第六章 参数估计与假设检验

第一节 参数估计

一、抽样误差

一句话本质

> **抽样误差不是错误，而是“抽样必然带来的随机波动”。**

- 只要不是普查，只要是随机抽样：
  - 样本统计量 ≠ 总体参数
- 这种差异 **不可避免、但可量化**

📌 关键区分：

- 抽样误差 ✅（随机、可分析）
- 系统误差 ❌（偏倚、不可忽略）

（一）、均数的标准误（SE of Mean）——“均数靠不靠谱”的刻度

1️⃣ 抽样分布的核心结论（一定要记住）

如果总体为：

$$X \sim N(\mu, \sigma^2)$$

那么从中反复抽取样本量为 n 的样本，其**样本均数** $\bar{X}$ 的分布为：

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

👉 **均数的均数 = 总体均数**
👉 **均数的变异 = $\sigma/\sqrt{n}$**

2️⃣ 中心极限定理（现实世界的底气）

当：

- $n \geq 50$

即使原始数据 **不服从正态分布**，
样本均数的抽样分布也**近似正态。**

📌 这就是医学统计里敢用正态近似的根本理由。

3️⃣ **什么是“均数的标准误”**

- **标准差（SD）**：个体之间差异有多大
- **标准误（SE）**：样本均数之间差异有多大

  SE 描述的是“估计的不确定性”

4️⃣ **均数标准误公式（直觉版）**

总体标准差已知：

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

实际更常见（σ未知）：

$$S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

5️⃣ 公式背后的三条铁律

1️⃣ **变异越大（σ 或 S 大）**
→ 均数越不稳定

2 样本量越大（n 大）

→ 均数越精确

3 SE < SD 永远成立

→ 均数比个体稳定

---

6 例 6-1 的真正含义

【例 6-1】在某地随机抽查成年男性 140 人，得红细胞均数 4.77×10^{12}/L，标准差 0.38×10^{12}/L，试计算其标准误。

按公式（6-2）计算得：

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{0.38}{\sqrt{140}} = 0.032\ (\times 10^{12}/\mathrm{L})$$

(李康,贺佳主编, 2024, p. 57) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

算出来的：

$$S_{\bar{X}} = 0.032$$

不是为了“算数”，而是告诉你：

这个样本均数大约在 ±0.03 的量级内波动

这就是后面：

- 置信区间
- 假设检验

的全部基础。

（二）、率的标准误（SE of Rate）——“比例稳不稳”的刻度

1️⃣ 样本率的本质

- 每个人只有两种状态：是 / 否
- 本质是 **二项分布**
- 样本率：

$$p = \frac{X}{n}$$

2️⃣ 率的标准误（总体已知）

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

含义非常直观：

- 发生概率越接近 0.5 → 波动越大
- 样本量越大 → 波动越小

3️⃣ 实际使用的估计式（最常用）

因为 π 通常未知，用 p 代替：

$$S_p = \sqrt{\frac{p(1-p)}{n}}$$

📌 **这是所有率的区间估计、比较检验的起点。**

均数 SE vs 率 SE（放在一起看）

指标	描述对象	标准误反映什么
均数 SE	定量数据	均数估计的不确定性
率 SE	定性数据	比例估计的不确定性

👉 **SE 不是描述数据本身，而是描述“估计质量”**

这一节真正想让你记住的 4 句话

1️⃣ 抽样误差 **一定存在，不是错误**
2️⃣ 标准误 = 抽样误差的量化表达
3️⃣ 样本量 ↑ → 标准误 ↓
4️⃣ SE 是 **置信区间和假设检验的地基**

