



① STATA基础

知识点概要

- ◆ STATA简介
- ◆ STATA软件安装
- ◆ STATA统计分析热身小案例
- ◆ STATA中的不同文件
- ◆ STATA的帮助文档
- ◆ STATA的内置数据集
- ◆ STATA数据预处理与数据管理（重点）
- ◆ STATA描述统计与统计制表(重点)
- ◆ STATA统计制图(重点)

- stata是用于windows, Mac, Unix系统上的一种功能完备的统计软件包（跨平台）；
- stata容易操作，运算速度快，包括一整套预先编好的分析与数据管理功能（基于内存计算）；
- stata也允许用户根据需要来创建自己的程序添加更多的功能（可使用第三方插件）；
- 大部分操作既可以通过下拉菜单系统来完成，也可以更直接的通过键入命令来完成（菜单操作+程序命令）；
- status命令有很强的一致性和直观意义，可以使有经验的用户更高效的工作（程序命令一般为相应专业术语名次）；
- stata对更复杂或需要多次重复执行的任务进行编程十分容易（使用do文件记录分析处理过程）；
- 在用stata时，还可以混用菜单方法和命令方法（用菜单操作，用程序记录）；

- ▣ **Windows安装**

- ▣ 学术用途长期使用方法：通过第三方程序，修改针对stata程序运行时的系统时间，系统时间和其它软件不受影响。

- ▣ **macOS安装**

- ▣ 学术用途长期使用方法：使用网络上分享的序列号和注册文件即可使用。

- 案例数据来自于《应用STATA做统计分析》，是关于气候变化的相关数据。数据集名称为 [Arctic9.dta](#)
 - 数据记录了1979年~2011年间9月份北冰洋的冰情观测。
- 分析任务
 - 从指定目录下导入stata数据文件，知道如何设置和查看工作目录
 - 查看数据集的基本描述信息
 - 查看数据集的内容
 - 计算变量的描述性统计量
 - 计算变量之间的相关系数
 - 绘制多个变量之间两两相关矩阵散点图
 - 绘制时间序列变量的时间序列图

▣ 数据文件

- ▣ 数据文件是用来存储数据的文件，stata有自己的数据文件格式，.dta后缀
- ▣ 支持导入多种其它格式的文件，如.xls .xlsx .csv

▣ do文件

- ▣ do文件中的内容是Stata命令的集合，方便一次性执行多条stata命令；
- ▣ do文件使数据分析工作能够自动重复；
- ▣ （建议）在实际的统计分析工作中，将命令保存到do文件中；

▣ log文件

- ▣ log文件可以记录分析过程。
- ▣ 打开日志文件进行记录，① 菜单：File→Log → Begin ② 命令：log using log_file_name.
- ▣ log文件有自己默认的格式 .smcl，也可以转换成文本格式的 .log文件

- ▣ **软件自带PDF帮助文档**

- ▣ STATA自带帮助文档内容相当丰富，多达25+个PDF文档。
- ▣ 可在软件的安装目录下找到它们；也可以通过菜单 `help` → `PDF documentation`查看。

- ▣ **使用help命令**

- ▣ 在使用某个具体的命令时，可以通过`help` + 命令名称查看，比如：`help summarize` 可用于查看 `summarize` 命令的帮助信息。

