MAT220 第1讲：统计学概念

目录

这是MAT220 的第一讲，所以我们的目标是建立统计学的基本语言。我们首先会问：统计学是什么，以及为什么它不仅仅是一堆公式。然后，我们会讨论数据的类型，包括四种测量层次：定类、定序、定距和定比。之后，我们会介绍抽样，比较几种抽样方法，并解释为什么不同的方法可能会导致不同的结论。最后，我们会讨论数据误导我们的主要方式：抽样问题、非抽样误差和推断错误。

---

1 什么是统计学?

统计学是对数据进行数学研究的学科。我们可以把数据理解为从现实世界中的某些事物那里收集到的信息。统计学的目的，是收集、整理、分析、解释并呈现这些信息，从而帮助我们做出更好的决策。

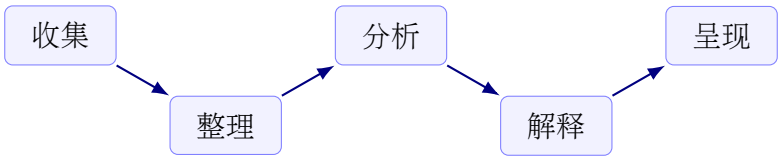
统计学广泛用于科学、商业、政府、医学、体育、教育和技术之中。只要有人试图用数据理解一个复杂的现实系统，他们就在使用统计思维。

定义

统计学是对数据的研究，尤其是对数据的收集、整理、分析、解释和呈现。

1.1 统计研究的工作流程

一项好的统计研究会把现实背景和数学技术结合起来，从而发现关于某个系统的新信息。统计研究的一种基本形式如下：



这里：

- **收集**是指测量现实世界中的某个系统，并从中记录数据。
- **整理**是指把数据重新安排成有用的格式。
- **分析**是指使用合适的数学方法来描述数据集中的有用特征。
- **解释**是指说明结果在现实背景中的含义。
- **呈现**是指清楚、简洁、诚实地传达结果。

注意，上面的工作流程并不仅仅是一组数学工具。相反，一项统计研究是数学技术与现实的实践性考虑之间的一种有根据的结合。

1.2 为什么统计学有时显得枯燥

统计学有时会显得枯燥、干巴巴，因为它具有公式化的特点。事实上，统计学中使用的许多公式背后都有深刻的数学理论，但实用统计学很少直接使用这些理论。从数学复杂性的角度来看，统

计学常常显得枯燥，因为它看起来非常简单，而且许多最常见的公式只使用基本算术。

实际上，统计学也正是因为同样的原因而非常有趣：即使从最基本的数学运算出发，统计学也能创造出一系列巧妙的公式，用来描述数据中的整体结构。统计学的有趣之处在于，它从如此朴素的起点出发，却成为了一种如此强大而有用的工具。统计学的实际用途再怎么强调也不为过：世界由信息组成，而关于你和社会的许多性质都可以被转化为数据并加以分析。

1.3 描述统计与推断统计

统计学有两大主要支柱：描述统计和推断统计。

- **描述统计**是描述我们已经拥有的数据的技术。这包括表格、图表、平均数、百分比、变异性以及其他概括。
- **推断统计**是根据少量数据对更大群体作出结论的科学。这包括置信区间、假设检验和统计建模。

例子。 假设我们测量了50 名学生的通勤时间。

- 如果我们计算这50 名学生的平均通勤时间，那么我们是在做**描述统计**。
- 如果我们用这50 名学生来估计全校学生的平均通勤时间，那么我们是在做**推断统计**。

2 数据的类型

2.1 数据获取

在现实中，我们并不总是能够获得完整的数据集。通常，我们只能获得我们能够收集到的数据。推断统计建立了一些工具，帮助我们在并不完全了解一个较大群体的情况下，推断关于它的信息。这些方法极其有力，也很有洞察力；然而，数据的质量取决于我们获取数据的方法。如果一开始收集到的就是糟糕的数据，那么再高级的数学分析也不会有帮助。不需要的影响可能会意外进入统计分析之中，从而导致错误的结论。我们的任务是正确识别这些潜在影响，并尽量减少它们。

