

第二章 数据的搜集

李德山

四川师范大学商学院

2022 年 3 月 5 日

Contents

- ① 数据的来源
- ② 调查方法
- ③ 实验方法
- ④ 数据的误差

问题的提出

- 下图自零点研究咨询集团发布的一项关系城市居民“快乐”指数的调查结果分析。该调查是采用多阶段随机抽样方式于 2008 年 5 月针对北京、上海、广州、武汉、沈阳、西安、成都 7 个城市共 1883 名 18-60 岁居民的入户访问。
- 这样的抽样调查，我们常在电视和报刊上见到。但其中所表示的意义是什么？

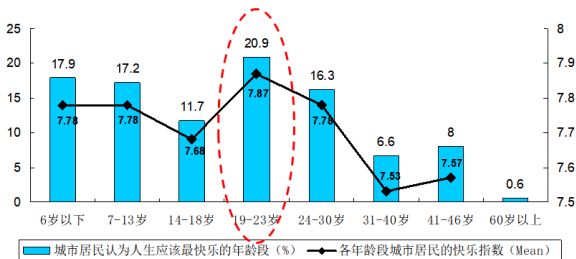


Figure: 城市居民对快乐的认识

问题的提出

- 1. 国家统计局每年公布的人口规模、GDP 数据是如何调查统计得到的？
- 2. 你收集的是一手数据还是二手数据？
- 3. 你有没有参与或组织过微观数据的调研活动？

问题的提出

- 1. 国家统计局每年公布的人口规模、GDP 数据是如何调查统计得到的？
- 2. 你收集的是一手数据还是二手数据？
- 3. 你有没有参与或组织过微观数据的调研活动？

问题的提出

- 1. 国家统计局每年公布的人口规模、GDP 数据是如何调查统计得到的？
- 2. 你收集的是一手数据还是二手数据？
- 3. 你有没有参与或组织过微观数据的调研活动？

① 数据的来源

② 调查方法

③ 实验方法

④ 数据的误差

数据的来源

- 数据的来源包括间接来源和直接来源

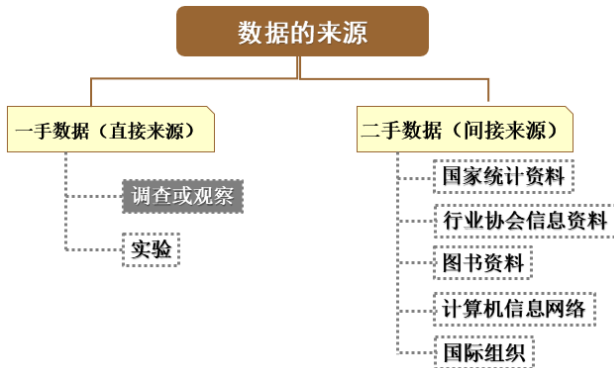


Figure: 统计数据的来源

数据的来源

- 一般而言，统计调查是获取数据的主要形式，收集到的主要是第一手资料
- 查阅文献、年鉴，上因特网等是获取统计资料的辅助形式，收集到的主要是第二手资料

一些常见数据来源

- 世界银行的数据库 <http://devdata.worldbank.org/data-query/>
- 美国普查局 <http://www.census.gov/>
- 国家统计局普查中心 <http://www.stats.gov.cn/tjfx/ztfx/decjbdwpc/>
- 中国统计年鉴 <http://www.stats.gov.cn/tjsj/ndsjs/>
- 国研网 <http://www.drcnet.com.cn/www/int/>

二手数据

- 二手数据的特点
 - 1. 搜集容易，采集成本低
 - 2. 作用广泛
 - 分析所要研究的问题
 - 提供研究问题的背景
 - 帮助研究者更好地定义问题
 - 检验和回答某些疑问和假设
 - 寻找研究问题的思路和途径
 - 3. 搜集二手资料在研究中应优先考虑

数据的来源

- 数据是谁搜集的？可信度！
- 为什么目的而搜集的？
- 数据是怎样搜集的？
- 什么时候搜集的？
- 使用二手数据应注意的问题：
 - 1. 数据的统计口径和计算方法。
 - 2. 二手数据的时效性。
 - 3. 应充分了解二手数据的来源和可靠度。
 - 4. 使用二手数据，应该注明数据的出处。