- ▣ **随STATA一起安装的数据集**

- ▣ 使用这些数据，电脑不需要连接网络，它们随软件一起安装到本地。
- ▣ 点击File → Example Datasets可查看这些数据。

- ▣ **STATA提供的可以联网在线下载的数据集**

- ▣ 同样在File → Example Datasets下访问这些数据。但电脑需要连接网络。

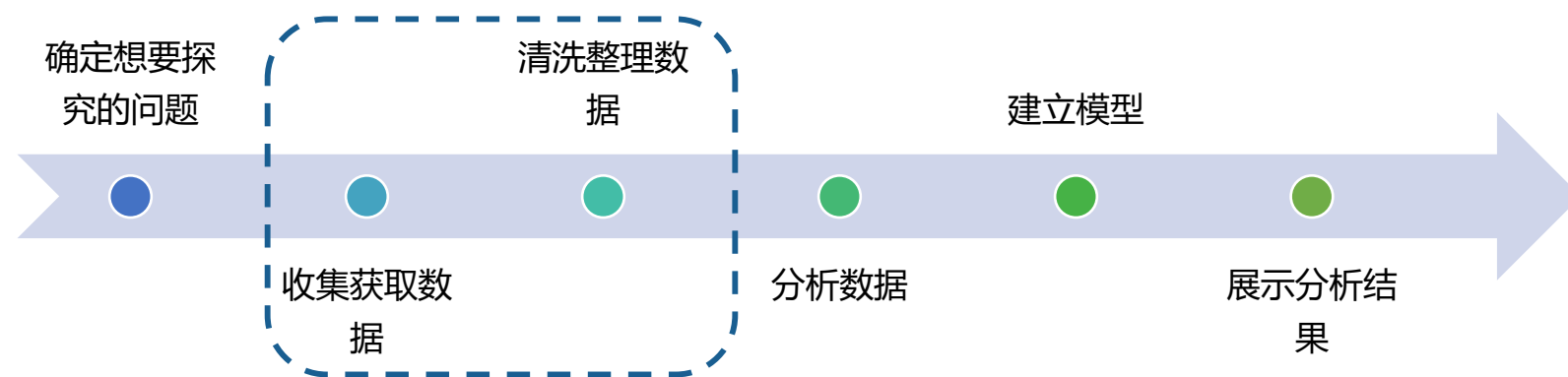
- ▣ **内置数据集的用处**

- ▣ 用于练习各种数据分析方法。
- ▣ 官方帮助文档，案例演示，一般都是用stata内置数据集，可方便的照着进行练习。

内容包括：

- 录入和导入数据
- 存储与导出数据
- 查看数据资料
- 基本统计量
- 定义数据的子集
- 创建和替换已有变量
- 缺失值编码
- 使用函数
- 分类变量编码
- 合并stata数据文件
- 分类汇总
- 重组数据结构
- 生成随机数据和随机样本
- 编制数据管理程序