重要观点

糟糕的数据通常不能靠好的数学来拯救。如果数据是以混乱、有偏或粗心的方式收集的，那么即使数学上技术正确，最终的统计结论也可能不可靠。

2.2 定量数据与分类数据

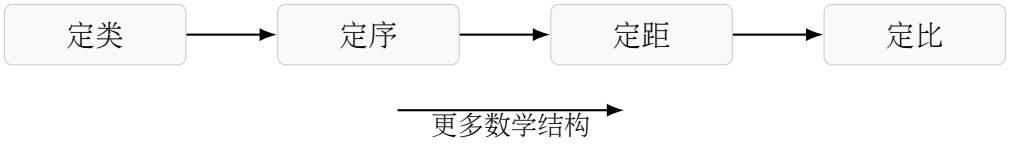
数据大致可以分为两大类：

- **定量数据**是数值型数据。
- **分类数据**是非数值型数据。

定量数据具有数值结构，因此通常可以使用算术运算进行分析。分类数据更像是一组标签、名称或类别，因此可供操作的数学结构要少得多。

2.3 测量层次

现实世界中我们可能收集到许多不同类型的数据。因此，根据我们能够获得的数据类型，我们可能可以，也可能不可以，对它们有意义地执行某些数学运算。处理这个潜在混乱问题的一种方法，是使用Stevens 于1946 年提出的测量层次。测量有四种基本层次：



粗略地说，数学结构越多，我们就能对数据执行越多运算。下面我们逐一回顾这些测量层次。

重点是什么?

测量层次帮助我们直接而明确地说明自己的数据。通过对数据类型进行分类，我们告诉读者哪些数学运算是有意义的，哪些主张是不恰当的。

2.3.1 定类尺度

定类数据由没有内在顺序的标签或类别组成；它是分类的。这一测量层次没有数值结构，所以我们能做的只是判断两个观测是否相同或不同，并统计它们出现的频数。但是，我们没有规定好的顺序，也没有可以使用的算术结构。

定类尺度

定类数据由没有自然顺序的质性类别组成。例子包括：

- 最喜欢的颜色，
- 血型，
- 国籍，
- 电影类型，
- 通勤使用的铁路线路。

对于定类数据，检验是否相等以及统计频数是有意义的。

2.3.2 定序尺度

英文中的“ordinal” 与“order” 意思相近。因此，定序数据是具有有意义顺序或排名的数据。然而，排名之间的间隔并不假定为相等，也不假定在任何意义上可测量。

定序尺度

定序数据由具有有意义顺序的类别或数值组成，但相邻数值之间的距离并不清楚。
例子：

- 参赛者排名，
- 疼痛量表，
- 满意度回答，例如“非常不满意”、“不满意”、“中立”、“满意”、“非常满意”，
- 年级身份，例如一年级、二年级、三年级和四年级。

对于定序数据，中位数和百分位数通常是有意义的。均值在实践中有时也会使用，但应当小心，因为相邻类别之间的距离未必相等。

2.3.3 定距尺度

定距数据具有有意义的顺序，并且数值之间的差异相等。然而，零并不表示该数量的绝对不存在。正因为如此，像“两倍多”这样的比率陈述通常没有意义。

定距尺度

定距数据具有有意义的顺序和有意义的差异，但没有真正的零点。
例子：

- 用摄氏度测量的温度，
- 出生年份。

例如，从物理意义上说，20°C 并不是10°C 的“两倍热”，因为0°C 并不表示温度不存在。然而，20°C 与30°C 之间的差异，和30°C 与40°C 之间的差异一样大。

2.3.4 定比尺度

定比数据具有顺序、相等间隔，以及表示该数量不存在的真正零点。因此，差异和比率都是有意义的。

定比尺度

定比数据具有有意义的顺序、有意义的差异，以及真正的零点。
例子：

- 体重，
- 身高，
- 长度，
- 时间长度，
- 教室中的学生人数，
- 薪水。

例如，60 分钟的通勤时间确实是30 分钟通勤时间的两倍长，因为存在一个可作为基准的绝对零点，即0 分钟。这种比较是有意义的，因为0 分钟意味着没有时间经过。

2.3.5 测量层次：总结

层次	主要结构	允许的 比较	例子
定类	只有类别	相同或不同	血型
定序	有顺序的类别	大于或小于	满意度评分
定距	差异相等，没有真正零点	差异	摄氏温度
定比	差异相等且有真正零点	差异和比率	身高

