第 2 章　数据的搜集

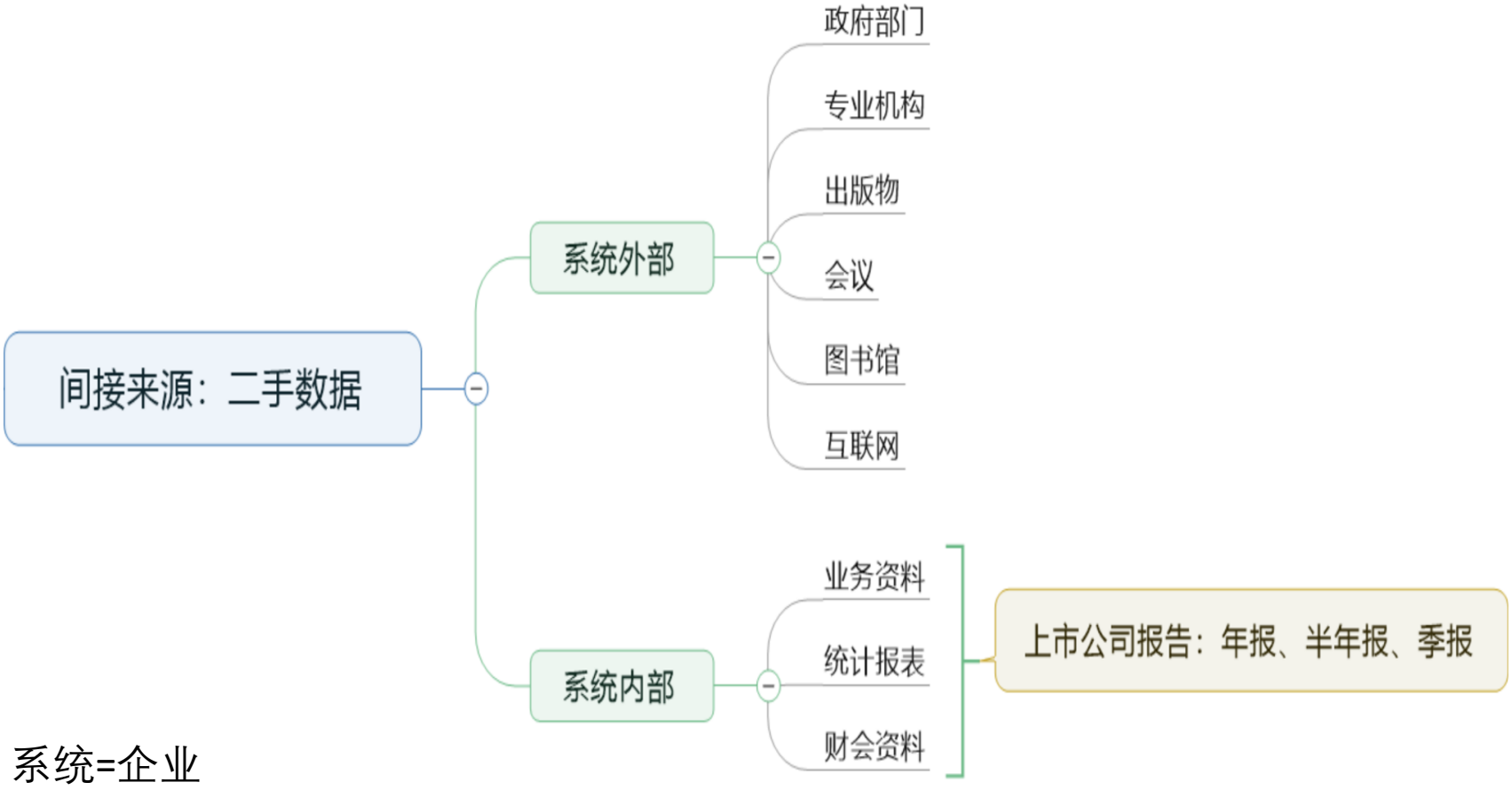
内容提要

2.1 数据来源

数据的间接来源

数据的直接来源

数据的间接来源



系统=企业

数据的间接来源：系统外部数据

- **政府部门**
  - 统计局：宏观数据（统计公报，统计年鉴）　国家统计局：如何获得数据
  - 行业监管机构：行业数据（行业报告）
- **专业机构：**信息中心、咨询机构、调查机构
  - 政府所属：宏观数据、行业数据
  - 私营部门：行业数据（行业报告）
- **出版物：**期刊、报纸、书籍
- **会议（专业数据或指标）：**博览会、展销会、交易会、学术研讨会
- **图书馆**
- **互联网**
  - **行业网站：行业数据**
  - **企业网站：企业数据**

