第02章 数据收集、整理和清洗

# 2.1 数据目标

## 数据收集的步骤

### 定义研究目标

在开始数据收集过程之前，需要准确确定要实现的目标。可以从以下方面入手：

* 编写现实难题（problems）陈述：说明要解决的实际或科学难题及其重要性
* 提出研究问题（questions）：精确定义要查找的内容
* 根据研究问题选择数据类型：
  + 定量数据：用于检验假设、精确测量或获得大规模统计见解
  + 定性数据：用于探索想法、了解经验或获得特定环境的详细见解
  + 混合方法：当有多个目标时，可同时收集两种类型的数据

### 选择数据收集方法

根据要收集的数据类型，确定最适合的研究方法。常见的数据收集方法包括：

* 问卷调查
* 访谈
* 观察
* 实验
* 文献研究
* 档案研究
* 案例研究

### 规划资料收集程序

需要准确计划如何实现选定的方法：

* 决定问题形式（针对调查或访谈）
* 制定实验设计（针对实验研究）
* 实现可操作化：将抽象概念转变为可测量的观察结果
* 设计采样方式：定义总体和样本
* 编写标准化程序：确保多个研究人员的一致性
* 制定数据管理计划：包括数据组织、存储和保护措施

### 动手收集资料

在实施数据收集时需要注意：

* 记录所有相关信息
* 仔细检查手动数据输入的错误
* 评估数据的可靠性和有效性（针对定量数据）

## 陈述现实难点（problem）

### 基本内容

陈述现实难点需要包含以下要素：

* 将难点放在特定背景之中（已知内容）
* 描述研究将要解决的确切难点（需要知道的内容）
* 显示难点的相关性（为什么需要知道）
* 设定研究目标（将采取的行动）
* 提出研究难点（精确定义要聚焦或解决的难点）

### 陈述方法

1. **将难点具体化**：
   * 对于实际研究难点：关注何时何地出现、影响对象、已有尝试
   * 对于理论研究难点：关注科学、社会、地理或历史背景
2. **说明其重要性**：
   * 对于实际研究难点：说明对组织、机构或社会的影响
   * 对于理论研究难点：说明对理论理解或未来研究的贡献
3. **设定目的和目标**：
   * 目的（aim）：研究的总体目的，通常以不定式形式编写
   * 目标（objectives）：实现目的的具体步骤

## 提出研究问题（question）

### 基本要求

好的研究问题应具备以下特征：

* 专注于单个问题
* 可使用主要和/或次要来源进行研究
* 在时限和实际限制条件下可行回答
* 具体到足以彻底回答
* 足够复杂，可以在论文或论文的范围内得出答案
* 与学习或社会有广泛相关性

### 研究问题类型

常见的研究问题类型包括：

* 描述性研究：X的特征是什么？
* 比较研究：X和Y之间有什么区别和相似之处？
* 相关研究：变量X和变量Y之间有什么关系？
* 探索性研究：X的主要因素是什么？Y在Z中的作用是什么？
* 解释性研究：X对Y有影响吗？Y对Z的影响是什么？X的原因是什么？
* 评估研究：X的优缺点是什么？Y工作得如何？Z有多有效或理想？
* 行为研究：如何实现X？改善Y的最有效策略是什么？

### 好的研究问题的特征

1. **专注性和可研究**：
   * 专注于单个主题和问题
   * 不要求主观价值判断
   * 可使用主要或次要数据
   * 避免使用”为什么”的问题
2. **可行而具体**：
   * 在限定条件下完成
   * 使用明确的特定概念
   * 不要求最终的解决方案/政策或行动方案
3. **复杂而有争议**：
   * 无法用”是”或”否”回答
   * 无法用容易找到的事实和数字回答
   * 提供辩论和审议的范围
4. **与现实相关且具有原创性**：
   * 解决与所在领域或学科相关的问题
   * 有助于进行话题性的社会或学术辩论
   * 问题尚未得到回答