练习

将每个变量分类为定类、定序、定距或定比。

(a) 学生最喜欢的便利店。
(b) 从1 到5 的满意度评分。
(c) 用摄氏度测量的温度。
(d) 通勤到校园所用的时间。

解答

(a) 定类，因为回答是类别标签。
(b) 定序，因为回答有顺序，但评分之间的间隔未必相等。
(c) 定距，因为差异有意义，但摄氏零度并不是真正的温度不存在。
(d) 定比，因为时间长度有真正的零点，并且比率有意义。

3 作为概念的抽样

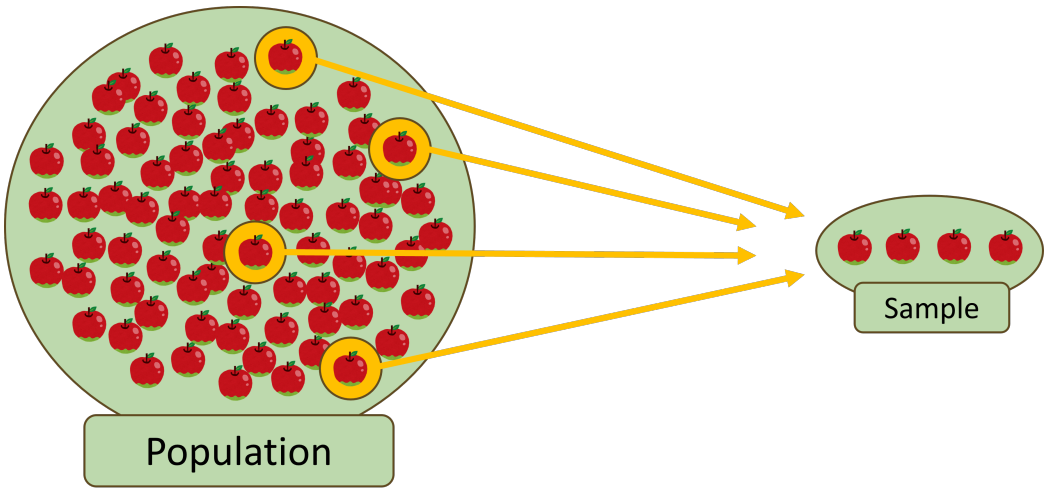
3.1 核心定义

如前所述，在研究现实世界时，我们想研究的对象集合常常太大，无法完整处理。因此，在实践中，我们经常只取其中较小的一部分，并研究这一部分。这种现实的、实践性的考虑引出了下面的术语。

定义

- **总体：** 你想了解的完整人群或事物群体。
- **样本：** 你实际从中收集数据的较小群体，它是从总体中选出的。
- **总体参数：** 描述整个总体的真实数值，即使你并不准确知道它。
- **样本统计量：** 从样本中计算出来的数值，常用于估计总体参数。

换句话说，总体就是我们真正想研究的集合。我们想发现的事实称为总体参数，但我们通常无法直接发现它们，因为总体太大了。因此，我们转而取一个子集，也就是从整体总体中选出的较小群体。然后，我们可以在这个较小群体上进行更容易的计算，由此得到某些部分信息，称为样本统计量。下面是样本的一个图示例子：



3.2 例题：东京的平均薪水

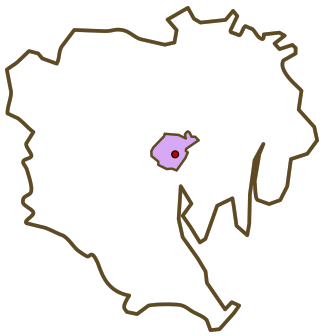
从总体中取得样本的过程称为抽样。抽样是统计学中极其重要的一部分，本课程的大部分内容都会围绕样本与总体之间的关系展开。然而，在抽样时我们必须非常小心，因为我们可能会意外引入某些不能代表真实总体的影响。下面我们用一个例子来说明这个想法。