数据的间接来源：系统内部数据

- **来源**
  - **业务资料**：与业务经营活动有关的各种单据，记录
  - **统计报表**：经营活动过程中的各种统计报表
  - **财务资料**：各种财务，会计核算和分析资料等
- **上市公司 vs 非上市公司**
  - **非上市公司**：内部数据通常不能公开
  - **上市公司**：信息披露义务上市公司季报、半年报和年报
    - 公告：重大事项的决议，必须公告披露
    - 定期报告：季报、半年报、年报、其他报告
    - 例子：中国银行2016年报（A股）

数据的间接来源

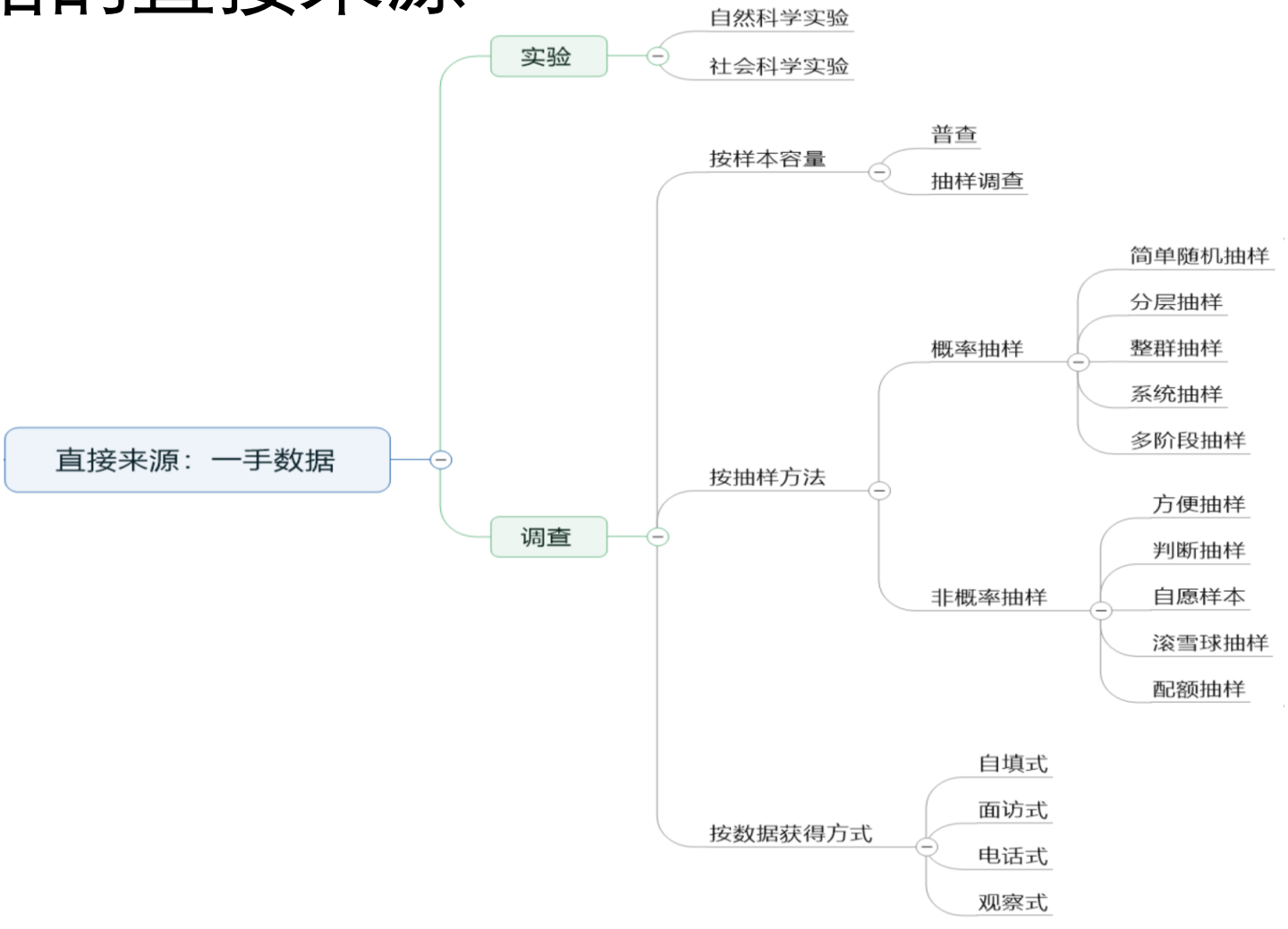
二手数据的特点

- **搜集容易，采集成本低**
- **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
- **搜集二手资料在研究中应优先考虑**

二手数据的评估

- **评估原则**
  - 客观性
  - 可信性
  - 可用性
- **评估内容**
  - **主体**：数据是谁搜集的？
  - **用途**：为什么目的而搜集的？
  - **收集方法**：数据是怎样搜集的？
  - **时效性**：什么时候搜集的？

数据的直接来源


直接来源：一手数据
实验
自然科学实验
社会科学实验
调查
按样本容量
普查
抽样调查
按抽样方法
概率抽样
简单随机抽样
分层抽样
整群抽样
系统抽样
多阶段抽样
非概率抽样
方便抽样
判断抽样
自愿样本
滚雪球抽样
配额抽样
按数据获得方式
自填式
面访式
电话式
观察式

数据的直接来源：实验vs调查

实验

- 以**自然现象**为对象
- **自然科学**常用的研究方法
- 通常取自**无限总体**
- **抽样独立**

调查

- 以**社会现象**为对象
- **社会科学**常用的研究方法
- 通常取自**有限总体**
- **抽样不独立**

调查

普查

- **定义：**调查总体中**所有**的个体。
- 中国的普查：国家统计局：普查数据
- **三大普查**：
  - 经济普查：每5年1次，尾数逢3或8的年度
