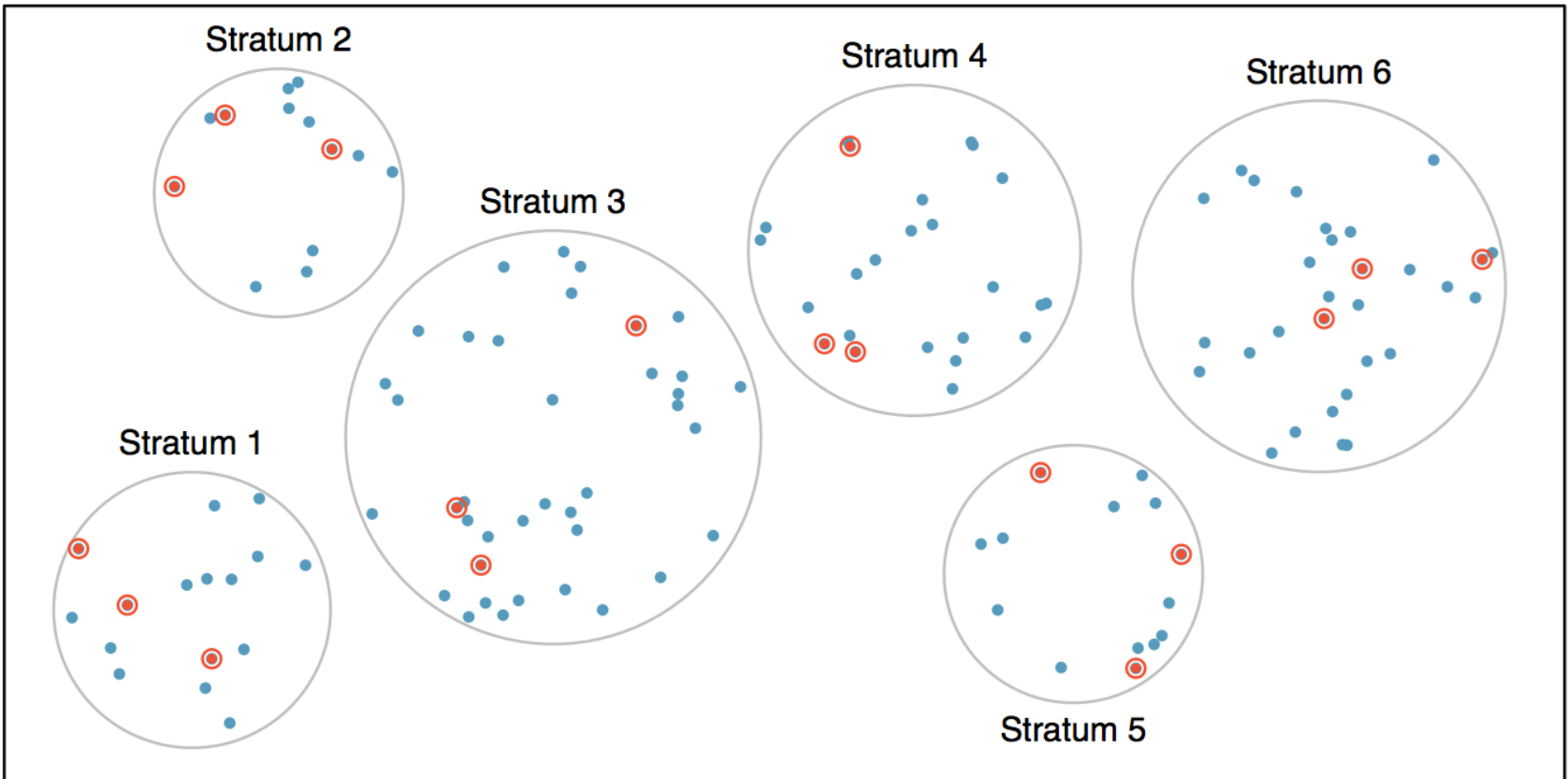
便利抽样通常是一种不良的推断总体结论的基础，因为容易获取的个体更可能被纳入样本。



