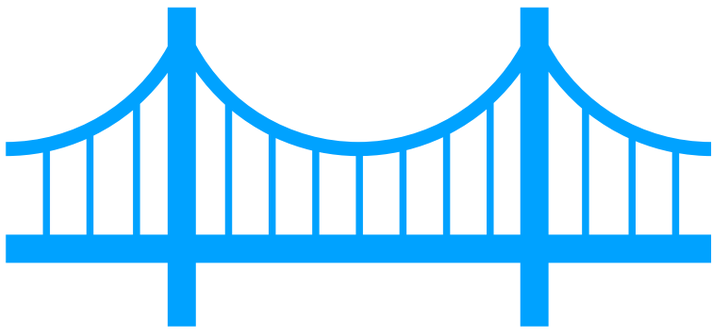
定量分析：数据思维与商业统计

陈文波
cwb@whu.edu.cn
2021年10月

课程梗概

描述性分析与EDA

- 数值方法：三类指标
- 图形方法：五种图形



- 抽样分布、CLT
- 正态分布
- T分布
- 卡方分布
- F分布

推断性分析

- 区间估计
- 假设检验
- 应用：
  - 分类 vs. 分类
  - 分类 vs. 数值
  - 数值 vs. 数值

R

课程梗概

- Session 1：内容简介、基本概念、描述性统计1
- Session 2：描述性统计2
- Session 3：抽样、中心极限定理、正态分布
- Session 4：其他三个分布，区间估计
- Session 5：假设检验，两个总体均值和比例的推断
- Session 6：总体方差和方差比的推断、分类数据 Vs. 分类数据： 卡方检验与列联表分析
- Session 7：分类数据 Vs. 数值型数据： 方差分析；数值型数据：回归分析
- Session 8：课程总结

描述性统计与EDA

- 三类指标
- 五种图形

分清类型、三类指标、五种图形

三类指标

- 集中趋势
- 离散趋势
- 形状

集中趋势指标

- Mean: 均值
- Median：中位数
- Mode：分位数
- Quartile：四分位数
- Percentile：百分位数

集中趋势指标：计算

- mode: 众数
  - 针对分类数据

集中趋势指标：计算

- median: 中位数
  - 有序分类及以上的数据

集中趋势指标：计算

- Mean: 均值
  - 几何平均数
  - 定基、环比
  - 加权平均数

五类图形

- Bar graph
- Histogram
- Linegraph
- Boxplot
- Scatterplot

讨论

- 统计不关心个别点（一定意义上）
- 计算、画图的目的不是为了为计算、为画图而画图
- 理解数据的规律

附加项

数据分析报告的写作

陈文波

2021年10月

基本原则

面向	最终用户	自己
强调	产品思维	报告思维
重点	结论第一、清晰、美观	可复现
数字时代	结构化、自动化（中等）	结构化、自动化（强）
工具	Word，pages	支持markdown的工具

主要框架：数据产品视角

- 第一部分：主要结论
- 第二部分：数据收集、总体介绍
- 第三部分：重要的分结论
- 第四部分：未来的发展建议
- 最后部分：致谢、附录等

主要框架：数据分析报告

- 第一部分：主要结论
- 第二部分：载入、数据清洗
- 第四部分：数据EDA
- 第五部分：模型
- 第六部分：结论

Wrap-up

- 描述性统计与EDA
- 三类指标
- 五种图形

分清类型、三类指标、五种图形

课堂练习-1

- 学习通资料里，Data 目录下：WE.xlsx数据集
- Description:
- WE 公司是一家较为成功的互联网公司，其主要业务是通过订阅服务帮助中小企业管理线上的业绩表现。目前管理层意识到需要对一些关键的业务流程进行更深入的分析，客户留存率就是其中一个很重要的方面。WE与客户签订的是有固定期限的合约，期限为单月、半年或一年。
- 要求：利用你现有的数据分析工具（R，Excel等），对于WE公司的客户流失数据进行探索，根据WE公司的数据思考：流失率和客户生命周期是否相关？如何预测用户流失？

课堂练习-2

- 附件`lj_sh_2019.Rdata`数据是来自某房地产网站2019年6月份二手房的10%的抽样数据。各变量名定义与说明如下：
  - line: 临近的地铁线路
  - station: 临近的地铁站
  - property_name: 小区名称
  - bedrooms: 房间个数
  - livingrooms: 厅的个数
  - building_area: 建筑面积
  - direction1： 主要朝向
  - direction2: 次要朝向
  - decoration：装修程度
  - has_elevator: 是否有电梯，0为没有，1为有
  - hml: 位于一栋楼的高中低区
  - building_height: 建筑总层数
  - building_year: 建筑年代
  - building_style: 建筑风格
  - Building_location: 所处板块
  - price_sqm: 每平米价格（元）
  - price_ttl: 总价（万元）
  - house: 是否是别墅
  - house2: 是否是别墅

课堂练习

- 请使用你掌握的数据分析方法，对以上数据集进行探索性数据分析（描述性方法为主）。可以讨论的问题包括（但不限于）：价格特点、年代特点、区位特点；价格与年代、年代与房产供应，等等。希望大家发现更有趣的现象。
- 要求：
  - 使用markdown工具（例如Rmarkdown, Typora)撰写数据分析报告。
  - 要有完整的形式：
    - 主要结论
    - 数据分析思路与方法
    - 分析过程