

第二章 定量数据的统计描述

第一节 频数分布

频数分布的目的：把“一堆散乱数字”，压缩成“一眼能看出形态的分布结构”。

当原始数据是**定量数据**且例数较多时，直接看原始值几乎没有信息密度；
通过**分组** → **频数表** → **直方图**，才能显露数据的分布规律。

一、频数表 (frequency table)

频数表是“分布的文字版压缩表达”。

定义

- 频数表同时列出：
 - 变量的**取值区间 (组段)**
 - 各区间内的**观测次数 (频数)**
 - 因为能系统反映分布规律，也称 **频数分布表**
-

二、频数表的编制步骤 (核心操作流程)

① 确定组数 (number of classes)

组数是“分辨率”的问题。

- 组数太少 → 细节被抹平，分布看不清
- 组数太多 → 表格臃肿，噪声反而变多
- **经验范围：8-15 组** (以“看清分布”为准)

✦ 这是经验法则，不是硬公式。

② 确定组距 (class width)

组距决定 “每一格看多宽” 。

- 全距 (range)

$R = \text{最大值} - \text{最小值}$

- 参考组距

$\text{组距} \approx R/k$

🔗 例 2-1 (红细胞计数)

【例 2-1】某地用随机抽样方法抽取了 140 名正常成年男性检测其红细胞计数 ($\times 10^{12}/L$), 检测结果如下所示。

| | | | | | | | | | | | |
|------|------|------|------|------|------|------|------|------|------|------|------|
| 4.76 | 5.26 | 5.61 | 5.95 | 4.46 | 4.57 | 4.31 | 5.18 | 4.92 | 4.27 | 4.77 | 4.88 |
| 5.00 | 4.73 | 4.47 | 5.34 | 4.70 | 4.81 | 4.93 | 5.04 | 4.40 | 5.27 | 4.63 | 5.50 |
| 5.24 | 4.97 | 4.71 | 4.44 | 4.94 | 5.05 | 4.78 | 4.52 | 4.63 | 5.51 | 5.24 | 4.98 |
| 4.33 | 4.83 | 4.56 | 5.44 | 4.79 | 4.91 | 4.26 | 4.38 | 4.87 | 4.99 | 5.60 | 4.46 |
| 4.95 | 5.07 | 4.80 | 5.30 | 4.65 | 4.77 | 4.50 | 5.37 | 5.49 | 5.22 | 4.58 | 5.07 |
| 4.81 | 4.54 | 3.82 | 4.01 | 4.89 | 4.62 | 5.12 | 4.85 | 4.59 | 5.08 | 4.82 | 4.93 |
| 5.05 | 4.40 | 4.14 | 5.01 | 4.37 | 5.24 | 4.60 | 4.71 | 4.82 | 4.94 | 5.05 | 4.79 |
| 4.52 | 4.64 | 4.37 | 4.87 | 4.60 | 4.72 | 4.83 | 5.33 | 4.68 | 4.80 | 4.15 | 4.65 |
| 4.76 | 4.88 | 4.61 | 3.97 | 4.08 | 4.58 | 4.31 | 4.05 | 4.16 | 5.04 | 5.15 | 4.50 |
| 4.62 | 4.73 | 4.47 | 4.58 | 4.70 | 4.81 | 4.55 | 4.28 | 4.78 | 4.51 | 4.63 | 4.36 |
| 4.48 | 4.59 | 5.09 | 5.20 | 5.32 | 5.05 | 4.41 | 4.52 | 4.64 | 4.75 | 4.49 | 4.22 |
| 4.71 | 5.21 | 4.94 | 4.68 | 5.17 | 4.91 | 5.02 | 4.76 | | | | |

(李康,贺佳主编, 2024, p. 9) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

- 最大值: 5.95
- 最小值: 3.82
- 全距: 2.13
- 拟分 10 组 → 参考组距 ≈ 0.21

👉 结合专业习惯, 选择 **0.20** 或 **0.25**

- 0.20 → 11 组 (更精细)
- 0.25 → 9 组 (更粗)

