ZOE's MindMap
https://wozielf.github.io/
描述性统计：数值方法

总结 → 数值方法

单变量

- 平均数
  - 算数平均数
  - 调整平均数
  - 加权平均数
  - 几何平均数
- 中心位置
  - 中位数 & 众数
  - 分位数
    - 百分位数
    - 四分位数
- 变异程度
  - 极差 & 四分位数间距
  - 方差 & 标准差
  - 标准差（变异）系数
  - 平均绝对偏差
- 分布形态
  - 偏度
  - 峰度
- 相对位置（z 分数）
  - 切比雪夫定理
  - 经验法则（延伸）
  - 异常值

双变量

- 线性关系
  - 协方差
  - 相关系数

度量值

- 来自
  - 样本 → 样本统计量
  - 总体 → 总体参数
- 用来汇总
  - 样本度量结果
  - 总体度量结果
- 「样本统计量」是相应「总体参数」的「点估计量」

中心位置的度量

数据「集中趋势」的度量

平均数

- 平均数（算数平均数）：$\bar{x} = \dfrac{\sum x_i}{n}$ $\quad \mu = \dfrac{\sum x_i}{N}$
  - 调整平均数 → p% 调整平均数 → 删除 p% 的最大、小值后计算「均值」
- 加权平均数：$\bar{x} = \dfrac{\sum w_i x_i}{\sum w_i}$
- ★ 几何平均数：$\bar{x}_g = \sqrt[n]{(x_1)(x_2)\cdots(x_n)}$ → n 个数值「乘积」的「n 次方根」
  - 应用 → 确定几个「连续时期」的「平均变化率」→ ★ 乘法过程
  - e.g. → 分析财务数据的「增长率」

中位数 median → 包含「异常值」时，比均值更合适

众数 mode → 可能有多个，但「多众数」无意义

分位数

- 百分数
  - 定义 → 第 p 百分位数 → 至少有 p% 的观测值 <= 该值 且 至少有 (1 − p%) 的观测值 >= 该值
  - 计算
    - 数据按「升序」排列
    - $i = \left(\dfrac{p}{100}\right)n$ → p 为百分位数 / n 为观测值个数
    - 不是整数 → 向上取整（大于 i 的下一个整数）
    - 是整数 → 第 i 和 (i+1) 项的平均值
  - 常用
    - 四分位数 → 第25、50、75百分位数
    - 五分位数 → 第20、40、60、80百分位数
    - 十分位数 → 第10、20 … 90百分位数
    - p.s. → 第50百分位数 = = 中位数
- 四分位数
  - Q1, Q2, Q3
  - 第一、二、三四分位数 即 → 第25、50、75百分位数

变异程度的度量

数据「离散程度」的度量

- 极差 range → max − min
- 四分位数间距 IQR → $IQR = Q_3 - Q_1$ → 中间 50% 数据的极差
- 方差
  - 「离差」平方求和的均值
    - 总体方差：$\sigma^2 = \dfrac{\sum(x_i - \mu)^2}{N}$
    - 样本方差：$s^2 = \dfrac{\sum(x_i - \bar{x})^2}{n-1}$
  - ★ 样本方差 → 总体方差 → 无偏估计（点估计） → 除以 (n−1) → 总体方差总是会比样本方差「大一点」
  - 替代计算方法：$s^2 = \dfrac{\sum x_i^2 - n\bar{x}^2}{n-1}$
- 标准差
  - 方差的正平方根：$s = \sqrt{s^2}$ $\quad \sigma = \sqrt{\sigma^2}$
  - ★ 与原始数据的「单位量纲」相同 → 易于比较和解释 → 更常用
- ★ 标准差系数（变异系数）：$\left(\dfrac{标准差}{平均数} \times 100\right)\%$ → 对「变异程度」的「相对度量」→ 比较 → 具有「不同标准差」和「不同均值」的 → 变量的「变异程度」
- ★ 平均绝对偏差 MAE：$MAE = \dfrac{\sum |x_i - \bar{x}|}{n}$ → 用于时间序列