# 2.2 数据收集

## 数据的来源

### 研究方法产生的数据类型

* 观察数据
* 调查数据
* 实验数据

### 数据产生的方式方法

* 问卷数据
* 访谈数据
* 文献数据
* 痕迹数据（注意不是痕迹证据）
* 日志数据：记录数据获得过程、来源、基本特征等

## 数据的载体

### 数字化程度

* 数字化的数据
* 非数字化的数据

### 数值化程度

* 数值数据
* 非数值数据

### 具体形态

1. **文本数据**：
   * 访问、观察中的文字记录
   * 数字化的字符形态数据
   * 文字加载于各种载体的数据
2. **图片数据**：
   * 访谈照片、搜集图片等
   * 数字化为像素点形态的图片数据
   * 图形加载于各种载体的数据
3. **音频数据**：
   * 访问录音、语音日志等
   * 数字化为波形形态的音频数据
   * 音频加载于各种载体的数据
4. **视频数据**：
   * 访谈录像、搜集视频等
   * 数字化为像素点加波形形态的视频数据
   * 视频加载于各种载体的数据
5. **实物数据**：
   * 需要实物才能完整保存信息的载体数据
   * 访谈中搜集的实物、观察到的实物等

## 数据的状态

### 研究可用性

* **原始数据**：一般不能直接用于研究
* **研究数据**：处理为结构化的、有变量、数值、属性标签的数据

### 数据持续性

1. **已经存在的数据**：
   * 政府各类统计数据
   * 上市公司公开数据
   * 研究机构或个人公开的数据
2. **将要产生的数据**：系统采集的、不断推进补充的数据

### 数据产生者

* **一手数据**：研究者自己调查获取的数据
* **二手数据**：已被使用过的数据，可再次分析

使用二手数据的优势： - 数据的可靠性已被检验 - 研究成果具有可比性 - 节省调查资源和时间

### 数据获取权限

1. **无需授权**：
   * 正式出版物提供的数据
   * 仅需在使用说明中注明出处
2. **需要申请授权**：
   * 公开的学术研究数据
   * 需要申请并获授权
3. **需要授权且未公开**：
   * 政府数据
   * 银行数据
   * 电信数据等

## 二手数据收集

### 搜索引擎工具

* 谷歌搜索
* 谷歌学术
* 谷歌图书
* 必应搜索

### 综合型数据平台

1. **国内文献和数据**：
   * 中国知网（含统计年鉴资源）
   * 搜数网
   * 人大经济论坛
2. **国外文献和数据**：
   * SpringerLink电子期刊及电子图书
   * Wiley Online Library
   * ScienceDirect
   * Emerald
   * ProQuest学位论文全文库

### 专项型数据平台

1. **国外平台**：
   * 美国大学联盟数据集成中心(ICPSR)
   * 美国芝加哥大学广泛社会调查(GSS)
   * 美国芝加哥大学收入动态调查面板数据(PSID)
   * 美国密歇根大学健康和退休调查数据(HRS)
   * 英国艾塞克斯大学认识社会调查数据库
2. **国内平台**：
   * 北京大学中国社会科学调查中心(ISSS)
   * 中国人民大学中国调查与数据中心(NSRC)
   * 西南财经大学中国家庭金融调查与研究中心
   * 中国疾病控制中心(CDC)

### 互联网公开数据

1. **开放形式**：
   * 网站零散化页面
   * 网站结构化页面
   * 网站API数据接口
2. **所需技能**：
   * 网络爬虫编程能力
   * HTML、SQL和API知识

## 调查数据收集

### 自填式问卷调查

**特点**： - 无调查员协助 - 问卷递送方式多样（调查员分发、邮寄、网络、媒体）

**优点**： - 问卷结构严谨 - 说明清晰

**缺点**： - 返回率较低 - 不适合复杂问卷 - 调查周期长 - 问题难以及时调整

### 面访式问卷调查

**特点**： - 调查员与被调查者面对面 - 提问-回答形式

**优点**： - 回答率高 - 数据质量好 - 时间可控

**缺点**： - 成本高 - 质量控制难度大

### 电话式问卷调查

**特点**： - 通过电话实施调查

**优势**： - 速度快 - 适合样本分散情况

**局限性**： - 需要被调查者有电话 - 访问时间受限 - 问卷需简单 - 难以说服拒绝者

# 2.3 资料整理和数据清洗

## 资料整理的流程

### 分类存储

* 依据数据的载体类型、研究的时间需求进行分类
* 采用合适的存放工具进行存放
* 纸版问卷需要按照分类标准存放，便于后续工作

### 建立目录

* 目录用于检索和利用
* 便于数据的使用和查找
* 对调查获得的数据建立目录，方便检索

### 编制索引

* 对于复杂数据，在目录与存储之间建立关联
* 便于快速定位和查找数据

## 资料整理的记录

### 记录内容

1. **数据来源信息**：
   * 调查项目
   * 调查人
   * 采集人
   * 采集时间
   * 地点
   * 对象
2. **数据载体类型信息**：
   * 具体载体类型（如纸张、数字等）
3. **数据描述信息**：
   * 数据规模
   * 内容
   * 关联主题等
4. **数据分类信息**：
   * 根据载体形态或其他标准分类
   * 大型项目需要建立基本分类
5. **数据存储信息**：
   * 存储载体
   * 存储方式
   * 存储位置
   * 与数据安全相关的信息（版本、份数、时间变化等）

