Data Collection

Applied Statistics

Fall 2025

目录

1   数据来源 Sources of Data

好的统计分析始于对数据来源的仔细研究。

A good statistical analysis starts with a careful study of the source of the data.

- 轶事数据 Anecdotal data

- 现有数据 Available data

- 抽样调查数据 Sample surveys data

- 实验数据 Experiments data

1.1   轶事数据 Anecdotal data

- 基于随意选择的个别案例，这些案例往往因为某种方式与众不同而被我们记住。
  Based on haphazardly selected individual cases, which we tend to remember because they are unusual in some way.

- 例如：社交媒体上分享的"经验"、新闻事件等。
  e.g. 'experiences' shared on social media, news events, etc.

- 但它们可能不能代表任何更大的群体。从轶事数据中得出一般性结论常常会产生误导。
  But they may not be representative of any larger group of cases. It is often misleading to draw a general conclusion from anecdotal data.

- 并非所有轶事数据都是坏的。个人或一小群案例的经验可能会引发精心设计的研究。
  Not all anecdotal data are bad. The experiences of an individual or a small group of cases might lead to carefully designed studies.

1.2   现有数据 Available data

- 现有数据是为其他目的而产生的，但可能有助于廉价地回答感兴趣的问题。
  Available data are data that were produced for some other purpose but that may help answer a question of interest inexpensively.

- 行政数据：犯罪数据、税收数据、出生和死亡证明数据。
  Administrative data: crime data, tax data, birth and death certificate data.

- 互联网活动产生的"大数据"也符合这一定义。
  'Big data' generated by internet activities also fits the definition.

- 示例：通过 CDC WONDER 访问最新的临时出生数据。
  Example: Access the newest provisional natality data through CDC WONDER.

1.3 抽样调查数据 Sample surveys data

- 抽样调查通过从总体中选择和测量样本来收集关于总体的信息。
  A sample survey collects information about a population by selecting and measuring a sample from the population.

- 检查总体中的每个项目（普查）的费用（和耗时）使得抽样成为实际需要。
  The expense (and time-consuming) of examining every item in the population (census) makes sampling a practical necessity.

- 设计和执行良好的抽样调查有助于了解总体特征。
  Well-designed and conducted sample surveys are good for learning the characteristics of a population.

2 抽样设计 Sampling design

好的抽样调查数据依赖于适当的抽样设计：从总体中选择样本的方法。
Good sample survey data rely on appropriate sampling design: the method used to choose the sample from the population.

- 自愿回应样本由通过回应一般呼吁而自我选择的人组成。
  A voluntary response sample consists of people who choose themselves by responding to a general appeal.

- 自愿回应样本常常显示偏差，因为有强烈意见（特别是负面意见）的人可能更有可能回应。
  Voluntary response samples often show bias because people with strong opinions (especially negative opinions) may be more likely to respond.

- 我们需要代表性样本：样本应与目标总体相似。
  We need representative samples: the sample should be similar to the targeted population.

2.1 概率样本 Probability samples

为确保代表性，我们需要概率样本：

- 通过机会选择的样本。
  A sample chosen by chance.

- 总体中的每个项目必须有相等或已知的被选概率。
  Each item in the population has to have an equal or known probability of being chosen.

  四种常见的抽样设计：

- 简单随机抽样 Simple random sampling

- 分层随机抽样 Stratified random sampling

- 整群抽样 Cluster sampling

- 多阶段抽样 Multistage sampling

2.1.1 简单随机抽样 Simple Random Sampling

- 简单随机抽样：总体中的每个元素被选中的概率相等。
  Simple random sampling: each element of the population is equally likely to be chosen.

- 示例：政府税务审计员从 1000 家公司中抽取 40 家进行审计。每家公司被选中的概率为 0.04。
  Example: a government tax auditor chooses a sample of 40 of 1,000 firms to audit. Each firm has a probability of 0.04 to be chosen.

- 用于具有完整总体名单的简单调查（更多用于自助重抽样）。
  Used in simple surveys with a complete list of the population (more often in bootstrap re-sampling).

- 系统抽样：每 k 个成员中选择一个，直到达到所需的样本量。当抽样名单未知但存在内在顺序时很有用，例如商店中的顾客。
  Systematic sampling: Select every k-th member until we have the desired sample size. Useful when the sampling list is unknown but there is an intrinsic order, such as customers in a shop.

