# 第四章 定性数据的统计描述

## 第一节 常用相对数

### 一、 率

1️⃣ 一句话定义（先定住）

> **率，是在一定时间或空间范围内，某现象发生的“强度或频率”。**

数学形式永远是：

$$率 = \frac{某现象实际发生数}{可能发生该现象的总数} \times 比例基数$$

👉 核心不是“比”，而是**“发生的频率感”**。

---

2️⃣ **率 ≠ 比例 ≠ 构成比**

你现在学“率”，脑子里必须立刻多一句判断：

> **它是不是在描述“一定时间内发生了多少”？**

如果答案是“是”，那它大概率是**率**。

---

3️⃣ **比例基数不是随便乘的**

比例基数的唯一目的：**让结果“好读、好用”**

常见习惯搭配：

| 指标 | 常用表示 |
| --- | --- |
| **治愈率、感染率** | % |

| | |
|---|---|
| **出生率、死亡率** | ‰ |
| **某些疾病死亡率** | 1/万、1/10 万 |

📌 经验原则：

- 让结果**整数部分保留 1–2 位**
- 否则再“精确”也没可读性

---

4️⃣ 总体率 vs 样本率（符号要敏感）

- **总体率**：π
- **样本率**：p

👉 这是后面：

- 率的估计
- 率的假设检验
- 区间估计

的**符号基础**，不是装饰。

---

5️⃣ 率几乎一定“带时间”

这是**考试和理解都非常关键的一点**：

> **率，通常隐含时间概念。**

- 出生率
- 死亡率
- 发病率

📌 默认口径：

- **1 年内发生的频率**

👉 如果不带时间，那你就要警惕：
**它可能不是“率”，而是比例。**

---

6️⃣ 率真正想回答的问题

不是“多不多”，而是：

> **“在可发生的背景下，这种现象出现得有多频繁？”**

这也是为什么：

- 分母必须是**“可能发生的人群”**
- 分母选错，率就彻底没意义

【例 4-1】某单位在 2022 年有 3 128 名职工，该单位每年对职工进行体检，在这一年新发现高血压患者 12 例，则

$$\text{高血压发病率}=\frac{12}{3\,128}\times 1\,000‰=3.84‰$$

(李康,贺佳主编, 2024, p. 33) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

## 二、构成比

## 1️⃣ 一句话定义（定性）

> **构成比，表示一个整体内部，各组成部分“占整体的比例结构”。**

它回答的问题是：

> **“这堆东西是怎么分的？”**

而不是“发生得多不多”。

---

## 2️⃣ 数学定义（非常干净）

$$\text{构成比} = \frac{\text{某一组成部分的例数}}{\text{该事物内部所有例数之和}} \times 100\%$$

关键点只有一个：

👉 **分子 + 分母，必须来自同一个“整体”**

---

## 3️⃣ 构成比的三个“硬性质”（必记）

### （1）构成比之和 = 100%

- 各部分互相制约
- 一个部分↑，必然挤压其他部分↓

👉 **零和结构**

---

### （2）不带时间概念

- 构成比：
  - 不关心“多久发生一次”
  - 只关心“怎么分布”

👉 这是它和“率”的**根本差别**

---

### （3）只能描述“结构”，不能描述“风险 / 强度”

这点非常容易被误用。

---

## 4️⃣ 构成比 vs 率（正面对照，必须会）

| 维度 | 构成比 | 率 |
| --- | --- | --- |
| 核心问题 | 内部怎么分 | 发生得多频繁 |
| 分母 | 内部总数 | 可能发生的总数 |
| 时间概念 | ❌ 无 | ✅ 有 |
| 各部分关系 | 互相制约 | 互不影响 |
| 总和 | =100% | 无此要求 |

📌 **一句判断口诀**：

问“比例结构” → 构成比
问“发生强度” → 率

---

## 5️⃣ 教材中特别提醒的“高频误区”

### ❌ 误用 1：用死因构成比判断“致死严重程度”

- 死因构成比说明的是：**某病死亡人数在“所有死亡”中占多大比例**
- 它**不能说明**：
  - 这个病“有多容易致死”