分层样本

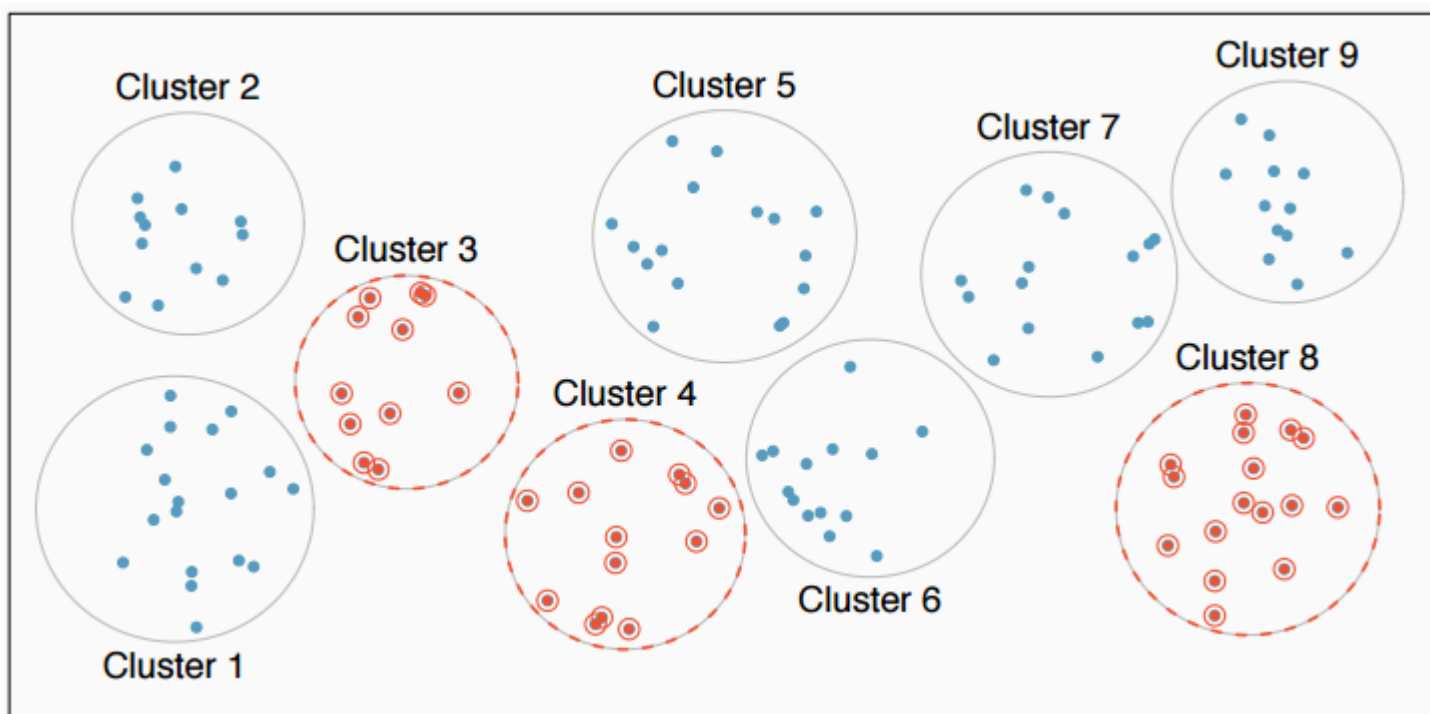
层由相似的观测值构成。我们从每一层中进行简单随机抽样。

。



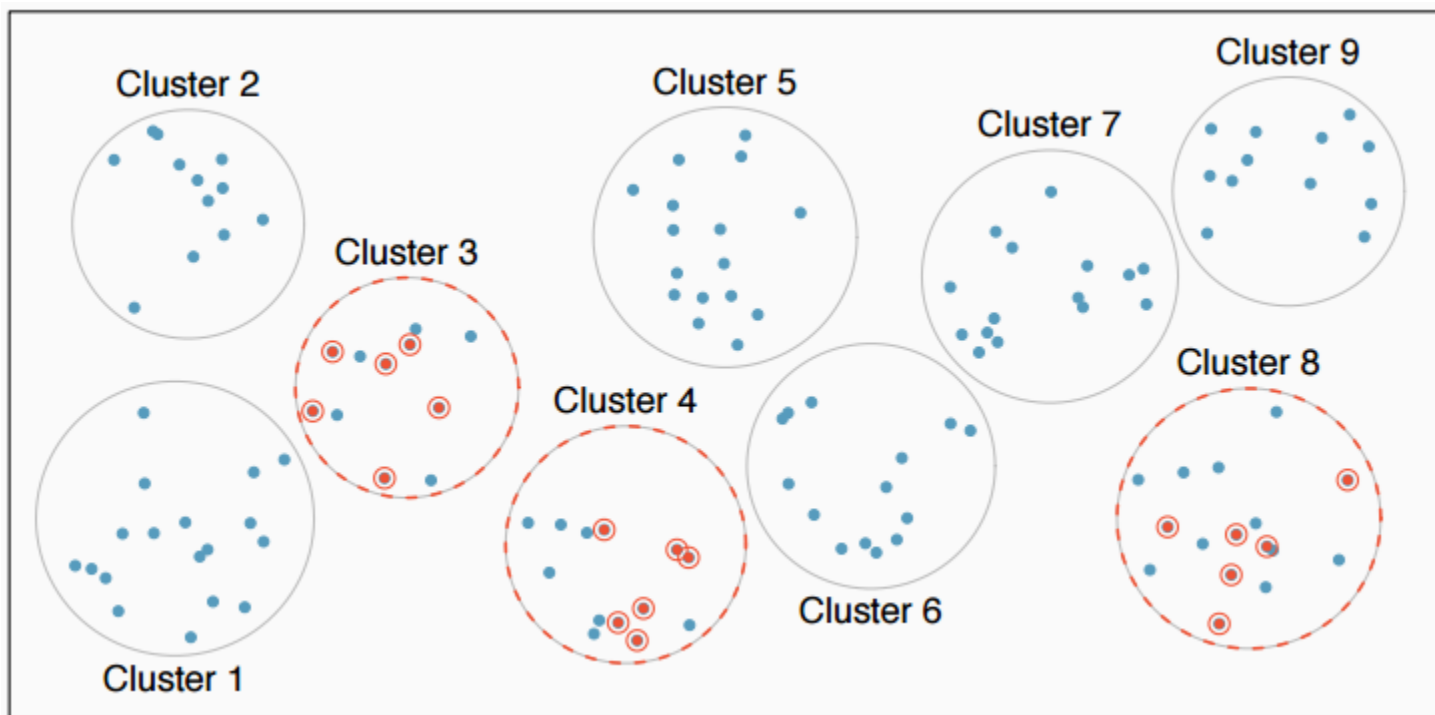
整群样本

群通常不是由同质观测值构成的。我们抽取简单的随机群样本，然后对该群中的所有观测值进行抽样。通常出于经济原因而首选这种方法。



多阶段样本

群通常不是由同质观测值构成的。我们抽取简单的随机群样本，然后从被抽中的群中再抽取简单随机样本作为观测值。



练习题

某市议会已要求对其城市郊区进行一项家庭调查。该区域划分为许多独特而鲜明的社区，有的包含大型住宅，有的只有公寓。以下哪种方法可能最不有效？

- (a) 简单随机抽样
- (b) 整群抽样
- (c) 分层抽样
- (d) 分块抽样

实验

实验设计原则

1. 对照：将关注的处理与对照组进行比较。
2. 随机化：随机分配受试者到各个处理组，并在可能的情况下从总体中随机抽样。
3. 重复：在研究内部，通过收集足够大的样本来重复。或对整个研究进行重复。
4. 区组化：如果存在已知或怀疑会影响响应变量的变量，首先将这些变量分组成区组，然后分别在每个区组内将个体随机分配到各个处理组。

关于分区的更多信息



- 我们想要设计一个实验来探究能量凝胶是否能让跑得更快：
 - 处理：能量凝胶
 - 对照：无能量凝胶（或使用安慰剂以规避心理效应）
- 假设能量凝胶对专业运动员和业余爱好者的影响可能不同，因此我们针对专业状态进行分区：
 - 将样本分为专业和业余
 - 随机将专业运动员分配到处理组和对照组
 - 随机将业余运动员分配到处理组和对照组
 - 在最终得到的处理组和对照组中，专业/业余状态的比例相同

为什么这很重要？你能想到其他需要分区的变量吗？

练习

一项研究旨在测试光照水平和噪声水平对学生考试成绩的影响。研究人员还认为，光照和噪声水平可能对男性和女性产生不同的影响，因此希望确保在每个组中两种性别都能得到平等代表。以下哪项是正确的？

- A. 有3个解释变量（光照、噪声、性别）和1个响应变量（考试成绩）
- B. 有2个解释变量（光照和噪声），1个分区变量（性别），和1个响应变量（考试成绩）