最终选择 **0.20**

🔴 **专业直觉很重要：**

组距不是数学问题，是“医学数据怎么看更合理”。

③ 确定组限 (class limits)

原则：每个数据“有且只能进一个组”。

- 实际操作中：
 - **包含下限，不包含上限**
 - 例如：
 - $3.80 \sim <4.00$
 - $4.00 \sim <4.20$

🔴 最后一组通常**标记上限值**以覆盖最大值。

④ 确定频数、频率与累积量

频数表常包含以下信息：

- **频数**：该组内观测个数
- **频率**：频数 / 总例数
- **累积频数**：到该组为止的总个数
- **累积频率**：到该组为止的比例

👉 频数看“局部”，累积频率看“整体位置”。

表 2-1 某地 140 名正常成年男性红细胞计数的频数表

| 红细胞计数/(× 10 ¹² /L) (1) | 组中值 (2) | 频数 (3) | 累积频数 (4) | 频率/% (5) | 累积频率/% (6) |
|--|--------------|-------------|---------------|---------------|-----------------|
| 3.80 ~ < 4.00 | 3.9 | 2 | 2 | 1.43 | 1.43 |
| 4.00 ~ < 4.20 | 4.1 | 6 | 8 | 4.29 | 5.71 |
| 4.20 ~ < 4.40 | 4.3 | 11 | 19 | 7.86 | 13.57 |
| 4.40 ~ < 4.60 | 4.5 | 25 | 44 | 17.86 | 31.43 |
| 4.60 ~ < 4.80 | 4.7 | 32 | 76 | 22.86 | 54.29 |
| 4.80 ~ < 5.00 | 4.9 | 27 | 103 | 19.29 | 73.57 |
| 5.00 ~ < 5.20 | 5.1 | 17 | 120 | 12.14 | 85.71 |
| 5.20 ~ < 5.40 | 5.3 | 13 | 133 | 9.29 | 95.00 |
| 5.40 ~ < 5.60 | 5.5 | 4 | 137 | 2.86 | 97.86 |
| 5.60 ~ < 5.80 | 5.7 | 2 | 139 | 1.43 | 99.29 |
| 5.80 ~ 6.00 | 5.9 | 1 | 140 | 0.71 | 100.00 |

(李康,贺佳主编, 2024, p. 10) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

三、直方图 (histogram)

直方图是“频数表的可视化”。

 构成规则

- 横轴：组限（反映组距宽度）
- 纵轴：频数或频率
- 每一组对应一个**相邻无缝的矩形条段**

 关键区别

- 条形图 (bar chart)：类别型数据，条之间有空隙
- **直方图**：连续型定量数据，**条之间无空隙**

⚠ 不等距分组的特殊处理

- 若使用不等组距：
 - 需先把频数折算成等距频数
 - 再绘制直方图
 - 👉 否则柱高会误导分布判断

四、通过直方图识别分布形态

直方图真正的价值：判断数据“长什么样”。

1 对称分布 (symmetric distribution)

- 中间最多
- 两侧对称下降
- 医学中最重要的一种：**正态分布 (normal distribution)**

🔗 例 2-1 红细胞计数：

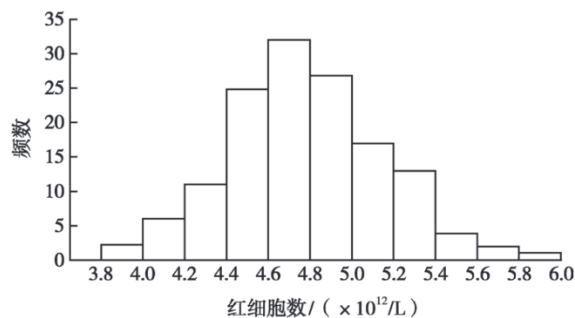


图 2-1 140 名正常成年男性红细胞计数的直方图

(李康,贺佳主编, 2024, p. 10) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

- 峰值约在 $4.8 \times 10^{12}/L$
- 左右近似对称
- 近似正态分布

2 偏态分布 (skewed distribution)