2.1.2 分层随机抽样 Stratified Random Sampling

- 分层随机抽样：首先根据某些分类标准将总体划分为互斥的集合或层。
  Stratified random sampling: first dividing the population into mutually exclusive sets, or strata, based on some classification criteria.

- 层可以基于人口统计信息（如年龄和性别）或地理区域（省/县）。
  Strata might be based on demographic information like age and gender or geographic strata by province/county.

- 在每个层内进行简单随机抽样以生成样本。
  Simple random sampling within each stratum to generate the sample.

- 从每个层中抽取的简单随机样本的大小与该层的相对大小成比例。
  The simple random sample drawn from each stratum is in sizes proportional to the relative size of each stratum.

表 1: 分层抽样示例Stratified Sampling Example

| 收入类别 | 总体比例 | 样本大小 $n = 400$ | 样本大小 $n = 1000$ |
Income Category	Population Proportion	Sample Size $n = 400$	Sample Size $n = 1000$
under \$25,000	25%	100	250
\$25,000 - \$39,999	40%	160	400
\$40,000 - \$60,000	30%	120	300
over \$60,000	5%	20	50

2.1.3 整群抽样

Cluster Sampling

- 整群样本是元素组或群的简单随机样本（与单个对象的简单随机样本相对）。
  A cluster sample is a simple random sample of groups or clusters of elements (vs. a simple random sample of individual objects).

- 当难以或成本高昂地开发完整的总体成员列表，或者总体元素在地理上广泛分散时，此方法很有用。
  This method is useful when it is difficult or costly to develop a complete list of the population members or when the population elements are widely dispersed geographically.

- 示例：在中学生调查中选择整个班级的学生。
  Example: select the entire classroom of students in a survey for middle school students.

2.1.4 多阶段抽样
Multistage Samples

- 用于大型社会调查，结合多阶段分层抽样和整群抽样等。
  Used in large social surveys, a combination of multi-stage stratified sampling and cluster sampling, etc.

- 例如：美国就业和失业调查——当前人口调查（CPS）每月采访约 60,000 户家庭。CPS 使用多阶段抽样设计以降低成本并确保州/县级别的代表性：按州和县以及人口统计信息进行多阶段分层；在层内进行整群抽样。最终样本由访问员可以轻松访问的附近家庭群组成。
  For example, US employment and unemployment survey the Current Population Survey (CPS) which conducts interviews in about 60,000 households each month.CPS uses a multistage sampling design to reduce cost and also ensure representativeness in the state/county level: multistage stratify by state and county, as well as demographic information; cluster sampling within the strata. The final sample consists of clusters of nearby households that an interviewer can easily visit.

多阶段抽样步骤

1. 第一阶段：将美国划分为 2007 个称为初级抽样单位（PSU）的地理区域。PSU 不跨越州界。选择 754 个 PSU 的样本。该样本包括人口最多的 428 个 PSU 和其他 326 个 PSU 的分层样本。
   Stage 1. Divide the United States into 2007 geographical areas called Primary Sampling Units, or PSUs. PSUs do not cross state lines. Select a sample of 754 PSUs. This sample includes the 428 PSUs with the largest population and a stratified sample of 326 of the others.

2. 第二阶段：将每个选定的 PSU 划分为称为"街区"的较小区域。使用种族和其他信息对街区进行分层，并对每个 PSU 中的街区进行分层抽样。
   Stage 2. Divide each PSU selected into smaller areas called "blocks." Stratify the blocks using ethnic and other information and take a stratified sample of the blocks in each PSU.

3. 第三阶段：将每个街区中的住房单元按附近单元聚类排序。
   Stage3. Sort the housing units in each block into clusters of our nearby units.

2.2 关于抽样调查的注意事项 Cautions about sample surveys

2.2.1 覆盖不足 Undercoverage

- 当总体中的某些群体在选择样本的过程中被遗漏时发生。
  When some groups in the population are left out of the process of choosing the sample.

- 大型人群的完整列表很少可用；大多数调查都有一定程度的覆盖不足。
  Complete lists of large human populations are rarely available; most surveys suffer from some degree of undercoverage.

- 示例：城市调查中的农民工、被监禁人口。
  Example: migrant workers in urban surveys, incarcerated population.