👉 真正反映“致死性”的指标是：

### ✅ 病死率（case fatality rate）

$$\text{病死率} = \frac{\text{某病死亡人数}}{\text{该病患者总数}}$$

📌 **这是考试、论文、现实中都极易踩雷的一点**

“【例 4-2】某医院某月住院病人数及死亡人数如表 4-1 所示,其中第(4)栏为构成比,是由第(3) 栏数据计算而得。第(5)栏为率,是由第(2)与第(3)栏数据计算而得。” (李康,贺佳主编, 2024, p. 33) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

表 4-1　某医疗机构某月住院病人数及死亡人数统计

| 疾病类型<br>（1） | 病人数<br>（2） | 病死人数<br>（3） | 死亡构成比/%<br>（4） | 病死率/‰<br>（5） |
|---|---|---|---|---|
| 呼吸系统疾病 | 620 | 25 | 23.81 | 40.32 |
| 循环系统疾病 | 1 030 | 35 | 33.33 | 33.98 |
| 消化系统疾病 | 540 | 20 | 19.05 | 37.04 |
| 恶性肿瘤 | 300 | 25 | 23.81 | 83.33 |
| 合计 | 2 490 | 105 | 100.00 | 42.17 |

(李康,贺佳主编, 2024, p. 34) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

## 三、相对比

1️⃣ 相对比的“母概念”

一句话定义

> **相对比，是两个有关联指标 A 与 B 的比值，用来描述“一个是另一个的多少倍”。**

$$\text{相对比} = \frac{A}{B}$$

它回答的问题永远是：

> **A 相对于 B，大还是小？大多少倍？**

📌 注意：

- 不要求时间概念
- 不要求“发生概率”
- 只是**量级对比**

---

2️⃣ 相对比的三种典型形态（必须区分）

教材这里其实讲了 **三种层级完全不同的“比”**。

---

（一）两类别例数之比（最朴素的相对比）

本质

- **数量对数量**
- 同一总体内的两个类别

例子：性别比

$$\text{男女性别比} = \frac{\text{男性人口数}}{\text{女性人口数}}$$

👉 结论只表示：

- 男性是女性的多少倍

❌ 不表示：

- 风险
- 发生概率
- 因果关系

📌 **这是纯描述性指标**

---

（二）相对危险度 RR（Relative Risk）

这是相对比里**最容易被滥用、但最有解释力的一个。**

---

定义一句话

> **RR 表示：暴露组发生某疾病的风险，是非暴露组的多少倍。**

$$RR = \frac{P_1}{P_0}$$

- $P_1$：暴露组的发病率 / 患病率
- $P_0$：非暴露组的发病率 / 患病率

---

RR 在逻辑上的三个“前提”

1️⃣ 有**暴露 vs 非暴露**
2️⃣ 有**发生概率（率）**
3️⃣ 通常来自**前瞻性研究（队列研究）**

👉 **RR 是“风险层面”的倍数比较**

---

RR 的解释非常直接

- RR = 1 → 无关联
- RR > 1 → 暴露是危险因素
- RR < 1 → 暴露是保护因素

📌 这也是 RR 在公共卫生里极受欢迎的原因：**好解释、接近直觉**

【例 4-3】为了解吲达帕胺片治疗原发性高血压的疗效，将 80 名高血压患者随机分为两组，试验组用吲达帕胺片加辅助治疗，对照组用安慰剂加辅助治疗，结果见表 4-2，试计算两组疗效的相对危险度。

表 4-2　两种疗法治疗原发性高血压的疗效

| 组别 | 合计 | 有效例数 | 有效率/% |
|---|---|---|---|
| 对照组 | 40 | 22 | 55.00 |
| 试验组 | 40 | 31 | 77.50 |

