# 统 计 学 Statistics

陈 灯 塔

Econ, XJTUCC

CDT.WISE@G

2021 年 4 月 28 日

# 内容梗概

# 为谁平均

张家有财一千万
九个邻居穷光蛋
若把平均算一算
每家每户一百万

评论：平均给谁？“被平均”

需要完整的分布，没有人家财产为均值，均值只是一个概要统计量

现代观点：把分布 (核密度估计, kernel density) 的图形给出，再描述这些特征 (位置、离散度，形态)，不是更好吗？

# 平均数

集中趋势: 一组数据向某一中心值靠拢的倾向 ✗

不恰切，只是代表中心点的位置，并没有什么趋势。　　寻找数据一般水平的代表值或中心值

average 平均数　　In statistics the word “average” is used in three different senses: mean, median, and mode. 适合不同类型的数据

# 众数和中位数

# 众数

$M_o$ mode, 一组数据中出现次数最多的变量值 (可能不是数，但可以数字化): 例子 p39–40, E3-1 to E3-4

- 可能不存在，可能多个
- 不受极端值影响
- 分组数据，先确定众数组

$$M_o = X_L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot d$$

其中 $d$ 表示所在组的组距，$\Delta_1$, $\Delta_2$ 为众数组与前后组的次数之差。等价地

$$M_o = X_U - \frac{\Delta_2}{\Delta_1 + \Delta_2} \cdot d$$

# 中位数

$M_e$ median, 排序后居于中间位置的标志值: 例子 p41–2, E3-5 to E3-6

- 位置: $1+\frac{n-1}{2}=\frac{n+1}{2}$, 分奇偶 (非整数时插值)

$$M_e=\begin{cases} x_{\frac{n+1}{2}}=x_{l+1} & n=2l+1 \\ \frac{1}{2}(x_{\frac{n}{2}}+x_{\frac{n}{2}+1})=\frac{1}{2}(x_l+x_{l+1}) & n=2l \end{cases}$$

- 组距数列: 中位数位置近似取为 $\frac{1}{2}\sum f_i$, 所在的组为 $m$. 正向 (向上) 累计

$$M_e=X_L+\frac{\frac{1}{2}\sum f_i-S_{m-1}}{f_m}\cdot d$$

反向 (向下) 累计

$$M_e=X_U-\frac{\frac{1}{2}\sum f_i-S_{m+1}}{f_m}\cdot d$$

提示: 中位数的位置取通常的 $\frac{1}{2}(1+\sum f_i)$, 公式变复杂些, 仍然只是估算而已

问题: 如果刚好落在两个组之间, 如何计算? 例如 E3-6 中 70~80 的人数改为 7 人, 则 $\frac{1}{2}\sum f_i=20.5$ 直接使用组限 (反正都是近似计算, 用简单的方法)

# 均值

# 算术均值

样本均值

$$\bar{x} = \frac{\text{变量值总和}}{\text{个体数目}} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

加权平均: 某些值重复出现多次，先举例子

设各组的组中值为: $x_1, x_2, \cdots, x_k$, 相应的频数为: $f_1, f_2, \cdots, f_k$, 且 $\sum_{j=1}^{k} f_j = n$

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_k f_k}{f_1 + f_2 + \cdots + f_k} = \frac{1}{n}\sum_{i=1}^{k} x_i f_i$$

记 $\frac{f_i}{\sum f_j} = w_i$ 为相对比例 (比重、百分比、权重)

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \sum x_i \frac{f_i}{\sum f_j} = \sum x_i w_i$$

提醒：组距分组数据要计算组中值, 本身已经是近似

建议：分组数据提供更详细信息，如各组的均值和方差，或者分组数据提供全样本的均值、中位数等信息。当然，最好是使用更原始数据。

# 调和均值

例 3-11: 蔬菜价格早、中和晚市单价为 0.5, 0.4, 0.2 元/kg, 若早、中和晚市各买 1 元钱，均价是多少?

$$\text{均价} = \frac{\text{总金额}}{\text{总质量}} = \frac{1+1+1}{\frac{1}{0.5}+\frac{1}{0.4}+\frac{1}{0.2}} = \frac{3}{9.5} = 0.315\,79 = 0.32$$