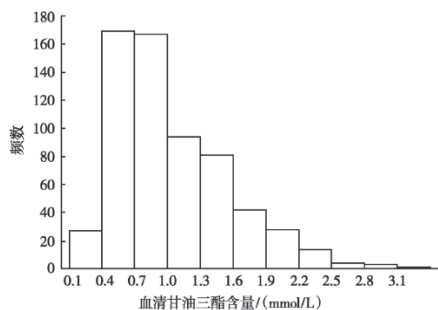


图 2-2 630 名正常女性血清甘油三酯含量的频数分布

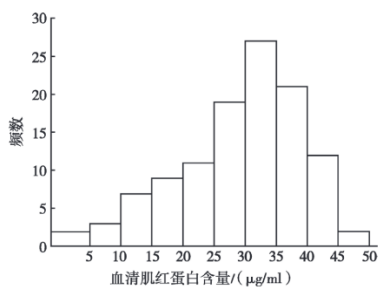


图 2-3 100 名女性血清肌红蛋白含量的频数分布

(李康, 贺佳主编, 2024, p. 11) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

正偏态 (right-skewed)

- - 峰在左
 - 长尾向右
 - 常见于：甘油三酯、住院天数、费用
- **负偏态 (left-skewed)**
 - 峰在右
 - 长尾向左
 - 相对少见

🔴 这一步非常关键：

后续选择

- 均值 vs 中位数
 - t 检验 vs 非参数检验
- 都以分布形态为前提。

五、这一节在“统计全局”中的位置

频数分布 = 描述统计的起点，不是终点。

它解决的问题只有一个：

👉 **数据整体分布形态是什么？**

但它直接决定：

- 能不能假设正态分布
 - 用什么集中趋势指标
 - 后面该不该做参数检验
-

🎯 **一句话总结本节**

频数分布不是“做表画图”，而是为后续一切统计推断建立“对数据形态的第一直觉”。

第二节 描述集中趋势的统计学指标

集中趋势指标回答的只有一件事：这一组数据“典型值”大概在哪儿。不同指标对分布形态与极端值的敏感度不同。

📌 四个“代表值”怎么选（先给决策直觉）

- **算术均数 (arithmetic mean)**：最常用，但怕偏态/极端值
 - **几何均数 (geometric mean, GM)**：适合**倍数变化**（滴度/菌落计数/效价等）
 - **中位数 (median)**：对偏态与极端值更稳健 (robust)
 - **众数 (mode)**：出现次数最多的值/直方图的峰
-

一、算术均数 (arithmetic mean)

均数=把每个观测值都算进去的“平均水平”。

总体均数记作 μ ，样本均数记作 \bar{x} 。

(一) 直接法 (raw data)

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum X}{n} \tag{2-1}$$

🔗 例 2-1 (红细胞计数)

