Summary Statistics

Applied Statistics

Fall 2025

目录

1 中心趋势度量 Measures of Center

1. 均值 Mean

- **均值（算术平均）**：将所有观测值相加后除以观测总数。
- **Mean (arithmetic average)**: computed by adding up all the observations and dividing by the total number of observations.
- population mean (总体均值) 用 μ 表示，sample mean (样本均值) 用 $\bar{x}$ 表示：

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- 示例：6 家公司的市盈率（P/E）：35, 30, 22, 18, 15, 12。
- 均值与每个观测值之间的距离称为 deviation (偏差)。

1.1 中位数 Median

- **中位数**：将数据排序后位于中间的值。
- **Median**: the middle observation when data are sorted.
- 奇数个观测：取中间值；偶数个观测：取中间两个值的平均。
- 示例：每周上网小时数：{0, 7, 12, 5, 14, 8, 0, 9, 22}，中位数为 8.5。

1.2 均值与中位数的比较 Comparing Mean and Median

- **均值对极端值敏感**，中位数对异常值更稳健。
- **Mean is sensitive to outliers; median is robust.**
- 当分布呈 symmetric(对称) 状态时，均值与中位数应非常接近。
- 通过比较均值与中位数，还可判断 skewness(偏度) 的程度。
- 在右偏分布中，均值 $>$ 中位数；左偏分布中，均值 $<$ 中位数。
- 均值可谨慎应用于有序数据 (categorical Data)，例如消费者评分均值为 3.4 分（5 分为最高满意度）常见于电影、餐厅等场景的评分数据。然而，仅当两组数据的问题设计一致时，其均值才具有可比性。

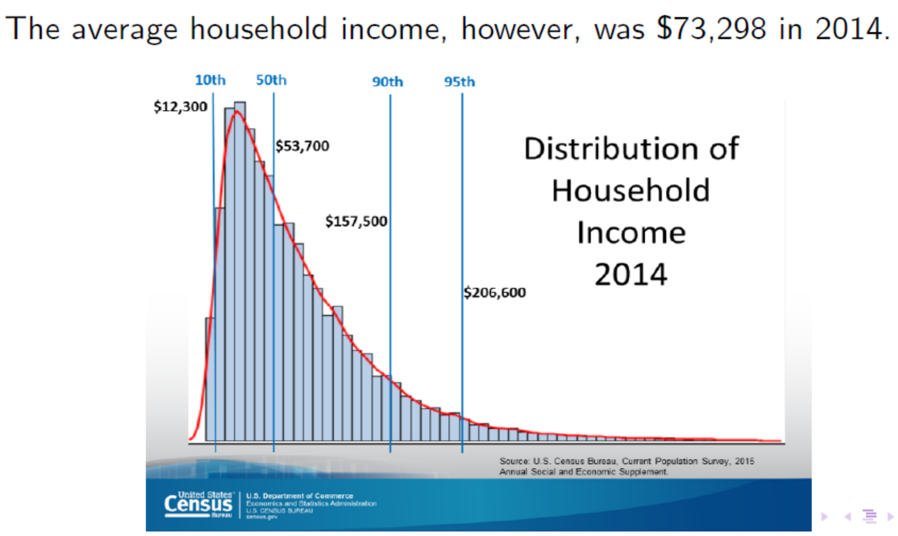


图 1: Mean & Median example

1.3 加权均值 Weighted Mean

- 公式：

$$\bar{x}_w = \sum_{i=1}^{n} w_i x_i, \quad \sum w_i = 1$$

- 常用于投资组合回报计算, 在利用分组数据估算均值时也十分实用，例如分组收入数据。
- 示例：70% 股票 + 30% 债券的组合回报。

1.4 几何平均 Geometric Mean

- 用于计算增长率或复合回报率：

$$G = \sqrt[n]{\prod_{i=1}^{n} x_i}, \quad x_i \geq 0$$

- 金融回报中的几何平均：

$$R_G = \left[\prod_{t=1}^{T} (1 + R_t) \right]^{\frac{1}{T}} - 1$$

- 几何均值始终小于或等于算术均值，仅当所有观测值完全相同时，两者才相等。
- 几何平均回报率反映的是一项投资的复合回报率。算术平均回报率聚焦于单期表现的平均水平。
- 示例：某只股票初始价格为 100 美元，一年后涨到 200 美元，第二年末价格回落至 100 美元（无股息）。试比较该股票的算术平均年回报率与几何平均年回报率。