数据的来源

- 数据是谁搜集的？可信度！
- 为什么目的而搜集的？
- 数据是怎样搜集的？
- 什么时候搜集的？
- 使用二手数据应注意的问题：
 - 1. 数据的统计口径和计算方法。
 - 2. 二手数据的时效性。
 - 3. 应充分了解二手数据的来源和可靠度。
 - 4. 使用二手数据，应该注明数据的出处。

数据的来源

- 数据是谁搜集的？可信度！
- 为什么目的而搜集的？
- 数据是怎样搜集的？
- 什么时候搜集的？
- 使用二手数据应注意的问题：
 - 1. 数据的统计口径和计算方法。
 - 2. 二手数据的时效性。
 - 3. 应充分了解二手数据的来源和可靠度。
 - 4. 使用二手数据，应该注明数据的出处。

数据的来源

- 数据是谁搜集的？可信度！
- 为什么目的而搜集的？
- 数据是怎样搜集的？
- 什么时候搜集的？
- 使用二手数据应注意的问题：
 - 1. 数据的统计口径和计算方法。
 - 2. 二手数据的时效性。
 - 3. 应充分了解二手数据的来源和可靠度。
 - 4. 使用二手数据，应该注明数据的出处。

数据的来源

- 数据是谁搜集的？可信度！
- 为什么目的而搜集的？
- 数据是怎样搜集的？
- 什么时候搜集的？
- 使用二手数据应注意的问题：
 - 1. 数据的统计口径和计算方法。
 - 2. 二手数据的时效性。
 - 3. 应充分了解二手数据的来源和可靠度。
 - 4. 使用二手数据，应该注明数据的出处。

数据的来源

- 1. 调查数据
 - 通过调查方法获得的数据
 - 通常是对社会现象而言
 - 通常取自有限总体
- 2. 实验数据
 - 通过实验方法得到的数据
 - 通常是对自然现象而言
 - 也被广泛运用到社会科学中

① 数据的来源

② 调查方法

③ 实验方法

④ 数据的误差

统计调查方式

- 统计调查方式：普查、抽样调查、统计报表、重点调查和典型调查等。
- 以上调查方式的优缺点。

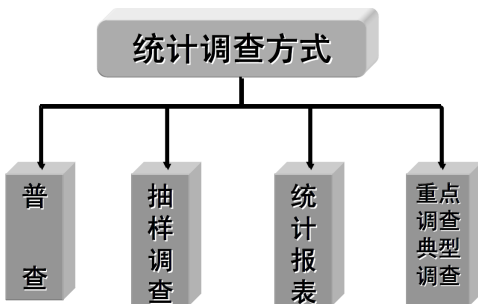



Figure: 统计调查方式

统计调查方式

- 常见普查：人口普查、经济普查、农业普查、工业普查等。

人口普查	第六次人口普查数据	第五次人口普查数据	
经济普查	第三次经济普查	第二次经济普查	
	第一次经济普查		
农业普查	第二次农业普查	第一次农业普查	
	(经普 农业卷 农村卷 农民卷 综合卷)		
RAD资源调查	第二次RAD资源调查		<p>人口普查：每十年一次，逢“0”进行 第六次人口普查，2010</p> <p>经济普查：每五年一次，逢“3”、“8”进行 第三次经济普查，2013</p>
工业普查	第三次工业普查		
三产普查	第一次三产普查		
基本单位普查	第二次基本单位普查		

二、标准时点

第六次全国人口普查的标准时点为：2010年11月1日零时。普查员在掌握普查标准时点时，应该注意以下两点：

(一) 2010年11月1日零时以后出生的人不登记；2010年11月1日零时以后死亡的人仍要登记普查表短表或普查表长表。

(二) 2010年11月1日零时以后发生迁移的人，仍在原地登记。

调查方法

- 抽样主要考虑： 调查的效率和精确性
- 常见抽样方法： 概率抽样与非概率抽样

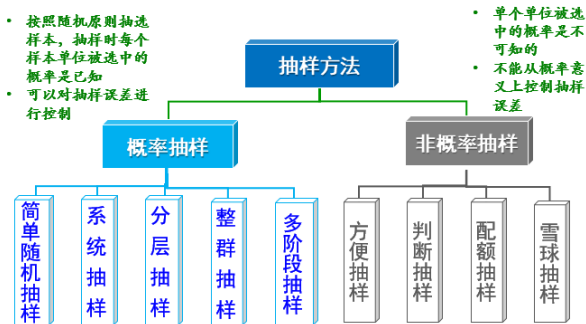


Figure: 抽样方法

调查方法

- 抽样单位 (Sampling unit): 在抽样调查中可以把总体分成若干个互不重叠又穷尽的有限个部分
- 抽样单位可以是一个总体单位, 也可以包含多个个体
- 抽样框 (Sampling Frame): 抽样单位的名单
- 抽样框应尽可能与目标总体相一致, 例如名单抽样框、区域抽样框、时间表抽样框

从5000名学生中抽选500名学生进行调查, 抽样单位和抽样框?

