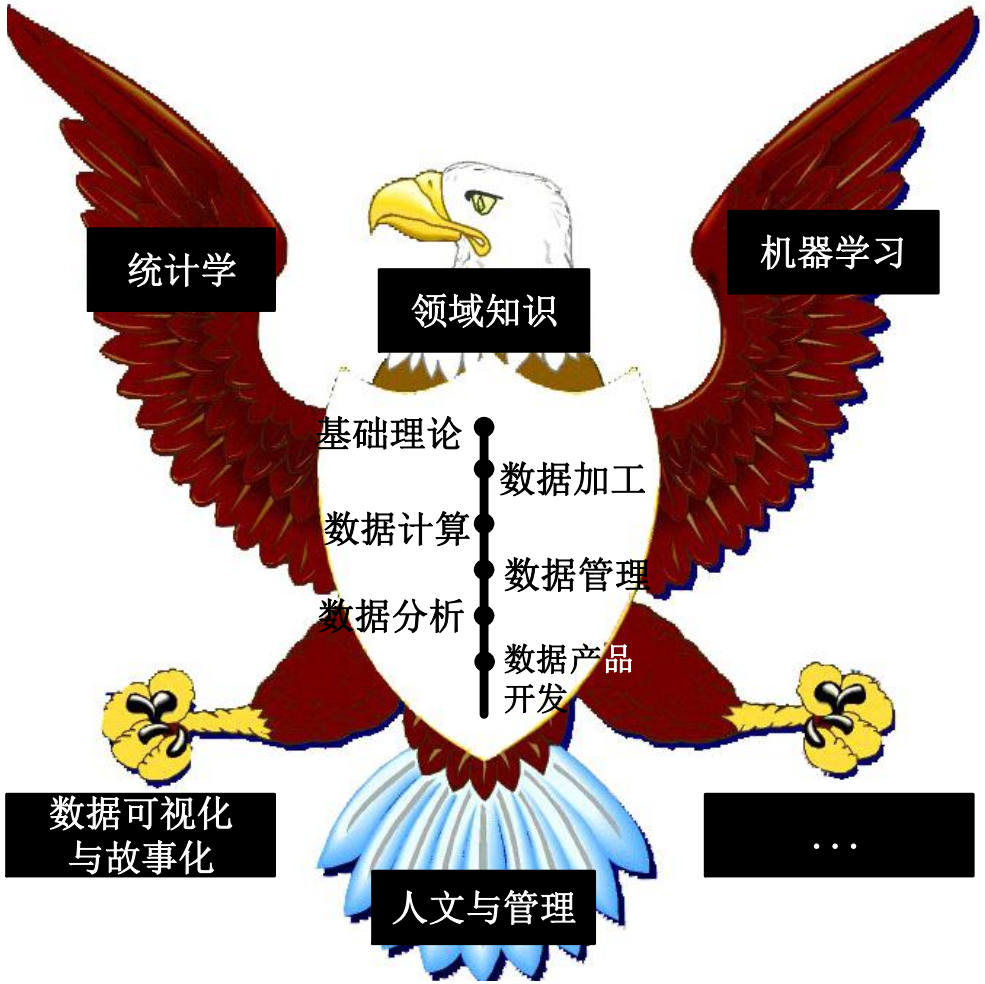
第2章 统计学与模型

编写思路

学习方法、要求及建议

疑难知识点的解读

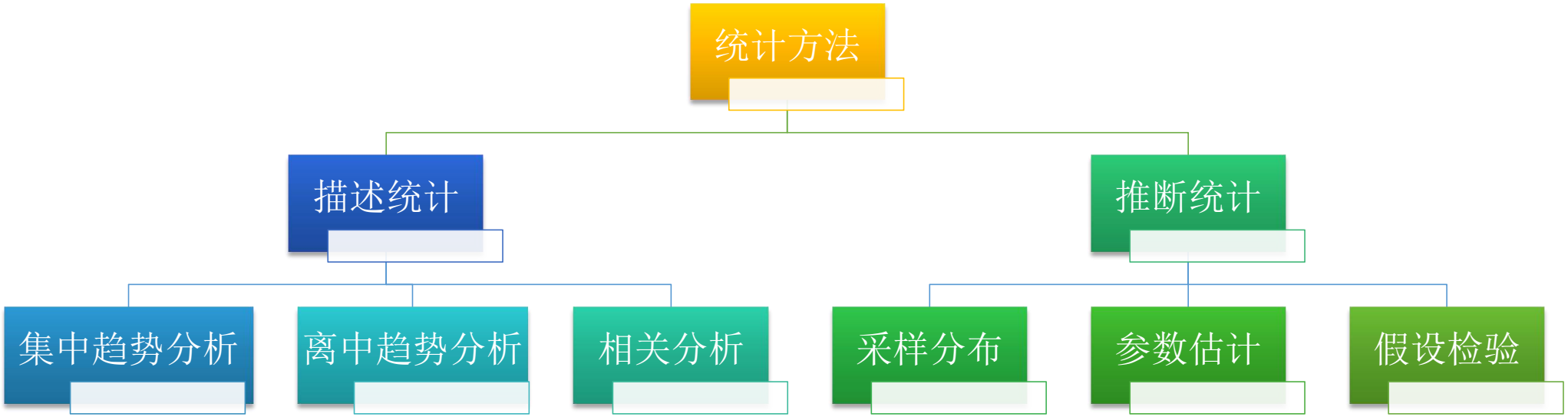
1.本章定位与内容简介



2.本章学习提示及要求

了解	理解	掌握	熟练掌握
• 统计学与数据科学的区别与联系 • 大数据环境下统计学面临的主要挑战	• 数据科学中应用统计学知识的基本步骤 • 统计学方法的类型及选择方法	• 面向统计学的数据划分及准备方法 • 统计学中对模型的评估方法	• 基于Python的统计学编程实践

