定量分析方法入门 (SPSS)

> 文中链接仅用于初次接触快速通过网页了解，请通过查阅书籍学习详情。

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

（一）读取、选择与录入

- **读取数据 Read 选择数据 Select**
- 录入：名称、类型、缺失值、测量（、标签、值）

SPSS 测量	测量对应
标度	定距(+-<>=≠)、定比(*/+-<>=≠)
有序	定序(=≠<>)
名义	定类(=≠)

*另一种分类：连续、离散、分类

- 通常，缺失值以 7777, 8888, 9999 为标记

（二）清理与转换

数据清理 Data cleaning

是整个数据流程本身的 60%工作甚至更多

- **缺失值**：分析中排除个案；建模中可以考虑排除或者使用均值填充
- **异常值**：箱线图、z-score $z = \frac{x-\mu}{\sigma}$ > 2.5/3/3.5

转换数据 Transformation

- 重编码 Recode（不推荐，会覆盖原始数据）： 转换 - 重编码为相同变量-选择变量- 旧值和新值- 定义对应关系 -确定
- 重编码为不同变量 Recode into different variables：转换 - 重编码为不同同变量-选择变量- 旧值和新值- 定义对应关系-确定
- * 可视分箱？

* 合并数据 Merge

纵向添加个案、横向添加变量

计算数据 Compute

转换 - 计算变量-目标变量- 对应关系 -（如果...）-

例如求绝对值(`Abs`)、求平均数(`Mean`)、求对数(`Log10`, `Ln`)等

（三）分析 Analyze

分析的类型

- 单变量：集中和离散的趋势
- 双变量：是否有关系、关系的方向、关系的强度
- 多变量
- 等等

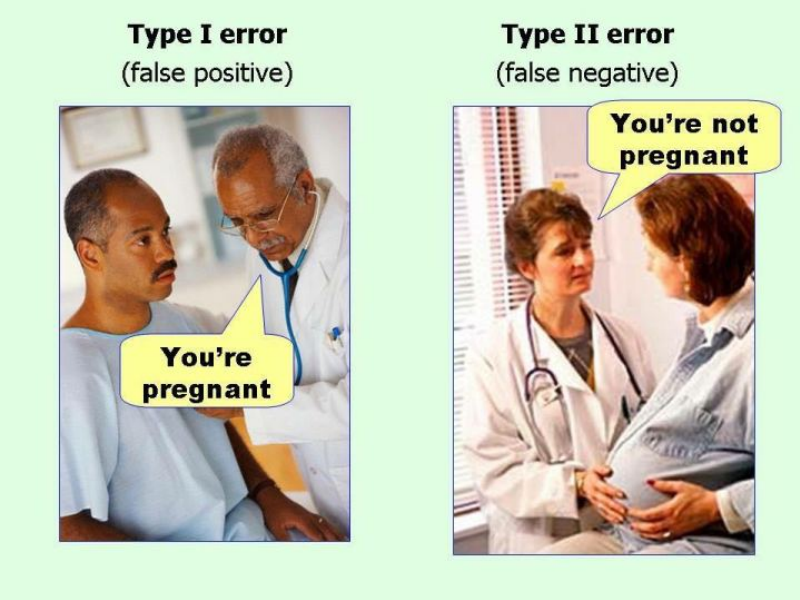
描述统计 Descriptive statistics

- **频率 Frequencies**：分析 - 描述统计- 频率-选择变量 -统计量-图表-格式（按计数的个数排序）
- **描述统计量 Descriptives**：分析 - 描述统计- 描述/探索
  - 集中趋势：均值、中位数、众数
  - 离散趋势：标准差、方差、最大值、最小值、范围(range)、分布（对称、左偏、右偏）
  - 样本与总体

推断统计 Inferential statistics

- 两种假设与两类错误

- 假设检验的逻辑
- **零/虚无/原假设 (Null/default hypothesis, H_0)** 希望证伪或推翻的假设，一般是一种虚无、没有关系的状态
- **备择/对立假设 (Alternative hypothesis, H_1)** 希望证实或支持的假设
- **一类错误(type I error)**：“弃真”，原假设正确但被错误拒绝，男人怀孕了
- **二类错误(type II error)**：“取伪”，原假设错误但被错误接受，孕妇没怀孕



- 显著性水平 α
  < 0.05 *, < 0.01 **, < 0.001***

均值 t 检验 T-test

统计课上已经学过

- 类型
  1. 单样本 t 检验 (One-Sample T Test)：单个样本总体均值和我们指定的值是否有显著差异。
  2. 两独立样本 t 检验 (Independent- Samples T Test)：两个互相独立的样本的总体均值是否有显著差异。
     “独立”指两个样本相互独立，即从一个总体中抽取样本，对从另一个总体中抽取样本，没有任何影响（样本容量可以不同）。如：男性和女性的收入。
  3. 两配对样本 t 检验 (Paired-Samples T Test)：两个互相有联系的样本总体的总体均值是否有显著差异。
     “互相有联系”指抽样不独立而有关联：个案前后的两种属性、某个事物的两个侧面。如：服用药品前后的效果。
- 操作
  - 单样本 t 检验 `carsales.sav`
    1. 分析 - 比较均值 - 单样本 t 检验
    2. 检验变量 - 检验值 输入检验数值