抽样单位: 每一个学生, 抽样框: 全校5000名学生的名单

从全校100个班级中抽选10个班进行调查, 抽样单位和抽样框?

抽样单位: 每一个班级, 抽样框: 全校100个班级的名单

Figure: 抽样方法

调查方法

- 调查实践中经常采用的概率抽样方式：
 - (1) 简单随机抽样 (simple random sampling)
 - (2) 分层抽样 (stratified sampling)
 - (3) 整群抽样 (cluster sampling)
 - (4) 系统抽样 (systematic sampling)
 - (5) 多阶段抽样 (multi stage sampling)

简单随机抽样

- 简单随机抽样 (simple random sampling)：又叫纯随机抽样，是最简单、最普遍的抽样组织方法。从总体 N 个单位中随机地抽取 n 个单位作为样本，每个单位入抽样本的概率是相等的。

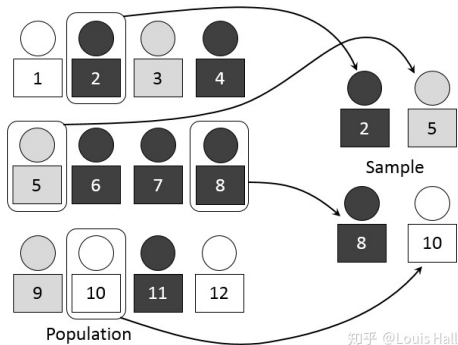


Figure: 简单随机抽样

简单随机抽样

- 有放回抽样和无放回抽样
- 通常有抽签法和随机数法两种抽选方法
- 特点：
 - (1) 简单、直观，在抽样框完整时，可直接从中抽取样本；
 - (2) 用样本统计量对目标量进行估计比较方便。
- 局限性：
 - (1) 当 N 很大时，不易构造抽样框；
 - (2) 抽出的单位很分散，给实施调查增加了困难；
 - (3) 没有利用其它辅助信息以提高估计的效率。

分层抽样

- 分层抽样 (stratified sampling)**：是将抽样单位按某种特征或某种规则分为不同的层，然后从不同的层中独立、随机地抽取样本。将各层的样本结合起来，对总体的目标量进行估计。

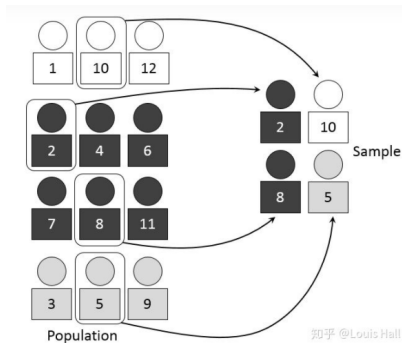


Figure: 分层抽样

分层抽样

- 优点

- (1) 保证样本的结构与总体的结构比较相近，从而提高估计的精度；
- (2) 组织实施调查方便；
- (3) 既可以对总体参数进行估计，也可以对各层的目标量进行估计。

整群抽样

- **整群抽样** (cluster sampling) : 将总体中若干个单位合并为组 (群), 抽样时直接抽取群, 然后对中选群中的所有单位全部实施调查。

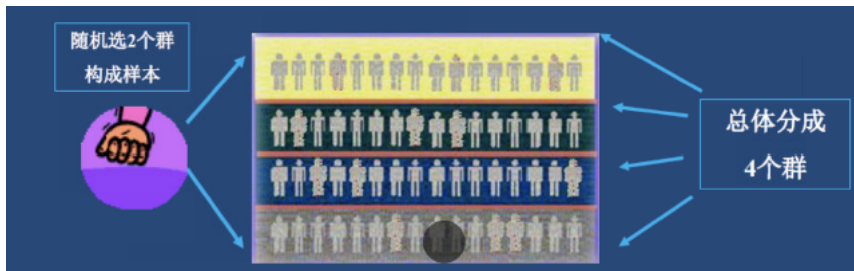


Figure: 整群抽样

整群抽样

- 特点

- (1) 抽样时只需群的抽样框，可简化工作量
- (2) 调查的地点相对集中，节省调查费用，方便调查的实施；
- (3) 缺点是估计的精度较差。

系统抽样

- 系统抽样** (systematic sampling): 又叫等距抽样或机械抽样。将总体中的所有单位 (抽样单位) 按一定顺序排列, 在规定的范围内随机地抽取一个单位作为初始单位, 然后按事先规定好的规则确定其它样本单位。