相对危险度为

$$RR = \frac{77.50}{55.00} = 1.41$$

说明试验组治疗原发性高血压的有效率是对照组的 1.41 倍。

(李康,贺佳主编, 2024, p. 34) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

---

（三）比数比 OR（Odds Ratio）

这是**最容易被误解、但在某些研究中“不得不用”的指标。**

---

定义一句话

> **OR 比的是“暴露优势（odds）”，而不是“发生概率（risk）”。**

$$\mathrm{OR} = \frac{P_1/(1-P_1)}{P_0/(1-P_0)}$$

- $P_1$：病例组的暴露比例
- $P_0$：对照组的暴露比例

---

OR 出现的背景（非常关键）

- **病例–对照研究**
- 研究设计中：
  - 无法直接算发病率
  - 只能比较“暴露比例”

👉 所以：

- **不能直接算 RR**
- 只能用 OR 作为替代指标

---

OR 的正确理解方式

- OR ≈ RR
  👉 **仅在疾病较少见（低发生率）时成立**
- 疾病不罕见时：

- OR 会 **夸大关联强度**
- 解释必须非常谨慎

📌 这是流行病学和临床论文里的**高频踩雷点**

---

3️⃣ RR vs OR ——必须能一句话分清

| 维度 | RR | OR |
| --- | --- | --- |
| 比较对象 | 风险（概率） | 优势（odds） |
| 常见研究 | 队列研究 | 病例-对照研究 |
| 可读性 | 非常直观 | 不直观 |
| 是否易被误解 | 较少 | 非常容易 |
| 能否直接算发病率 | 可以 | 不可以 |

【例 4-4】母亲围孕期是否有发热或感冒病史与婴儿神经血管畸形关系的病例对照研究的资料如表 4-3 所示，试计算母亲围孕期是否有发热或感冒病史引起婴儿神经血管畸形的比数比。

表 4-3　母亲围孕期是否有发热或感冒病史与婴儿神经血管畸形的关系

| 发热或感冒病史 | 神经血管畸形组 | 对照组 | 合计 |
|---|---|---|---|
| 有 | 40（$a$） | 20（$b$） | 60 |
| 无 | 112（$c$） | 203（$d$） | 315 |
| 合计 | 152（$a+c$） | 223（$b+d$） | 375 |

病例组中的暴露比例与非暴露比例分别为

$$P_1=\frac{a}{a+c},\quad 1-P_1=\frac{c}{a+c}$$

对照组的暴露比例与非暴露比例分别为

$$P_0=\frac{b}{b+d},\quad 1-P_0=\frac{d}{b+d}$$

由公式（4-5）可以得出

$$OR=\frac{P_1/(1-P_1)}{P_0/(1-P_0)}=\frac{[a/(a+c)]/[c/(a+c)]}{[b/(b+d)]/[d/(b+d)]}=\frac{ad}{bc} \tag{4-6}$$

本例

$$OR=\frac{40\times 203}{20\times 112}=3.63$$

即母亲围孕期是否有发热或感冒病史引起婴儿神经血管畸形的优势比为 3.63。

(李康,贺佳主编, 2024, p. 35) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

## 四、标准化率

### 1 为什么要用标准化率？（问题先行）

**典型陷阱（教材里的例子）**

【例 4-5】甲、乙两医院治愈率比较的资料如表 4-4 所示。

表 4-4　甲、乙两医院治愈率的比较

| 科室 | 甲医院 | | | 乙医院 | | |
|---|---|---|---|---|---|---|
| | 入院人数 | 治愈人数 | 治愈率/% | 入院人数 | 治愈人数 | 治愈率/% |
| 内科 | 1 500 | 975 | 65.0 | 500 | 315 | 63.0 |
| 外科 | 500 | 470 | 94.0 | 1 500 | 1 365 | 91.0 |
| 传染病科 | 500 | 475 | 95.0 | 500 | 460 | 92.0 |
| 合计 | 2 500 | 1 920 | 76.8 | 2 500 | 2 140 | 85.6 |