2.2.2 无回答 Nonresponse

- 当被选入样本的个人无法完成调查时发生。
  Nonresponse occurs when an individual chosen for the sample can't complete the survey.

- 回答偏差：人们在决定是否回答时自我选择。
  Response bias: people self-selection in the decision of response.

- 例如，在民意调查中，有强烈意见（特别是负面意见）的人最有可能回答。
  For example, people with strong opinions, especially negative opinions, may be most likely to respond in an opinion poll.

- 如果无回答占显著比例，代表性就成问题。
  If the non-response represents a significant proportion, the representativeness is in question.

2.3 一些潜在的数据来源 Some Potential Data Sources

2.3.1 国际比较 International comparison:

- IPUMPS (https://www.ipums.org/)

- ICPSR (https://www.icpsr.umich.edu/icpsrweb/)

2.3.2 中国社会调查 Social surveys in China:

- 中国家庭追踪调查（CFPS）
  China Family Panel Studies, CFPS (http://www.isss.pku.edu.cn/cfps/)

- 中国健康与养老追踪调查（CHARLS）

  China Health and Retirement Longitudinal Study, CHARLS (http://charls.pku.edu.cn/zh-CN)

- 中国健康与营养调查（CHNS）

  China Health and Nutrition Survey, CHNS (http://www.cpc.unc.edu/projects/china)

- 中国家庭收入项目（CHIP）

  Chinese Household Income Project, CHIP (http://ciid.bnu.edu.cn/chip/index.asp?lang=EN)

3   Learning about relations

因果关系是指变量之间直接的因果联系，例如父母的身高和孩子的身高。

Causation: a direct cause-and-effect link between these variables, such as the height of parents and children.

- 虚线表示关联。

  The dashed lines show an association.

- 实线箭头表示因果联系。

  The solid arrows show a cause-and-effect link.

- x 是解释变量，y 是响应变量，z 是潜在变量。

  x is an explanatory variable, y is a response variable, and z is a lurking variable.

Causation
(a)

Common response
(b)

Confounding
(c)
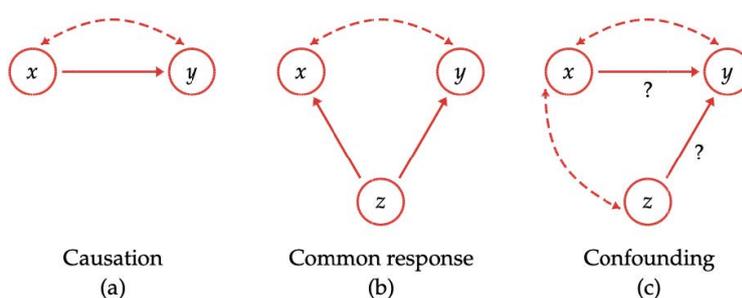
3.1   共同反应 Common Response

- 潜在变量（Lurking variable）是指不在解释变量或响应变量之中，但对所研究的变量有影响的变量。

  A lurking variable is a variable that is not among the explanatory or response variables but does have an effect on the variables studied.

- 潜在变量可能错误地暗示一种关系。
  Lurking variables can falsely suggest a relationship.

- 共同反应（Common response）：观察到的关联可以由潜在变量 z 解释。
  Common response: the observed association is explained by a lurking variable z.

- 示例：冰淇淋销量和溺水率。可能的潜在变量是什么？
  Example: ice-cream sales and drowning rates. What is a possible lurking variable?
  (例如：天气温度)

3.2 混杂 Confounding

- 即使存在直接因果关系，它也鲜能完全解释两个变量之间的关联。
  Even when direct causation is present, it is rarely a complete explanation of an association between two variables.

- 当两个变量对响应变量的影响无法区分时，它们就是混杂的。
  Two variables are confounded when their effects on a response variable cannot be distinguished.

- 混杂变量可能是解释变量或潜在变量。
  The confounded variables may be either explanatory variables or lurking variables.

- 示例：收入、健康和幸福感。
  Example: income, health and happiness.

3.3 因果效应 Causal effects

- 数据的代表性不足够。
  Representativeness of data is not enough.