  - 人口普查：每10年1次，尾数逢0的年度
  - 农业普查：每10年1次，尾数逢6的年度
- 普查的实施者：带你认识统计员家族

抽样调查

- **定义**：调查总体中**部分**个体。
- **两种类型**：
  - **宏观抽样调查**：统计局等部门开展，获得**宏观数据**，反映国家或地区的总体情况；（国家统计局：1%人口抽样调查）
  - **微观抽样调查**：民间机构或个人进行，获得的**微观数据**，反映局部个体的情况。

2.2 调查数据

概率抽样与非概率抽样

搜集数据的基本方法

抽样的基本问题

什么是**好的样本**?

- **性价比**高的样本
  - 性能：满足研究的需要
  - 价格：成本可控

如何抽选出好的样本?

- 关键：抽样成本与收益的权衡
  - 收益（性能）=数据精确度
  - 成本（价格）=调查费用

哪种抽样更好?

- 概率抽样：数据精确度高；调查费用高
- 非概率抽样：数据精确度低，调查费用低

概率抽样

- **定义**

  概率抽样，又称随机抽样，是指遵循**随机**原则进行抽样，总体中每一个单位都有一定的概率被选入样本。

- **特点**
  - 按一定的**概率**以**随机原则**抽取样本
  - 每个单位被抽中的概率是已知的，或是可以计算出来
  - 当用样本对总体目标量进行估计时，要考虑到每个样本单位被抽中的概率。

- **类型**
  - 简单随机抽样
  - 分层抽样
  - 整群抽样
  - 系统抽样
  - 多阶段抽样

概率抽样1：简单随机抽样

- **定义**

简单随机抽样，是指从总体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相等的抽样方法。

- **重要性**
  - 最基本的抽样方法
  - 其它抽样方法的基础
- **应用范围**：小规模抽样
- **优点**
  - 简单、直观，在**抽样框**完整时，可直接从中抽取样本
  - 用样本统计量对目标量进行估计比较方便
- **缺点**
  - 当N很大时，不易构造抽样框；
  - 抽出的单位很分散，给实施调查增加了困难；
  - 没有利用其它辅助信息以提高估计的效率。