假设Jenny 是东京市中心千代田区一家金融公司的实习生。老板要求她弄清楚东京居民的平均薪水是多少。因此，Jenny 需要进行一项统计研究来找出答案。由于Jenny 并不为市政府工作，她无法获得关于东京人口的完整薪水数据。于是，她决定抽取30 名东京居民作为样本，然后计算这些人的平均薪水。

在进行统计研究时，Jenny 意识到她的结果会强烈依赖于她询问的是哪些人。例如，样本会因以下因素而不同：

- 她抽样的时间，
- 她进行调查的位置，
- 参与者的年龄，

等等。下面是Jenny 可能抽取30 人样本的四种不同方式，以及每种方式内在的问题。

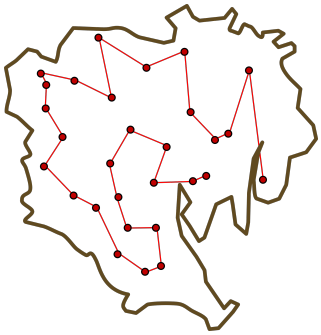


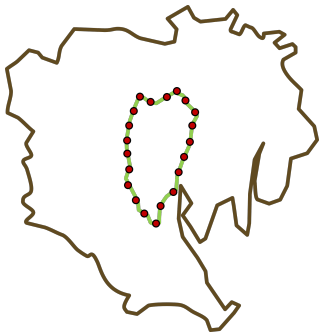
方法1： 从她在千代田区的办公室中选择30 名参与者。

缺乏多样性： 从她附近的人群中抽样当然很方便。然而，她得到的数字可能并不能代表整个东京。有些地区更富或更穷，居民更老或更年轻，工作类型也不同。因此，即使她询问了30 人，她也可能只是意外地测量了那个地区的特征，而不是整个东京的特征。以千代田区为例，她的样本很可能会高估平均薪水。

方法2： 从东京任意地点随机选择30 名参与者。

执行上的困难： 从全市随机抽样可能看起来没有偏见，但实际操作会很昂贵且耗时。她不仅需要前往随机选中的各个地区，而且还会遇到其他潜在偏差，这取决于她如何选择要询问的代表。尽管如此，一个在纸面上看起来没有偏见的计划，在实践中可能会因为资源限制而失败。



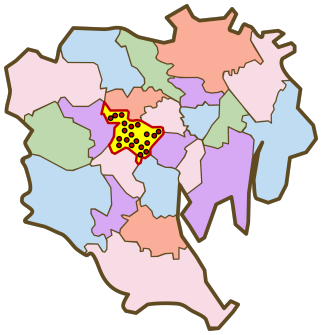


方法3： 沿山手线移动，在30 个车站的每一站停下，并在每站询问一人。

缺乏可达性： “每站一人”的抽样可能会包括游客和不住在东京的外来者。它也可能意外地偏向旅游车站、通勤枢纽或商业区，从而使样本偏向某些类型的工作者。换句话说，这个样本可能偏向一种特定类型的东京居民：在山手线沿线工作或出行的人。

方法4： 从东京23 个特别区中随机选择一个区，然后在该区内随机选择30 名参与者。

缺乏多样性： 虽然这比走遍整个东京更方便，也比Jenny 在千代田区附近抽样更少偏见，但这种样本中仍然会缺乏多样性。随机性并不能保证小样本能够完整代表总体的多样性。它只是让选择的第一阶段变成随机的。



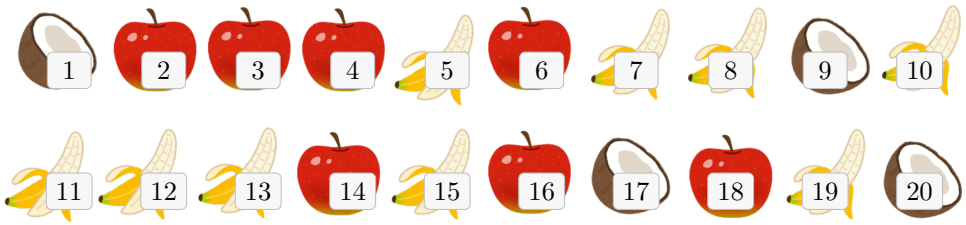
那么，这里最好的抽样方法是什么？事实上，这取决于Jenny 想做什么。在实践中，通常不存在唯一的“最好方法”。相反，抽样往往是在资源和误差之间取得平衡，有时我们必须妥协。理想情况下，进行研究的研究者应该理解这些偏差，并且在传达结果时尽可能明确说明这些偏差。