- 示例：学校质量与学生学业表现
  Example: School quality and student academic performance

  - 我们通过一项关于中学生学业表现及其学校特征的具有代表性的调查获得数据。
    We obtain data from a representative survey about middle school students about their academic performance and the characteristics of their schools.

  - 假设我们发现就读明星学校的学生在考试和认知测试中表现更好。
    Suppose we find that students who attended star schools perform much better in exams and cognitive tests.

– 这个结果能解释为因果效应吗？即，任何就读明星学校的学生平均都会在学业表现上有如此大的提高吗？

Can this result be interpreted as causal effects? i.e. any students attending star schools would on average expect to see such a large improvement in academic performances?

3.4 观察性研究 Observational Studies

- 在观察性研究中，我们观察因变量和自变量的状态，但无法控制因素和响应的分配。

  In an observational study we observe the state of the dependent variable and independent variable but have no control over the assignment of the factors and responses.

- 抽样调查产生观察性数据。

  Sample surveys generate observational data.

- 观察性研究是确定因果关系的较差方式:

  An observational study is a poor way to determine causality:

  - 使用统计模型（如多元线性回归或匹配）来控制一些重要的混杂变量可能有帮助。

    Using statistical models such as multi-linear regression or matching to control for some important confounding variables can be helpful.

  - 然而，总是存在遗漏变量偏差：我们很难控制所有重要的混杂变量：有时我们无法测量；其他时候可能超出我们当前的理解。

    However, there always remains omitted variable bias: we can hardly control for all important confounding variables: sometimes we cannot measure; other times it may be beyond our current understanding.

3.5 实验 Experiments

- 实验是检验一个变量对另一个变量影响的首选方法。

  Experiments are the preferred method for examining the effect of one variable on another.

- 实验是一种我们故意对人群、动物或物体施加某些条件以观察反应的研究。

  An experiment is a study in which we deliberately impose some conditions on people, animals, or objects to observe the response.

- 工作原理：由于我们可以控制施加的条件，我们可以通过随机化确保不同条件的组在可观察和不可观察的方面具有可比性。

  How it works: since we can control the conditions imposed, we can make sure the groups with different conditions are comparable in observed and unobserved ways using randomization.

- 所施加的条件称为处理/干预。

  The condition imposed is called a treatment/intervention.

3.6 实验设计 Design of Experiments

实验：我们故意对个体施加某种处理并观察他们的反应。

Experiment: we deliberately impose some treatment on individuals and observe their responses.

- 实验对象（Subjects）：进行实验的个体。

  Subjects: individuals on whom the experiment is done.

- 处理（Treatments，x）：应用于实验对象的实验条件，也称为因素。

  Treatments (x): experimental conditions applied to subjects, also called factors.

- 结果（Outcomes，y）：用于比较处理的测量变量。

  Outcomes (y): the measured variables that are used to compare the treatments.

3.6.1 示例 Example

在一项关于某种疾病的研究中，150 名患者被给予新的止痛药 A，另外 150 名患者被给予假药（维生素片）。研究人员记录了每个患者的疼痛发作次数。确定：

- 实验对象：300 名患者

  Subjects: 300 patients.

- 处理（x）：止痛药 A 和假药

  Treatments (x): painkiller A and dummy drug.

- 结果（y）：疼痛发作次数

  Outcomes (y): episodes of pain.

3.7 实验设计原则 The design of experiments

实验设计的基本原则是控制、随机化和重复。

The basic principles of the design of experiments are control, randomization, and repetition.

- 比较性（Comparative）：实验是比较性的：将处理组与对照组进行比较。
  Comparative: Experiments are comparative: compare treatment groups with a control group.

- 随机化（Randomize）：使用随机方法将实验单位分配到处理中。
  Randomize: Use chances to assign experimental units to treatments.

- 重复（Replication）：在每个单位上重复处理以减少结果的变异机会。
  Replication: Repeat each treatment on many units to reduce the chance of variation in the results.

3.7.1 比较性

- 对照组（Control group）是接受无效处理（安慰剂）、基线处理（当前标准）或不同处理以进行比较的一组实验对象。
  A control group is a group of subjects that receive an inactive treatment (placebo), a baseline treatment (the current standard), or a different treatment for comparison.

- 它作为处理组的参考标准。
  It serves as a reference mark for a treatment group.

- 对照组有助于消除潜在变量对响应变量的影响。
  The control group helps eliminate the effects of lurking variables on the response variable.