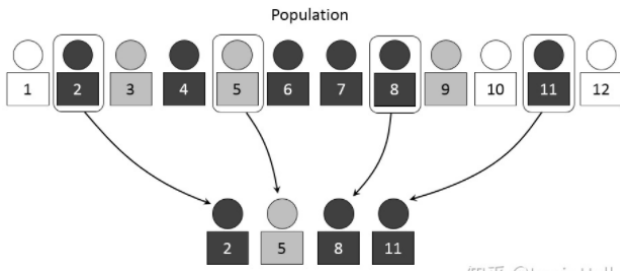


Figure: 系统抽样

系统抽样

- 优点：操作简便，可提高估计的精度。
- 缺点：对估计量方差的估计比较困难。

多阶段抽样

- **多阶段抽样** (multi stage sampling)：类似整群抽样，首先抽取群，但并不是调查群内的所有单位，而是再进一步抽样，从选中的群众中抽取若干单位进行调查。
- 例如，对成都市就业者”过劳”现状及成因进行调查：
 - 第一阶段：从成都市 11 个区（不包括县和代管县市）抽取 6 个辖区；
 - 第二阶段：从被抽取的 6 个辖区中各抽取 5 个街道；
 - 第三阶段：从被抽取的 30 个街道中抽取样本单位。

多阶段抽样

- 具有整群抽样的优点，保证样本相对集中，节约调查费用
- 需要包含所有低阶段抽样单位的抽样框；同时由于实行了再抽样，使调查单位在更广泛的范围内展开
- 在大规模的抽样调查中，是经常被采用的方法
- 注意：在大规模的抽样调查中，抽取样本的阶段应尽可能少。因为每增加一个抽样阶段就会增添一份估计误差，用样本对总体进行估计也就更加复杂。

非概率抽样

- 非概率抽样 (non-probability sampling), 相对于概率抽样而言。抽取样本时不是依据随机原则, 而是根据研究目的对数据的要求, 采用某种方式从总体中抽出部分单位对其实施调查。
- 有方便抽样、判断抽样、自愿样本、滚雪球抽样、配额抽样等方式。

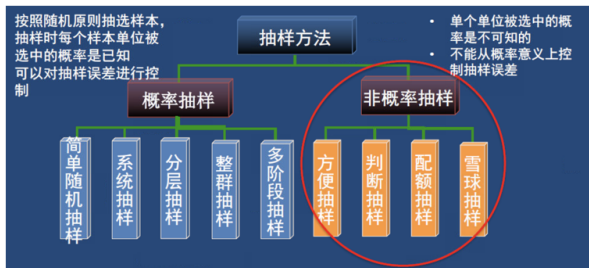


Figure: 非概率抽样

方便抽样

- 方便抽样：调查过程中由调查员依据方便的原则，自行确定入抽样本的单位。调查员在街头、公园、商店等公共场所进行拦截调查。
 - 优点：容易实施，调查的成本低。
 - 缺点：样本单位的确定带有随意性，样本无法代表有明确定义的总体，调查结果不宜推断总体。

判断抽样

- 判断抽样：研究人员根据经验、判断和对研究对象的了解，有目的选择一些单位作为样本。
 - 判断抽样是主观的，样本选择的好坏取决于调研者的判断、经验、专业程度和创造性。
 - 抽样成本比较低，容易操作。
 - 样本是人为确定的，没有依据随机的原则，调查结果不能用于推断总体。

自愿抽样

- 自愿抽样：被调查者自愿参加，成为样本中的一分子，向调查人员提供有关信息。例如，参与报刊上和互联网上刊登的调查问卷活动，向某类节目拨打热线电话等，都属于自愿样本。
 - 自愿样本与抽样的随机性无关
 - 样本是有偏的，不能依据样本的信息推断总体