## 资料的安全

### 版本保存

* 保存多个版本
* 纸笔传统版本：便于翻阅
* 数字化可检索版本：便于使用

### 数字化数据安全

1. **数据存储**：
   * 随时保持多个备份
   * 注意存储介质的可靠性
   * 考虑数据安全性
2. **数据安全**：
   * 防范使用者的误操作
   * 防范内部或外部的有意攻击
   * 离线保存的重要性

### 不同类型数据的安全

1. **文本数据安全**：
   * 防范自然灾害
   * 防范人为因素（如错误识别）
2. **非数字化数据安全**：
   * 根据介质特征进行科学整理
   * 采取适当的保管措施（如防潮）

## 数据清洗的内容

### 真实性评估

* 确认数据的真实性
* 避免道听途说、张冠李戴、杜撰臆想
* 防范”假新闻”现象

### 完整性评估

* 确保数据与研究目标相符
* 研究不需要的数据不应出现
* 研究需要的数据不应缺失
* 必要时补充收集数据

### 可用性评估

1. **数据可用性**：
   * 确认数据是否可以用于数据库化
   * 必要时进行数据加工
2. **数据处理**：
   * 图片、音频、视频数据的数字化
   * 痕迹数据的抽样处理
   * 数据脱敏化处理

### 错误性评估

* 评估可能的错误来源
* 评估可能的错误大小
* 评估对数据质量的影响

## 数据清洗的记录

### 清洗工作记录

* 清洗步骤的做法
* 参与人
* 时间
* 地点
* 过程信息

### 清洗内容记录

1. **数据真实性信息**：
   * 是否真实
   * 是否存在编造、作弊嫌疑
   * 不真实的部分和原因
2. **数据完整性信息**：
   * 是否完整
   * 是否有缺失
   * 缺失的具体部分
3. **数据可用性信息**：
   * 问卷数据是否加权
   * 痕迹数据是否数据化
   * 大数据处理方式
4. **数据错误性信息**：
   * 问卷数据中的缺失
   * 文献数据中的差错

## 数据清洗的安全

### 笔记清洗

* 区分私用笔记和公用笔记
* 将个人笔记清洗为数据笔记
* 确保笔记的可读性和可用性

### 音视频清洗

1. **音频清洗**：
   * 将语音文档抄录为文字
   * 处理格式问题
   * 转换为通用格式
2. **视频清洗**：
   * 非数字录像的数字化
   * 数字录像的时间记录码处理

### 清洗注意事项

* 不要轻易丢弃信息
* 清洗的目的是让数据清晰化
* 保留原始观察记录
* 原始问卷至少保留十年以上
* 访谈记录和观察笔记要求永久保留

## 数据清洗操作

### 观测性数据清洗

* 处理观察性数据的差异性
* 将个性化观察数据转变为标准化记录
* 处理不同观察者的特异性

### 文献数据清洗

1. **笔记清洗**：
   * 格式化清洗
   * 将素材转化为数据
   * 必要时进行编码
2. **文献清洗**：
   * 数字版本与编目信息关联
   * 建立个人档案馆
   * 处理非数字化文献

### 痕迹数据清洗

1. **数据来源确认**：
   * 网络爬取数据
   * 数据拥有者机构提供的数据
   * 平行数据
2. **数据格式化**：
   * 将非格式化数据清洗为格式化数据
   * 分门别类整理数据
3. **数据结构化**：
   * 建立多维度关联
   * 形成样本数据

