# 第 2 章　数据的搜集

PowerPoint

# 第 2 章　数据的搜集

2.1　数据的来源
2.2　调查方法
2.3　实验方法
2.4　数据的误差

# 学习目标

1. 数据的来源
2. 搜集数据的调查方法
3. 搜集数据的实验方法
4. 数据的误差

# 2.1 数据的来源

## 2.1.1 数据的间接来源

## 2.1.2 数据的直接来源

# 数据的间接来源

# 二手数据

1. 统计部门和政府部门公布的有关资料，如各类统计年鉴
2. 各类经济信息中心、信息咨询机构、专业调查机构等提供的数据
3. 各类专业期刊、报纸、书籍所提供的资料
4. 各种会议，如博览会、展销会、交易会及专业性、学术性研讨会上交流的有关资料
5. 从互联网或图书馆查阅到的相关资料

# 二手数据

1. 业务资料，如与业务经营活动有关的各种单据，记录
2. 经营活动过程中的各种统计报表
3. 各种财务，会计核算和分析资料等

# 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

# 二手数据的评估

1. 数据是谁搜集的？
   - 可信度评估
2. 为什么目的而搜集的？
3. 数据是怎样搜集的？
4. 什么时候搜集的？

# 数据的间接来源

# 数据的直接来源
# (原始数据)

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
     - 如心理学、教育学、社会学、经济学、管理学等

# 2.2 调查方法

## 2.2.1 概率抽样与非概率抽样
## 2.2.2 搜集数据的基本方法


作者：贾俊平，中国人民大学统计学院

# 概率抽样和非概率抽样

# 概率抽样 (probability sampling)

1. 也称随机抽样
2. 特点
   - 按一定的概率以随机原则抽取样本
     - 抽取样本时使每个单位都有一定的机会被抽中
   - 每个单位被抽中的概率是已知的，或是可以计算出来的
   - 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

# 简单随机抽样
# (simple random sampling)

1. 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
2. 最基本的抽样方法，是其它抽样方法的基础
3. 特点
   - 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
4. 局限性
   - 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

# 分层抽样
# (stratified sampling)

1. 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
2. 优点
   - 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
   - 组织实施调查方便
   - 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
# (cluster sampling)

1. 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
2. 特点
   - 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
   - 缺点是估计的精度较差

# 系统抽样 (systematic sampling)

1. 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
   - 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
2. 优点：操作简便，可提高估计的精度
3. 缺点：对估计量方差的估计比较困难

# 多阶段抽样
# (multi-stage sampling)

1. 先抽取群，但并不是调查群内的所有单位，而是再进行一步抽样，从选中的群中抽取出若干个单位进行调查
   - 二阶抽样中群是初级抽样单位，第二阶段抽取的是最终抽样单位。将该方法推广，使抽样的阶段数增多，就称为多阶段抽样
2. 具有整群抽样的优点，保证样本相对集中，节约调查费用
3. 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
4. 在大规模的抽样调查中，是经常被采用的方法

# 非概率抽样 (non-probability sampling)

1. 相对于概率抽样而言
2. 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
3. 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

# 方便抽样

1. 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
   - 厂家在出售产品柜台前对路过顾客进行的调查
2. 优点：容易实施，调查的成本低
3. 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

1. 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
   - 有重点抽样，典型抽样，代表抽样等方式
2. 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
3. 抽样成本比较低，容易操作
4. 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

# 自愿样本

1. 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
2. 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

1. 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
2. 适合于对稀少群体和特定群体研究
3. 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