滚雪球抽样

- 滚雪球抽样：先选择一组调查单位，对其实施调查之后，再请他们提供另外一些属于研究总体的调查对象，调查人员根据所提供的线索，进行此后的调查。这个过程持续下去，就会形成滚雪球效应。
 - 适合于对稀少群体和特定群体研究
 - 优点：容易找到那些属于特定群体的被调查者，调查的成本也比较低

配额抽样

- 配额抽样：先将总体中的所有单位按一定的标志（变量）分为若干类，然后在每个类中采用方便抽样或判断抽样的方式选取样本单位。
 - 操作简单，可以保证总体中不同类别的单位都能包括在所抽的样本之中，使得样本的结构和总体的结构类似
 - 抽取具体样本单位时，不是依据随机原则，属于非概率抽样

调查方法

- 概率抽样与非概率抽样的比较
- 概率抽样
 - 依据随机原则抽选样本
 - 样本统计量的理论分布存在
 - 可根据调查的结果推断总体
- 非概率抽样：
 - 不是依据随机原则抽选样本
 - 样本统计量的分布是不确定的
 - 无法使用样本的结果推断总体

调查方法

- “面对互联网的海量信息，大数据是导航仪。基于抽样调查 + 人口学特征的‘小样本模式’不再具有指航性。”
- 任何一个网站的数据都只是互联网行为数据的一个子集，而非全集——看起来的全数据恰恰是残缺数据
- 注：调查数据只要抽样合理，同样可以得到准确的结论



大数据+小数据？

调查方法

- 大数据对于传统经济统计，是补充，而非替代
- 横向来看，传统统计方法在经济增长、税收、贸易、收入分配等领域的统计上具有主导优势
- 大数据在物价、通货膨胀、失业率、消费等方面的统计上更具有优势



Figure: 大数据时代

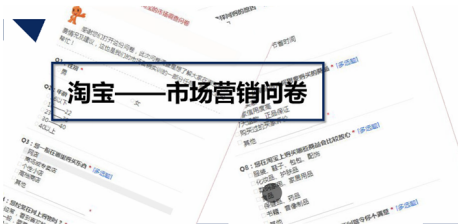
搜集数据的基本方法

- 常见的搜集数据的本方法有：自填式、面访式、电话式。
 - 自填式问卷调查：没有调查员协助的情况下由被调查者自己完成调查问卷。问卷递送方法有：调查员分发、邮寄、网络、媒体。
 - 面访式问卷调查：调查员与被调查者面对面提问、被调查者回答的一种调查方式。包括个别深度访谈和座谈会。
 - 电话式问卷调查：通过电话向被调查者实施调查。

项目	自填式	面访式	电话式
调查时间	慢	中等	快
调查费用	低	高	低
问卷难度	要求容易	可以复杂	要求容易
有形辅助物的使用	中等利用	充分利用	无法利用
调查过程控制	简单	复杂	容易
调查员作用的发挥	无法发挥	充分发挥	一般发挥
回答率	最低	较高	一般

搜集数据的基本方法

- 数据收集的其他方法：
 - 直接观察法
 - 通讯法
 - 网络调查法
 - 卫星遥感法
 - 随机田野视野法



在线交流调查法



✓BBS

✓网络实时交谈

✓网络会议等



搜集数据的基本方法



卫星遥感法



① 数据的来源

② 调查方法

③ 实验方法

④ 数据的误差

实验方法

- 将研究对象分为两组：实验组和对照组。
- 实验组和对照组的产生应遵循随机原则，而且应该匹配。
- 匹配指对实验单位的背景材料进行分析比较，将情况类似的每对单位分别随机地分配到实验组和对照组。
- 例子：在甲装置的烧杯中加入（“NaOH”或“NaHCO₃”）溶液。将两装置放在适宜的光照下照射 1 小时，测定红墨水滴的移动距离。