调和均值没有必要单独记忆: 各组的合计 $m_i = x_i f_i$

$$\bar{x}_H = \bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{\sum m_i}{\sum \frac{m_i}{x_i}}$$

调和均值与算术均值本质上一致，区别在于频数已知还是待求解

# 几何均值

$n$ 个变量连乘积的 $n$ 次方根

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdots x_n} = \sqrt[n]{\prod x_i}$$

加权几何均值

$$\bar{x}_G = \sqrt[f_1+f_2+\cdots+f_n]{x_1^{f_1} \cdot x_2^{f_2} \cdots x_n^{f_n}} = \sqrt[\sum f_i]{\prod x_i^{f_i}} = \left(\prod x_i^{f_i}\right)^{\frac{1}{\sum f_i}}$$

有

$$\ln \bar{x}_G = \frac{\sum f_i \ln x_i}{\sum f_i}$$

即对数的加权算术均值

例 3-14

# 对比

众数、中位数和算术均值：假设众数存在且唯一 [不是严格证明，而是单峰的情况下做图演示归纳，即经验观察而已]

- 且频数分布对称时三者相等
- 右偏 $\bar{x} > M_e > M_o$
- 左偏 $\bar{x} < M_e < M_o$

当频数分布的偏斜程度不是很大时 (如何度量未给出标准): $|M_e - M_o| = 2\,|M_e - \bar{x}|$, 以及 $M_o = 3M_e - 2\bar{x}$

- In some situations, the mode is the most useful average: three-bedroom house
- On other occasions a mean average is the most useful: plane that has nine passenger seats and a maximum carrying capacity of 1,350 pounds
- Sometimes a median average is the most meaningful: a product that appeals to people under age 35

# 对比

中位数和算术均值

- 各变量值与中位数的离差绝对值之和最小，即

$$\sum |x_i - M_e| = \min_x \sum |x_i - x|$$

画图证明很容易，偏离 $M_e$ 后, 总和多出了 $|x - M_e|$

- 各变量值与平均数的离差之和等于零

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

各变量值与平均数的离差平方和最小

$$\sum (x_i - \bar{x})^2 = \min_x \sum (x_i - x)^2$$

中位数和平均数数学性质的验证 ppt.34

# 总结

1. 平均数 (位置) 测度值的计算方法 (掌握)
2. 平均数 (位置) 的特点及应用场合 (理解)

**Keyword**: 平均数 (集中趋势), 众数, 中位数, 算术均值, 加权平均, 调和均值, 几何均值

**Homework**: ($*$ by group)

- Time loss of smart phones, pads and computers in classroom
- 基本概念要牢记，牢记，牢记!
- 先预习 S3.2, 然后小作文“浅议平均数的应用及其陷阱”, 800-1200 字
- Read
  1. Textbook: S3.2
  2. Reference book: corresponding chapters

## 2 离散程度

- 极差和四分位差
- 均值为中心
- 估算数据范围
- 总结

# 离散程度

平均数作为中心位置的代表，好坏取决于数据的离散水平

变异指标：异众比率、极差、四分位差、绝对离差 (平均差)、方差和标准差

- 变异指标值越大，平均数的代表性越差
- 产品质量的稳定性，经济活动的平稳性 (均衡、协调)

分类数据：异众比率是非众数组的频数占总频数的比例，用于衡量众数的代表性

$$v_r = \frac{\sum f_i - f_m}{\sum f_i} = 1 - \frac{f_m}{\sum f_i}$$

$f_m$ 是众数组的频数

# 极差和四分位差

# 极差

一组数据的最大值与最小值之差，用 $R$ 表示

$$R = \max(x_i) - \min(x_i)$$

反映总体标志值的变动范围，即

$$R = \text{极大值} - \text{极小值}$$

组距数列

$$R = \text{最大组上限} - \text{最小组下限}$$

反映一组数据的绝对变异程度

- 离散程度的最简单测度值
- 易受极端值影响
- 未考虑数据的分布，不能准确描述出数据的离散程度

# 四分位数

第 $i$ 四分位数 $Q_i$ 的位置为 [$n=25$ 得 7 13 19, 非课本 p50 的 $\frac{i}{4}(n+1)$]

$$h_i = 1 + \frac{i}{4}(n-1) \qquad i = 1, 2, 3$$

$Q_L = Q_1$, $Q_U = Q_3$, 中位数 $M_e = Q_2$ (wiki:Quantile 更多计算方法)