二、置信区间的概念

1️⃣ 参数估计的两种方式（先把地基立住）

（1）点估计（Point estimation）

用一个数，直接顶替总体参数。

- 例子：
  - 用 $\bar{X}$ 估计 μ
  - 用 p 估计 π

✅ 优点：简单、直观
❌ 致命缺点：**完全不告诉你“准不准”**

👉 点估计 **无不确定性信息。**

（2）区间估计（Interval estimation）

给一个“可能包含真实参数”的范围。

- 形式：

$$\text{参数} \in (\text{下限}, \text{上限})$$

- 并配一个概率保证。

👉 这才是**统计推断真正有用的形式。**

2️⃣ 什么是置信区间（CI）

核心定义（一句话版）

> **在给定置信度下，用样本构造的、可能包含总体参数的区间。**

三个关键词必须同时出现

- **区间**
- **概率**
- **重复抽样意义**

3️⃣ 置信度（Confidence level, 1-\alpha）

常用取值

- 95%（最常用）
- 99%（更保守）

正确理解（非常重要）

> **95% 置信区间 ≠ “μ 有 95% 的概率在这个区间里”**

真正的含义是：

> **如果无限次重复抽样并构造区间，**
> **那么其中约 95% 的区间会包含真实的 μ。**

📌 μ 是固定的，
📌 **“随机”的是区间，不是 μ。**

这是统计学里最容易被误解的一点。

4️⃣ 置信区间的组成

- **下限（Lower limit）**
- **上限（Upper limit）**

两者合在一起，构成 CI。

例子正确解读

> 某地成年男性红细胞均数的 95% CI 为
> $$4.71 \sim 4.83 \times 10^{12}/\mathrm{L}$$

正确理解是：

- 在当前抽样方案和方法下
- 这是一个 **95% 可靠的估计区间**
- 区间**越窄**，说明估计越精密

5️⃣ 置信区间的两个评价维度（考试高频）

① **准确度**（Accuracy）

- 由 **置信度 $1-\alpha$** 决定
- 越接近 1 → 覆盖真实参数的概率越大

---

② **精密度**（Precision）

- 由 **区间宽度** 决定
- 区间越窄 → 估计越精确

6️⃣ 一个不可回避的“矛盾”

在样本量固定时：

想要	结果
提高置信度	区间变宽
缩小区间	置信度下降

👉 **准确度与精密度此消彼长。**

怎么解决？

增加样本量 n

这是唯一同时：

- 不降低置信度
- 又能缩小区间

的方法。

7️⃣ 这一节真正想让你建立的直觉

1️⃣ 点估计 = “一个猜测”
2️⃣ 置信区间 = “带风险控制的猜测”
3️⃣ 标准误决定区间“宽不宽”
4️⃣ 样本量决定你“敢不敢说得更窄”

三、总体均数的区间估计

（一）、先给你一张“用什么公式”的决策表（非常重要）

情况	用什么分布	关键条件
σ 已知	**z 分布**	实际中较少见
σ 未知，n 较小	**t 分布**	原始数据近似正态
σ 未知，n ≥ 50	**z 近似**	中心极限定理

👉 **99% 的医学研究：σ 未知 → t 或 z 近似**

（二）、σ 已知时：z 置信区间（理论型）

1️⃣ 标准化思路

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

2️⃣ 双侧 95% 置信区间

$$\mu \in (\bar{X} - 1.96\,\sigma_{\bar{X}}, \bar{X} + 1.96\,\sigma_{\bar{X}})$$

更一般形式：

$$(\bar{X} - z_{\alpha/2}\sigma_{\bar{X}}, \bar{X} + z_{\alpha/2}\sigma_{\bar{X}})$$