### 大数据清洗

* 清洗前测试清洗策略
* 采用大数据清洗模式
* 运用Hadoop框架下的Map Reduce
* 处理云计算逻辑

# 2.4 数据的数据库化

## 数据库化的必要性

### 计算机处理需求

* 数据数量和复杂程度超出人工处理能力
* 计算机处理是最有效和快捷的方式
* 需要将数据转换为计算机可读取的数据库格式

### 数据库化的目的

* 便于数据分析和使用
* 使调查数据格式化、结构化
* 符合统计分析、计算的要求

## 数据库化的类型

### 计算机网络系统数据库化

* 主要用于数据存储
* 使用各种类型的数据库应用程序
* 如开源的MySQL等SQL数据库

### 分析计算用数据库化

* 通过建立数据库用于统计分析软件计算
* 主要学习单机统计计算与分析用的数据库化
* 大数据数据库化有不同特点和需求

## 调查数据的数据库化

### 主要步骤

1. **编码**：
   * 在清理工作中完成
   * 数据入库前需要审核
2. **数据录入与转化**：
   * 纸版问卷：使用专门录入软件
   * 计算机辅助调查：进行数据转化
   * 避免录入和转化中的差错
3. **检验和清理**：
   * 检查错行、错列问题
   * 确保数据准确性

### 编码方法

1. **原始编码**：
   * 直接运用问卷的编码
   * 适用于访题数量极少、应答简单的情况
2. **先编码**：
   * 调查开始前完成编码
   * 适用于基本是封闭访题的情况
3. **后编码**：
   * 问卷调查完成后进行编码
   * 适用于有开放访题的情况

### 编码部

* 相当于问卷数据的索引
* 关联变量、变量值和变量标签
* 类似于问卷数据字典

### 录入工具

1. **简单问卷调查**：
   * MS Office Excel
   * Mac Numbers
   * SPSS
   * Stata、statistica、R等
2. **复杂问卷调查**：
   * SPSS Data Entry模块
   * EpiData等专门录入软件

### 检验和清洗

1. **录入错误清理**：
   * 核对双录入的冲突数据
2. **编码清理**：
   * 检查变量值是否在编码范围内
3. **逻辑清理**：
   * 检查基本事实逻辑
   * 处理离群值、极大值和极小值
   * 处理无应答情况
   * 进行变量再编码

## 访谈调查数据的数据库化

### 主要步骤

1. **编码**：
   * 记录信息的编码
   * 记录内容的编码（用于文本分析）
2. **录入**：
   * 录入访谈记录信息
   * 录入访谈内容（用于内容分析）
3. **清理**：
   * 逐行核查
   * 确保内容准确性

### 编码内容

1. **访谈记录信息编码**：
   * 记录编号
   * 访谈时间
   * 地点
   * 人物
   * 主题
   * 位置图
   * 日志信息
2. **访谈记录编码**：
   * 用于内容分析软件的编码
   * 不同软件有不同编码要求

### 录入工具

1. **数字数据**：
   * Excel、SPSS、Stata等
2. **文本数据**：
   * Word、Numbers、Pages等
3. **内容分析软件**：
   * Nvivo、Aquad、ATLAS.ti等

## 观察数据的数据库化

### 主要步骤

1. **编码**：
   * 观察记录信息的编码
   * 观察记录内容的编码
2. **录入**：
   * 录入观察记录信息
   * 录入观察记录内容
3. **清理**：
   * 核查录入数据
   * 确保内容准确性

### 编码内容

1. **观察记录信息编码**：
   * 记录编号
   * 观察时间
   * 地点
   * 事件
   * 主题
   * 观察媒体
   * 日志信息
2. **观察记录内容编码**：
   * 录入为数据化文本文件
   * 便于交流和使用

### 录入工具

1. **文本和数字数据**：
   * Word或Pages
2. **图片数据**：
   * Adobe Lightroom等
3. **视频数据**：
   * Adobe Premier等
4. **音频数据**：
   * 专用音频数据库

## 文献数据的数据库化

### 主要步骤

1. **编码**：
   * 文献信息的编码
   * 编目信息的处理
2. **录入**：
   * 将原数据库信息抄录到研究用数据库
3. **清理**：
   * 核查录入数据
   * 完整性检查

### 编码内容

1. **文献记录信息**：
   * 作者
   * 篇名
   * 时间
   * 载体
   * 存放
   * DOI/ISBN/ISSN等
2. **文献内容信息**：
   * 文献内容
   * 阅读笔记
   * 思路图谱
   * 总结要点

### 管理软件

1. **文献管理**：
   * Zotero
   * Endnote
   * Papers
2. **内容管理**：
   * OneNote
   * MindManager
   * 印象笔记