140 名正常成年男性红细胞计数均值：

【例 2-1】某地用随机抽样方法抽取了 140 名正常成年男性检测其红细胞计数 ($\times 10^{12}/L$), 检测结果如下所示。

| | | | | | | | | | | | |
|------|------|------|------|------|------|------|------|------|------|------|------|
| 4.76 | 5.26 | 5.61 | 5.95 | 4.46 | 4.57 | 4.31 | 5.18 | 4.92 | 4.27 | 4.77 | 4.88 |
| 5.00 | 4.73 | 4.47 | 5.34 | 4.70 | 4.81 | 4.93 | 5.04 | 4.40 | 5.27 | 4.63 | 5.50 |
| 5.24 | 4.97 | 4.71 | 4.44 | 4.94 | 5.05 | 4.78 | 4.52 | 4.63 | 5.51 | 5.24 | 4.98 |
| 4.33 | 4.83 | 4.56 | 5.44 | 4.79 | 4.91 | 4.26 | 4.38 | 4.87 | 4.99 | 5.60 | 4.46 |
| 4.95 | 5.07 | 4.80 | 5.30 | 4.65 | 4.77 | 4.50 | 5.37 | 5.49 | 5.22 | 4.58 | 5.07 |
| 4.81 | 4.54 | 3.82 | 4.01 | 4.89 | 4.62 | 5.12 | 4.85 | 4.59 | 5.08 | 4.82 | 4.93 |
| 5.05 | 4.40 | 4.14 | 5.01 | 4.37 | 5.24 | 4.60 | 4.71 | 4.82 | 4.94 | 5.05 | 4.79 |
| 4.52 | 4.64 | 4.37 | 4.87 | 4.60 | 4.72 | 4.83 | 5.33 | 4.68 | 4.80 | 4.15 | 4.65 |
| 4.76 | 4.88 | 4.61 | 3.97 | 4.08 | 4.58 | 4.31 | 4.05 | 4.16 | 5.04 | 5.15 | 4.50 |
| 4.62 | 4.73 | 4.47 | 4.58 | 4.70 | 4.81 | 4.55 | 4.28 | 4.78 | 4.51 | 4.63 | 4.36 |
| 4.48 | 4.59 | 5.09 | 5.20 | 5.32 | 5.05 | 4.41 | 4.52 | 4.64 | 4.75 | 4.49 | 4.22 |
| 4.71 | 5.21 | 4.94 | 4.68 | 5.17 | 4.91 | 5.02 | 4.76 | | | | |

(李康,贺佳主编, 2024, p. 9) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

$$\bar{X} \approx 4.78(\times 10^{12}/L)$$

(二) 加权法 (frequency table)

当你只有频数表，没有原始数据时，用 “组中值×频数” 近似代替。

- 组中值：

$$x_i = \frac{\text{下限} + \text{上限}}{2}$$

- 加权均数：

$$\bar{X} = \frac{\sum fx}{n} \tag{2-2}$$

📌 重要提醒

- 样本量大、分组合理时：加权法≈直接法（结果接近）
- 但原则上：**能用原始数据就别用分组近似**（信息损失）

表 2-1 某地 140 名正常成年男性红细胞计数的频数表

| 红细胞计数/($\times 10^{12}/L$) (1) | 组中值 (2) | 频数 (3) | 累积频数 (4) | 频率/% (5) | 累积频率/% (6) |
|-------------------------------------|------------|-----------|-------------|-------------|---------------|
| 3.80~<4.00 | 3.9 | 2 | 2 | 1.43 | 1.43 |
| 4.00~<4.20 | 4.1 | 6 | 8 | 4.29 | 5.71 |
| 4.20~<4.40 | 4.3 | 11 | 19 | 7.86 | 13.57 |
| 4.40~<4.60 | 4.5 | 25 | 44 | 17.86 | 31.43 |
| 4.60~<4.80 | 4.7 | 32 | 76 | 22.86 | 54.29 |
| 4.80~<5.00 | 4.9 | 27 | 103 | 19.29 | 73.57 |
| 5.00~<5.20 | 5.1 | 17 | 120 | 12.14 | 85.71 |
| 5.20~<5.40 | 5.3 | 13 | 133 | 9.29 | 95.00 |
| 5.40~<5.60 | 5.5 | 4 | 137 | 2.86 | 97.86 |
| 5.60~<5.80 | 5.7 | 2 | 139 | 1.43 | 99.29 |
| 5.80~6.00 | 5.9 | 1 | 140 | 0.71 | 100.00 |

(李康,贺佳主编, 2024, p. 10) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

将表 2 - 1 的数据代入公式 (2 - 2) , 有

$$\bar{X} = \frac{2 \times 3.90 + 6 \times 4.10 + 11 \times 4.30 + \cdots + 2 \times 5.70 + 1 \times 5.90}{140} = 4.78(\times 10^{12}/L)$$

由此可见, 在样本例数较多的情况下, 加权法与直接法算得的结果很相近。

尽管如此, 在实际应用中, 还是**提倡基于原始数据**, 采用直接法计算算术均数, 仅在没有原始数据而只有频数表资料时, 才考虑用加权法计算算术均数。

(三) 均数的应用边界 (非常关键)