概率抽样2：分层抽样

- **定义**

　　分层抽样，是指将抽样单位按某种特征或某种规则划分为不同的**层**，然后从不同的层中独立、随机地抽取样本。

- **优点**
  - 保证样本的结构与总体的结构比较相近，提高估计的精度
  - 组织实施调查方便
  - 既能对总体参数进行估计，也能对各层的目标量进行估计。

概率抽样3：整群抽样

- 定义

整群抽样，是指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

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

概率抽样4：系统抽样

- **定义**

系统抽样，也称等距抽样、机械抽样、SYS抽样，是指将总体中的所有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。

- **步骤**

1. 对总体N进行排序并编号；
2. 确定抽样距离：k=N/n
3. 确定抽样的起点r：r介于编号1~k之间
4. 先从数字1到k之间随机抽取一个数字r作为初始单位，以后依次取r+k，r+2k…等单位。

- **优点**
  - 操作简便
  - 可提高估计的精度
- **缺点**
  - 排序不当时，会导致无效抽样

概率抽样5：多阶段抽样

- **定义**

多阶段抽样，类似整群抽样，先抽取**群**，随后从选中的群中随机抽取出若干个单位。

- **类型**
  - 两阶段抽样：第1阶段抽取群，第2阶段抽取的单位
  - 多阶段抽样：三阶段抽样、四阶段抽样等
- **应用：大规模抽样**

- **优点**
  - 具有整群抽样的优点，保证样本相对集中，节约调查费用
  - 不需要包含所有单位的抽样框，易于实施
  - 实行再抽样，可以覆盖更大的抽样范围

非概率抽样

- **定义**

非概率抽样，是指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。

- **类型**
  - 方便抽样
  - 判断抽样
  - 自愿样本
  - 滚雪球抽样
  - 配额抽样

非概率抽样1：方便抽样

- **定义**

  方便抽样，是指调查过程中由调查员依据方便原则，自行确定入选样本单位的方法。

- **场景**
  - 在街头、公园、商店等公共场所进行**拦截调查**
  - 厂家在出售产品柜台前对路过顾客进行的调查

- **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

非概率抽样2：判断抽样

- **定义**

判断抽样，是指研究人员根据经验、判断和对研究对象的了解，有目的选择一些单位作为样本的抽样方法。

- **类型**
  - **重点抽样**：依据比重
  - **典型抽样**：依据类型
  - **代表抽样**：兼具前两者的特点

- **优点**
  - 抽样成本比较低，容易操作
- **缺点**
  - 抽样具有主观性
  - 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；
  - 未遵循随机原则，不能用于推断总体。

非概率抽样3：自愿样本

- **定义**

  自愿样本，是指被调查者自愿参加，成为被选中的样本，向调查人员提供有关信息的调查方法。

- **场景**
  - 期刊，或互联网的问卷调查
  - 被调查者向某类节目拨打热线电话

- **优点**
  - 抽样成本低，易于操作

- **缺点**
  - 选择性样本，有偏的
  - 缺乏随机性，对总体代表性有限

非概率抽样4：滚雪球抽样

- **定义**

  滚雪球抽样，是指调查人员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，对应的抽样方法就是滚雪球抽样。

- **场景**
  - 对**稀少群体**的调研
  - 对**特定群体**的调研
- **例子**
  - 暑期实习中的农村入户调研
  - 针对特定疾病群体的调研
- **优点**
  - 顺藤摸瓜，易于找到被调查者
  - 实施成本低

非概率抽样5：配额抽样

- **定义**

配额抽样，是指先将总体中的所有单位按特定的标志(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的抽样方法。

- **步骤：**先分类，后抽样
- **类型**
  - 单一标志分类：性别，年龄
  - 多个标志分类：性别和年龄
  - 例子：单一标志分类 vs 多个标志分类

- **优点**
  - 操作简单
  - 样本分布均匀
  - 可以保证样本的结构和总体的结构类似
- **缺点**
  - 缺乏随机性
  - 无法推断总体

概率抽样 vs 非概率抽样

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（**随意**性）
- 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**

搜集数据的基本方法

- 自填式
- 面访式
- 电话式
- 观察式

搜集数据的基本方法1：自填式

- **定义**

  自填式，是指没有调查员协助的情况下由被调查者自己完成调查问卷。

- **问卷分发**
  - 调查员
  - 邮寄
  - 网络：问卷星 问卷网 ……
  - 媒体
- **要求**
  - 问卷结构必须严谨，有清楚的说明
  - 被调研者必须具备良好的阅读和理解能力
- **优点**
  - 调研管理组织容易
  - 成本最低
- **缺点**
  - 问卷的返回率比较低
  - 不适合结构复杂的问卷
  - 调查周期比较长
  - 数据搜集过程中出现的问题难于及时采取调改措施