(李康,贺佳主编, 2024, p. 35) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

- 甲医院：**各科室治愈率都高于乙医院**
- 但：**总治愈率却低于乙医院**

👉 直觉冲突的根源只有一个：

> **两医院“内部构成”不同（科室、年龄、性别等）**

📌 所以结论是：

> **未经校正的总率，不能直接比较。**

---

## 2️⃣ 标准化率解决的是什么问题？

### 一句话定义

> **标准化率，是在统一“内部构成”条件下，对不同人群率水平进行可比性校正后的结果。**

它的目标不是：

- 反映真实发生水平
  而是：
- **消除构成差异**

- **实现公平比较**

---

## 3️⃣ 什么时候“必须”考虑标准化?

只要满足下面任一条，就要警惕：

- 比较不同人群的：
  - 患病率
  - 发病率
  - 死亡率
  - 治愈率
- 且这些人群在：
  - 年龄
  - 性别
  - 工龄
  - 病程
  - 病情轻重

等方面**构成明显不同**

👉 **不标准化，比较就不成立。**

---

## 4️⃣ 直接标准化法（本节核心方法）

### 核心思想（先理解）

> **“假设各组有同样的内部构成，再看谁的率高。”**

---

### 计算公式（记住结构，不死背）

$$P' = \frac{\sum N_i P_i}{N}$$

或等价写成：

$$P' = \sum C_i P_i$$

---

**各符号的直观含义**

- $P'$：**标准化率**
- $P_i$：第 i 层的**原始率**
- $N_i$：标准构成中第 i 层的例数
- $C_i = N_i/N$：**标准构成比**
- $N$：标准构成总例数

👉 本质一句话：

> **用“统一权重”，去加权各层的原始率。**

---

## 5️⃣ 标准构成怎么选？（这是关键中的关键）

标准化结果**高度依赖标准构成**，常见三种选法：

### （1）大而稳定的人群（最推荐）

- 全国
- 全省
- 权威调查数据

👉 **代表性最好**

---

### （2）比较对象的合并总体

- 把各组例数合在一起作为标准

👉 **中性、常用**

### （3）任选一组作为标准

- 操作简单
- 但**主观性强**

📌 考试与论文都必须意识到：

**不同标准构成 → 不同标准化率**

【例 4-6】试根据表 4-4 资料计算甲乙两个医院的标准化率。

将比较各组的各层例数的合计作为标准构成，即将两医院各科室的人数之和作为标准构成，根据两医院各层的治愈率，计算两医院各层的预期治愈数，最后得到两组标准化后的预期治愈数，其计算结果如表 4-5 所示。

表 4-5　消除构成影响后两医院治愈率的比较

| 科室 | 标准构成人数 | 甲医院 | | 乙医院 | |
|---|---|---|---|---|---|
| | | 原治愈率/% | 预期治愈数 | 原治愈率/% | 预期治愈数 |
| 内科 | 2 000 | 65.0 | 1 300 | 63.0 | 1 260 |
| 外科 | 2 000 | 94.0 | 1 880 | 91.0 | 1 820 |
| 传染病科 | 1 000 | 95.0 | 950 | 92.0 | 920 |
| 合计 | 5 000 | | 4 130 | | 4 000 |