- 适合: **对称分布/偏度不大**, 尤其 **正态分布 (normal)**
- 不适合: **明显偏态**或存在极端值 → 均数会被“尾部”拉走

- 这时更应考虑 **中位数/几何均数**

二、几何均数 (geometric mean, GM)

当数据按“倍数关系”变化时 (log-scale) , 几何均数比算术均数更像“典型水平”。

典型场景：抗体滴度、细菌计数、凝集效价、部分浓度数据。

(一) 定义式

$$G = \sqrt[n]{X_1 X_2 \cdots X_n} \quad (2-3)$$

(二) 对数计算式 (实际常用)

$$G = 10^{\left(\frac{\sum \lg X}{n}\right)} \quad (2-4)$$

(等价理解：先算对数的均值，再取反对数)

(三) 频数表形式 (有分组/频数时)

$$G = 10^{\left(\frac{\sum n \lg x}{n}\right)} \quad (2-5)$$

⚠ 硬性条件

- 观测值不能有 **0 或负数**
- 同一组数据里：通常 **几何均数 < 算术均数**

🧪 例 2-2 (原始滴度, 24 人)

【例 2-2】测得 24 名健康志愿者接种 2 剂试验疫苗后 28 天中和抗体滴度的倒数分别为 64, 128, 128, 128, 512, 64, 128, 256, 128, 256, 256, 512, 512, 512, 256, 128, 128, 512, 256, 256, 64, 512, 256, 256, 求抗体的几何平均滴度。

(李康, 贺佳主编, 2024, p. 13) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

- 计算得到几何平均滴度 (GMT):

将数据代入公式 (2 - 4) , 计算几何均数:

$$G = \lg^{-1} \left(\frac{\lg 64 + \lg 128 + \lg 128 + \lg 128 + \dots + \lg 64 + \lg 512 + \lg 256 + \lg 256}{24} \right) \approx 209$$

👉 表达: 平均滴度约为 **1:209**

📌 例 2-3 (频数表滴度, 表 2-2; 120 人)

表 2-2 接种 2 剂试验疫苗后 28 天血清中和抗体滴度

| 中和抗体滴度倒数 | 频数 | 中和抗体滴度倒数 | 频数 |
|----------|----|----------|----|
| 16 | 4 | 128 | 35 |
| 32 | 4 | 256 | 30 |
| 64 | 21 | 512 | 21 |
| 96 | 3 | 1 024 | 2 |

(李康,贺佳主编, 2024, p. 13) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

频数表代入:

$$G = \lg^{-1} \left(\frac{4\lg 16 + 4\lg 32 + 21\lg 64 + 3\lg 96 + 35\lg 128 + 30\lg 256 + 21\lg 512 + 2\lg 1024}{120} \right) \approx 157$$

👉 平均滴度约为 **1:157**

三、中位数与百分位数 (median & percentile)

中位数只看“位置”，不看“极端值有多极端”。

把数据从小到大排序: $X_1 \leq \dots \leq X_n$

(一) 中位数 (median, M)

