

模块三: 从脏数据到干净数据

可用数据文件(含教育测量特色)

本模块提供专门设计的脏数据和测试题数据:

- dirty_student_grades.csv - 含各类数据质量问题的成绩表
- test_items_raw.csv - 原始测试题内容
- test_items_coded.csv - 编码后的测试题(知识点/能力维度)
- item_features_calculated.csv - 题目统计特征(难度/区分度)
- student_responses_raw.csv - 学生答题原始记录

特色内容1: 测试题编码与标注

这是教育测量领域的核心技能!

编码维度:

- 题型编码: SC(单选)、FB(填空)、SA(简答)等
- 知识点层级: L1学科 -> L2章节 -> L3具体知识点
- 能力维度: 布鲁姆分类法(1记忆->6创造)
- 预估难度: 基于教师经验的难度预测

特色内容2: CTT指标计算

经典测量理论核心指标:

- 难度P值 = 答对人数/总人数(0-1之间,越高越简单)
- 区分度r = 高分组答对率 - 低分组答对率(>0.3为良好)
- 干扰项有效性分析: 每个选项的选择分布

基础内容: 脏数据识别

dirty_student_grades.csv中的问题类型:

- 缺失值: S003的成绩标记为"-"
- 重复记录: S001出现两次
- 异常值: S022数学成绩999分(超出满分)
- 格式错误: S018数学成绩为"abc"
- 计算错误: total_score与各科成绩之和不符