按公式（4-7）计算得到甲医院标准化后的总治愈率为

$$P'_{甲}=\frac{4\,130}{5\,000}\times 100\%=82.6\%$$

乙医院标准化后的总治愈率为

$$P'_{乙}=\frac{4\,000}{5\,000}\times 100\%=80.0\%$$

(李康,贺佳主编, 2024, p. 36) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

---

6️⃣ 标准化率的“使用红线”（非常重要）

❗ 两条必须牢记的限制

1️⃣ **标准化率只用于“相互比较”**

- 不能解释为真实发生水平

2️⃣ **不能跨标准比较**

- 不同研究、不同标准
- 标准化率不能直接对比

👉 否则就是**统计性误读**

## 第二节 医学中常用的相对数指标

### 一、死亡统计指标

1️⃣ 死亡率（粗死亡率，Crude Death Rate）

一句话定义

> **某年某地，每千人口中“死了多少人”。**

$$\text{死亡率} = \frac{\text{某年某地死亡总人数}}{\text{同年该地年平均人口数}} \times 1000‰$$

关键理解点

- 反映的是：**总体死亡水平**
- 分母必须是**同一年的平均人口数**
- 年平均人口数 ≈（年初人口 + 年末人口）/ 2

局限性（非常重要）

- 强烈受以下因素影响：
  - 年龄结构（老人多 → 死亡率高）
  - 性别结构（男性多 → 死亡率高）

📌 **不同地区粗死亡率不能直接比**
👉 年龄/性别结构不同 → **必须先标准化**

---

## 2️⃣ 年龄别死亡率（Age-specific Death Rate）

一句话定义

> **在“同一年龄组”里，每千人中死了多少人。**

$$\text{年龄别死亡率} = \frac{\text{某年某地某年龄组死亡人数}}{\text{同年该地该年龄组平均人口数}} \times 1000‰$$

核心作用

- **消除年龄构成差异的影响**
- 用于：
  - 年龄间比较
  - 地区间更公平比较
  - 老龄化分析

📌 逻辑上：

> 年龄别死亡率 ≠ 总死亡率
> 它是“分层后”的死亡强度

---

## 3️⃣ 死因别死亡率（Cause-specific Death Rate）

一句话定义

> **某年某地，每 10 万人中，有多少人死于某一特定原因。**

$$\text{某病死亡率} = \frac{\text{某年某地某病死亡人数}}{\text{同年该地平均人口数}} \times 100000$$

它回答的问题

- **这种病对人群“杀伤力”有多大**

- 是：
  - 死因分析
  - 公共卫生决策
  - 疾病负担评估

的核心指标

📌 注意：

- 分母是**全人群**
- 不是该病患者

---

4️⃣ 死因构成（Proportion of Dying of a Specific Cause）

一句话定义

> **在所有死亡中，有多少比例是“死于这种原因”。**

$$\text{某死因构成比} = \frac{\text{因该死因死亡的人数}}{\text{总死亡人数}} \times 100\%$$

它真正反映的是

- **各种死因的“相对重要性”**
- 用于回答：

> “死亡主要是被什么夺走的？”

但它**不能说明**

- 该病是否“更致命”
- 该病死亡风险是否更高

👉 想回答“致死严重程度”，**必须用死因别死亡率或病死率**

---

5️⃣ 四个指标的终极对照（强烈建议记住）

| 指标 | 分母是谁 | 回答的问题 |
| --- | --- | --- |
| 死亡率 | 全部人口 | 总体死亡水平 |
| 年龄别死亡率 | 同年龄组人口 | 某年龄段的死亡强度 |
| 死因别死亡率 | 全部人口 | 某病对人群的致死危害 |
| 死因构成 | 全部死亡人数 | 死亡“结构” |

## 二、疾病统计指标

1️⃣ 发病率（Incidence rate）

一句话本质

**发病率 = 新病例出现的速度**

它回答的是：

**“这个病在多快地‘冒出来’？”**

$$\text{发病率} = \frac{\text{某时期新发病例数}}{\text{同期间可能患病的人群数}} \times \text{比例基数}$$

必须死记的三点

1. **只算“新病例”**
2. 分母是：**有患病风险的人**
   - 已患病者 ❌
   - 已有效免疫者 ❌
3. **强时间属性**（一年、一个月、随访期）

📌 关键词：**风险、发生、动态**

---

2️⃣ 患病率（Prevalence rate / 现患率）

一句话本质

**患病率 = 某一时点“有多少人在带病生存”**

$$\text{患病率} = \frac{\text{某时期现患病例数}}{\text{同期平均人口数（或调查人数）}} \times \text{比例基数}$$

核心理解

- 包括：
  - 旧病例
  - 新病例
- **时间点或短时间段**
- 特别适合描述：
  - 慢性病
  - 长病程疾病

📌 关键词：**负担、存量、静态**

---

## 3️⃣ 发病率 vs 患病率（必须会分）

| 维度 | 发病率 | 患病率 |
| --- | --- | --- |
| 分子 | 新病例 | 全部现患病例 |
| 时间 | 一段时间 | 某一时点 / 时期 |
| 反映 | 发生风险 | 疾病负担 |
| 典型疾病 | 急性感染 | 慢性病 |