- n 为奇数：第 $(n + 1)/2$ 个
- n 为偶数：第 $n/2$ 与第 $n/2 + 1$ 个的平均

✅ **优点**：抗极端值、适合偏态

❌ **缺点**：不便于代数运算（例如两组中位数很难合并推出总体中位数）

（二）频数表下的中位数（分组资料的近似计算）

核心思路：先找“中位数落在哪个组段”，再在该组段内做线性插值（假设组内均匀分布）。

通用公式

$$M = L + \frac{i_M}{f_M}(n \times 50\% - f_L) \quad (2-6)$$

- L ：中位数组段下限
 - i_M ：组距
 - f_M ：中位数组段频数
 - f_L ：中位数组段之前的累积频数
-

例 2-4

表 2-3 某地 630 名 50~60 岁正常女性血清甘油三酯含量的频数表

| 甘油三酯/(mmol/L) (1) | 频数 (2) | 累积频数 (3) | 累积频率/% (4) |
|----------------------|-----------|-------------|---------------|
| 0.10~<0.40 | 27 | 27 | 4.29 |
| 0.40~<0.70 | 169 | 196 | 31.11 |
| 0.70~<1.00 | 167 | 363 | 57.62 |
| 1.00~<1.30 | 94 | 457 | 72.54 |
| 1.30~<1.60 | 81 | 538 | 85.40 |
| 1.60~<1.90 | 42 | 580 | 92.06 |
| 1.90~<2.20 | 28 | 608 | 96.51 |
| 2.20~<2.50 | 14 | 622 | 98.73 |
| 2.50~<2.80 | 4 | 626 | 99.37 |
| 2.80~<3.10 | 3 | 629 | 99.84 |
| ≥3.10 | 1 | 630 | 100.00 |
| 合计 | 630 | — | — |

(李康,贺佳主编, 2024, p. 14) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

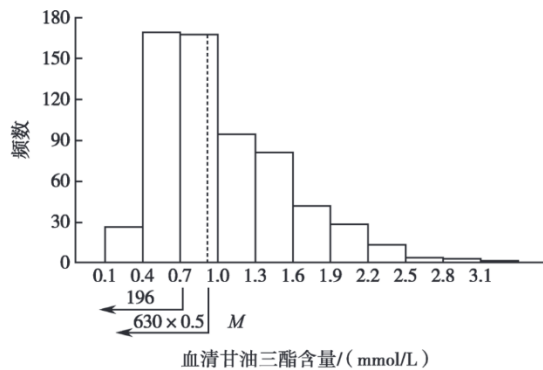


图 2-4 中位数计算方法示意图

(李康,贺佳主编, 2024, p. 14) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

- 总例数 $n = 630$
- 中位数位置: $630 \times 0.5 = 315$
- 315 落在 $[0.70, 1.00)$ 组段
 - 该组前累积频数 $f_L = 196$

- 该组频数 $f_M = 167$
- 组距 $i_M = 0.30$
- 下限 $L = 0.70$

$$M = 0.70 + \frac{0.30}{167} (315 - 196) = 0.914 \text{ mmol/L}$$

(三) 百分位数 (percentile, P_x)

P_x : 使得 **x% 的数据 \leq 该值**, 其余 \geq 该值。

特别地: $P_{50} = M$

分组资料公式

$$P_x = L + \frac{i_x}{f_x} (n \times x\% - f_L) \quad (2-7)$$

注意

- 用频数表算的 P_x 是**近似值**
- 靠近两端 (如 $P_{2.5}$ 、 $P_{97.5}$) 需要较大样本才稳定 (常见经验: $n > 100$)

例 2-5 (基于例 2-4 求百分位数)

$$P_{25} = 0.40 + \frac{0.30}{169} \times (630 \times 0.25 - 27) = 0.632 \text{ (mmol/L)}$$

$$P_{75} = 1.30 + \frac{0.30}{81} \times (630 \times 0.75 - 457) = 1.357 \text{ (mmol/L)}$$

$$P_{90} = 1.60 + \frac{0.30}{42} \times (630 \times 0.90 - 538) = 1.807 \text{ (mmol/L)}$$

(四) 中位数/百分位数的应用 (比公式更重要)

- **偏态分布**: 用中位数更稳健 (mean 会被长尾拽走)
- **分散程度**: 用 $P_{75} - P_{25}$ (即 IQR) 描述
- **参考范围**: 用 $P_{2.5}$ 与 $P_{97.5}$ 定义 95% 医学参考区间
- **生长发育分级**: 用百分位数划分等级 (P_3 、 P_{10} 、 P_{50} 、 P_{90} ...)