- 未分组: 如果 $h_i$ 非整数, 插值 (按比例分配相邻值 $x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)(x_{\lceil h \rceil} - x_{\lfloor h \rfloor})$)
- 分组: 先确定 $Q_i$ 所在的分组 $g_i$, 即 $S_{g_i-1} < 1 + \frac{i}{4}(\sum f_j - 1) \leqslant S_{g_i}$ ($S_0 = 0$)
  - 单变量分组 (课本 p50 式 3-18 与 3-19 的形式需还原成未分组)

    $$Q_i = x_{g_i} \qquad i = 1, 2, 3$$

    这是近似计算, 处于两组间归入下一组. 更精确的计算将其还原成未分组再计算
  - 组距分组 ($Q_i$ 的位置近似为 $\frac{i}{4}\sum f_j$)

    $$Q_i = L_i + \frac{\frac{i}{4}\sum f_j - S_{g_i-1}}{f_{g_i}} \cdot d_i \qquad i = 1, 2, 3$$

# 四分位差

interquartile range (IQR)

$$Q_R = Q_3 - Q_1$$

The quartile deviation $Q_d$ or semi-interquartile range is defined as half the IQR

- 反映了中间 50% 数据的离散程度
- 不受极端值的影响
- 用于衡量中位数的代表性

# 均值为中心

在描述数据离散程度时，若以均值 (算术平均数) 为中心 [均值代表某种中心位置，而不是什么集中或者趋势], 反映所有变量值平均值的离散状况，这样的测量值会比极差、异众比率、四分位差有更优良的性质。

# 绝对离差

mean absolute deviation $M_D$

$$M_D = \frac{1}{n}\sum |x_i - \bar{x}|$$

加权

$$M_D = \frac{\sum |x_i - \bar{x}| f_i}{\sum f_i}$$

“应用上有较大的局限性”，应该是错了，是数学上的处理比较麻烦，当数学处理或者软件算法改进后，应用就突破了，比如神经网络的深度学习算法

# 标准差

$\sigma$ 又称均方差，离差平方和的均值的方根，与变量同量纲

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

加权

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}}$$

通过平方消除正负号，数学上方便处理

样本标准差

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

加权

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i - 1}}$$

# 自由度

degree of freedom

- 自由度是指附加给独立的观测值的约束或限制的个数
- 从字面涵义来看，自由度是指一组数据中可以自由取值的个数
- 当样本数据的个数为 n 时，若样本平均数确定后，则附加给 n 个观测值的约束个数就是 1 个，因此只有 n-1 个数据可以自由取值，其中必有一个数据不能自由取值
- 按着这一逻辑，如果对 n 个观测值附加的约束个数为 k 个，自由度则为 n-k

# 离散系数

离散系数 (coefficient of variation) 又称变异系数

$$V_\sigma = \frac{\sigma}{\bar{x}} \times 100\%$$

当两个群体的平均数不等时 (不同水平), 就不能直接用标准差来比较，还必须计算离散系数才能测定标志变动程度。提醒：$\bar{x} \approx 0$ 时，有问题。

金融资产的 Sharpe 比

$$\frac{\bar{x} - R_0}{\sigma}$$

其中 $R_0$ 为无风险资产的收益率

# 估算数据范围

# 标准化值

标准化值, $z$ 值

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$

因为 $\bar{z} = 0$, $\sigma_z^2 = 1$

- 正态分布 68–95–99.7 rule, also known as the empirical rule. more precisely, 68.27%, 95.45% and 99.73% of the values lie within one, two and three standard deviations of the mean, respectively.
- Chebyshev's inequality: Let $X$ (integrable) be a random variable with finite expected value $\mu$ and finite non-zero variance $\sigma^2$. Then for any real number $k > 1$

$$\Pr\left(\frac{|X - \mu|}{\sigma} \geqslant k\right) \leqslant \frac{1}{k^2}$$

$k = 2$, 75%; $k = 3$, 88.89%

# 总结

1. 离散程度各测度值的计算方法 (掌握)
2. 离散程度各测度值的特点及应用场合 (理解)