3.统计方法的基础知识



注：Fisher提出的“推断统计”的三个中心

相关关系分析

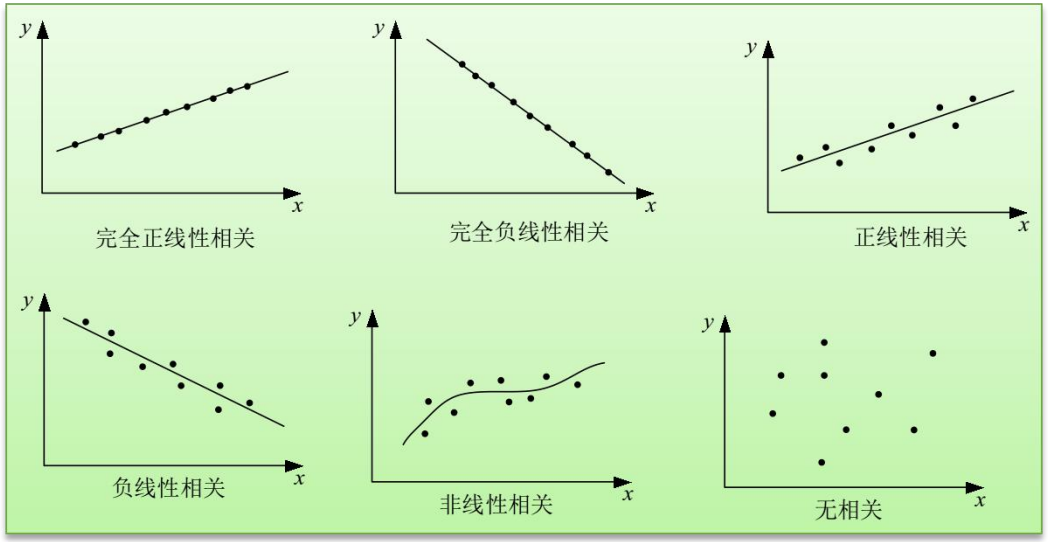
完全负相关　　无相关　　完全正相关

-1.0　　-0.5　　0　　0.5　　-1.0

r

负相关程度增加　　正相关程度增加