四、众数 (mode)

众数=出现次数最多的值；对连续型数据可理解为直方图的“最高峰位置”。

- 可能一个众数，也可能多个（多峰分布）
 - 在分布形态判断（是否双峰、是否异常亚群）时很有用
-

本节一句话总结

均数擅长“对称世界”，几何均数擅长“倍数世界”，中位数/百分位数擅长“偏态与极端值世界”，众数擅长“峰与亚群”。

第三节 描述变异程度的统计学指标

均数告诉你“典型水平”，变异指标告诉你“波动有多大”。两者缺一不可。

先用一个例子把“为什么要学变异”钉死


例 2-6：两位高血压患者 5 天收缩压

【例 2-6】对甲乙两名高血压患者连续观察 5 天，测得的收缩压分别如下：

甲患者 (mmHg) 162 145 178 142 186 ($\bar{X}_{\text{甲}} = 162.6$)

乙患者 (mmHg) 164 160 163 159 166 ($\bar{X}_{\text{乙}} = 162.4$)

(李康,贺佳主编, 2024, p. 15) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

 两人均数几乎一样，但：

- 甲：大起大落（不稳定）
- 乙：很平稳

结论：描述一组数据，必须同时给出

- **集中趋势 (center) + 变异程度 (spread)**
-

变异指标两大家族（背这个就够）

- **间距类 (distance-based)：**极差、四分位数间距 (IQR)
 - **平均差距类 (average deviation-based)：**方差、标准差、变异系数 (CV)
-

一、极差 (range, R)

极差只看两端：最大值 - 最小值。

$$R = X_{\max} - X_{\min}$$

例 2-6

- 甲： $R = 186 - 142 = 44\text{mmHg}$
- 乙： $R = 166 - 159 = 7\text{mmHg}$

✅ **优点：**简单直观，适合“最短/最长”类问题（如潜伏期范围）

❌ **缺点：**

- 只用到两个点，浪费绝大多数信息
- 样本量越大，出现极端值概率越高，R 越不稳定
- 偏态分布时更容易被极端值带偏

👉 **定位：**极差适合“快速说明范围”，不适合作为严谨变异指标。

二、四分位数间距 (IQR, interquartile range)

IQR = 去掉两端 25%，只看中间 50% 的“稳定波动范围”。

$$IQR = P_{75} - P_{25}$$

🔗 承接上一节例 2-4 (甘油三酯)

- $P_{25} = 0.632 \text{ mmol/L}$
- $P_{75} = 1.357 \text{ mmol/L}$

$$IQR = 1.357 - 0.632 = 0.725 \text{ mmol/L}$$

解释：约 **50%** 女性甘油三酯在 $[0.632, 1.357]$ mmol/L 之间。

✅ **优点**：不容易被极端值影响，适合**偏态分布**

❌ **缺点**：仍没用到每个观测值，信息利用不充分

👉 常配合箱线图/分位数一起用

三、方差 (variance,)

方差的思想：把每个点离均差“平方”后平均，避免正负抵消。

总体方差： σ^2 ；样本方差： S^2

$$S^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

- $\sum (X - \bar{X})^2$ ：**离均差平方和** (sum of squares, SS)
- 分母 $n - 1$ ：**自由度 (degree of freedom)**
 - 因为 \bar{X} 已经“消耗”了一个约束，所以只有 $n - 1$ 个独立信息

常用等价计算（更方便）：

$$\sum (X - \bar{X})^2 = \sum X^2 - \frac{(\sum X)^2}{n}$$

🔴 **单位陷阱**：方差单位是“原单位的平方” (mmHg^2)，不直观。

四、标准差

标准差 = 方差开方，回到原单位，最常用的波动指标。

$$S = \sqrt{\frac{\sum(X - \bar{X})^2}{n - 1}}$$

或等价式：

$$S = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n - 1}}$$


例 2-6

甲患者：

$$\begin{aligned}\sum X &= 162 + 145 + 178 + 142 + 186 = 813 \\ \sum X^2 &= 162^2 + 145^2 + 178^2 + 142^2 + 186^2 = 133713 \\ S &= \sqrt{\frac{133713 - 813^2/5}{5 - 1}} = 19.49(\text{mmHg})\end{aligned}$$