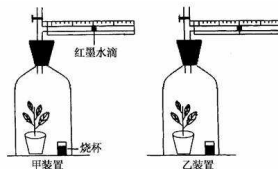


Figure: 测定某植物的光合作用强度

实验方法

- 社会科学中的因果推断

例子：医疗的效果

通过观察法得到的医院效果 = 那些去了医院的人的健康状况 - 那些没去医院的人的健康状况

实验方法

- 但是，这两群人也许是不同的，去医院的人群健康状况往往较差。因此，通过观察获得的效果可能不是真实效果！
- 因果推断的根本问题：没有足够的“反事实”。没法避免，只能减少推论的不确定性。
- 我们用数学公式来重新表达一下基于观察法的医院效果：

$$\overbrace{E(Y_{1i}|D_i = 1)}^{\text{该去医院也的确去了的那些人}} - \overbrace{E(Y_{0i}|D_i = 0)}^{\text{不该去医院的人且实际上也没去的那些人}} \quad (1)$$

$$\underbrace{\overbrace{E(Y_{1i}|D_i = 1)}^{\text{该去医院实际上也去了的人的身体状况}} - \overbrace{E(Y_{0i}|D_i = 1)}^{\text{应该去医院但是没去的身体状况}}}_{\text{真实的因果关系 (ATE)}} + \underbrace{\overbrace{E(Y_{0i}|D_i = 1)}^{\text{应该去医院但是没去的身体状况}} - \overbrace{E(Y_{0i}|D_i = 0)}^{\text{没去医院的人的身体状况}}}_{\text{样本选择偏误}} \quad (2)$$

实验方法

- 但是，这两群人也许是不同的，去医院的人群健康状况往往较差。因此，通过观察获得的效果可能不是真实效果！
- 因果推断的根本问题：没有足够的“反事实”。没法避免，只能减少推论的不确定性。
- 我们用数学公式来重新表达一下基于观察法的医院效果：

$$\overbrace{E(Y_{1i}|D_i = 1)}^{\text{该去医院也的确去了的那些人}} - \overbrace{E(Y_{0i}|D_i = 0)}^{\text{不该去医院的人且实际上也没去的那些人}} \quad (1)$$

$$\underbrace{\overbrace{E(Y_{1i}|D_i = 1)}^{\text{该去医院实际上也去了的人的身体状况}} - \overbrace{E(Y_{0i}|D_i = 1)}^{\text{应该去医院但是没去的身体状况}}}_{\text{真实的因果关系 (ATE)}} + \underbrace{\overbrace{E(Y_{0i}|D_i = 1)}^{\text{应该去医院但是没去的身体状况}} - \overbrace{E(Y_{0i}|D_i = 0)}^{\text{没去医院的人的身体状况}}}_{\text{样本选择偏误}} \quad (2)$$

实验方法

- 实验中的若干问题

- 1 人的意愿。研究的对象是人的时候，在划分实验组和对照组时的随机原则将面临挑战
- 2 心理问题。人们对被研究非常敏感，这使得他们更加注意自我，从而走到事物的另一个极端
- 3 道德问题。当某种实验涉及道德问题时，人们会处于进退两难的尴尬境地

实验方法

- 实验中的若干问题

- 1 人的意愿。研究的对象是人的时候，在划分实验组和对照组时的随机原则将面临挑战
- 2 心理问题。人们对被研究非常敏感，这使得他们更加注意自我，从而走到事物的另一个极端
- 3 道德问题。当某种实验涉及道德问题时，人们会处于进退两难的尴尬境地

实验方法

- 实验中的若干问题

- 1 人的意愿。研究的对象是人的时候，在划分实验组和对照组时的随机原则将面临挑战
- 2 心理问题。人们对被研究非常敏感，这使得他们更加注意自我，从而走到事物的另一个极端
- 3 道德问题。当某种实验涉及道德问题时，人们会处于进退两难的尴尬境地

实验方法

- 实验中的若干问题

- 1 人的意愿。研究的对象是人的时候，在划分实验组和对照组时的随机原则将面临挑战
- 2 心理问题。人们对被研究非常敏感，这使得他们更加注意自我，从而走到事物的另一个极端
- 3 道德问题。当某种实验涉及道德问题时，人们会处于进退两难的尴尬境地

实验方法

- 随机实验的局限性：昂贵、周期长且操作困难。
- 例子：Tennessee STAR experiment
 - 研究目的：估计班级大小对学生成绩的作用
 - 费用：1200 万美元
 - 研究对象：1985/86 学年入学的 11600 个小学生
 - 时间跨度：跟踪 4 年

实验方法

- 随机实验的局限性：昂贵、周期长且操作困难。
- 例子：Tennessee STAR experiment
 - 研究目的：估计班级大小对学生成绩的作用
 - 费用：1200 万美元
 - 研究对象：1985/86 学年入学的 11600 个小学生
 - 时间跨度：跟踪 4 年