1.5 百分位数 Percentiles

- **百分位数 Percentile** P_y：一个数值，使得有 $y\%$ 的数据小于或等于该值，同时有 $(100 - y)\%$ 的数据大于该值。
- **Percentile** P_y: the value for which y percent are less than or equal to that value and $(100 - y)\%$ are greater than that value.
- Suppose you scored in the 60th percentile on the GMAT, that means 60% of the other scores were below yours, while 40% of the scores were above yours.
- **Quartiles 四分位数**: Q1 (25%), Q2 (50%, median), Q3 (75%).
- **Quintiles 五分位数**: 20%, 40%, 60%, 80%.

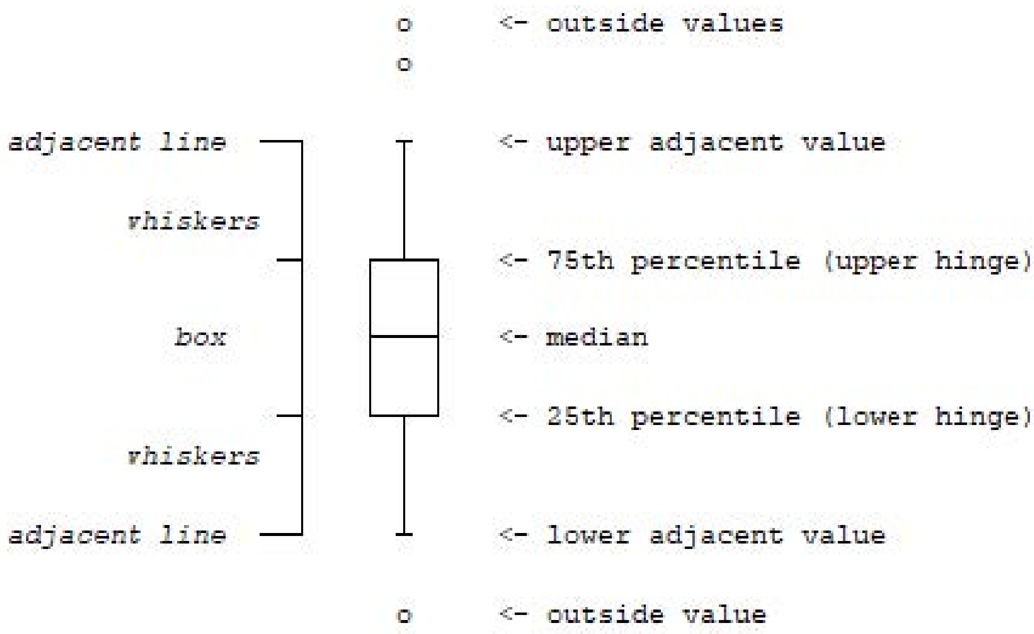
Table 2 Concentration

Variable	Gini Index	Coefficient of Variation	Top 1% to Bottom 40% Ratio	99th to 40th Percentile Ratio
Earnings	0.710	2.94	10.71	44.06
Income	0.664	3.20	4.47	25.07
Wealth	0.761	4.54	9.45	53.13

图 2: percentiles example table

1.6 箱线图 Box Plot

- 基于“五数概括”：最小值、第一四分位数（Q1）、中位数、第三四分位数（Q3）、最大值。
- **箱体**：Q1 到 Q3；**箱体内部的线段**：中位数（Q2）；**须线**：延伸至非异常值的最小/最大值。
- 异常值 outliner：落在 Q1 - 1.5×IQR 以下或 Q3 + 1.5×IQR 以上的值。
- **IQR** = Q3 - Q1。
- 背对背箱线图（Back-to-back box plots）在对比多个分布时尤为实用。



- Practice: the three quartiles of a variable are 60, 65, 70 and the maximum and minimum are 35, 90. Are there outliers in the data?