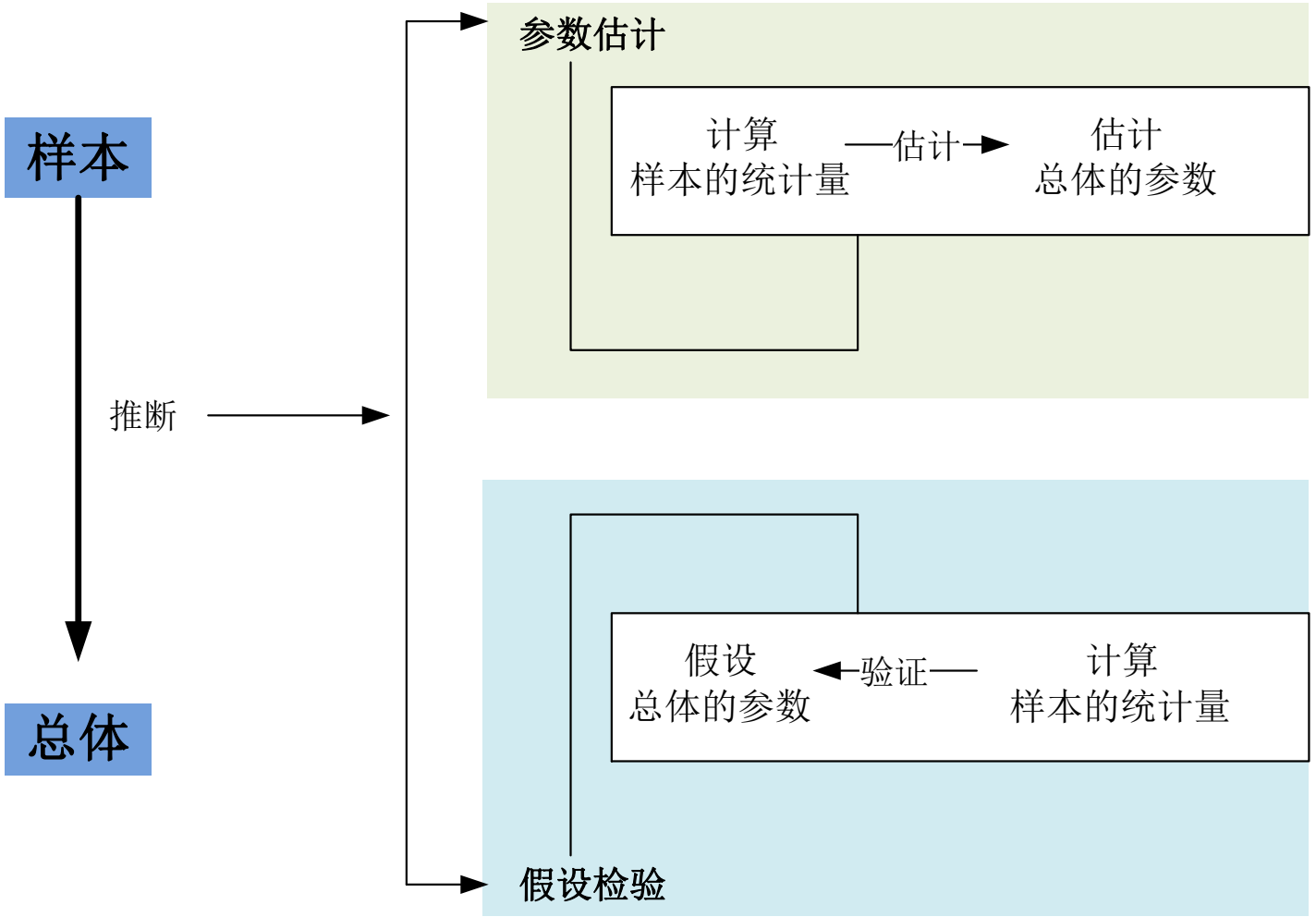
总体相关系数（ρ）
样本相关系数（r）



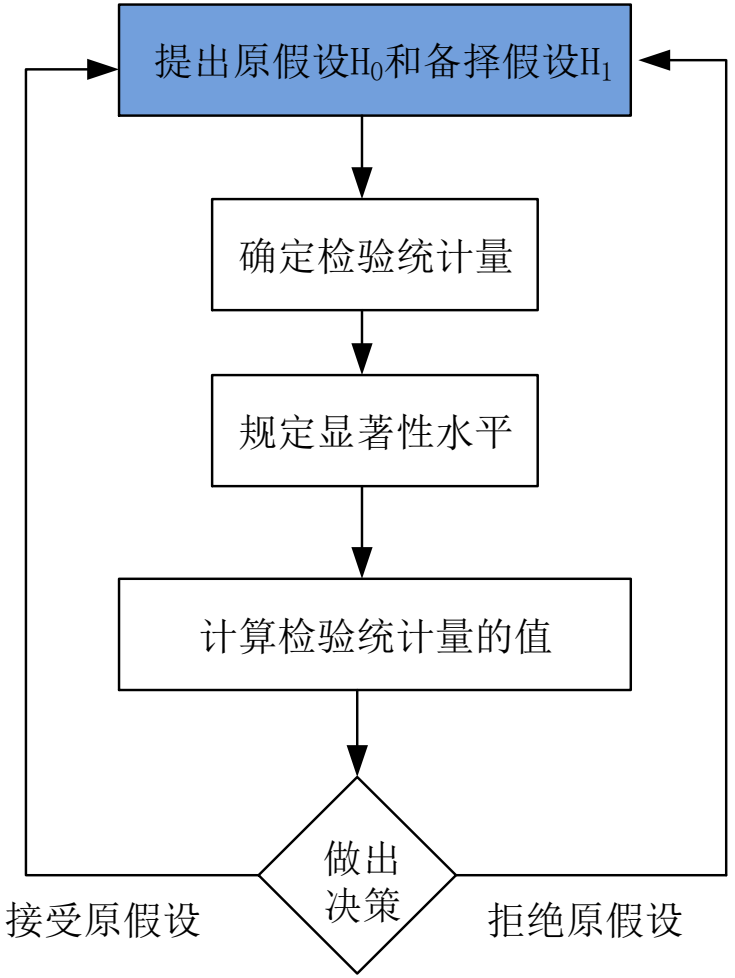
$$r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\cdot\sum(y-\bar{y})^2}}$$