## 痕迹数据的数据库化

### 主要特点

* 基于变量的数据
* 需要转换为基于样本的数据
* 主要依靠脚本完成

### 主要步骤

1. **编码**：
   * 确认使用原有编码
   * 必要时重新编码
2. **清理**：
   * 确认数据的可计算性
   * 检查格式化、结构化问题
   * 确认是否可用于分布式并行计算或单机计算

]

# 2.5 数据质量

## 数据质量评判原则

### 基本原则

1. **真实性**：
   * 数据确实来源于调查
   * 数据产生过程真实存在
   * 调查对象真实存在
   * 访问、观察真实存在
   * 应答、场景、文献真实存在
2. **准确性**：
   * 调查人员准确执行研究设计
   * 准确处理调查对象和反馈
   * 准确转录原始数据
3. **时效性**：
   * 调查实施过程符合规定时间要求
   * 满足时效要求的数据

### 评判维度

1. **正向评估**：
   * 与标准要求的距离
   * 符合性问题
2. **反向评估**：
   * 误差的大小

## 误差分类

### 按来源分类

1. **覆盖性误差**：
   * 涉及调查对象的备选机会
   * 存在于所有调查方法中
   * 抽样框与目标总体不一致
2. **测量性误差**：
   * 来源于测量工具
   * 来源于工具使用过程
   * 反映在信度和效度上
3. **应答性误差**：
   * 调查对象未回应或未完全应答
   * 导致数据缺失
   * 存在于所有调查类型中
4. **抽样性误差**：
   * 仅出现在抽样问卷调查中
   * 来源于抽样框制作
   * 来源于抽样过程

### 按可改进性分类

1. **随机误差**：
   * 调查活动中随机产生
   * 可通过规范行为减少
   * 增大变量测量的方差
2. **系统误差**：
   * 由设计因素影响产生
   * 测量工具带来的误差
   * 增大测量的偏移量（bias）

### 调查总误差

* 所有数据收集、整理、清洗活动产生的误差综合
* 用均方误（MSE）表示
* MSE = 偏差平方 + 方差

## 覆盖性误差

### 基本概念

* 又称抽样框误差
* 目标总体与抽样框总体不一致
* 导致调查对象错位

### 相关概念

1. **目标总体**：
   * 调查对象总体
   * 有明确的调查对象所指
2. **抽样框总体**：
   * 用于抽样的所有调查对象的集合
   * 简称框总体
3. **样本总体**：
   * 被抽中的调查对象集合

### 误差来源

1. **丢失或重叠目标总体要素**：
   * 框总体小于或大于目标总体
   * 部分要素失去或获得多次被抽中机会
2. **包含非目标总体要素**：
   * 框总体大于目标总体
   * 目标总体备选概率降低
3. **不正确的辅助信息**：
   * 分层抽样中使用错误信息
   * 影响层要素的备选概率

## 测量性误差

### 信度检验

1. **概念**：
   * 测量工具的可靠性
   * 重复测量得到相同结果的概率
   * 测量工具的稳定性
2. **实践类型**：
   * 垂直重复信度（前-后测信度）
   * 水平重复信度（复本信度）
3. **计算方法**：
   * 折半信度法
   * 克隆巴赫系数法（Cronbach α）

### 效度检验

1. **概念**：
   * 测量工具的正确性和有效性
2. **类型**：
   * 预测效度
   * 同时效度
   * 结构效度
   * 内容效度

## 应答性误差

### 基本概念

* 访员发出调查请求，调查对象未回应或未完全应答
* 导致数据缺失
* 可理解为广义覆盖性误差的一种

### 无应答类型

1. **对象无应答**：
   * 抽样调查中称样本无应答
   * 非抽样调查中称失访
2. **议题无应答**：
   * 抽样调查中称选项无应答
   * 非抽样调查中称议题未访到

### 应答率

* 应答率 = 应答样本数 / 样本总数 × 100%
* 分子：完全应答或部分应答
* 分母：有效样本、未接触样本、无应答样本等

## 抽样性误差

### 误差来源

1. **抽样框制作**：
   * 样本覆盖性问题
   * 抽样框与目标总体不一致
2. **抽样过程**：
   * 分层、多阶段抽样
   * 末端抽样方法
   * 抽样人员操作

### 误差计算

1. **均值的变异系数**：
   * 样本均值除以标准误
   * 比例值：
   * 经验上小于50%可接受
2. **样本均值的相对方差**：
   * 样本方差除以均值的平方
   * 比例值：