分析“给定分类变量取值时，数值变量的条件分布”，可判断这两个变量是否存在关联。若这些条件分布均高度相似，则我们没有理由认为该分类变量与数值变量之间存在关联。

2 离散度度量 Measures of Dispersion

2.1 极差 Range

- **极差**：最大值与最小值之差。
- **Range**: Largest observation minus smallest observation.
- 示例：{4, 4, 4, 4, 50} 与 {4, 8, 15, 24, 39, 50} 的极差相同，但离散程度不同。
- 极差对离散度的信息有限，但对检查异常值有用。

2.2 方差与标准差 Variance and Standard Deviation

- **方差**：衡量数据与均值的平均平方离差。
- **Variance**: Average of squared deviations from the mean.

- 总体方差用 σ^2 表示，样本方差用 s^2 表示：

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

- **标准差**：方差的平方根，与原始数据单位相同。
- **Standard deviation**: $s = \sqrt{s^2}$.
- 方差和标准差对异常值敏感，是非稳健统计量。
- 方差的单位是原度量单位的平方，因此难以直观解读，而标准差与样本观测值的单位一致（便于解释）。
- 标准差（s）衡量数据围绕均值的离散程度，且仅当均值被选为中心度量指标时才适用。
- 标准差（s）等于 0 的唯一情况是数据无离散性，即所有观测值完全相同时；否则，标准差（s）均大于 0。
- 标准差（s）不具备抗干扰性：少数异常值即可导致其数值大幅增大。

2.2.1 Why divide by $n-1$?

- 因为离差和为零，只有 $n-1$ 个离差可以自由变化，自由度是 $n-1$。
- 使用 $n-1$ 可使样本方差 s^2 成为总体方差 σ^2 的无偏估计。
- 从经验分布角度，标准正态数据的方差服从自由度为 $n-1$ 的卡方分布。

2.3 变异系数 Coefficient of Variation (CV)

- 当比较均值或单位不同的数据集的离散度时，可使用变异系数：

$$CV = \frac{s}{\bar{x}}$$

- 示例：小公司（平均销售额 7000 万美元）和大公司（平均销售额 8.2 亿美元）的标准差都是 1680 万美元，但变异系数不同。

2.4 经验法则 The Empirical Rule

- 对于钟形分布（正态分布）：
  - 约 68% 的数据落在 $\bar{x} \pm 1s$ 内。
  - 约 95% 的数据落在 $\bar{x} \pm 2s$ 内。

  - 约 99.7% 的数据落在 $\bar{x} \pm 3s$ 内。

- 示例：中国男性身高均值为 172 cm，标准差为 7.5 cm，则：

  - 68% 在 164.5 cm 至 179.5 cm 之间。
  - 95% 在 157 cm 至 187 cm 之间。
  - 99.7% 在 149.5 cm 至 194.5 cm 之间。

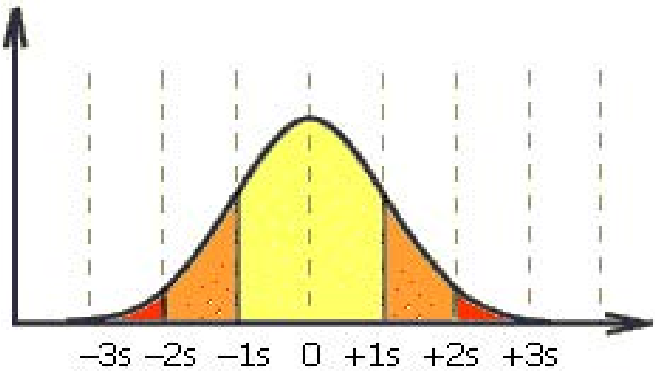


图 3: empirical law

2.5 切比雪夫不等式 Chebyshev's Inequality

一般形式

设随机变量 X 具有有限的数学期望 $\mathbb{E}[X] = \mu$ 和方差 $\mathrm{Var}(X) = \sigma^2$，则对任意 $t > 0$，有：