$$t=|r|\frac{\sqrt{n-2}}{\sqrt{1-r^2}}\sim t(n-2)$$

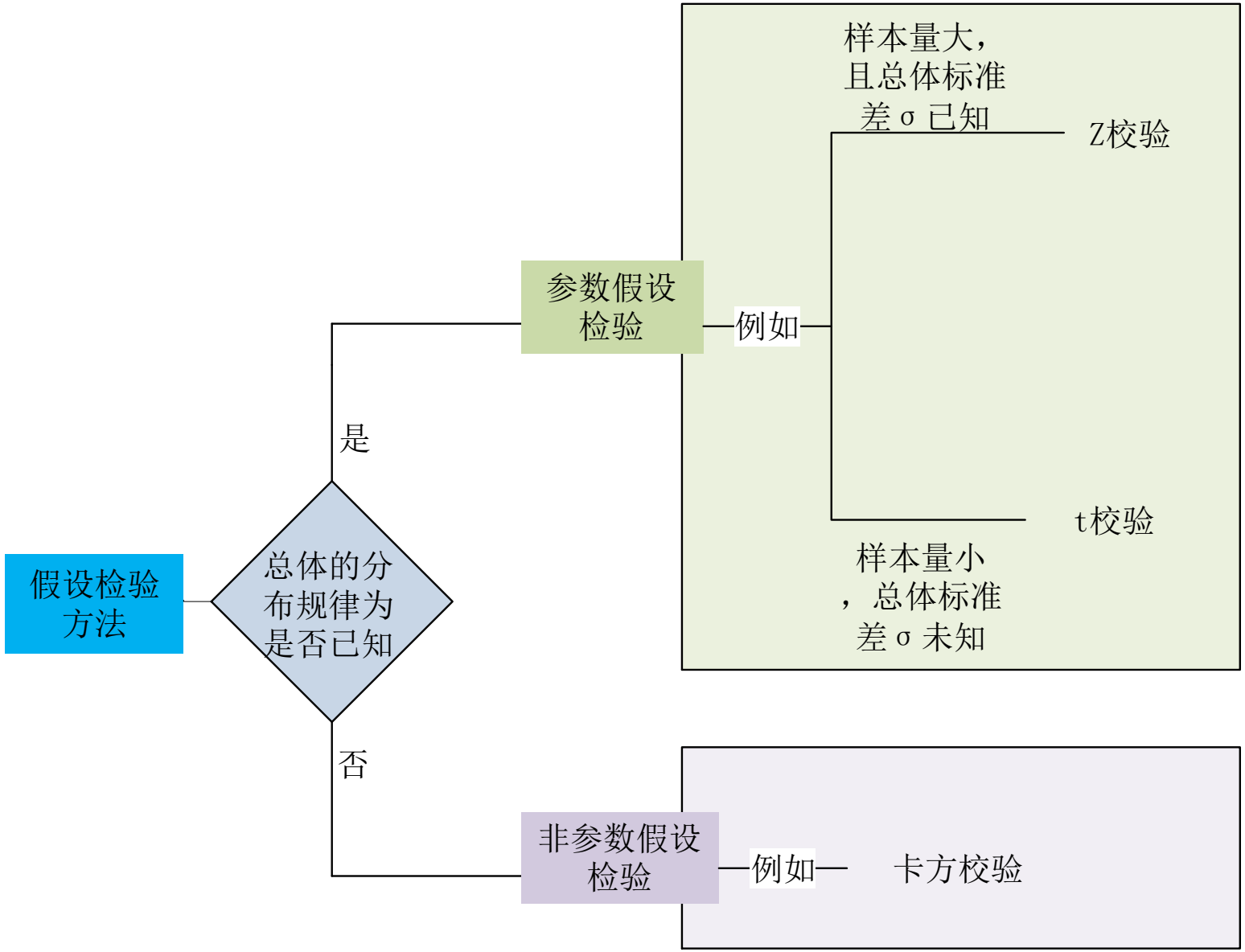
统计推断



假设检验



项目	没有拒绝原假设H_0	拒绝原假设H_0
H_0为真	1-α（正确决策）	α（弃真错误）
H_0为假	β（取伪错误）	1-β（正确决策）


假设检验方法
总体的分布规律为是否已知
是
否
参数假设检验
例如
样本量大，且总体标准差σ已知
Z校验
样本量小，总体标准差σ未知
t校验
非参数假设检验
例如
卡方校验

