定量分析方法入门 (SPSS)

一、SPSS 软件简介

SPSS Statistics is a software package used for interactive, or batched, statistical analysis. Long produced by SPSS Inc., it was acquired by IBM in 2009. The current versions (2015) are named IBM SPSS Statistics. The software name originally stood for **Statistical Package for the Social Sciences**(SPSS), reflecting the original market, although the software is now popular in other fields as well, including the health sciences and marketing. (*Wikipedia*)

- 常用统计软件

软件	收费	可视化	优势
SPSS	是	是	社会科学统计分析；简单制图
Excel / Tableau	是	是	较直观的数据处理；简单制图
R	否	否*	拓展性；灵活制图；较大数据
Python	否	否*	拓展性；通用编程；大数据处理

* 有对应的 Jupyter Notebook / Rstudio，但主要操作均以编程为主

- Syntax：命令、指示请参见 IBM 社区教程
- 清洗数据（data cleaning）；数据挖掘（data mining）
- 处理、冶炼、提纯、设计、打造、包装
- 数据视图 Data view 与变量视图 Variable view (name, type, width, decimals, label, values, missing, columns, align, measure, role)

二、SPSS 具体操作

读取、选择与录入

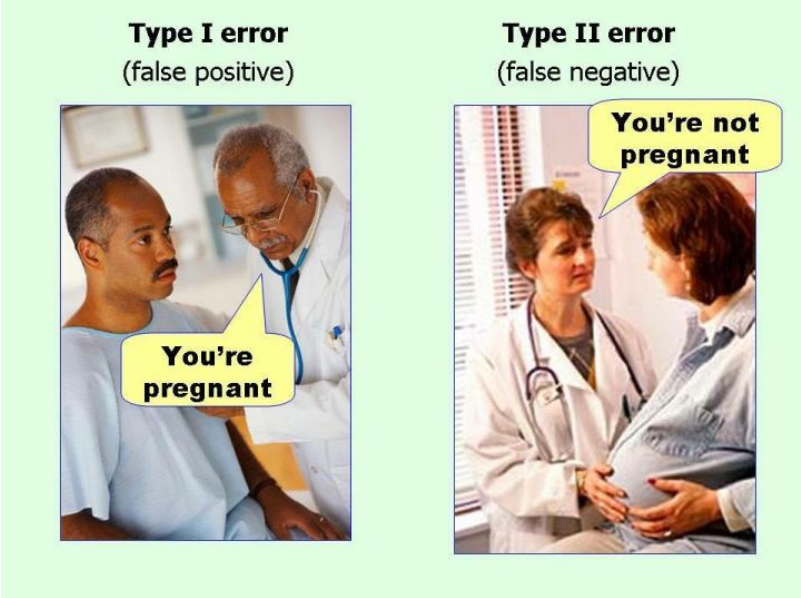
- 读取数据 **Read**
- 选择数据 **Select**

清理与转换

- 转换数据
  - 重编码 Recode
  - 重编码为不同变量 Recode into different variables
  - 转换 Transform
- 合并数据 **Merge**
- 计算数据 **Compute**

分析 Analyze

- 描述统计 **Descriptive statistics**
  - 频率 Frequencies
  - 描述统计量 Descriptives
  - 交叉表 Crosstabs
- 推断统计 **Inferential statistics**
  - 两种假设与两类错误
    - 零/原假设(Null/default hypothesis) 与备择假设(Alternative hypothesis)
    - 一类错误(type I error)：“弃真”，原假设正确但被错误地拒绝——男人怀孕了
    - 二类错误(type II error)：“取伪”，原假设错误但被错误地接受——孕妇没怀孕



  - 显著性水平 α
    0.05 * 0.01 **
- 比较均值 统计课上已经学过
  - 类型
    1. 单样本 t 检验 (One-Sample T Test)
    2. 独立样本 t 检验 (Independent- Samples T Test)
    3. 配对样本 t 检验 (Paired-Samples T Test)
  - 操作
  - 假定
  - 报告
  - * 单尾 t 检验？
- 交叉表 **Crosstab**
  - 定类-定类：lambda, tau-y
  - 定序-定序：gamma，斯皮尔曼 rho
  - 定距-定距：peason
  - 定类-定序：lambda, tau-y

    - 定类-定距：eta
    - 定序-定距：相关比率 E2
    - *方差分析？

- 相关与回归 **Correlation and regression**
    - 相关
        - 定义：两事物之间非一一对应的统计关系（薛薇）
        - 操作：
            1. 绘制散点图查看变量间是否可能存在相关关系
            2. 分析 - 相关 - 双变量
            3. 添加需要分析的变量到 变量
            4. 在 相关系数 选择需要计算的相关系数
            5. 在 显著性检验 选择 双侧检验 还是 单侧检验
            6. 勾选 标记显著性相关 以 * 形式展现显著性结果
            7. 在 选项 中 勾选 双积偏差和协方差 以输出利差平方和、样本方差、两变量的叉积利差和协方差。
        - 判断：
            1. 相关系数
                - 正负：正相关与负相关
                - 大小：经验值，r<0.2 极弱或无；0.2-0.4 弱；0.4-0.6 中；0.6-0.8 强；0.9-1 极强
            2. p 值
               $p < 0.05$ ($p < \alpha, \alpha$ 也可以是 0.01) 即显著（我们拒绝“两个变量不存在相关关系”的零假设，而接受“两个变量存在相关关系”备择假设，这种判断与事实不符的概率）
        - 报告：
            - 相关系数 / r
            - 显著度 / p 值 / p-value
            - 根据显著性水平标注 * 号
        - 假定：
            1. Pearson 相关系数要求两个样本均符合正态分布（以参数检验）
            2. 样本独立性
            3. *极值已经处理
        - * 三种相关系数？ Pearson 相关系数用于定距定比变量之间的相关关系（参数检验）；Spearman 相关系数用于定序变量间相关关系（参数检验）；Kendall 相关系数用于度量定序变量间线性相关关系（非参数检验）。
        - *不符合正态分布？ 开根号、对数、倒数等
    - 回归
- 降维 **Dimension reduction**

    - 因子分析 Factor analysis
    - scale

绘图 Graphs

- 条形图 Histogram
- *散点图、折线图、饼图、箱线图

三、参考资料

- 李连江《戏说统计：文科生的量化方法》：觉得太枯燥可看的入门；
- 薛薇《SPSS 统计分析方法与应用》：高效，每种分析方法的实际操作。

四、拓展阅读

- 邱泽奇《社会统计学》：对社会统计分析方法的系统介绍