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

等价形式

令 $t = k\sigma$，其中 $k > 0$，则：

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

切比雪夫不等式也可以写成其互补形式：

$$P(|X - \mu| < ks) \geq 1 - \frac{1}{k^2}$$

- 适用于任何形状的分布，更保守的估计。
- 至少 $1 - 1/k^2$ 的数据落在 $\bar{x} \pm ks$ 内（$k > 1$）。

- 示例：
  - $k=2$：至少 75% 的数据在 $\bar{x} \pm 2s$ 内（经验法则为 95%）。
  - $k=3$：至少 88.9% 的数据在 $\bar{x} \pm 3s$ 内（经验法则为 99.7%）。

表 1: 切比雪夫不等式下的比例 Proportions from Chebyshev's inequality

k	区间 Interval	至少比例 Proportion (%)
1.25	$\bar{x} \pm 1.25s$	36
1.5	$\bar{x} \pm 1.5s$	56
2	$\bar{x} \pm 2s$	75
2.5	$\bar{x} \pm 2.5s$	84
3	$\bar{x} \pm 3s$	89
4	$\bar{x} \pm 4s$	94

3 分布形状度量 Measures of Shape

3.1 偏度 Skewness

- 偏度系数：
$$Skewness = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$
- 正偏（右偏）：均值 $>$ 中位数，分布右侧有长尾。
- 负偏（左偏）：均值 $<$ 中位数，分布左侧有长尾。
- 偏度的绝对值越大，数据的偏斜程度越严重。
- 目前存在其他偏度度量指标（如皮尔逊偏度系数），因此需明确所使用的具体指标类型。
- 示例：中国收入与财富分布（Tan 等，2017）：

3.1.1 偏度练习 Practice: Skewness

- 两个投资组合的回报分布均为单峰。组合 1 偏度为 0.77，组合 2 偏度为-1.11。以下哪项正确？

  A) 对于组合 1，中位数小于均值。

  B) 对于组合 1，众数大于均值。

C) 对于组合 2，均值大于中位数。

- 答案：A（正偏时均值 > 中位数）。

表 2: 中国收入与财富分布的偏度 Skewness of Income Distribution, China

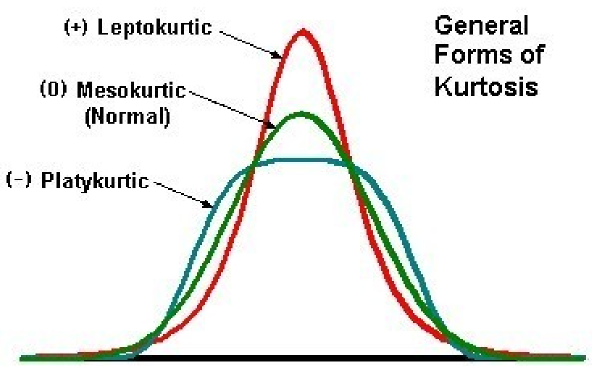
变量 Variable	均值所在百分位 Location of Mean (Percentile)	均值与中位数之比 Ratio of Mean to Median	偏度 Skewness
收入	77	2.12	17.74
财富	80	3.48	33.83