4.Python数据科学中常用的包

基础库

- **Pandas，Numpy，Scipy**

绘图及可视化

- **Matplotlib，Seaborn，Bokeh，Basemap，Plotly，NetworkX**

机器学习

- **SciKit-Learn，TensorFlow，Theano，Keras**

统计建模

- **Statsmodels**

自然语言处理、数据挖掘及其他

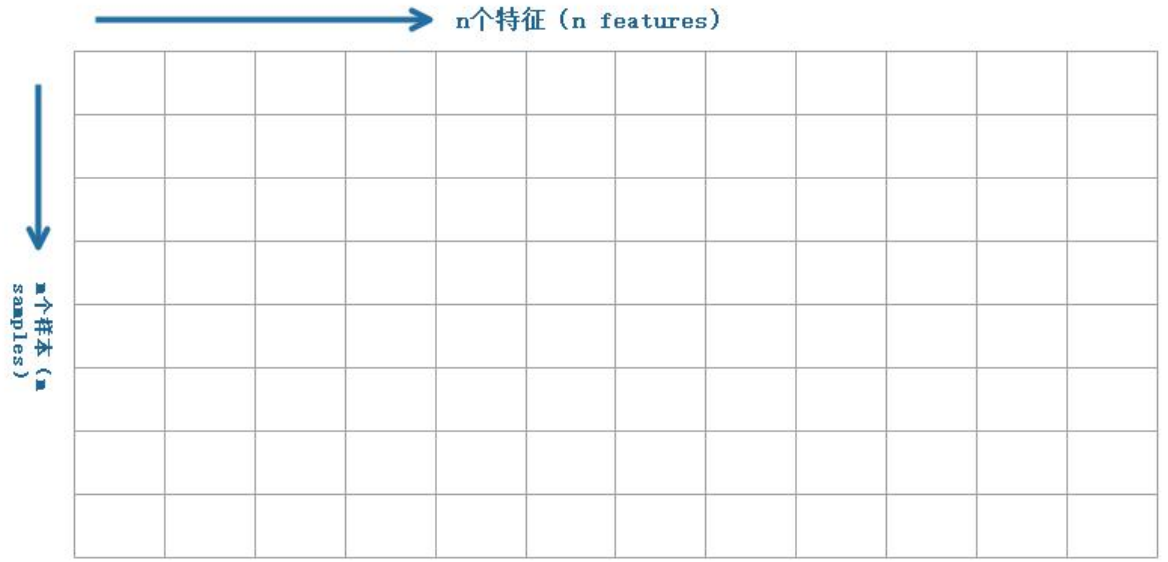
- **NLTK，Gensim，Scrapy，Pattern**

数据科学领域常用的Python库

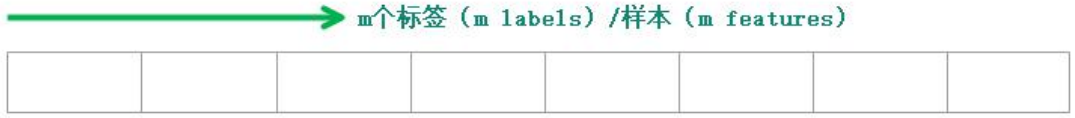
5.统计分析中的数据加工

y=f（X）

X为特征矩阵（Feature_Matrix）



y为目标向量（Target Vector）



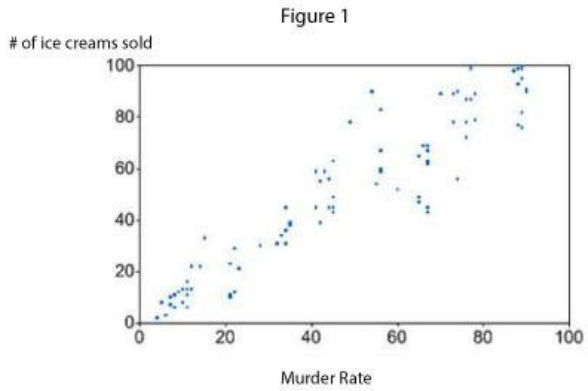
6.统计学面临的挑战与趋势

Mayer-Schönberger V, Cukier K. Big data: A revolution that will transform how we live, work, and think[M]. Houghton Mifflin Harcourt, 2013.