实践中的抽样

抽样往往是在资源和误差之间取得平衡。最准确的数据收集方法通常更昂贵、更困难或更耗时。有时我们需要妥协，但在呈现结果时应当明确说明这种妥协。

4 抽样方法

受上一例中Jenny 的启发，我们现在回顾一些常见的抽样方法。在本节中，我们假设总体由20 个水果组成：7 个苹果、9 根香蕉和4 个椰子。



我们的目标是使用不同的抽样方法，从这个总体中选择一个大小为5 的样本。

4.1 简单随机样本

定义：简单随机样本

简单随机样本是这样取得的：给定大小的每一个可能样本都有相同的可能性被选中。

例如，从我们的水果总体中，一个可能的大小为5 的简单随机样本是：

样本 5 9 10 12 18

另一个可能的大小为5 的简单随机样本是：

样本 2 4 8 14 17

在简单随机样本中，我们并不刻意控制样本中有多少苹果、香蕉和椰子。样本在选择规则上是公平的，但它仍然可能因为偶然而不平衡。

4.2 系统样本

定义：系统样本

系统样本是通过选择一个随机起点，然后选择总体中此后每第k 个成员而取得的。

例如，如果我们从1 开始，并选择每第4 个成员，就得到：

样本 1 5 9 13 17

如果我们从3 开始，并选择每第4 个成员，就得到：

样本 3 7 11 15 19

系统抽样可能很方便，特别是当总体已经排列成列表时。然而，如果列表本身存在重复模式，它也可能很危险。例如，如果工厂列表中每第4 个物品都来自同一台机器，那么选择每第4 个物品可能会意外地只抽到一台机器的产品。

4.3 分层样本

定义：分层样本

分层样本是这样取得的：先根据某个相关特征把总体分成若干组，这些组称为**层**，然后从每一层中随机抽样。

在我们的水果例子中，自然的层是3 种水果：

$$\text{苹果}, \qquad \text{香蕉}, \qquad \text{椰子}.$$

一个分层样本可能会刻意选择一些苹果、一些香蕉和一些椰子，从而使样本代表所有三种水果类型。

大小为5 的分层样本例子：

样本 2 6 9 14 19

这里我们从每种水果类型中都进行了抽样。目标不仅仅是随机性，而且是重要群体之间的代表性。

当总体具有重要子群体，并且我们想确保每个子群体都被代表时，分层抽样很有用。

4.4 整群样本

定义：整群样本

整群样本是这样取得的：把总体分成若干群，通常根据自然或方便的分组方式，然后随机选择一个或多个完整的群。

例如，假设我们的水果列表被分成大小为5 的群：

$$\{1,2,3,4,5\}, \quad \{6,7,8,9,10\}, \quad \{11,12,13,14,15\}, \quad \{16,17,18,19,20\}.$$

如果我们随机选择第二群，那么样本是：

样本 6 7 8 9 10

如果我们随机选择第三群，那么样本是：

样本 11 12 13 14 15

当从分散在整个总体中的个体抽样既昂贵又困难时，整群抽样很有用。例如，研究者可以随机选择几个教室，并调查这些教室里的每一名学生。

4.5 常见混淆：分层抽样与整群抽样

分层抽样和整群抽样可能看起来相似，因为它们都会把总体分成若干组。区别在于这些组扮演的角色不同。

方法	组如何形成	样本如何选择
分层抽样	根据重要特征形成组，例如水果类型、性别、年级或行政区。	从每一层中随机抽样。
整群抽样	组通常是自然或方便的群，例如教室、社区或列表区块。	随机选择完整的群。