3. 在 选项 选择 缺失值 的处理办法（ 分析顺序排除 分析设计变量有缺失值时剔除；列表排除 剔除任一变量上含缺失值变量）
    - 判断： $p < 0.05$ ($p < \alpha, \alpha$ 也可以是 0.01) 即存在显著差异
    - 假定：正态总体、随机样本
- 两独立样本 t 检验 `carsales.sav`
    1. 分析 - 比较均值 - 独立样本 t 检验
    2. 选择 检验变量，检验什么变量
    3. 选择 分组变量，用什么变量分成两类
    4. 定义组，怎么分成两类
    5. 选项 同上
    - 判断：
        1. 判断两个总体方差是否相等：莱文方差等同性检验 p>0.05 看第一行假定等方差；p<0.05 看第二行不假定方差
        2. 判断平均值：$p < 0.05$ ($p < \alpha, \alpha$ 也可以是 0.01) 即存在显著差异
    - 假定：正态总体、随机样本、两样本总体方差相等（方差齐性）
- 两配对样本 t 检验
    1. 分析 - 比较均值 - 配对样本 t 检验
    2. 选择一对或若干 成对变量，检验什么变量
    3. 选项 同上
    - 判断： $p < 0.05$ ($p < \alpha, \alpha$ 也可以是 0.01) 即存在显著差异
    - 假定：正态总体、随机样本、两样本总体方差相等（方差齐性）

- 报告：均值、p 值
- * 单尾 t 检验如何实现？ 大于、小于问题

***非参数检验**

当总体参数不可知的情况下进行推断统计的方法，网上比较散，薛薇书籍

***方差分析 ANOVA**

很重要（可以用于多个组别之间的对比）但暂不涉及，约科奇或薛薇书籍，需要注意方差分析的假定

- 方差分析 ANOVA、协方差分析 ANCOVA、多元方差分析 MANOVA

***聚类分析**

根据数据特征，在没有先验知识的情况下自动分类（机器学习：无监督学习 unsupervised learning）。SPSS 能够完成层次聚类、K-means 聚类

相关与回归 Correlation and regression

- 定距以下变量间相关：交叉表 **Crosstab /** 列联表 **contingency tables** 与相关系数
    - 列联表常用于表示两个变量间的关系
    - 期望：假设无关的得到的预测
- 操作：

1. 分析-描述统计-交叉表
2. 定义行、列变量
3.

- 交叉表的卡方检验：一个分类变量在另一个分类变量上是否存在显著差异，如性别（男性和女性）在工作情况（工作与不工作）上是否存在显著差异
- 皮尔逊卡方 $\chi^2 = \sum(\frac{(观测值-预测值)^2}{预测值})$
  - 自由度 $= (行数-1)*(列数-1)$

- 相关性衡量系数：
  - 定类-定类：Phi（范围从-1 到 1，记作 -1~1 下同；适用于 2*2）,V（对 Phi 的修正，不止适用于 2*2） 列联系数（0~1）, Cramer's
    - 定序-定序：Kendall's Tau-b(-1~1，适用于方形), Kendall's Tau-c(-1~1，任何形式列联表), Gamma(适用于 2*2, 0 ~1 ，0 为独立，1 为很好地相关)
    - 定类-定序：lambda(Goodman and Kruskal tau-y), 0~1, 0 为独立，1 为很好地相关
    - 定类/定序-定距：相关比例 Eta(0-1)
    - *定距（及以上）-定距（及以上）：Pearson
  - 课上没有展示，操作

- 定距以上变量间相关`carsales.sav`
  - 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 以输出离差平方和、样本方差、两变量的叉积离差和协方差。
  - 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

- p 值
  $p < 0.05$ ($p < \alpha, \alpha$ 也可以是 0.01) 即显著（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）

- 报告：

1. 相关系数 / r
2. 显著度 / p 值 / p-value
3. 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(优先于 Spearman 参数检验)；Spearman 相关系数用于两定距/定比变量之间或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

- 线性回归
  - 变量等级：两个均为定距以上变量（*少量定类也可以，构建哑变量 dummy variable）
  - 多元回归方程：

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + + \hat{\beta}_n x_n$$

  - 操作

1. 根据 **理论假设** 确定自变量(independent variable, IV) 和因变量(dependent variable, DV)
2. 建立方程：分析 - 回归 - 线性 - 定义 自变量 和因变量

> *逐步回归的争议？[刘德寰](http://media.people.com.cn/GB/40606/12102091.html：
>
> 传播学定量分析方法与技术的反思。..... 模型的简单与复杂。简单是不是就好？列联表的错误率。逐步回归有对的吗？案例：谁会对广告有较好的容忍度。如果我们用逐步回归的方式做，结果非常简单，哪个显著，哪个不显著，但是有用吗？实际的结果是这样的（PPT），如果我们把它变成图是这样的（PPT），这是简单与复杂的关系吗？不是，它是对与错的关系。复杂的模型真的是很漂亮，这是关于每周读书时间的研究，它的结论一张表得不出来，我们快速看它的结论，能看到人们在阅读过程当中发展的趋势，小学、初中、高中、大专、本科、研究生，当我们找到这个趋势的时候，能够发现原来它是这种规律性，这种规律靠我们用数字去表述是非常非常困难的，我们只能用模型去分析。实际上，简单与复杂不是量的差异，是质的差别，简单的模型可以简单的处理，但是复杂的，如果你没有对问题的真正思考是出不来的。