📌 经典关系式（理解即可）：

**患病率 ≈ 发病率 × 病程**

---

4️⃣ 病死率（Case Fatality Rate）

一句话本质

**病死率 = 得了这个病，有多大概率会死于它**

$$\text{病死率} = \frac{\text{因该病死亡人数}}{\text{该病患者总人数}} \times 100\%$$

关键点（高频考点）

- 分母是：**患者**
- 不是全人群
- 反映的是：
  - 疾病严重程度
  - 医疗救治水平

📌 常用于：

- 急性传染病
- 暴发疫情评估

---

5️⃣ 治愈率（Cure rate）

一句话本质

**治愈率 = 治疗后“成功的人占多少”**

$$\text{治愈率} = \frac{\text{治愈人数}}{\text{接受治疗人数}} \times 100\%$$

**使用边界要清楚**

- 强依赖：
  - “治愈”的定义
  - 随访是否完整
- 反映：
  - 治疗效果
  - 医疗质量

📌 **不能直接等价为“疾病轻”**

---

6️⃣ 四个指标的终极对照（强烈建议记住）

| 指标 | 分母是谁 | 回答的核心问题 |
| --- | --- | --- |
| 发病率 | 有风险人群 | 病出现得多快 |
| 患病率 | 总人群 | 病的社会负担 |
| 病死率 | 患者 | 病有多致命 |
| 治愈率 | 接受治疗者 | 治疗有多有效 |

## 第三节 相对数指标使用的注意问题

### 1️⃣ 构成比 ≠ 率

**常见误用**

- 用**构成比**去解释：
  - 发生强度
  - 风险大小
  - 致死严重程度

**正确理解**

- **构成比**：说明“内部结构”
- **率**：说明“发生频率 / 强度”

📌 教材中的典型对比：

- 循环系统疾病
  - **病死人数构成比最高** → 说明“死得多”
- 恶性肿瘤
  - **病死率最高** → 说明“更容易致死”

👉 **“死得多” ≠ “更致命”**

## 2️⃣ 分母过小的相对数 ≠ 可靠结论

**问题本质**

- 分母小 → **一个例数的变化就会导致剧烈波动**
- 数值“看起来很高”，但**极不稳定**

**典型例子**

- 5 例中 4 例有效 → 80%
- 5 例中 3 例有效 → 60%

👉 只差 1 例，差了 20%

📌 操作原则：

> **例数过少时，用绝对数，不要硬算率。**

---

## 3️⃣ 合计率不能“直接平均”

### ❌ 错误做法

$$\text{合计率} \neq \frac{P_1 + P_2 + \cdots}{k}$$

（当各组例数不等时，这样算是错的）

---

### ✅ 正确做法

$$\text{合计率} = \frac{\sum \text{分子}}{\sum \text{分母}}$$

📌 本质一句话：

**合计率一定是“先合人，再算率”，**
**不是“先算率，再平均”。**

---

## 4️⃣ 比较相对数之前，先问一句：可比吗？

### “可比性”检查清单

在比较之前，至少要确认：

- 观察对象是否一致？
- 研究方法是否一致？
- 观察时间是否一致？
- 诊断标准是否一致？

⚠️ 尤其要注意：

- **内部构成是否相同**
  - 年龄
  - 性别
  - 病情分层

👉 构成不同：

- 要么 **分层比较**
- 要么 **标准化后再比**

---

## 5️⃣ 样本率 / 构成比都有抽样误差

### 最容易犯的逻辑错误

> “A 是 12%，B 是 10%，所以 A 比 B 高。”

❌ 这是不成立的。

**正确态度**

- 样本率、构成比都是：
  - **估计值**
  - 有抽样误差

👉 是否“真有差别”，必须：

- 做**假设检验**
- 或看**置信区间**

📌 **“看数字大小下结论”是统计学大忌。**

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