- 示例：为了了解"翻转课堂"教学法的效果，对于给定课程，选择一些部分作为处理组，其余作为对照组。
  Example: to learn the effect of a teaching method "flipped classroom", for a given course some sections are selected as the treatment group and the rest are the control group.

3.7.2 随机化 Randomization

- 如果处理组与对照组差异很大，将会导致偏差。
  If the treatment groups differ greatly from the control group, bias will result.

- 解决偏差问题的方法是随机分配：使用某种随机过程将实验单位随机分配到处理中。
  The solution to the problem of bias is random assignment: experimental units are assigned to treatments at random using some sort of chance process.

- 随机化确保处理组和对照组在所有可观察和不可观察的方面相似。
  Randomization ensures that treatment and control groups are similar in all observed and unobserved ways.

- 如果结果出现显著差异，我们可以确信这是由于处理造成的。
  If there appear to be significant differences in outcomes, we can be confident that it is because of the treatment.
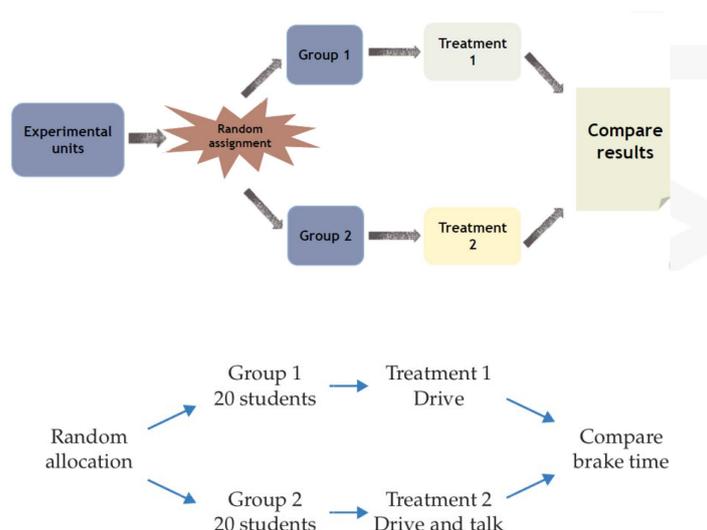

图 1: driving while talking example

3.7.3 练习 Practice

一位统计学教师想知道哪条路线能让她最快到达学校。从 10 月 2 日到 11 月 15 日，每天当她到达转折点时，她检查车上的里程表。如果显示偶数，她走高速公路；如果显示奇数，她走市区路线。她记录每天的总时间。

这项研究是：

a. 观察性研究。
   observational study.

b. 样本。
   sample.

c. 实验。
   experiment.

答案：c. 实验。

社会科学中的经典实验 One classic experiment in social science

- 田纳西 STAR 项目是一项关于班级规模影响的实验。
  The Tennessee STAR program was an experiment on the effects of class size.

- 实验对象是 6385 名刚开始上幼儿园的学生。每个学生被分配到三种处理之一：常规班级（22 至 25 名学生）配一名教师，常规班级配一名教师和一名全职助教，小班（13 至 17 名学生）。
  The subjects were 6385 students who were beginning kindergarten. Each student was assigned to one of three treatments: regular class (22 to 25 students) with one teacher, regular class with a teacher and a full-time teacher's aide, and small class (13 to 17 students).

- 结果表明，小班确实改善了学生的结果，包括行为发展，并且这种影响是持久的。
  Shows that small classes do improve student outcomes, including behavioral development and the effect is persistent.

3.8 自然实验和准实验设计 Natural Experiments and Quasi-Experimental Designs

社会科学中的实验往往不可行。Experiments in social science are often not possible.

在经济学中，我们依赖自然实验。例如，我们跟踪刚好高于和低于进入明星学校分数的学生，并比较他们的结果。

像断点回归设计这样的方法利用这种任意规则来近似随机分配。

Methods like Regression Discontinuity Design (using a cutoff score) exploit such arbitrary rules to approximate random assignment.

3.9 安慰剂效应 Placebo effect

- 随机比较实验的逻辑取决于我们能够以完全相同的方式对待所有实验对象，除了被比较的实际处理。
  The logic of a randomized comparative experiment depends on our ability to treat all the subjects in exactly the same way, except for the actual treatments being compared.