📌 **本质**：

用“标准正态的尾部概率”来包住 μ。

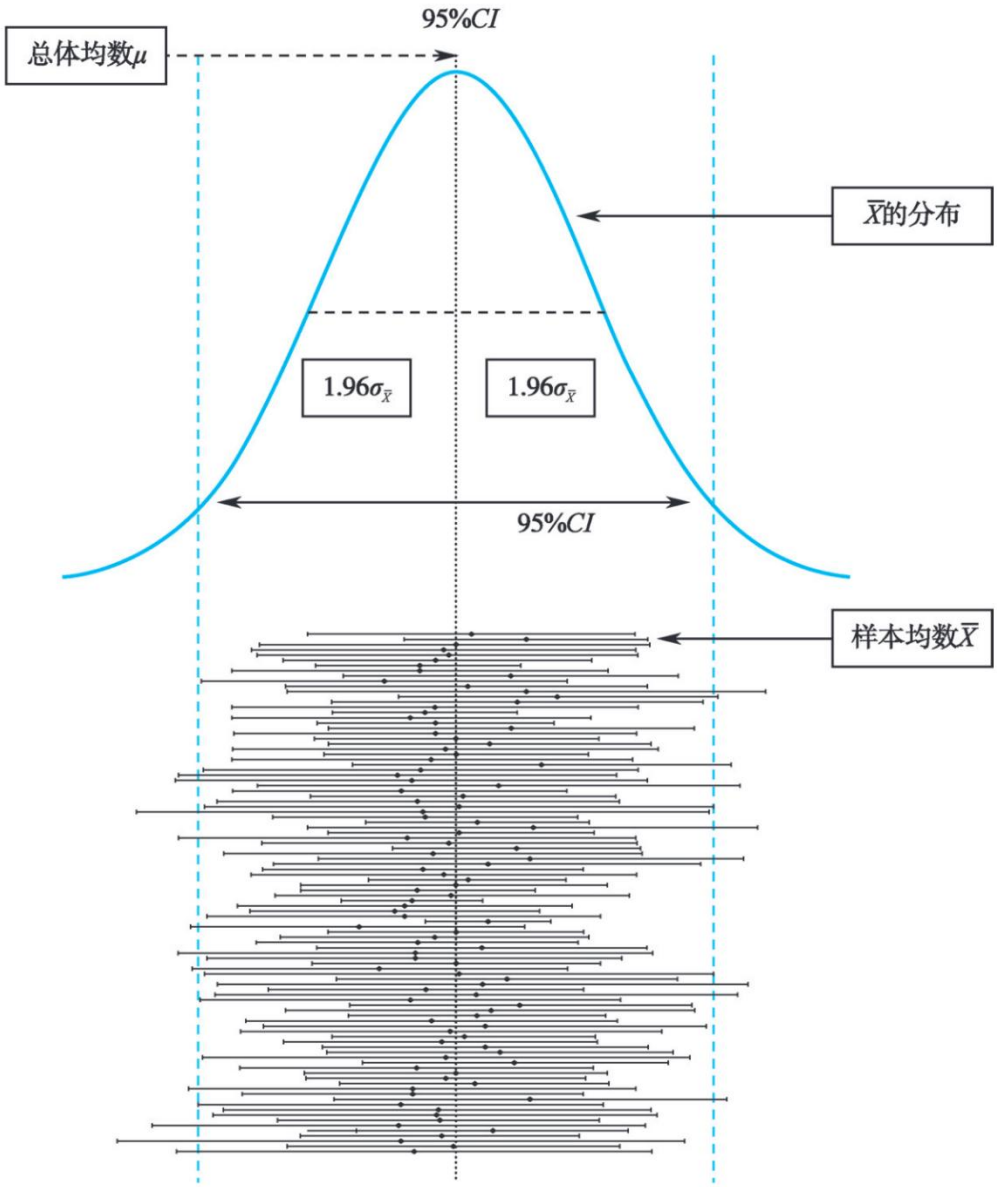

图 6-1　总体均数 95%*CI*

(李康,贺佳主编, 2024, p. 58) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

（三）、σ 未知时：t 置信区间（真实世界主力）

1️⃣ 为什么要用 t 分布?

因为:

$$\frac{\bar{X}-\mu}{S/\sqrt{n}}$$

不再服从标准正态，而服从:

$$t(v=n-1)$$

2️⃣ t 分布你必须知道的 3 点

1️⃣ 以 0 为中心、对称
2️⃣ 自由度 $v=n-1$ 越小 → 尾巴越厚
3️⃣ $v \to \infty \rightarrow t \rightarrow z$

👉 小样本更“保守”，区间更宽

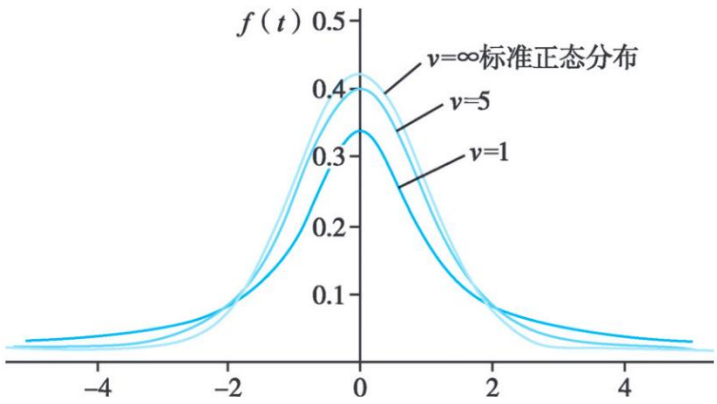

图 6-2　不同自由度的 t 分布图

(李康,贺佳主编, 2024, p. 59) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

3️⃣ 双侧置信区间公式（核心）

$$\mu \in (\bar{X} - t_{\alpha/2,v}\, S_{\bar{X}}, \bar{X} + t_{\alpha/2,v}\, S_{\bar{X}})$$

其中：

- $v = n - 1$
- $t_{\alpha/2,v}$：查表得到

🔽 使用前提（考试爱考）

- **小样本（n<50）**：
  - 原始变量 X **应近似正态**
- **大样本（n≥50）**：
  - 可直接用 z 近似

（四）、两个例题“真正想教你的是什么”

【例 6-2】小样本 + σ 未知 → t 区间

【例 6-2】某医生测得 25 名动脉粥样硬化患者血浆纤维蛋白原含量的均数为 3.32g/L，标准差为 0.57g/L，试计算该种病人血浆纤维蛋白原含量总体均数的 95% 置信区间。

本例 $n=25$，$\bar{X}=3.32$，$S=0.57$，$\nu=n-1=25-1=24$，$\alpha=0.05$，查 t 值表，$t_{0.05/2,24}=2.064$，按公式（6-12）计算得：

下限：$\bar{X}-t_{\alpha/2,\nu}S_{\bar{X}}=3.32-2.064\times0.57/\sqrt{25}=3.085$（g/L）

