第01章 导论

# 为什么学习统计学

## 辛普森悖论的警示

* 辛普森悖论表明，在分析数据时需要注意以下几点：
  + 数据的分组方式会影响结论
  + 总体数据和分组数据可能得出相反的结论
  + 需要考虑潜在的混杂因素

## 信念偏见诅咒

* 信念偏见效应：人们在判断论点的逻辑有效性时，往往会受到结论可信度的影响
* 统计分析需要克服主观偏见，保持客观中立
* 需要基于数据事实，而不是个人信念做出判断

# 变量和数据

## 变量的类型

1. **按性质分类**：
   * 定性变量：非数值化的变量
   * 定量变量：数值化的变量
     + 离散变量：可能的取值有限且可列举
     + 连续变量：可能的取值较多，以特定微小数值间隔
2. **按随机性分类**：
   * 随机变量：在事件集合上具有随机分布状态
   * 非随机变量：也称为确定性变量
3. **按抽象化程度分类**：
   * 经验变量：可观察到的实际事物
   * 理论变量：统计学家构造的数学变量（如z统计量、t统计量等）

## 数据的类型

1. **按性质分类**：
   * 定性数据：定性变量的取值
   * 定量数据：定量变量的取值
     + 离散数据：离散变量的取值
     + 连续数据：连续变量的取值
2. **按获取方式分类**：
   * 观察性数据：被动记录的数据
   * 实验性数据：通过主动控制获得的数据

# 数据的计量层次

## 名义数据

* 仅用于区分类别的定性数据
* 数据表现为类别，通常用文字表述
* 例如：性别、婚姻状况等

## 顺序数据

* 变量取值具有自然顺序
* 取值的差值没有实际意义
* 例如：五分量表、满意度评级等

## 区间数据

* 取值有自然顺序且差值有意义
* 取值的比率没有实际意义
* 例如：温度、年份等

## 比率数据

* 存在真实”零点”
* 取值有自然顺序
* 取值的差值和比率都有实际意义
* 例如：GDP、个人收入等

# 数据的时间状态

## 时间序列数据

* 对变量在不同时间点的观测结果
* 分为高频数据和低频数据
* 需要考虑平稳性问题

## 截面数据

* 在同一时间点对不同对象的观测
* 需要考虑异质性问题
* 可能存在尺度效应

## 面板数据

* 兼具时间序列和截面数据特征
* 分为平衡面板和非平衡面板
* 观测数 = 截面单元数 × 时期数

# 统计学的体系

## 描述性统计

* 研究数据收集、处理、汇总等方法
* 目的是描述数据特征，找出基本规律
* 主要内容包括：
  + 搜集数据
  + 整理数据
  + 展示数据
  + 描述性分析

## 推断性统计

* 研究如何利用样本推断总体特征
* 主要内容包括：
  + 参数估计
  + 假设检验

## 总体和样本

* 总体：研究对象的全部个体集合
  + 有限总体
  + 无限总体
* 样本：从总体中抽取的部分个体
  + 样本容量：样本中的个体数量

## 参数和统计量

* 参数：描述总体特征的数字度量
  + 用希腊字母表示（如μ, σ²）
* 统计量：描述样本特征的数字度量
  + 用英文字母表示（如x̄, s²）

# 统计分析的基本过程

## 基本步骤

1. **实际问题**：发现问题
2. **收集数据**：取得数据
3. **处理数据**：整理与图表展示
4. **分析数据**：利用统计方法
5. **解释数据**：结果说明
6. **得出结论**：客观结论

## 统计分析能力要求

1. **基础背景教育**：
   * 数学、概率论等基础知识
   * 前沿阅读能力
   * 知识管理能力
2. **数据获取与管理**：
   * 实验设计
   * 数据库管理
   * 数据清洗能力
3. **分析与展示**：
   * 数据可视化
   * 统计建模
   * 结果解释
4. **报告与交流**：
   * 学术写作
   * 演讲汇报
   * 团队协作