第九讲（第13章）资料处理

资料处理的一般过程

现场控制

↓

逻辑处理

↓

数学处理

↓

系统研究

↓

编写

现场控制

对不合格问卷的处理：

- 返回现场工作**Returning to the Field**
- 指派缺失值**Assigning Missing Values**
- 丢弃不合格问卷**Discarding Unsatisfactory Respondents**

逻辑处理

- （1）鉴别
- （2）分类
- （3）编码
- （4）合并
- （5）比较
- （6）筛选

不真实信息资料的形式

- 1.欺骗　　2.误传
- 3.虚构　　4.添加
- 5.拼凑　　6.混淆
- 7.夸张　　8.偏颇
- 9.孤证　　10.回避
- 11.捉影　　12.假象

信息资料鉴别的具体方法

- 1.分析判断法
- 2.核对法
- 3.比较法
- 4.佐证法

数据编码

编码是为每个问题的每种可能的答案分配一个代码。

编码字典记录编码结构和有关变量的必要信息，编码字典包括如果下内容：

- 栏数column number
- 记录数record number
- 变量数variable number
- 变量名 variable name
- 问题数question number
- 编码说明 instructions for coding

编码方法

- 1.顺序编码法
- 2.分组编码法
- 3.信息组码编码法
- 4.表意式文字编码法
- 5.缩写编码法

数据清理

1.一致性检查

Consistency checks identify data that are out of range, logically inconsistent, or have extreme values.

- 超出正常范围的取值，如备选答案为1-5，在数据中出现了6,7。
- 逻辑上不合理的取值。
- 仔细检查极端值（ extreme values ）。

异常值(outlier)：离最近分位数的距离在四分位差的1.5-3倍之间,SPSS中用O表示。

极端值(extreme values)：离最近分位数的距离在四分位差的3倍以上,SPSS中用＊表示。

例：aflatoxin.sav

2.缺失值处理

- 以中间值代替**Substitute a Neutral Value**.

  用整个序列的均值替换缺失值

  线性插值，缺失值前的最后一个有效值和缺失值后的第一个有效值的平均数
- 用估计值代替**Substitute an Imputed Response**.

  点处的线性趋势，以1－n为自变量，作回归后，用相应的预测值代替。
- 整例删除**casewise deletion**.
- 成对删除 **pairwise deletion**.

对数据进行转换

1.赋权 **weighting**

2. 量纲转换Scale transformation.

例如，常用标准化 **standardization**. Z_i, :

$$Z_i = (X_i - \bar{X})/s_x$$