上限：$\bar{X}+t_{\alpha/2,\nu}S_{\bar{X}}=3.32+2.064\times0.57/\sqrt{25}=3.555$（g/L）

根据该资料计算得到，动脉粥样硬化病人血浆纤维蛋白原含量总体均数的 95% 置信区间为 3.085~3.555g/L。

(李康,贺佳主编, 2024, p. 59) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

- n = 25 → 小样本
- σ 未知 → 用 S

- 查 $t_{0.025,24} = 2.064$

得到：

$$95\%CI = (3.085,3.555)g/L$$

📌 **含义不是“患者大多在这个范围”，**
而是：

> **总体均数 μ 在这个区间内是 95% 可靠的估计。**

【例 6-3】大样本 → z 近似

> 【例 6-3】试计算例 6-1 中该地成年男性红细胞总体均数的 95% 置信区间。
> 本例属于大样本，可采用正态近似的方法计算置信区间（$\alpha=0.05$）。因为 $\bar{X}=4.77, S=0.38, n=140$，则 95% 置信区间为
> 下限：$\bar{X}-z_{\alpha/2}S_{\bar{X}}=4.77-1.96\times0.38/\sqrt{140}=4.707(\times10^{12}/L)$
> 上限：$\bar{X}+z_{\alpha/2}S_{\bar{X}}=4.77+1.96\times0.38/\sqrt{140}=4.833(\times10^{12}/L)$
> 估计该地成年男性红细胞总体均数的 95% 置信区间为 $4.707\times10^{12}/L\sim4.833\times10^{12}/L$。

(李康,贺佳主编, 2024, p. 60) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

- n = 140 ≥ 50
- 用：

$$\bar{X} \pm 1.96\frac{S}{\sqrt{n}}$$

得到一个更窄的区间。

📌 **关键直觉**：

> 样本量↑ → 标准误↓ → 置信区间变窄

（五）、单侧置信区间（什么时候才用？）

核心前提

> **研究问题本身就是单向的。**

不是“事后只看一边”，而是**研究设计阶段就只关心一侧。**

典型场景

- 是否**优于**标准?
- 是否**至少达到**某阈值?

单侧区间公式（以 t 为例）

只改两点：

1️⃣ 用 **单侧界值** $t_{\alpha,v}$

2️⃣ 只算 **下限或上限**

$$\text{下限} = \bar{X} - t_{\alpha,v} S_{\bar{X}}$$

例子的真正逻辑

- 关心：平均降压是否 **>10 mmHg**
- 只看 **下限**
- 下限 = 11.770 mmHg > 10

👉 **结论成立**

（六）、这一节你必须真正记住的 6 句话

1. 置信区间 = 点估计 ± 临界值 × 标准误
2. σ 已知 → z；σ 未知 → t
3. 自由度小 → t 大 → 区间宽
4. 大样本 → t ≈ z
5. 区间宽度反映**精密度**
6. 单侧区间只能在**研究设计允许**时用

四、两总体均数差值的区间估计

> **两独立样本均数差的置信区间 = 均数差 ± 临界值 × 标准误**
> 小样本 → **t + 合并方差**
> 大样本 → **z + 分别方差**

（一）研究目的与适用场景

目的：
估计两个总体均数之差 $\mu_1 - \mu_2$ 的可能取值范围，用区间而不是单点判断差异大小。

常见应用：

- 试验组 vs 对照组
- 两种药物或两种处理方法的平均效果比较

（二）基本思想与总体框架

两总体均数差值的置信区间统一形式为：

$$(\bar{X}_1 - \bar{X}_2) \pm (\text{临界值}) \times (\text{标准误})$$

其中：

- $\bar{X}_1 - \bar{X}_2$：样本均数差（点估计）
- 标准误：反映抽样波动
- 临界值：由置信度和分布类型决定（t 或 z）

（三）小样本且两总体方差相等时的区间估计

1️⃣ 置信区间公式

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2,\,v}\, S_{\bar{X}_1 - \bar{X}_2}$$

其中自由度为：

$$v = n_1 + n_2 - 2$$

2️⃣ 均数差的标准误计算

合并方差：

$$S_c^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

均数差的标准误：

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{S_c^2 (\frac{1}{n_1} + \frac{1}{n_2})}$$

📌 **适用前提**：

- 两总体近似正态
- 方差齐性
- 样本量较小

（四）大样本时的区间估计（近似法）

当 $n_1, n_2 \geq 50$ 时：

- 用 $z_{\alpha/2}$ 代替 $t_{\alpha/2,v}$
- 不再合并方差

标准误为：

$$S_{\bar{X}_1-\bar{X}_2}=\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

（五）例 6-4 的计算与结果

【例 6-4】评价复方缬沙坦胶囊与缬沙坦胶囊对照治疗轻中度高血压的有效性，将 102 名患者随机分为两组，其中试验组和对照组分别为 51 例和 51 例。经六周治疗后测量收缩压，试验组平均下降 15.77mmHg，标准差为 13.17mmHg；对照组平均下降 9.53mmHg，标准差为 13.55mmHg。试估计两组收缩压平均下降差值的 95% 置信区间。