# 2.6 抽样设计

## 抽样设计的基本概念

### 抽样调查的定义

* 从总体中抽取部分单位作为样本
* 对样本进行调查
* 用样本结果推断总体特征

### 抽样调查的特点

1. **经济性**：
   * 节省人力、物力和时间
   * 降低调查成本
2. **时效性**：
   * 快速获取数据
   * 及时反映总体情况
3. **准确性**：
   * 减少调查误差
   * 提高数据质量
4. **灵活性**：
   * 适应不同调查需求
   * 便于组织实施

## 抽样设计的基本原则

### 随机性原则

* 每个单位有相同的被抽中机会
* 保证样本的代表性
* 避免主观选择偏差

### 代表性原则

* 样本能够反映总体特征
* 样本结构与总体结构相似
* 保证推断的可靠性

### 可行性原则

* 考虑实际调查条件
* 便于组织实施
* 确保调查质量

## 抽样设计的基本步骤

### 确定调查目标

1. **明确研究问题**：
   * 确定研究目的
   * 界定研究范围
2. **确定调查内容**：
   * 选择调查指标
   * 确定调查项目

### 确定总体和抽样框

1. **界定总体**：
   * 明确总体范围
   * 确定总体单位
2. **建立抽样框**：
   * 列出总体单位清单
   * 确保抽样框的完整性

### 选择抽样方法

1. **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
2. **非概率抽样**：
   * 方便抽样
   * 判断抽样
   * 配额抽样
   * 滚雪球抽样

### 确定样本量

1. **影响因素**：
   * 总体规模
   * 总体异质性
   * 允许误差
   * 置信水平
   * 调查成本
2. **计算方法**：
   * 基于总体方差
   * 基于比例估计
   * 基于成本约束

### 实施抽样

1. **抽样过程**：
   * 严格按照抽样方案
   * 记录抽样过程
   * 确保抽样质量
2. **质量控制**：
   * 检查抽样结果
   * 评估样本代表性
   * 处理异常情况

## 抽样方法的选择

### 简单随机抽样

* 每个单位有相同被抽中机会
* 适用于总体单位较少
* 操作简单，易于理解

### 分层抽样

* 按特征将总体分层
* 在各层内随机抽样
* 提高估计精度

### 整群抽样

* 将总体分成若干群
* 随机抽取若干群
* 对选中群进行全面调查

### 系统抽样

* 按固定间隔抽取样本
* 操作简便
* 适用于有序总体

### 多阶段抽样

* 分阶段进行抽样
* 适用于大规模调查
* 降低调查成本

## 样本量的确定

### 影响因素

1. **总体特征**：
   * 总体规模
   * 总体异质性
   * 总体分布
2. **调查要求**：
   * 允许误差
   * 置信水平
   * 调查精度
3. **实际条件**：
   * 调查成本
   * 时间限制
   * 人力物力

### 计算方法

1. **基于总体方差**：

其中： - 为样本量 - 为总体规模 - 为总体方差 - 为允许误差

1. **基于比例估计**：

其中： - 为总体比例 - 其他符号同上

## 抽样误差的控制

### 误差来源

1. **抽样误差**：
   * 样本与总体差异
   * 不可避免的随机误差
2. **非抽样误差**：
   * 调查设计误差
   * 调查实施误差
   * 数据处理误差

### 控制方法

1. **抽样设计控制**：
   * 选择合适的抽样方法
   * 确定合理的样本量
   * 优化抽样方案
2. **调查实施控制**：
   * 规范调查程序
   * 加强质量控制
   * 减少非抽样误差
3. **数据处理控制**：
   * 严格数据审核
   * 规范数据处理
   * 保证数据质量

# 2.7 抽样分布和抽样误差

# 抽样误差

## 抽样误差概述

抽样误差是指由于抽样过程中样本与总体之间的差异而产生的误差。它是抽样调查中不可避免的误差来源之一，主要受以下因素影响：

* 总体异质性程度
* 抽样方法的选择
* 样本量的大小
* 抽样设计效率
* 应答率

## 系统抽样误差

系统抽样是一种等距抽样方法，其抽样误差计算公式如下：

其中： - 为总体单位数 - 为样本量 - 为平均群组大小 - 为第个群组的样本均值

## 整群抽样误差

### 无偏估计法

其中： - 为总体群组数 - 为总体单位数 - 为样本群组数 - 为群组间方差

### 比率估计法

其中： - 为第个群组的单位数 - 为比率估计值

## 多阶段抽样误差

### 无偏估计法

### 比率估计法

### 比例概率估计法（PPS）

## 必要样本量确定

必要样本量的确定需要考虑以下因素：

1. 总体特征
   * 总体规模
   * 总体异质性程度
   * 总体方差
2. 抽样设计
   * 抽样方法
   * 抽样效率
   * 设计效应
3. 估计精度要求
   * 可接受的误差水平
   * 置信度要求
   * 可靠性系数
4. 其他因素
   * 调查成本
   * 应答率
   * 可用资源