数据分析的一般流程



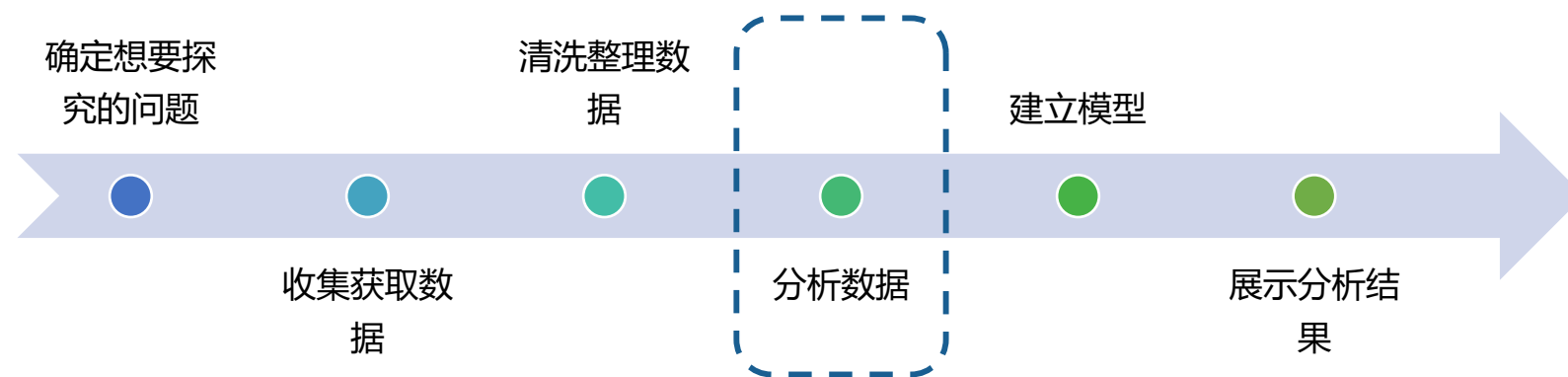
数据预处理的重要性：

- 数据分析，无论使用哪种软件，工具或者编程语言，数据的预处理过程都是极其重要的一个环节。极端情况下，90%以上的工作都在做数据预处理。一般情况下也有60%左右的时间在做数据清洗工作。

内容包括:

- 数值型变量的描述性统计
- 探索性数据分析
- 正态性检验和数据转换
- 频数表和交叉表
- 多个频数表和多维交叉表
- 频数加权

数据分析的一般流程



建模之前对数据进行探索分析的意义:

- 得到一些基本的描述性统计量, 以描述样本的统计学特征, 如均值, 中位数, 频数, 标准差, 偏度等;
- 探索数据的特征, 检查是否满足建立特定模型的要求;

内容包括:

- 直方图
- 散点图
- 箱线图
- 柱形图
- 折线图
- 面积图
- 绘制拟合回归直线

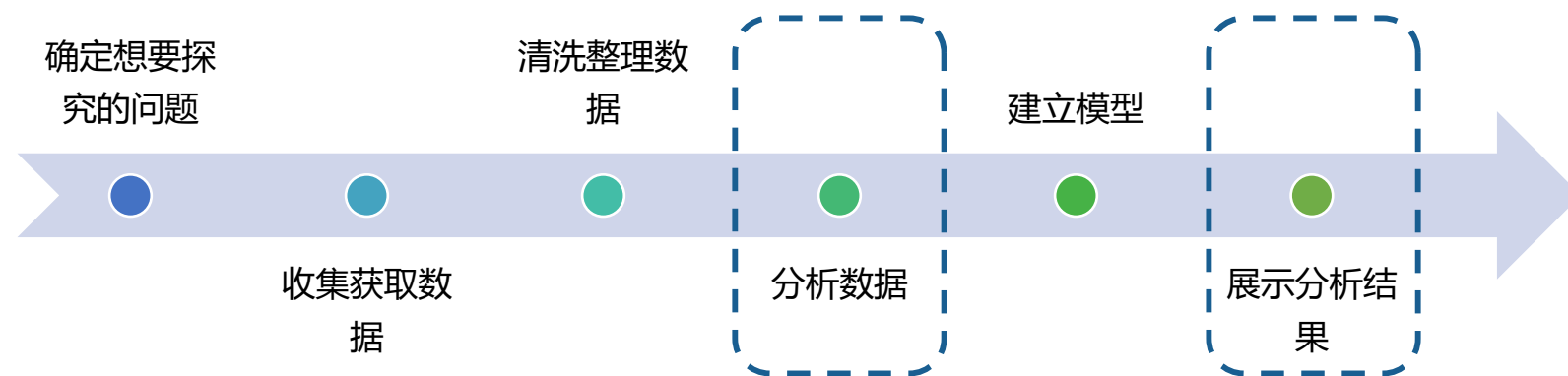
STATA统计制图官方案例:

- [stata-graphs](#)
- [example-graphs](#)

借助官方文档绘制图表:

- 绘制dot-charts
- 访问在线文档 (直接, 但速度慢)
- 访问本地文档 (需要自己找)

数据分析的一般流程



图表分析在数据分析/统计分析中的作用:

- 对数据进行探索分析, 如: 用直方图和箱线图了解数据的分布; 用散点图观察变量之间的相关性;
- 将统计指标与图形结合起来看, 如: 平均值/直方图, 相关系数/散点图。
- 将数据中的模式, 隐含的信息直观展示出来, 将数据分析结论以图表的形式展现出来;
- 注意, “一图胜千言”, 但 “图表也是具有欺骗性的”