由公式（6-15）和公式（6-16）计算

$$S_c^2=\frac{(51-1)\times13.17^2+(51-1)\times13.55^2}{51+51-2}=178.526$$

$$S_{\bar{X}_1-\bar{X}_2}=\sqrt{178.526\times\left(\frac{1}{51}+\frac{1}{51}\right)}=2.646$$

查附表 2 的 t 界值表得 $t_{0.05/2,100}=1.984$，由公式（6-14）算得两组总体均数之差的 95% 置信区间为

下限：$(\bar{X}_1-\bar{X}_2)-t_{\alpha/2}S_{\bar{X}_1-\bar{X}_2}=(15.77-9.53)-1.984\times2.646=0.990$（mmHg）

上限：$(\bar{X}_1-\bar{X}_2)-t_{\alpha/2}S_{\bar{X}_1-\bar{X}_2}=(15.77-9.53)+1.984\times2.646=11.490$（mmHg）

即两组收缩压平均下降差值的 95% 置信区间为 0.990~11.490mmHg。

(李康,贺佳主编, 2024, p. 60) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

（六）置信区间的统计学解释

- 95% 置信区间表示：按该方法重复抽样，**95% 的区间会包含真实均数差**
- 区间 **不包含 0**，说明两组平均降压效果存在统计学差异
- 且试验组的平均降压幅度更大

五、总体率的区间估计

（一）研究目的与基本概念

总体率（π）的点估计：
直接用样本率

$$p = \frac{X}{n}$$

作为总体率 π 的估计值。

问题在于：
点估计 **不反映抽样误差**，在医学研究中解释力不足，因此更常用 **总体率的区间估计。**

（二）小样本率的区间估计（二项分布法）

1 适用条件

- 样本量较小（通常 $n \leq 50$）
- 总体率未知，不能使用正态近似
- 基于 **二项分布的精确区间估计**

2 计算方法

- 给定置信度 $1-\alpha$（通常取 95%）
- **直接查附表 6（百分率的置信区间表）**
- 不需要代公式计算

3 例 6-5 的理解

【例 6-5】采用某康复疗法治疗脑卒中后吞咽功能障碍患者 38 例，治疗 1 个月后经评定吞咽功能障碍改善人数 14 例，求该康复疗法治疗 1 个月吞咽功能改善率的 95% 置信区间。

查附表 6，在 $n=38$，$X=14$ 的纵横交叉处的数值上行 22~54，下行 18~59，即吞咽功能改善率 95% 的置信区间为 22%~54%。

注意：附表 6 中的 X 只列出 $X \leqslant n/2$ 部分，当 $X > n/2$ 时，应以 $n-X$ 值查表，然后用 100 减去查表得到的数值，即为所求的置信区间。

(李康,贺佳主编, 2024, p. 61) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

已知：

- $n = 38$，改善人数 $X = 14$
- 样本率 $p = 14/38$

查表：

- 在 $n = 38, X = 14$ 对应位置
- 得到区间：$22\% \sim 54\%$

即：**吞咽功能改善率的 95% 置信区间为 $22\% \sim 54\%$。**

4️⃣ 特别注意（考试爱考）

> 附表 6 只列出 $X \le n/2$ 的情况。

- 若 $X > n/2$：
  - 用 $n - X$ 查表
  - 再用 100% 减去上下限
  - 得到原来的置信区间

（三）大样本率的区间估计（正态近似法）

1️⃣ 适用条件（必须同时满足）

- 样本量较大
- $np \ge 5$ 且 $n(1-p) \ge 5$

此时样本率 p **近似服从正态分布。**

2️⃣ 置信区间公式

$$p \pm z_{\alpha/2} S_p$$

其中：

$$S_p = \sqrt{\frac{p(1-p)}{n}}$$

当 $\alpha = 0.05$ 时：

$$z_{0.025} = 1.96$$

---

3 例 6-6 的计算结构

【例 6-6】某区疾病预防控制中心对该乡镇 250 名小学生进行贫血的检测，结果发现有 86 名贫血者，检出率为 34.40%，求贫血检出率 95% 的置信区间。

本例 $n=250$ 较大，且 $np=86$，$n(1-p)=164$ 均大于 5，可用公式（6-17）计算总体率 95% 的置信区间。

$$p \pm z_{\alpha/2} S_p = p \pm z_{0.05/2} \sqrt{\frac{p(1-p)}{n}} = 0.344\,0 \pm 1.96 \sqrt{\frac{0.344\,0(1-0.344\,0)}{250}} = (0.285, 0.403)$$