实验方法

- 随机实验的局限性：昂贵、周期长且操作困难。
- 例子：Tennessee STAR experiment
 - 研究目的：估计班级大小对学生成绩的作用
 - 费用：1200 万美元
 - 研究对象：1985/86 学年入学的 11600 个小学生
 - 时间跨度：跟踪 4 年

实验方法

- 随机实验的局限性：昂贵、周期长且操作困难。
- 例子：Tennessee STAR experiment
 - 研究目的：估计班级大小对学生成绩的作用
 - 费用：1200 万美元
 - 研究对象：1985/86 学年入学的 11600 个小学生
 - 时间跨度：跟踪 4 年

实验方法

- 随机实验的局限性：昂贵、周期长且操作困难。
- 例子：Tennessee STAR experiment
 - 研究目的：估计班级大小对学生成绩的作用
 - 费用：1200 万美元
 - 研究对象：1985/86 学年入学的 11600 个小学生
 - 时间跨度：跟踪 4 年

实验方法

- 随机实验的局限性：昂贵、周期长且操作困难。
- 例子：Tennessee STAR experiment
 - 研究目的：估计班级大小对学生成绩的作用
 - 费用：1200 万美元
 - 研究对象：1985/86 学年入学的 11600 个小学生
 - 时间跨度：跟踪 4 年

实验方法

- 实验中的统计
 - 1 实验设计本身就是一个统计问题
 - 2 确定进行实验所需要的单位的个数，以保证实验可以达到统计显著的结果
 - 3 将统计的思想融入到实验设计中，使实验设计符合统计分析的标准
 - 4 对实验数据进行分析时，统计可以提供最恰当的分析方法
- 内部有效性 VS. 外部有效性

① 数据的来源

② 调查方法

③ 实验方法

④ 数据的误差

数据的误差



数据的误差

- 统计数据的误差通常是指统计数据与客观现实之间的差距，误差的类型主要有抽样误差和非抽样误差两类。

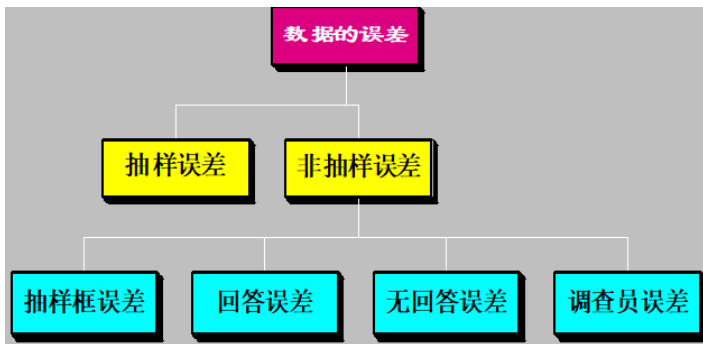


Figure: 数据的误差类型

数据的误差

- 抽样误差：由于抽样的随机性所带来的误差
 - 所有样本可能的结果与总体真值之间的平均性差异
 - 影响抽样误差的大小的因素：样本量的大小、总体的变异性

数据的误差

- 误差的控制
 - 调查员的挑选
 - 调查员的培训
 - 督导员的调查专业水平
 - 调查过程控制：调查结果进行检验、评估；现场调查人员进行奖惩的制度

数据的误差

- 案例：1936 年美国总统选举
 - 《文学摘要》预测总统选举结果：罗斯福 VS. 兰登
 - 经济背景：国家正努力从大萧条中恢复，失业人数高达九百万人
 - 民意测验：将问卷邮寄给一千万人，他们的名字和地址摘自电话簿或俱乐部会员名册，其中 240 万人寄回答案（回收率 24%）
 - 预测结果：罗斯福：43% 兰登：57%
 - 竞选结果：罗斯福：62% 兰登：38%

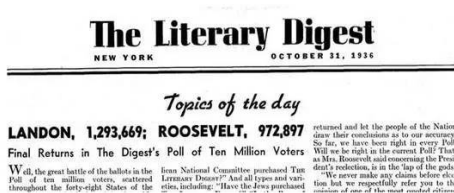


Figure: 《文学摘要》预测

数据的误差

- 案例：1936 年美国总统选举
 - 《文学摘要》预测总统选举结果：罗斯福 VS. 兰登
 - 经济背景：国家正努力从大萧条中恢复，失业人数高达九百万人
 - 民意测验：将问卷邮寄给一千万人，他们的名字和地址摘自电话簿或俱乐部会员名册，其中 240 万人寄回答案（回收率 24%）
 - 预测结果：罗斯福：43% 兰登：57%
 - 竞选结果：罗斯福：62% 兰登：38%