搜集数据的基本方法2：面访式

- 定义

面访式，是指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。

- 类型
  - 结构化访谈：有提纲
  - 非结构化访谈：无提纲

- 优点
  - 可提高调查的回答率
  - 可提高调查数据的质量
  - 能调节数据搜集所花费的时间

- 缺点
  - 调查的成本较高
  - 调查过程的质量控制有一定难度

搜集数据的基本方法3：电话式

- **定义**

  电话式，是指通过电话向被调查者实施调查。

- **优点**
  - 速度快，能在短时间内完成调查
  - 适合于样本单位十分分散的情况

- **缺点**
  - 如果被调查者没有电话，调查将无法实施
  - 访问的时间不能太长
  - 使用的问卷需要简单
  - 被访者不愿意接受调查时，难以说服

搜集数据的基本方法4：观察式

- **定义**

  观察式，是指调查人员通过直接观测的方式获取信息。

- **常见类型：**交通流量调查

- **优点**
  - 调查人员不是强行介入
  - 能够在被调查者不察觉的情况下获得资料

搜集数据的基本方法：比较

	自填式	面访式	电话式
调查时间	**慢**	**中等**	**快捷**
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**

2.3 实验数据

- 实验组与对照组
- 实验中的若干问题
- 实验中的统计
- 自然科学实验与社会科学实验

实验组与对照组

- **分组**
  - 实验组：重点关注，控制实验条件
  - 对照组：比较基准，不控制实验条件（安慰剂）
- **分组原则：**随机分组，匹配原则
- **分组方法：**

  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

实验中的若干问题

实验中的若干问题

- **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
- **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 （霍桑效应）
- **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。（艾滋病实验中的对照组）

实际应用：随机双盲实验

- **随机分组**：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
- **双盲**：被实验者和实验操作者都不清楚实验分组信息
- **例子**：药物和疫苗的临床实验

实验中的统计

实验设计本身就是一个统计问题

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

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

对实验数据进行分析时，统计可以提供最恰当的分析方法

自然科学实验 vs 社会科学实验

自然科学实验

- 实验条件**可控性强**
- 实验结果可以**重复**和**验证**
- 例子：詹姆斯·林德的败血症的实验

社会科学实验

- 实验条件**难以完全控制**
- 实验结果**难以完全重复和验证**
- 例子：霍桑效应

2.4 数据误差

- 抽样误差
- 非抽样误差
- 误差的控制

抽样误差

- **定义**

抽样误差，是指由于抽样的**随机性**所引起的所有样本可能的结果与总体真值之间的平均性差异。

- **存在范围**：概率抽样

- **影响因素**
  - 样本量

    样本量越大，抽样误差越小
  - 总体的变异性

    变异性越小，抽样误差越小

非抽样误差

- **定义**

  非抽样误差，是指抽样误差之外的，由于其他原因造成的样本观察结果与总体真值之间的差异

- **存在范围**
  - 概率抽样
  - 非概率抽样

- **类型**
  - 抽样框误差
  - 回答误差
  - 无回答误差
  - 调查员误差
  - 测量误差

非抽样误差1：抽样框误差

- **抽样框的定义**

  有关总体全部单位的名录。

- **抽样框评判标准**

  好的抽样框应该与总体一一对应

- **误差情形**
  - 抽样框不完整
  - 抽样框包含无效样本
- **误差处理**
  - 构造完整的抽样框；
  - 精心遴选样本，剔除无效样本

非抽样误差2：回答误差

- **定义**

回答误差，是指被调查者在接受调查时给出的回答与实际情况不符。

- **误差情形**
  - 理解误差
    - 问卷不够精确
    - 问卷排序不当
  - 记忆误差
  - 有意识误差
    - 涉及隐私

    例子：“婚外恋”问题的调查
    - 利益驱动

    例子：克强指数

- **误差处理**
  - 细化问卷设计，精确表述，巧妙排序
  - 周期性调研
  - 建立平等的信任关系，切忌逼问被调查者
  - 迂回提问，避免直接触及隐私问题
  - 识别并放弃无效样本