简而言之：在**分层抽样**中，我们从每个组中抽样；但在**整群抽样**中，我们抽取的是完整的组本身。

4.6 其他抽样方法

还有许多其他抽样方法。有些在实践中很有用，但它们通常也有严重限制。

方法	基本想法
方便抽样	抽取最容易接触到的人。
自愿回应抽样	人们自行决定是否参与。
面板抽样	随时间追踪同一组人。
随意抽样	没有清晰随机规则，以非正式或粗心的方式抽样。
滚雪球抽样	现有参与者帮助招募未来参与者。
极端案例抽样	刻意抽取极端案例，例如最小值和最大值。

这些方法并不自动就是错误的，但必须谨慎解释。例如，方便样本可能便宜而快速，但它告诉我们的可能更多是“谁容易接触到”，而不是完整总体的情况。

4.7 回到Jenny 的东京研究

现在，我们可以使用抽样语言描述Jenny 的方法：

- 方法1：**在千代田区办公室附近询问30 人。** 这最接近方便样本。主要问题是千代田区可能不能代表东京。
- 方法2：**从东京随机抽取30 人。** 这最接近简单随机样本。主要问题是原则上准确，但实践中困难。
- 方法3：**在每个山手线车站询问一人。** 这最接近基于路线的方便样本。主要问题是它可能抽到通勤者和游客，而不是居民。
- 方法4：**随机选择一个区，然后在其中抽样。** 这最接近整群式样本。主要问题是一个区可能不能代表所有区。

这些术语本身不能解决实践问题，但它们帮助我们清楚地描述研究的优点和弱点。

5 数据何时误导我们

统计结论可能失败，主要有三个原因。

统计结论可能失败的三种方式

1. **抽样问题：** 样本不能诚实地代表总体。
2. **非抽样误差：** 即使样本选择得很好，数据本身也被扭曲了。
3. **推断错误：** 结论超出了数据实际支持的范围。

一项好的统计研究必须处理所有这三点。拥有大样本还不够，使用正确公式也不够。整个推理链条都必须可信。

5.1 抽样误差与抽样偏差

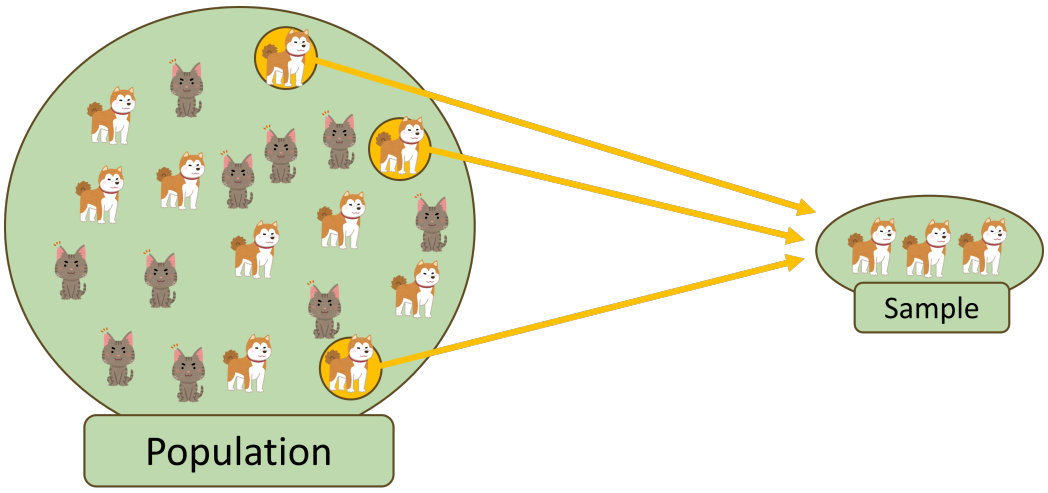
抽样问题有两种重要类型。

抽样误差与抽样偏差

- **抽样误差**是样本与总体之间的随机不匹配。即使抽样方法是公平的，这也可能发生。
- **抽样偏差**是抽样方法中的系统性不公平。方法本身使总体中的某些部分更容易或更不容易被包括进来。