3.2 峰度 Kurtosis

- **峰度**：衡量分布尖峭程度和尾部厚度。
- **Kurtosis**: Measure of how peaked and heavy-tailed a distribution is.
- 常用峰度系数（基于四阶矩）：

$$K_E = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4}$$

- 正态分布的峰度为 3。
- 金融数据常具有厚尾特征（峰度 > 3），即极端值更多。
- 注意：峰度高的分布不一定有更厚的尾部，但通常相关联。



4 线性关系度量 Measures of Linear Relationship

4.1 概述 Overview

- 线性关系度量用于衡量两个变量之间线性关系的强度和方向。
- **Measures of linear relationship**: Provide information on the strength and direction of a linear relationship between two variables.
- 常用指标：协方差（Covariance）、相关系数（Correlation Coefficient）、最小二乘回归线（Least Squares Line）、决定系数（Coefficient of Determination, R^2）。
- 数据形式：成对数据 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$。
- **注意**：这些度量仅捕获线性关系，其他关系（如二次关系）可能无法反映。

4.2 协方差 Covariance

- **样本协方差 Sample covariance**：

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- 总体协方差用 σ_{xy} 表示。
- **解释 Interpretation**：
  - 正值 $s_{xy} > 0$：x 和 y 同向变动（正相关）。
  - 负值 $s_{xy} < 0$：x 和 y 反向变动（负相关）。
  - “较大”的绝对值通常表示较强关系，但“大”的标准模糊，且受量纲影响。

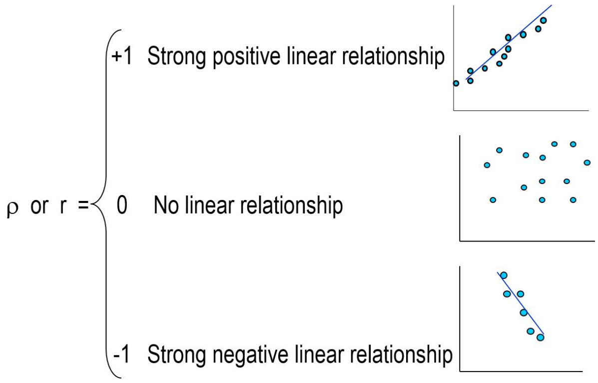
4.3 相关系数 Coefficient of Correlation

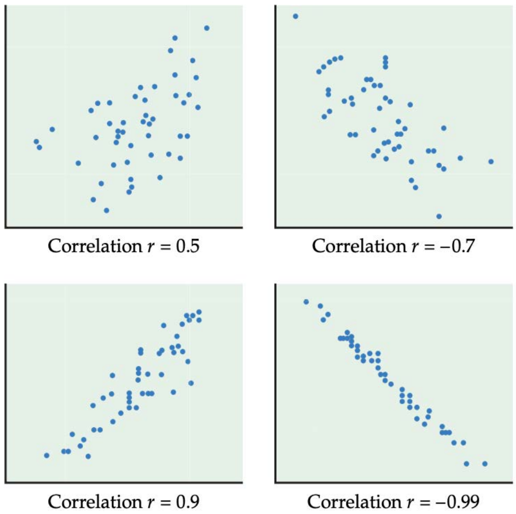
- **样本相关系数** r：将协方差标准化，消除量纲影响。

$$r = \frac{s_{xy}}{s_x s_y}$$

- 总体相关系数用 ρ 表示。
- **性质 Properties**：
  - 取值范围固定：$-1 \leq r \leq 1$。
  - $r = 1$：完全正线性关系。