非抽样误差3：无回答误差

- **定义**

  无回答误差，是指被调查者拒绝接受调查，调查员得到一份空白的问卷。

- **情形**
  - 电话无法接通
  - 问卷丢失
  - 访谈中止，或者被拒绝

  ……

- **分类**
  - 随机性的误差

    与调查内容无关，意外导致
  - 非随机误差

    与调查内容有关，非意外情形

- **误差处理**
  - 随机性的误差：增大样本容量
  - 非随机误差：有限的预防和补救

非抽样误差4：调查员误差

- **定义**

  调查员误差，是指由于调查员的原因而产生的误差。

- **情形**
  - 疏忽导致的记录错误
  - 诱导性的提问
  - 有意或无意识流露的看法或倾向

- **误差处理**
  - 强化个人责任：事前教育，事后追责
  - 问卷质量控制机制：问卷评分
  - 对调查员进行专业培训

非抽样误差5：测量误差

- **定义**

测量误差，是指由于测量工具，或者测量方法导致的误差。

- **情形**
  - **测量工具**不够精确
  - **测量方法**不够专业
- **误差处理**
  - 采用精确的测量工具
  - 采用科学的测量方法

- **量表 vs问卷**
  - 量表 团队测试量表

    潜在变量;间接测量;复杂，主观的对象;多个项目测量
  - 调查问卷

    观测变量;直接测量;简单、客观的对象;单一项目测量
- **例子：“满意度调查”的误区**
  - 满意度是否重要？
  - 满意度如何测量：量表还是调查问卷？
  - 不同主体满意度是否具有可比性？
  - 满意度调查是否具有可信性？

误差控制

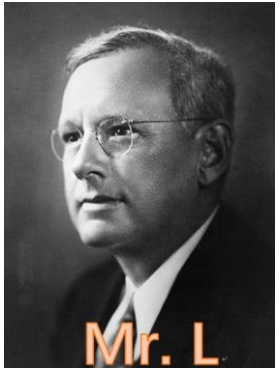
- **抽样误差**
  - 无法完全消除
  - 可计算和控制
- **非抽样误差**
  - **事前控制**
    - 抽样框选取
    - 问卷设计
  - **过程控制**
    - 调查员的挑选
    - 调查员的培训
    - 督导员的调查专业水平
    - 调查结果进行检验、评估
    - 调查人员进行奖惩的制度

抽样误差：案例

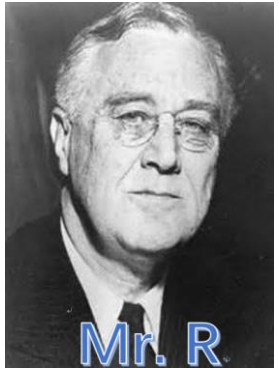
在20世纪30年代的一次美国年总统选举中，参与角逐分别是共和党候选人L和民主党候选人R。

- 民意调查1：根据盖洛普公司一份针对5万名选民的民意调查，候选人R的得票率为56%
- 民意调查2：根据著名杂志《文学文摘》对240万人的调查结果，候选人L的得票率为57%

请你来预测：谁会赢得大选？



VS



抽样误差：案例（续）

- **选举结果**

  得票

  候选人R赢得2770万张选票(62%)

  候选人L赢得1660万张选票(38%)