## 抽样误差的注意事项

1. 选择合适的抽样方法
   * 根据总体特征选择适当的抽样方法
   * 考虑抽样效率和成本效益
2. 正确使用方差计算公式
   * 不同抽样方法使用对应的方差计算公式
   * 避免错误使用简单随机抽样的方差公式
3. 考虑设计效应
   * 复杂抽样设计需要考虑设计效应
   * 设计效应通常为2左右较为理想
4. 样本量确定
   * 小规模研究最小样本量建议为30
   * 大规模研究需要综合考虑多种因素
5. 误差控制
   * 注意非抽样误差的控制
   * 考虑缺失值的影响
   * 提高应答率

## 抽样误差的实践应用

1. 在实际调查中，需要根据具体情况选择合适的抽样方法
2. 抽样误差的估计需要考虑调查的实际情况
3. 抽样误差的估计需要考虑调查的可行性
4. 抽样误差的估计需要考虑调查的经济性
5. 抽样误差的估计需要考虑调查的时效性

在实际应用中，需要根据具体情况选择最合适的估计方法，以获得更准确的估计结果。

# 2.8 问卷设计技术

## 问卷设计的主要步骤

问卷设计的主要步骤包括：

* 决定所需要的信息
* 根据所需资料，开发个别的问句
* 决定问卷的顺序(ordering)
* 拟定问卷初稿
* 问卷实体制作
* 检讨和修正
* 前测(pretest)
* 定稿

## 量表设计的基本流程

量表设计的基本流程包括：

A. 确定调查的目的

B. 定义量表结构及其关系 - 需要进行文献综述 - 制定一个概念框架/理论框架

C. 查询可供参考或采用的量表

E. 写出量表的条目

F. 对量表开展试点调查

G. 分析试调数据并相应修订量表

## 李科特量表

**李科特量表**（Likert Scale）是一组相互关联的调查项目，通过一组非常具体的一致的回答选项来测度一个特定的目标话题（潜在变量、理论变量）。

* 李科特量表可能是调查研究中最容易被误解、误用、滥用的一个术语或问卷设计技术
* 李科特形态的设问题目并不属于李科特量表
* 即使一个单独的问题可能以陈述的形式伴随着一个反应量表，其选项包括（5分制或7分制形式）非常同意或非常不同意，但这不是李克特量表

## 量表质量：信度

**信度（Reliability）**刻画量表的真实得分与实际得分之间的一致性程度（Consistency），同时将评估引起真实得分与实际得分不一致的多种误差来源。

### 信度类别

1. **内部信度(internal)**：基于量表内部的多个题项，使用克朗巴哈信度系数(Cronbach)计算
2. **评估者信度(inter-rater)**：基于不同评分者打分，使用科恩卡帕系数(Cohen)计算
3. **副本信度(parallel)**：基于量表设计的多个版本，使用等效系数(equivalence)计算
4. **重测信度(test–retest)**：基于不同调查者的多次测试，使用稳定系数(stability)计算

### 内部信度：克朗巴哈信度系数

克朗巴哈信度（Cronbach Alpha 或）是表达量表内部一致性程度的一种测度系数，具体计算公式为：

其中： - 题项方差（item variance，纵向方差） - 观测得分方差（total test variance，横向方差） - 题项之间的协方差

### 评估者信度

评估者信度用于测度不同评分者打分一致性，可以通过以下系数计算：

1. **评级共识系数（nominal agreement）**：
2. **科恩卡帕系数（Cohen Kappa）**：

## 量表质量：效度

**效度（Validity）**指的是证据和理论在多大程度上支持对量表得分的解释，以达到预期的研究目的。

### 效度类别

1. **内容效度(content relevance)**：评价量表内容与研究意图的相关程度
2. **流程效度(response process)**：评价受访者的调查认知和参与过程的真实性
3. **结构效度(internal structure)**：评价量表题项和量表成分之间的一致性
4. **变量效度(variables relation)**：评价不同量表之间的关联程度
5. **影响效度(consequence effect)**：评价调研后影响效果与预期的一致性