分布形态的度量

偏度 Skewness

- 度量 → 对称性 → 偏斜方向和程度
- 意义 → vs 正态分布
  - = 0 → 平均数 = 中位数
  - > 0 → 正偏或右偏 → 平均数 > 中位数 → 长尾巴拖在右边（有较多极端值）
  - < 0 → 负偏或左偏 → 平均数 < 中位数 → 长尾巴拖在左边（有较多极端值）
- 「偏度的绝对值」数值越大 → 「分布形态的偏斜程度」越大

峰度 Kurtosis

- 度量 → 陡缓程度
- 意义 → vs 正态分布
  - = 0 → 陡缓程度相同
  - > 0 → 更陡峭（尖顶峰）
  - < 0 → 更平缓（平顶峰）
- 「峰度的绝对值」数值越大 → 陡缓程度与「正态分布」的差异程度越大

相对位置的度量

★ z-score：$z_i = \dfrac{x_i - \bar{x}}{s}$

- 与「均值」的距离 → 度量 → 以「标准差」的个数
- 相同的「相对位置」→ 具有 → z 分数相同
- 又称「标准化数值」

延伸

切比雪夫定理

- 定义
  - 无论「数据分布」的形态 → 适用于「任何」数据集
  - 在「z 个标准差内」的数据 → 与「均值」的距离
  - $(1 - 1/z^2)$ → 「所占比例」至少为
  - z：> 1 / 实数 / 不一定是整数
- 应用
  - >=75% 的数据与均值的距离在 z=2 个标准差以内
  - >=89% 的数据与均值的距离在 z=3 个标准差以内
  - >=94% 的数据与均值的距离在 z=4 个标准差以内

经验法则

- 适用于「钟形/正态分布」数据
- 应用
  - 大约 68% 的数据与均值的距离在「1个标准差」以内
  - 大约 95% 的数据与均值的距离在「2个标准差」以内
  - ★ 「几乎所有」的数据与均值的距离在「3个标准差」以内

异常值

- z-score → < −3 or > 3
- 上下限 → < 下限 or > 上限
  - Lower Limit = $Q_1 - 1.5(IQR)$
  - Upper Limit = $Q_3 + 1.5(IQR)$

★ 两变量间「线性关系」的度量

协方差

- 定义
  - 样本协方差：$s_{xy} = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$ → 样本容量 n
  - 总体协方差：$\sigma_{xy} = \dfrac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$ → 总体容量 N
- 解释
  - 正相关 → 正值
  - 无线性相关 → 零
  - 负相关 → 负值
  - 线性关系越强 → 绝对值越大
- ★ 缺点
  - 结果受异常明显 → 不同「计量单位」→ 采用
  - 有偏差 → 用「数值」衡量「线性关系强度」
  - 改用「相关系数」→ 为规避这种缺陷

★ 相关系数

又称"皮尔积矩相关系数"

- 定义
  - 样本相关系数：$r_{xy} = \dfrac{s_{xy}}{s_x s_y}$ → 「样本相关系数」是对「总体相关系数」的「点估计量」
  - 总体相关系数：$\rho_{xy} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$ → 将「样本统计量」全部替换为「总体参数」即可
- 解释
  - 范围 → −1 到 +1
  - 线性关系越强 → 越接近 0
- ★ 注意 → 度量的是「线性关系」而不是「因果关系」

探索性数据分析

五数概括法

- 最小值
- 第一四分位数（Q1）
- 中位数（Q2）
- 第三四分位数（Q3）
- 最大值

箱形图

- 基于 → 五数概括法 → Q1, Q3, Q2
- 步骤 → 确定「上、下限」
  - Lower Limit = $Q_1 - 1.5(IQR)$
  - Upper Limit = $Q_3 + 1.5(IQR)$
- 延续线 / 触须线 → 「上下限以内」的最大、小值
- 异常值 → 用 * 标出 → 上下限以外