**Keyword**: 异众比率, 极差, 四分位差, 绝对离差 (平均差), 方差和标准差, 自由度, 变异系数

**Homework**: ($*$ by group)

- Time loss of smart phones, pads and computers in classroom
- 基本概念要牢记，牢记，牢记!
- $*$ 收集上次小组作业的股票至少一年的日收益率 ($\frac{P_t}{P_{t-1}} - 1$, 记得考虑分红送配股), 计算中位数, 算术均值, 几何均值, 极差、四分位差、绝对离差, 标准差和离散系数 (作业纸只需分布特征的计算结果和必要文字解说)
- 下载 FinScore-stu.xlsx
- Read
  1. Textbook: S3.3–4
  2. Reference book: corresponding chapters

## 3 分布的形态

- 偏态和峰态
- 总结

# Simpson's paradox

panel data (pooled vs individual effect)

a correlation present in different groups is reversed when the groups are combined.

| | Treatment A | Treatment B |
|---|---|---|
| Small Stones | 93% (81/87) | 87% (234/270) |
| Large Stones | 73% (192/263) | 69% (55/80) |
| Both | 78% (273/350) | 83% (289/350) |

Implications to decision making

# 偏态和峰态

# 矩

矩 (moment, 动差): 中心矩

$$m_i = \frac{1}{n}\sum (x_i - \bar{x})^i \qquad i = 1, 2, \cdots$$

显然

$$m_1 = 0 \qquad m_2 = \sigma^2$$

位置、分散程度和形态: ppt.5

# 偏态和峰态

偏态 (分组数据 $\frac{1}{n}\sum z_i^3 f_i$, $z_i = \frac{x_i - \bar{x}}{\sigma}$)

$$\alpha = \frac{m_3}{\sigma^3} = \frac{\frac{1}{n}\sum (x_i - \bar{x})^3}{\left[\frac{1}{n}\sum (x_i - \bar{x})^2\right]^{3/2}} = \frac{1}{n}\sum \left(\frac{x_i - \bar{x}}{\sigma}\right)^3 = \frac{1}{n}\sum z_i^3$$

峰态 (分组数据 $\frac{1}{n}\sum z_i^4 f_i$)

$$\beta = \frac{m_4}{\sigma^4} = \frac{\frac{1}{n}\sum (x_i - \bar{x})^4}{\left[\frac{1}{n}\sum (x_i - \bar{x})^2\right]^2} = \frac{1}{n}\sum z_i^4 \geqslant \left(\frac{m_3}{\sigma^3}\right)^2 + 1$$

经验观察 ✗, see AE p916

- A distribution with positive excess kurtosis is called leptokurtic, or leptokurtotic. “Lepto-” means “slender”. In terms of shape, a leptokurtic distribution has fatter tails.
- A distribution with negative excess kurtosis is called platykurtic, or platykurtotic. “Platy-” means “broad”. In terms of shape, a platykurtic distribution has thinner tails.

# Excel

用 Excel 计算描述统计量 FinScore-stu.xlsx

- MODE.SNGL 众数; MODE.MULT 返回多个值到列数组; 旧 MODE
- MEDIAN 中位数
- QUARTILE.INC 四分位数; 旧 QUARTILE
- AVERAGE 平均数; TRIMMEAN 切尾均值
- HARMEAN 简单调和平均数
- GEOMEAN 几何平均数
- AVEDEV 平均绝对离差
- STDEV.S 样本标准差; STDEV.P 总体标准差; 旧 STDEV STDEVP
- SKEW 偏度
- KURT 峰度

# 总结

1. 偏态与峰态 (了解)
2. 用 Excel 计算描述统计量并进行分析 (掌握)
3. 本章复习

**Keyword**: 偏态, 峰态, Excel 描述统计函数

**Homework**: ($*$ by group)

- Time loss of smart phones, pads and computers in classroom
- 基本概念要牢记，牢记，牢记!
- $*$Y97-stu.docx [里面的案例和指定为小组作业的为小组完成，其他为个人作业]
- $*$ 修改小作文“浅议平均数的应用及其陷阱” [小组作业了，小组合并改写]
- Read
  1. Textbook: S6.1–S6.2.1
  2. Reference book: corresponding chapters