即该乡镇小学生贫血检出率 95% 的置信区间为 28.511%~40.289%。

(李康,贺佳主编, 2024, p. 61) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

已知：

- $n = 250$
- 贫血人数 $X = 86$
- 样本率 $p = 0.3440$

条件检验：

$$np = 86, n(1-p) = 164 (均 > 5)$$

→ 可用正态近似法。

区间计算：

$$0.3440 \pm 1.96 \sqrt{\frac{0.3440(1-0.3440)}{250}}$$

得到：

$$(0.285, 0.403)$$

即：**贫血检出率的 95% 置信区间为 $28.5\% \sim 40.3\%$。**

（四）总体率区间估计的解释要点

- 置信区间反映的是 **估计方法的可靠性**
- 不是“真实率有 95% 概率在区间内”
- 而是：**长期重复抽样中，95% 的区间会包含真实总体率**

（五）考试速记对照（非常重要）

情况	方法	工具
小样本率	二项分布法	查附表 6
大样本率	正态近似法	$p \pm zS_p$
判断标准	是否可近似正态	$np, n(1-p) \geq 5$

六、两总体率差值的区间估计

两总体率差值的区间估计 = 两样本率之差 ± 正态临界值 × 合并标准误

1. 只能用于 **大样本率**
2. 标准误使用 **合并率 p_c**
3. 区间是否包含 0 → 判断是否有差异

（一）研究目的与基本问题

研究目的：
估计两个总体率之差

$$\pi_1 - \pi_2$$

的可能取值范围，用于判断两组事件发生概率是否存在差异及差异大小。

典型应用：

- 新药 vs 常规药的有效率比较
- 干预组 vs 对照组的发生率差异

（二）适用条件（正态近似前提）

当满足以下条件时，可采用 **正态近似法**：

- 两组样本量均较大
- 且同时满足

$$n_1 p_1, n_1(1-p_1), n_2 p_2, n_2(1-p_2) \geq 5$$

📌 本质：两样本率及其差值 **近似服从正态分布。**

（三）置信区间的计算公式

1️⃣ 基本形式

$$(p_1 - p_2) \pm z_{\alpha/2}\, S_{p_1-p_2}$$

2️⃣ 两率差的标准误

$$S_{p_1-p_2} = \sqrt{p_c(1-p_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

其中 **合并率** 为：

$$p_c = \frac{X_1 + X_2}{n_1 + n_2}$$

- X_1, X_2：两组中事件发生例数
- n_1, n_2：两组样本量

📌 注意：
这里使用 **合并率 p_c**，而不是分别用 p_1, p_2。

（四）例 6-7

【例 6-7】某医院口腔科医生用某新药治疗牙本质过敏症，以某常规药作对照，进行了 1 年的追踪观察，结果见表 6-1 所示，试估计两组有效率差别 95% 的置信区间。

表 6-1　治疗牙本质过敏症两组有效率的比较

组别	总牙数	有效数	有效率/%
试验组	77	61	79.22
对照组	69	38	55.07
合计	146	99	67.81

(李康,贺佳主编, 2024, p. 61) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

本例：

$$p_c=\frac{X_1+X_2}{n_1+n_2}=\frac{61+38}{77+69}=0.678\ 1$$

$$S_{p_1-p_2}=\sqrt{p_c(1-p_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}=\sqrt{0.678\ 1\times(1-0.678\ 1)\times\left(\frac{1}{77}+\frac{1}{69}\right)}$$

$$=0.077\ 5$$

两组总体率差别 95% 的置信区间为

$$(p_1-p_2)\pm z_{\alpha/2}S_{p_1-p_2}=(0.792\ 2-0.550\ 7)\pm 1.96\times 0.007\ 5$$

$$=(0.089\ 6,0.393\ 4)$$

即两组治疗效果的总体有效率之差的 95% 置信区间为 8.96%~39.34%。

(李康,贺佳主编, 2024, p. 62) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

（五）结果表达与统计学解释

- 两组总体有效率差的 **95% 置信区间** 为

$$8.96\% \sim 39.34\%$$

- 区间 **不包含 0**
  → 两组有效率存在统计学差异
- 新药有效率 **显著高于** 常规药

第二节 假设检验

一、基本原理

（一）、假设检验是什么（What)

假设检验（Hypothesis Test / Significance Test)
是一种：

> **用样本 → 概率性判断 → 总体是否真的存在差别**的方法。

它回答的不是：

- “这两个数哪个大？”

而是：

- **“我看到的差别，是不是只是抽样误差？”**

（二）、为什么不能直接比均数（Why)