## 量表设计的基本原则

### 量表标签与内容结构一致性

量表标签设定应该与量表内容结构保持一致，避免出现不一致的情况。

### 题项问题与应答选项一致

问题的表述应该与应答选项保持一致，确保被调查者能够准确理解并做出选择。

## 问卷设计的基本原则

问题设计的基本原则包括：

* 问题要让受访者充分了解，问句内容不可超出受访者之知识及能力之范围
* 问题是否切合研究假设之需要
* 要能引发受访者真实的反应，而非敷衍了事
* 问项是否含混不清，易引起受访者的误解
* 问题是否涉及社会禁忌、偏好
* 问题是否产生暗示作用
* 便于忠实的记录
* 便于数据处理及数据分析

## 问卷设计的用词原则

问题设计的用词原则有：

* 使用通用的词汇要浅显易懂
* 问题描写要简单明了
* 语句意义要清楚不能模糊
* 不能假设受测者都懂
* 不能用有偏差误导的字句
* 不要有暗示的作用
* 不要隐藏其它的方案
* 间接问题的利用
* 句子要短而集中，且一个问句只问一个事物、概念或事件

## 问卷设计的友善原则

卷首语相当于问卷的门面，好的卷首语可以增加被调查者填写问卷的可能性。卷首语主要包括：

* 表明身份
* 说明调查目的
* 作出承诺
* 必要时，可以加上问卷作答所需时间

## 问卷设计的常见误区

### 未穷尽所有选项

问题选项设计时，需要补齐所有可能性的选项。对于会出现多种情况的，可以添加一个【其他】，【以上情况都有】选项，保证被调查者有选项可以选择。

### 不符合惯例

问卷设计出来后，应当先找几个没有参与设计问卷的不同文化水平的人进行试填，大多数问题都可以在试填过程中被发现。

### 出现生僻用词

在设计选项时，尽可能避免生僻词语，更多的使用大众经常使用的词语。必要时，需在生词旁边进行简单的解释，让被调查者可以正确理解问卷选项的内容。

### 重点不明

必要时可以采用加”【】“或加着重号等，将重点标注。

### 未设置甄别式问题

有必要进行检验性题目的设置，剔除无效问卷，从而保证问卷数据的准确可靠。甄别性题目间的位置最好不要太接近，两道题之间选项的内容要打乱。

## 问卷设计的注意事项

### 避免提问笼统、抽象

问句的最终落脚点往往是被调查者最后接收到的信息。因此，问句应尽量把问句的主语、谓语、宾语等关系理顺。

### 避免答案设计重叠

答案选项之间不应存在重叠，而且应该包含所有可能性。

### 避免双重或多重涵义

问题的含糊往往是对某个容易产生歧义的要素，缺乏限定或限定不清引起的。可参考六要素对照法。

### 避免超出认知范围

换位思考、设身处地，站在被调查者的视角来审视研究者自己设定的问题。

### 避免问题与答案不一致

在问卷设计时，一定要注意提问与选项的范围一致、内容一致、口径一致、对应一致。

### 避免诱导性提问

引导性提问会导致不良后果： - 被调查者不加思考就同意所引导问题中暗示的结论 - 对于一些敏感性问题，在引导性提问下，被调查者不敢表达真实想法 - 引导性提问惯用权威或大多数人的态度，在引导性提问下，被调查者会不敢表达其真实想法

### 避免强制回答式设问

当预计某些问题回答者可能没有看法时，应在答案中区分出来：没有看法。另外也可以在问题中说明”许多人都没有固定的意见”，这样人们在作决定时的心理压力要小一些，减少臆测的可能性。

### 处理敏感性和伤害性话题

对于敏感性和伤害性问题的处理对策： - 去掉不必要的敏感性问题 - 通过技术手段降低提问的敏感程度 - 采用非直接、联想式提问 - 将所询问的行为或态度”大众化”

## 问卷设计的质量控制

### 甄别性问题

甄别性问题用于检验被调查者是否认真填写问卷，常见的甄别性问题包括： - 要求选择特定选项 - 设置虚构的选项 - 设置重复性问题

这些甄别性问题可以帮助识别无效问卷，提高问卷数据的质量。