5.1.1 例子：抽样误差

假设总体中猫和狗的数量相等。我们抽取一个大小为3 的简单随机样本，结果碰巧选中的3 只动物全都是狗。



这是抽样误差，因为方法是公平的，但这个具体样本不走运。

5.1.2 例子：抽样偏差

假设Jenny 只询问千代田区这样富裕地区的人关于薪水的问题。这不仅仅是运气不好。抽样方法本身就是系统性不公平的，因为它过度代表了富裕人群。

这是抽样偏差。如果Jenny 多次重复同样的方法，她会不断得到倾向于高估东京平均薪水的样本。

5.2 非抽样误差

即使样本选择得很好，数据本身仍然可能被扭曲。这些扭曲称为**非抽样误差**。

定义：非抽样误差

非抽样误差是统计研究中的一种误差，它不是由随机选择样本这一行为造成的。

常见的非抽样误差包括：

- 问题不清楚或具有诱导性，
- 测量不准确，
- 回答不诚实，
- 回应缺失，
- 记录错误，
- 数据输入错误，
- 设备故障，

等等。重要的是，即使抽样方法本身合理，这些误差也可能发生。

5.3 例子：带偏见的调查问题

假设你调查东京的通勤者，并问：

“你今天早上的电车之旅有多糟糕、多可怕？”

这是一个糟糕的问题，因为它假定了电车之旅是糟糕和/或可怕的，因此会把回答者推向负面答案。更好的版本是：

“你今天早上的电车之旅怎么样？”

非常差	差	一般	好	非常好

这里的措辞更加中立，不会鼓励回答者作出某种特定的价值判断。

5.4 结论变得有偏的常见方式

有些偏差不仅仅是技术错误。相反，它们有时是我们选择、记忆或解释证据时的失败。下面是这一现象的三个常见例子。

幸存者偏差

幸存者偏差发生在我们只观察那些经过某个选择过程后幸存下来的例子时。
例子：“过去的建筑太美了，而现代建筑太丑了。”
这可能具有误导性，因为许多丑陋的旧建筑已经被拆毁，而美丽的旧建筑被保存了下来。幸存下来的建筑并不是过去建筑的随机样本。

选择偏差

选择偏差发生在数据的选择方式偏向某个结论时。
例子：“我问了歌舞伎町的人东京是否吵闹混乱。他们说是。因此，东京是吵闹混乱的。”
歌舞伎町并不是东京的代表性样本。这个结论告诉我们的可能更多是关于歌舞伎町，而不是关于整个东京。

资金偏差或利益冲突

资金偏差可能在研究由某个会从特定结论中受益的团体资助时发生。
例如，如果某个行业资助关于其自身产品是否有害的研究，我们就应该仔细关注研究设计、数据、分析和解释。

5.5 推断错误

即使样本合理、数据记录准确，最终结论仍然可能是错误的。当推断超出证据所支持的范围时，这种情况就会发生。

常见的推断错误包括:

- 从小样本推广到巨大总体，
- 把相关性当作因果关系，
- 忽略潜在变量，
- 忽略混杂变量，
- 正确使用某个统计量但错误解释它，
- 作出比证据所能支持的更强的结论。

5.6 相关性与因果关系

两个变量可能相关，但其中一个并不导致另一个。如果两件事一起发生，这并不自动意味着其中一件事是另一件事的原因。

例子。 假设在东京非常炎热的日子里，便利店卖出更多瓶装水，同时更多人报告中暑。
如果因此得出“购买瓶装水导致中暑”的结论，那就很荒唐。更合理的解释是，炎热天气同时影响了这两个变量。

在这个例子中，瓶装水销量和中暑是相关的，但因果解释涉及另一个变量：温度。

5.7 判断统计主张是否可信的检查清单

在相信一个统计结论之前，问下面这些问题。

1. **总体**： 这个主张真正涉及的总体是什么?
2. **样本**： 实际被抽样的是谁或什么?
3. **抽样方法**： 样本成员是如何被选择的?
4. **数据质量**： 测量、问题和记录是否可靠?
5. **缺失数据**： 谁没有回应? 这是否可能重要?
6. **变量**： 是否存在潜在变量或混杂变量?
7. **结论**： 结论是否从数据中推出，还是过度延伸了?

总结

一个统计主张的强度取决于其背后的完整链条：

$$\text{收集} \longrightarrow \text{整理} \longrightarrow \text{分析} \longrightarrow \text{解释} \longrightarrow \text{呈现}.$$

如果链条中的某一部分薄弱，最终结论也可能薄弱。