样本=总体
效率>精准度
相关关系>因果关系

传统思维	大数据思维
随机样本	全体数据
精确性	混杂性
因果关系	相关关系

验证性数据分析 *VS* 探索性数据分析
基本分析 *VS* 元分析



7.数据科学中常用模型

（1）广义线性模型（是多数监督机器学习方法的基础，如采采回归和Tweedie回归）

（2）时间序列方法（ARIMA、SSA、基于机器学习的方法）

（3）结构方程建模（针对潜变量之间关系进行建模）

（4）因子分析（调查设计和验证的探索型分析）

（5）功效分析/试验设计（特别是基于仿真的试验设计，以避免分析过度）

（6）非参数检验（MCMC 等）

（7）k 均值聚类

（8）贝叶斯方法（朴素贝叶斯、贝叶斯模型平均、贝叶斯确应性试验等）

（9）惩罚性回归模型（弹性网络、Lasso、LARS 等）以及对通用模型（SVM、XGBoost 等）加罚分

（10）样条模型（MARS 等），主要用于流程建模

（11）马尔可夫链和随机过程（时间序列建模和预测建模的替代方法）

（12）缺失数据插补方法及其假设（missForest、MICE 等）

（13）生存分析（主要特点是考虑了每个观测出现某一结果的时间长短）

（14）混合建模

（15）统计推断和组群测试（A/B 测试以及用于营销活动的更复杂的方法）

8.如何继续学习本章知识

1.综合应用能力

- 在数据科学项目中，统计学方法往往与其他方法综合应用
- 例如：数据挖掘

2. 统计学基本功

- 统计学基础知识
- 数据科学中常用统计学模型

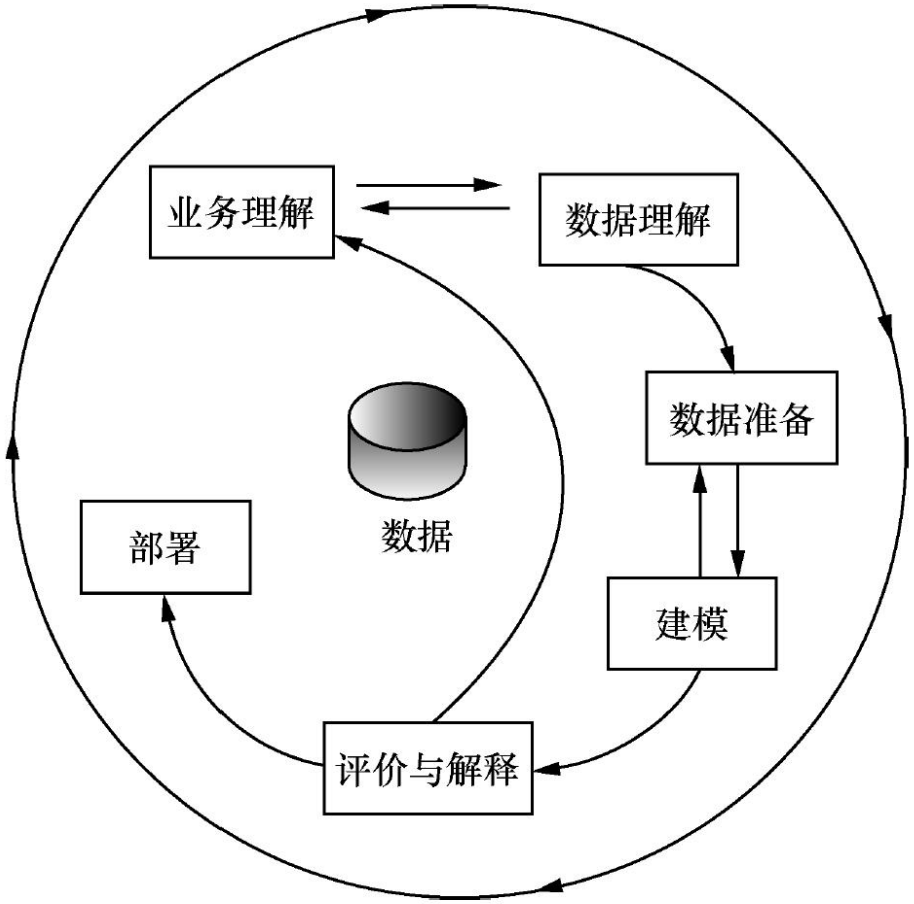


图 2-21　CRISP-DM 数据挖掘各阶段

小结

1.本章定位与内容简介

2.本章学习提示及要求

3.统计方法的基础知识

4.Python数据科学中常用的包

5.统计分析中的数据加工

6.统计学面临的挑战与趋势

7.数据科学中常用模型

8.如何继续学习本章知识