- C. 有1个解释变量（性别）和3个响应变量（光照、噪声、考试成绩）
- D. 有2个分区变量（光照和噪声），1个解释变量（性别），和1个响应变量（考试成绩）

分区变量与解释变量之间的区别

- 因子是我们能够施加在实验单元上的条件。
- 分区变量是实验单元本身具备的特征，我们希望对这类特征进行控制。
- 分区类似于分层，只不过在随机分配时使用于实验环境中，而非抽样时使用。

更多实验设计术语...

- 安慰剂：虚假的处理，通常在医学研究中用作对照组
- 安慰剂效应：实验单元仅仅因为相信自己正在接受特殊处理而表现出改善
- 盲法：实验单元不知道自己属于对照组还是处理组
- 双盲：实验单元和与患者互动的研究人员都不知道谁属于对照组、谁属于处理组

练习

观察性研究和实验的主要区别是什么？

- A. 实验在实验室中进行，而观察性研究则不需要。
- B. 在观察性研究中，我们仅观察过去发生的事情。
- C. 大多数实验使用随机分配，而观察性研究则不进行随机分配。
- D. 观察性研究完全无用，因为基于其发现无法做出因果推断。

随机分配 对比 随机抽样

<i>ideal experiment</i>	Random assignment	No random assignment	<i>most observational studies</i>
Random sampling	Causal conclusion, generalized to the whole population.	No causal conclusion, correlation statement generalized to the whole population.	Generalizability
No random sampling	Causal conclusion, only for the sample.	No causal conclusion, correlation statement only for the sample.	No generalizability
<i>most experiments</i>	Causation	Correlation	<i>bad observational studies</i>