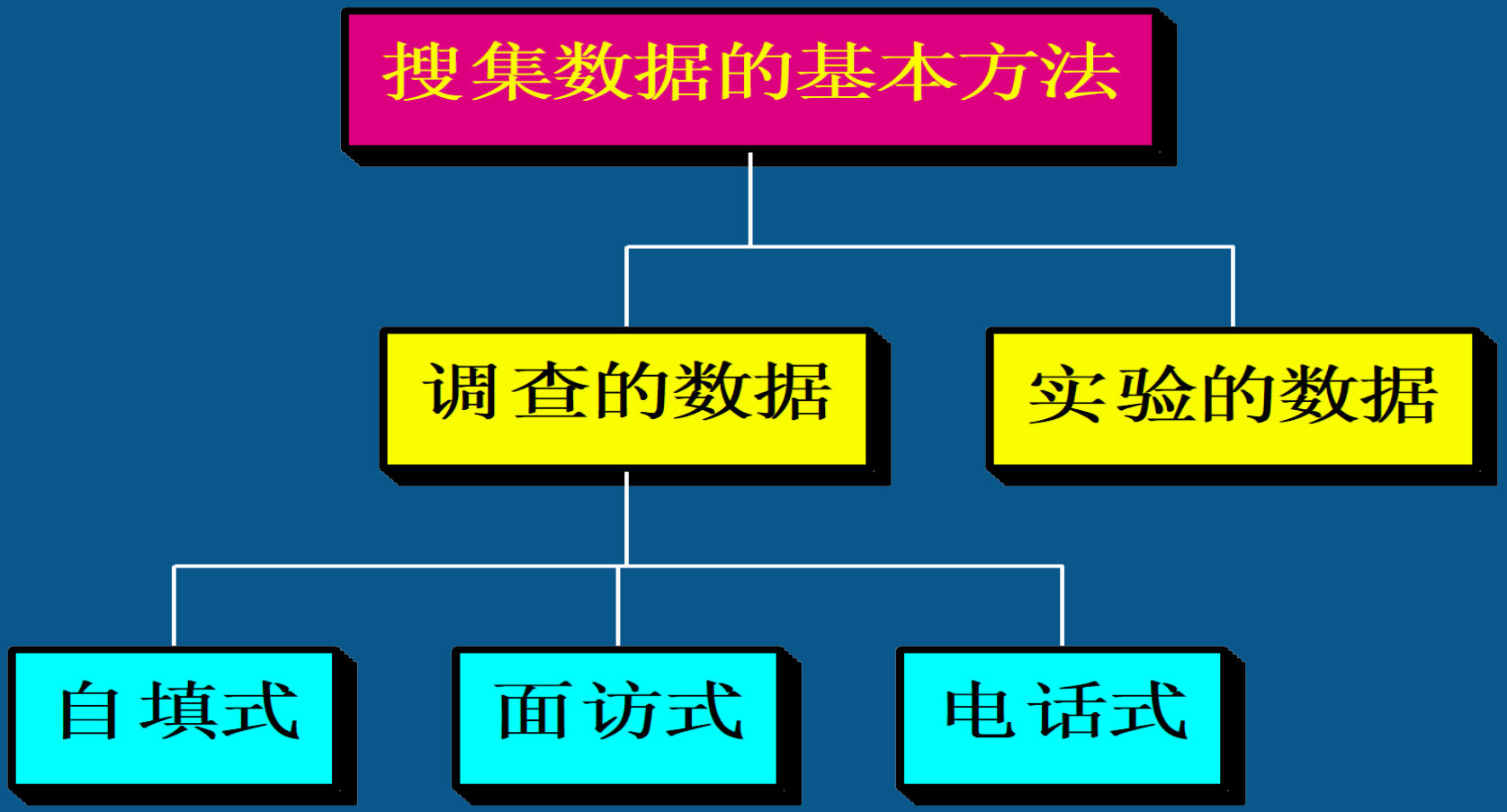
1. 先将总体中的所有单位按一定的标志(变量)分为若干类，然后在每个类中采用方便抽样或判断抽样的方式选取样本单位
2. 操作简单，可以保证总体中不同类别的单位都能包括在所抽的样本之中，使得样本的结构和总体的结构类似
3. 抽取具体样本单位时，不是依据随机原则，属于非概率抽样

# 概率抽样与非概率抽样的比较

1. 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
2. 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

# 搜集数据的基本方法

# 搜集数据的基本方法



# 自填式问卷调查

1. 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
2. 要求调查问卷结构严谨，有清楚的说明
3. 弱点
   - 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

# 面访式问卷调查

1. 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
2. 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
3. 弱点
   - 调查的成本较高
   - 调查过程的质量控制有一定难度

# 电话式问卷调查

1. 通过电话向被调查者实施调查
2. 特点
   - 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
3. 局限
   - 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

# 观察式调查

1. 就调查对象的行动和意识，调查人员边观察边记录以收集所需信息
2. 调查人员不是强行介入
3. 能够在被调查者不察觉的情况下获得资料
   - 如交通流量的调查

# 各调查方法的比较

| | 自k | b 访 | 5 话 |
|---|---|---|---|
| 查时间 | 慢 | 中等 | 快捷 |
| 查9( | 低 | 高 | 低 |
| 问w 难度 | 要求容易 | 可以复杂 | 要求容易 |
| b 助i ( | 中等利用 | 充分利用 | 无法利用 |
| 查过 控6 | 简单 | 复杂 | 容易 |
| 查X\( 发 % | 无法发挥 | 充分发挥 | 一般发挥 |
| 回T | 最低 | 较高 | 一般 |

# 2.3 实验方法

2.3.1 实验组和对照组
2.3.2 实验中的若干问题
2.3.3 实验中的统计
2.3.4 实验法案例


作者：贾俊平，中国人民大学统计学院

# 实验组和对照组

1. 将研究对象分为两组：实验组和对照组
2. 实验组和随机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