3. 选择关心的统计量，一般需要勾选回归系数的估计；模型拟合度、共线性诊断、Durbin-Waston，各项含义查阅薛薇书
4. 绘制-Y*ZRESID-X*ZPRED，勾选标准化残差图中直方图和正态概率图
5. 检验拟合效果:
   - 拟合的优度
     - 一元：$$R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}= \frac{\sum_{i=1}^{n}(\hat{y}i-\bar{y})^2}{\sum{i=1}^{n}(y_i-\bar{y})^2}=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}i)^2}{\sum{i=1}^{n}(y_i-\bar{y})^2}$$
     - 多元：调整 R^2 Adjusted $R^2 = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}}$
       - *为什么？书里都会有
     - p < 0.05

4. 预测

- 判断
  - 相关性：自变量和因变量相关的方向和强度
  - 模型摘要：$R^2/(R^2_{adjusted})$ 模型解释了多少
  - 方差分析(ANNOVA)：整个模型是否显著(p<0.05)
  - 系数：就每个相关关系而言，这种关系是否显著(p<0.05)
- 假定
  - 线性关系
  - 独立性：每一个 case /个案/观测值是互相独立的
  - 正态性
  - 方差齐性
  - 无多重共线性：自变量间没有严格线性关系
  - *随机误差与自变量不相关
  - *随机误差服从零均值、同方差的正态分布
  - *因变量 x 在所抽样本中具有变异性，而且随着样本量的增加，因变量 x 的方差趋近于一个非零常数
- 报告：$R^2/(R^2_{adjusted}), \beta, t, p-value$，$df$
- *对每个选项更详细的介绍
- *redundant variables(机器学习)?
  - feature selection：理论假设（前人经验）、前期探索
  - regularization(Lasso, ridge)：通过对方程的损失函数(lost function) 添加惩罚项以减少过度拟合但是对我们没用
- *regression tree(回归树，机器学习)：用一森林的树（回归方程）共同预测
- *OLS(Ordinary Least Square) 与 PLS(Partial Least Square) 回归估算方法
- 更详细的操作视频

- ***逻辑回归 logistic regression**：预测分类的可能性。如果 x1 今天天气, x2 地形, x3 季节, 那么 y 明天会下雨/不下雨
- ***中介分析、调节分析、中介调节分析**：自变量和因变量的关系的发生可能是经由一些变量产生影响(中介反应 mediation effect)或者取决于一些变量的影响（调节反应 moderation effect），SPSS AMOS, process in SPSS, lavaan in R

降维 **Dimension reduction**

- 意义：将多个变量合并，减少计算量、减少变量间相关性、避免之后建模的多重共线性
- **(探索性) 因子分析 (Exploratory) Factor analysis**
  - 前提：原有变量之间存在相关关系才能提取；样本量不能太小，至少为变量数的 5 倍。
  - 操作：
    1. 分析-降维-因子分析-添加变量

2. 描述指定结果：勾选原始分析结果、系数、反映像、KMO 和 Barlett 球形度检验
3. 抽取指定提取因子的方法，一般使用主成分分析法; 分析依据选择相关性矩阵；选择如何确定因子数目（基于特征值、设定因子个数）；在输出中勾选未旋转的因子解、碎石图
4. 旋转中选择因子旋转方法最大方差法，勾选输出旋转解；在方法中指定因子得分计算方法回归法
5. 可选：选项按顺序排列、禁止显示小因子

- 判断：
  1. 变量是否适合因子分析：相关系数高、Barlett 显著($p<0.05$)、KMO(>0.8 适合, >0.7 一般)
  2. 提取因子：解释的总方差高代表效果理想；旋转后的因子载荷矩阵载荷高意味着相关程度高；载荷高归为一个因子
  3. 因子命名：理论
  4. *计算因子得分：每个变量分别如何决定了所得因子
  5. 评价
- 课上来不及演示，请参考视频
- * 验证性因子分析(confirmatory factor analysis)？SPSS AMOS, lavaan in R
- * 商业分析中，联合分析法(conjoint analysis)

（四）绘图 Graphs

条形图 Histogram

图形-旧对话框-简单/柱状/堆积

- *排序？在输出结果中双击该条形图进入图表编辑器-双击那些柱子即可进入属性/类别-选择需要排序的变量-使用上/下箭头排序

***散点图、折线图、饼图、箱线图(小提琴图)**

三、参考资料

- 罗纳德·D·约科奇《SPSS 其实很简单》/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。
- 风笑天《现代社会调查方法》：方法和分析入门，也有 SPSS 具体操作；
- 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；

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
- 谢宇《回归分析》：回归