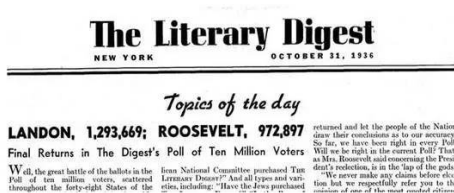


Figure: 《文学摘要》预测

数据的误差

- 案例：1936 年美国总统选举
 - 《文学摘要》预测总统选举结果：罗斯福 VS. 兰登
 - 经济背景：国家正努力从大萧条中恢复，失业人数高达九百万人
 - 民意测验：将问卷邮寄给一千万人，他们的名字和地址摘自电话簿或俱乐部会员名册，其中 240 万人寄回答案（回收率 24%）
 - 预测结果：罗斯福：43% 兰登：57%
 - 竞选结果：罗斯福：62% 兰登：38%

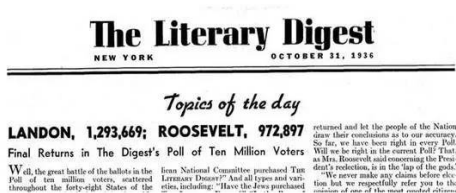


Figure: 《文学摘要》预测

数据的误差

- 案例：1936 年美国总统选举
 - 《文学摘要》预测总统选举结果：罗斯福 VS. 兰登
 - 经济背景：国家正努力从大萧条中恢复，失业人数高达九百万人
 - 民意测验：将问卷邮寄给一千万人，他们的名字和地址摘自电话簿或俱乐部会员名册，其中 240 万人寄回答案（回收率 24%）
 - 预测结果：罗斯福：43% 兰登：57%
 - 竞选结果：罗斯福：62% 兰登：38%



Figure: 《文学摘要》预测

数据的误差

- 案例：1936 年美国总统选举
 - 《文学摘要》预测总统选举结果：罗斯福 VS. 兰登
 - 经济背景：国家正努力从大萧条中恢复，失业人数高达九百万人
 - 民意测验：将问卷邮寄给一千万人，他们的名字和地址摘自电话簿或俱乐部会员名册，其中 240 万人寄回答案（回收率 24%）
 - 预测结果：罗斯福：43% 兰登：57%
 - 竞选结果：罗斯福：62% 兰登：38%

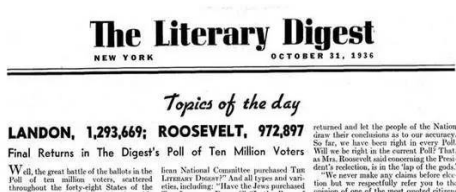


Figure: 《文学摘要》预测

数据的误差

- 案例：1936 年美国总统选举
 - 《文学摘要》预测总统选举结果：罗斯福 VS. 兰登
 - 经济背景：国家正努力从大萧条中恢复，失业人数高达九百万人
 - 民意测验：将问卷邮寄给一千万人，他们的名字和地址摘自电话簿或俱乐部会员名册，其中 240 万人寄回答案（回收率 24%）
 - 预测结果：罗斯福：43% 兰登：57%
 - 竞选结果：罗斯福：62% 兰登：38%

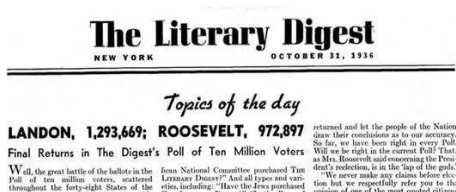


Figure: 《文学摘要》预测

数据的误差

- 失败的原因是抽样方法不正确：选择偏倚
 - 《文学摘要》为了寄送调查问卷，随机抽取了电话黄页和车辆注册系统的地址。可是在 1936 年的美国，富裕的家庭才有私人电话和汽车。
 - 正确的统计学抽样方法，样本不需要很大，只要能良好地反映总体，就能对总体进行准确的推测。

补充内容

- 统计调查问卷的设计
 - Slide 内容见《统计学 B》-课件 3.nb

参考资料

- 贾俊平.《统计学》(第八版) [M]. 北京: 中国人民大学出版社, 2021。
- 李金昌. 统计学 [M]. 北京: 高等教育出版社, 2018。

Q&A

THANK YOU