乙患者：

$$\begin{aligned}\sum X &= 164 + 160 + 163 + 159 + 166 = 812 \\ \sum X^2 &= 164^2 + 160^2 + 163^2 + 159^2 + 166^2 = 131902 \\ S &= \sqrt{\frac{131902 - 812^2/5}{5 - 1}} = 2.88(\text{mmHg})\end{aligned}$$

 甲波动远大于乙（这比“极差”更全面，因为用到了每个观测值）

频数表资料的标准差（分组近似）

$$S = \sqrt{\frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n - 1}}$$

- x: 组中值
- f: 该组频数

例 2-7

【例 2-7】根据第一节表 2-1 的频数表资料,计算成年男性红细胞数的标准差。计算结果如表 2-4:

(李康,贺佳主编, 2024, p. 17) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

表 2-4 140 名正常成年男性红细胞计数(× 10¹²/L)的标准差计算表

| 红细胞计数/(× 10 ¹² /L) (1) | 组中值(x) (2) | 频数(f) (3) | fx (4) | fx ² (5) |
|--|---------------|--------------|-----------|------------------------|
| 3.80~<4.00 | 3.90 | 2 | 7.80 | 30.42 |
| 4.00~<4.20 | 4.10 | 6 | 24.60 | 100.86 |
| 4.20~<4.40 | 4.30 | 11 | 47.30 | 203.39 |
| 4.40~<4.60 | 4.50 | 25 | 112.50 | 506.25 |
| 4.60~<4.80 | 4.70 | 32 | 150.40 | 706.88 |
| 4.80~<5.00 | 4.90 | 27 | 132.30 | 648.27 |
| 5.00~<5.20 | 5.10 | 17 | 86.70 | 442.17 |
| 5.20~<5.40 | 5.30 | 13 | 68.90 | 365.17 |
| 5.40~<5.60 | 5.50 | 4 | 22.00 | 121.00 |
| 5.60~<5.80 | 5.70 | 2 | 11.40 | 64.98 |
| 5.80~6.00 | 5.90 | 1 | 5.90 | 34.81 |
| 合计 | — | 140 | 669.80 | 3 224.20 |

(李康,贺佳主编, 2024, p. 18) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

教材计算得:

$$S = \sqrt{\frac{3224.20 - 669.80^2/140}{140 - 1}} = 0.38(\times 10^{12}/L)$$

为什么标准差/方差重要?

- 它们能“完整刻画”正态分布的形状：**均数 + 方差/标准差** ≈ 描述一个正态总体
- 方差还有一个工程级优点：多个样本方差可用于求**合并方差**（后面 t 检验会用）

五、变异系数 (coefficient of variation, CV)

当均数差很大或单位不同, 标准差不能直接比; 用 CV 做“相对波动”。

$$CV = \frac{S}{\bar{X}} \times 100\%$$

例 2-8: 舒张压 vs 收缩压

【例 2-8】测得某地成年人舒张压的均数为 77.5mmHg, 标准差为 10.7mmHg; 收缩压的均数为 122.9mmHg, 标准差为 17.1mmHg。试比较舒张压和收缩压的变异程度。


(李康, 贺佳主编, 2024, p. 18) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

- 舒张压: $\bar{X} = 77.5, S = 10.7$

$$CV = \frac{10.7}{77.5} \times 100\% = 13.81\%$$

- 收缩压: $\bar{X} = 122.9, S = 17.1$

$$CV = \frac{17.1}{122.9} \times 100\% = 13.91\%$$

 结论: 两者相对变异几乎相同 (仅看 SD 会误以为收缩压更“波动”)

使用警戒线 (教材经验)

- 若 $CV \geq 0.20$ (20%) 常提示变异偏大 → 需要追查原因 (人群异质、测量误差、分布偏态等)

CV 的硬伤

- 当 \bar{X} 接近 0, CV 会被放大, 解释容易失真。

本节一句话总结

极差看“边界”，IQR看“中间一半”，标准差看“整体波动”，CV看“相对波动”；选哪个取决于分布形态、极端值、以及是否需要跨尺度比较。

[BIBLIOGRAPHY] Please click Zotero - Refresh in Word/LibreOffice to update all fields