现实问题的对象是**总体**，但我们永远只能拿到**样本。**

即便：

- 甲药：均数下降 1.36
- 乙药：均数下降 1.12

也不能直接说甲药更好，因为：

- 下一次抽样，结果可能反过来
- 差别可能只是**随机波动（sampling error）**

👉 所以核心问题是：

> **观察到的差别，能不能用“抽样误差”合理解释？**

（三）、假设检验的底层逻辑（How it works）

1️⃣ 核心思想（两点）

① 小概率思想（small probability）

> 在某个假设成立的前提下，如果样本结果“极不可能出现”，那就怀疑这个假设。

② 反证法（proof by contradiction）

> 先假定“没有差别”，再看数据是否把这个假设“逼到绝境”。

2️⃣ 假设检验的“默认立场”

先假设：

$$H_0: \mu = \mu_0$$

也就是：

> 样本来自一个“和标准一样”的总体

然后问一句非常关键的话：

> **如果 H_0 真的成立，现在这组样本出现得合理吗？**

（四）、例 6-8 的逻辑拆解（Example → 思维）

【例 6-8】某研究者从某工厂从事铅作业男性工人中随机抽取了 25 人，测量了血红蛋白含量，结果如下。问该厂从事铅作业男性工人的血红蛋白是否不同于正常成年男性血红蛋白平均值 140g/L？

148.22	123.13	158.44	140.15	166.69	171.66	118.11	113.48
141.88	77.81	136.83	121.74	110.56	121.87	140.24	147.00
111.04	78.14	101.55	127.16	116.24	144.57	119.18	147.44
136.28							

(李康,贺佳主编, 2024, p. 62) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

🎯 研究问题

铅作业男性工人的血红蛋白
是否 **不同于** 正常成年男性 140 g/L？

📌 已知信息

- 样本量：n = 25
- 样本均数：$\bar{X} = 128.78$
- 假设总体均数：$\mu_0 = 140$
- 总体方差未知
- 变量近似正态分布

👉 **自然进入：单样本 t 检验（one-sample t test）**

🧠 核心判断链条

1️⃣ **原假设（H_0）**

$$\mu = 140$$

2️⃣ 在 H_0 成立前提下，构造：

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

3️⃣ t 的含义不是“差多少”

而是：

> **“这个差别，离 0 有多远？”**

📏 决策逻辑

- |t| **很小**
  → 抽样误差能解释
  → **不能拒绝 H_0**
- |t| **很大**
  → 在 H_0 成立下几乎不可能
  → **拒绝 H_0**

👉 判断标准：**是否超过事先规定的 t 临界值（critical value）**

（五）、假设检验的本质一句话版（重点）

> **假设检验 = 判断“差别”是随机的，还是结构性的。**

不是证明：

- “两个总体一定不同”

而是判断：

- **“在‘总体相同’这个假设下，这个样本是否难以接受”**

（六）、推广结论（别只记 t）

- t 检验：均数问题

- F 检验：方差 / 回归
- χ^2 检验：分类数据

统计量不同，但步骤完全一致：

1. 提出 H_0
2. 构造统计量
3. 确定分布
4. 比较临界值 / P 值
5. 决策是否拒绝 H_0

（七）、你现在应该形成的“直觉”

- 假设检验不是在“找差异”
- 而是在**给“差异”定罪或洗清嫌疑**

差别 ≠ 有意义
只有“难以用随机解释的差别”，才值得你认真对待。

二、假设检验的基本步骤

> **一句总纲**：
> 假设检验不是算数，是**按规则做判断。**

（一）建立假设 & 确定检验水准

这是“立场选择”，不是计算问题。

1️⃣ 两个假设，必须“互斥且完备”

- **原假设（H_0, null hypothesis）**

  默认成立、被质疑的对象
  通常是“**无差别 / 相等**”

$$H_0: \mu = \mu_0$$

- **备择假设（H_1, alternative hypothesis）**

  一旦 H_0 被拒绝，就接受它
  通常是“**有差别 / 不相等 / 更大 / 更小**”

$$H_1: \mu \neq \mu_0 (\text{双侧})$$

📌 **关键直觉**：
你不是在“证明 H_1”，
而是在**尝试推翻 H_0。**

2️⃣ 单侧 vs 双侧：不是数学问题，是科学问题

- **双侧检验**：

$$H_1: \mu \neq \mu_0$$

  → 不预设方向
  → 教材默认、最保守

- **单侧检验**：

$$H_1: \mu > \mu_0 \text{或} \mu < \mu_0$$

  → 必须**事前**有明确专业依据
  → 不能“算完再挑方向”