  胜出

  候选人R以绝对优势胜出，赢得了大选。

- **问题**
  - 为什么一个**大样本的预测**结果没有**小样本的预测**结果准确？
  - 预测错误的统计学原因是什么？

抽样误差：案例分析

- **预测错误的原因**

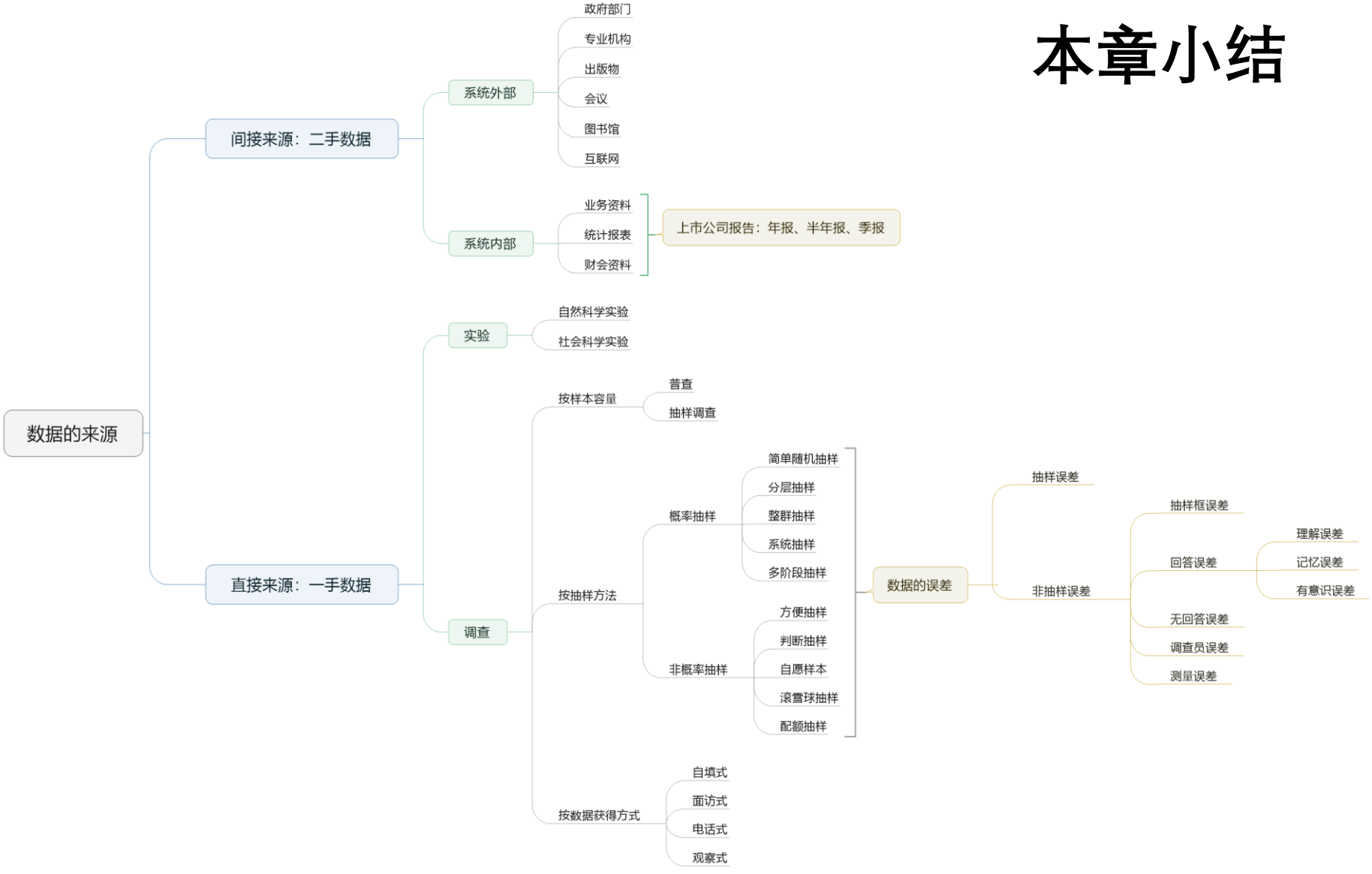
  民意调研存在抽样误差

- **民意调查的抽样误差**
  - 错误的抽样框

    以富裕阶层为主，忽略了失业的劳工阶层
  - 无回答误差

    《读者文摘》发放了1000万份问卷，仅回收了240万份，无回答比例过高

本章小结


数据的来源
间接来源：二手数据
系统外部
政府部门
专业机构
出版物
会议
图书馆
互联网
系统内部
业务资料
统计报表
财会资料
上市公司报告：年报、半年报、季报
直接来源：一手数据
实验
自然科学实验
社会科学实验
调查
按样本容量
普查
抽样调查
按抽样方法
概率抽样
简单随机抽样
分层抽样
整群抽样
系统抽样
多阶段抽样
非概率抽样
方便抽样
判断抽样
自愿样本
滚雪球抽样
配额抽样
按数据获得方式
自填式
面访式
电话式
观察式
数据的误差
抽样误差
非抽样误差
抽样框误差
回答误差
理解误差
记忆误差
有意识误差
无回答误差
调查员误差
测量误差