# 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

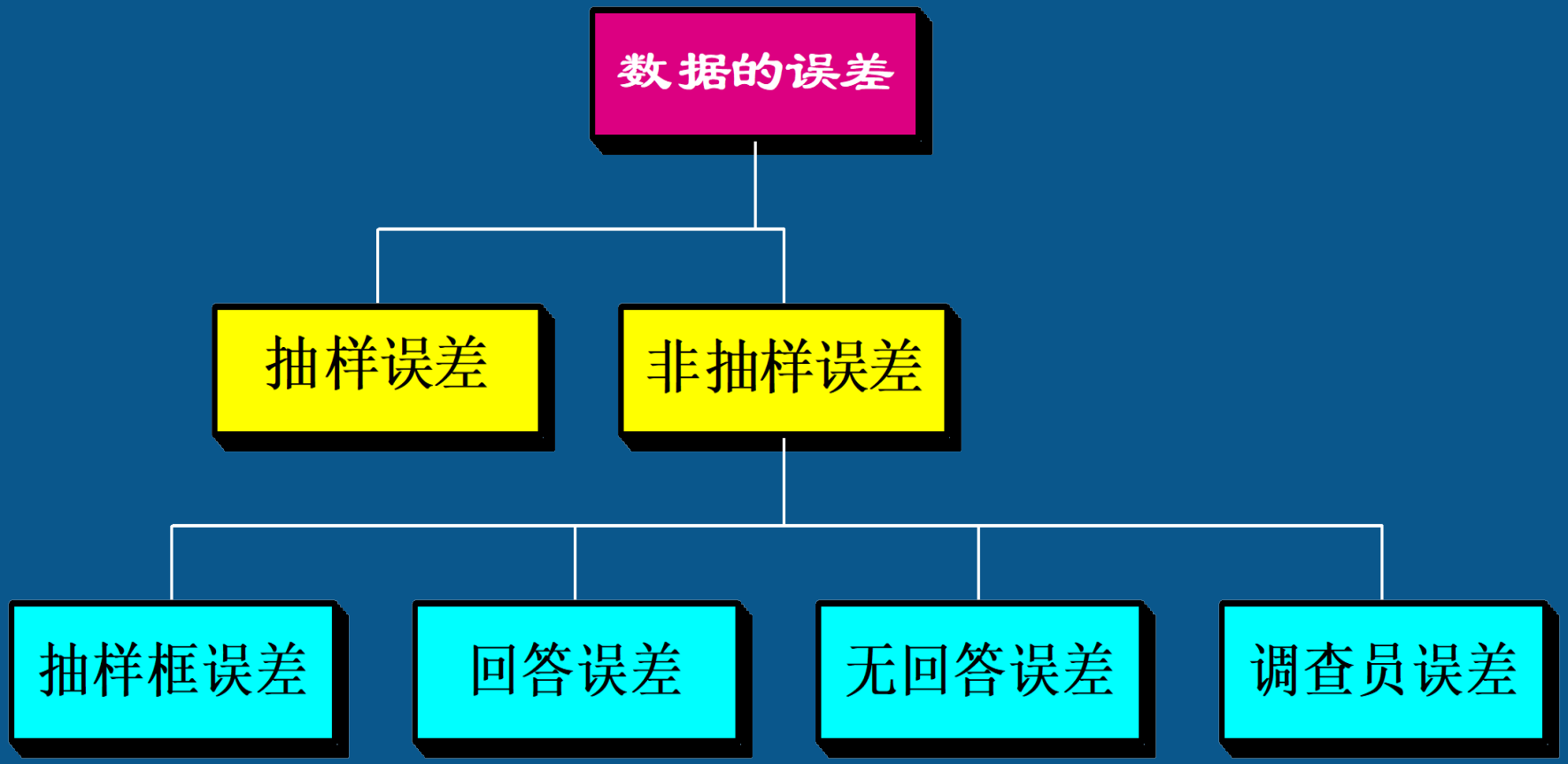
# 2.4 数据的误差

2.4.1 抽样误差

2.4.2 非抽样误差

2.4.3 误差的控制

# 数据的误差



# 抽样误差 (sampling error)

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

# 非抽样误差 (non-sampling error)

1. 相对抽样误差而言
2. 除抽样误差之外的，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
3. 存在于所有的调查之中
   - 概率抽样，非概率抽样，全面性调查
4. 有抽样框误差、回答误差、无回答误差、调查员误差、测量误差

# 误差的控制

1. 抽样误差可计算和控制
2. 非抽样误差的控制
   - 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

# 本章小结

1. 数据的来源
2. 调查数据
3. 实验数据
4. 数据的误差

# 结　束