- 安慰剂效应（Placebo effect）：当一个人服用安慰剂或"假"治疗后，其身体或 mental 健康似乎有所改善。
  Placebo effect: The placebo effect is when a person's physical or mental health appears to improve even after taking a placebo or 'dummy' treatment.

- 例如：即使给予糖丸，疼痛也可能比不给任何药丸更好。

  For example: even given a sugar pill, the pain may be better than given no pills at all.

- 为了避免安慰剂效应，对照组是必要的。

  A control group is necessary to avoid the placebo effect.

3.10   双盲 Double-blind

- 此外，实验对象不知道他们是在处理组还是对照组。

  Moreover, the subjects would not know if they were in a treatment or control group.

- 除了安慰剂效应，实验者可能存在无意识的偏差。

  In addition to the placebo effect, there may be unconscious bias by the experimenters.

- 理想情况是双盲（Double-blind）：实验对象自己和实验者都不知道任何实验对象接受了哪种处理。

  The ideal situation is double-blind: Neither the subjects themselves nor the experimenters know which treatment any subject has received.

3.11   关于实验的注意事项 Cautions about Experimentation

- 实验存在外部效度（External validity）差的问题：实验的对象、处理或环境可能无法真实地复制我们真正想要研究的条件。

  Experiments suffer from poor external validity: the subjects or treatments or setting of an experiment might not realistically duplicate the conditions we really want to study.

- 缺乏真实性也可能来自实验对象与一般人群不同：例如，育龄妇女通常被排除在临床试验之外，因为可能伤害胎儿。在男性身上测试的药物很可能在女性身上效果不同。

  Lack of realism can also come from the experimental subjects being different from the general population: for example, women of childbearing age are usually excluded from clinical trials because of the possibility of harming the fetus. Drugs tested on men may very likely work differently on women.

- 我们必须谨慎推广实验结论。

  We must take caution in generalizing about the conclusions of the experiment.

- 目标决定方法：描述总体？使用调查。确定原因？争取实验。
Goal Dictates Method: Describe a population? Use a survey. Establish a cause? Aim for an experiment.

4* Summary

- **数据来源（Sources of Data）：**

  - 轶事数据（Anecdotal data）：基于随意选择的个别案例，可能不具有代表性

  - 现有数据（Available data）：行政数据、大数据等，成本较低但可能非针对性

  - 抽样调查数据（Sample surveys data）：通过抽样了解总体特征，需注意抽样设计

  - 实验数据（Experiments data）：通过控制条件研究因果效应

- **抽样设计（Sampling Designs）：**

  - 简单随机抽样（Simple random sampling）：每个元素被选概率相等

  - 分层抽样（Stratified sampling）：先分层后抽样，确保各层代表性

  - 整群抽样（Cluster sampling）：以群为单位抽样，适用于分散总体

  - 多阶段抽样（Multistage sampling）：结合多种方法，常用于大型社会调查

- **因果关系与关联（Causation and Association）：**

  - 潜在变量（Lurking variable）：未被测量但影响研究变量的因素

  - 共同反应（Common response）：由潜在变量解释的虚假关联

  - 混杂（Confounding）：变量效应无法区分的问题

- **研究设计类型（Types of Research Design）：**

  - 观察性研究（Observational study）：仅观察不干预，难以确定因果关系

  - 实验设计（Experimental design）：主动施加处理，是确定因果关系的金标准

  - 自然实验与准实验（Natural and Quasi-experiments）：利用自然发生的变异近似随机化

- **实验设计原则（Principles of Experimental Design）：**

  - 控制（Control）：使用对照组作为比较基准

  - 随机化（Randomization）：随机分配消除偏差

  - 重复（Replication）：多次试验减少偶然变异

- 双盲设计（Double-blind）：避免实验者和受试者偏差

- 安慰剂效应控制（Placebo control）：区分实际处理效应和心理效应

- **注意事项与局限性（Cautions and Limitations）：**

  - 覆盖不足（Undercoverage）：某些群体未被纳入抽样框架

  - 无回答偏差（Nonresponse bias）：受访者拒绝参与影响代表性

  - 外部效度（External validity）：实验结论推广到真实世界的能力

  - 伦理限制（Ethical constraints）：某些实验设计不可行