- $r = -1$：完全负线性关系。
- $r = 0$：无线性关系（注意：可能仍存在非线性关系）。
- 可用于比较不同数据集之间关系的强度。





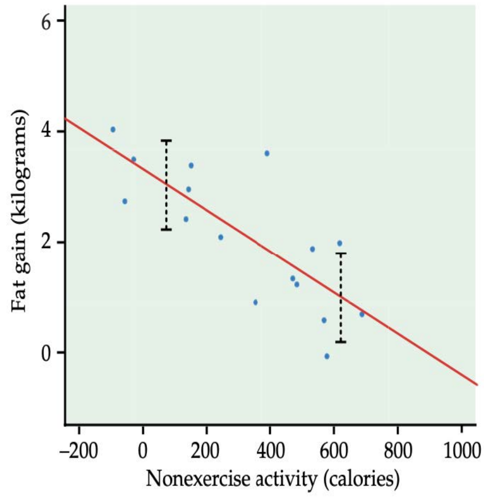
4.4 最小二乘回归线 Least Squares Line

- 目标：在散点图中拟合一条直线，描述 x 对 y 的边际效应。
- 方法：最小化残差平方和 $\sum(y_i - \hat{y}_i)^2$。
- 模型：$y = \beta_0 + \beta_1 x + \epsilon$
  - β_1：斜率（Slope），表示 x 每增加一单位，y 的平均变化量。
  - β_0：截距（Intercept），当 $x = 0$ 时 y 的预测值。
- **参数估计 Coefficient estimation：**

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- 斜率与相关系数的关系：x 变化一个标准差，预测的 y 变化 r 个标准差。



4.5 决定系数 Coefficient of Determination (R^2)

- 定义：回归模型所能解释的 y 的变异比例。

$$R^2 = \frac{\text{解释变异}}{\text{总变异}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

- 取值范围：0 到 1。
- 解释：R^2 越大，模型对响应变量的解释能力越强。
- 在单变量线性回归中：$R^2 = r^2$。
- 示例：脂肪增加模型的 $R^2 = 0.61$，意味着 61% 的脂肪增加变异可由非运动活动（NEA）的变异解释。

4.6 注意事项 Cautions

- **仅度量线性关系**：相关系数只反映线性关联，非线性关系可能被低估。
- **先绘图后计算**：在计算前务必绘制散点图，检查线性趋势和异常值。
- **外推风险**：在数据范围之外进行预测（外推）往往不可靠。
- **非稳健性**：相关系数和最小二乘线对异常值敏感，异常点可能对结果产生过大影响。
- **相关不等于因果**：即使存在强相关，也不能直接推断因果关系。

4.7 练习 Practice

- 题目：经济学家研究每日吸烟量（CIGAR）与月收入（INCOME，百美元）的关系。根据样本数据摘要，补全表格并回答问题。

- 已知部分数据：

统计量	CIGAR	INCOME
均值 Mean		42.4
中位数 Median	10	36
众数 Mode	6	–
标准差 Standard Deviation		19.16
样本方差 Sample Variance	260.23	
极差 Range	58	76
最小值 Minimum		
最大值 Maximum	60	86
总和 Sum	433	
观测数 Count		25

- 问题：

  1. 若样本中 80% 的人平均每天吸 17 支烟，剩余 20% 的人平均每天吸多少支？
  2. 哪个变量（CIGAR 或 INCOME）异质性更大？如何判断？

5 * Summary

- **中心趋势（Central Tendency）：**
  - 均值（Mean）：算术平均，对异常值敏感。
  - 中位数（Median）：基于位置，对异常值稳健。
  - 几何平均（Geometric Mean）：用于增长率和复合回报。
  - 加权平均（Weighted Mean）：用于组合数据或分组数据。

- **离散程度（Dispersion）：**
  - 极差（Range）：简单但不稳健。
  - 方差与标准差（Variance and Standard Deviation）：最常用的变异度量。
  - 变异系数（Coefficient of Variation）：用于比较不同尺度数据的离散度。

- **分布形状（Shape）：**
  - 偏度（Skewness）：衡量分布不对称性（左偏、右偏）。
  - 峰度（Kurtosis）：衡量分布尖峭和尾部厚度。
- **经验法则与切比雪夫不等式：**
  - 经验法则（Empirical Rule）：适用于钟形分布，快速估计比例。
  - 切比雪夫不等式（Chebyshev's Inequality）：适用于任意分布，保守估计比例。
- **线性关系（Linear Relationship）：**
  - 协方差（Covariance）：衡量共同变化的方向。
  - 相关系数（Correlation Coefficient）：标准化的协方差，衡量线性关系强度与方向。
  - 最小二乘回归线（Least Squares Line）：拟合最佳直线，描述边际效应。
  - 决定系数（R^2）：衡量模型解释变异的能力。