📌 **一句狠话**：

  单侧检验是“方向性承诺”，不是“为了更容易显著”。

3️⃣ 检验水准 α：你愿意冒多大的错？

- **α（significance level）**
  = 在 H_0 真实成立时，被你错杀的概率上限

常用：

- $\alpha = 0.05$
- $\alpha = 0.01$

📌 **一定要记清**：
α **不是** H_0 成立的概率
α 是：**你愿意承担的“误判风险”**

（二）选择检验方法 & 计算检验统计量

这是“工具选择 + 按假设算数”。

1️⃣ 检验方法取决于三件事

- 资料类型（均数 / 构成比 / 方差）
- 研究设计（单样本 / 两独立样本 / 配对）
- 推断目的（是否比较、是否有方向）

👉 所以才有：

- t 检验
- z 检验
- χ^2 检验
- F 检验

2️⃣ **检验统计量的“灵魂前提”**

所有检验统计量：

> **都是在 H_0 成立的前提下计算的**

以 t 为例：

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

⚠️ 注意一个常被忽略的点：

- 分子减的是 **μ_0（假设值）**
- 而不是未知的真实 μ

👉 这意味着：

> t 统计量本身就是在“假设 $\mu = \mu_0$ 成立”的世界里构造出来的。

（三）用 P 值做出统计推断

这是“裁决阶段”，不是再算一遍。

1️⃣ P 值到底是什么（非常重要）

定义（准确版）：

> **P 值 = 在 H_0 成立的前提下，**
> **得到“当前样本结果或更极端结果”的概率**

关键词：

- 前提：**H_0 成立**
- “更极端”：**朝着 H_1 指向的方向**

2️⃣ P 值 ≠ H_0 成立的概率（必须死记）

❌ 错误理解：

> P 小 → H_0 不成立的概率大

✅ 正确理解：

P 小 → 如果 H_0 真的成立，这个结果很难出现

所以逻辑是：

- **小概率事件原则**
- 不是“算概率真假”，而是“是否值得怀疑”

3️⃣ 决策规则（考试直接用）

以双侧 t 检验为例：

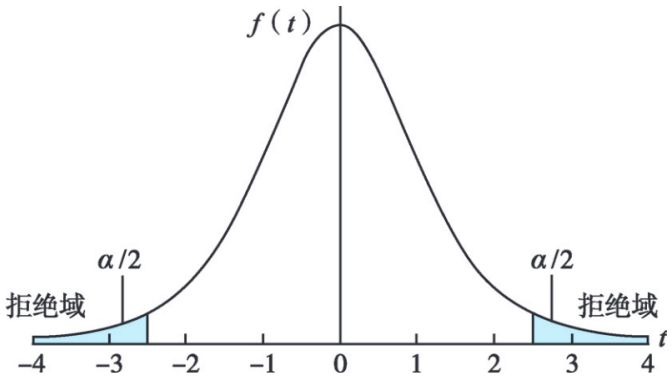

图 6-3　利用 t 分布进行假设检验原理示意图

(李康,贺佳主编, 2024, p. 64) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

- 若

$$P \leq \alpha (\text{或} \mid t \mid \geq t_{\alpha/2,\nu})$$

→ 拒绝 H_0，认为差异有统计学意义

- 若

$$P > \alpha$$

→ **不能拒绝 H_0**

📌 **关键表述**：

不能拒绝 H_0 ≠ 接受 H_0
只是：**证据不足**

4️⃣ 单侧 vs 双侧的“隐含关系”

- 同一数据、同一 α：
  - 单侧检验的界值 **更小**
  - 更“容易”拒绝 H_0

因此：

双侧能拒绝 → 单侧一定能拒绝
反之不成立

四、统计结论的规范表达（现实很重要）

- 只能说：
  “差异有统计学意义 / 无统计学意义”
- 不要说：
  ❌ “两总体相同 / 不同”
- 软件输出：
  - P = 0.000 → 写 **P < 0.001**
  - 最好报告**精确 P 值**

五、这一节真正的“思想核心”

假设检验的出发点是：
能不能安全地拒绝“无差别”这个假设

而不是：

- 去估计 H_0 成立的概率
- 去证明两总体“真的不同”

统计学永远更擅长证伪，而不是证明。

第三节 假设检验中的两类错误

> **一句总纲**：
> 假设检验不是对错判断，而是**在不确定性下做取舍。**

一、为什么一定会出错（先立现实）

假设检验是用**有限样本**去判断**无限总体。**
只要有抽样误差，就**不存在“零风险决策”**。

你不可能做到：

- 永远不错判
- 永远不漏判

你能做的，只有：
👉 **控制哪种错误更重要、概率多大**

二、两类错误的定义（必须极其清楚）

表 6-2　统计推断的两类错误及其概率

真实情况	假设检验结论	
	拒绝 H_0	不拒绝 H_0
H_0 成立	Ⅰ类错误（α）	推断正确（$1-\alpha$）
H_1 成立	推断正确（$1-\beta$）	Ⅱ类错误（β）

(李康,贺佳主编, 2024, p. 65) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

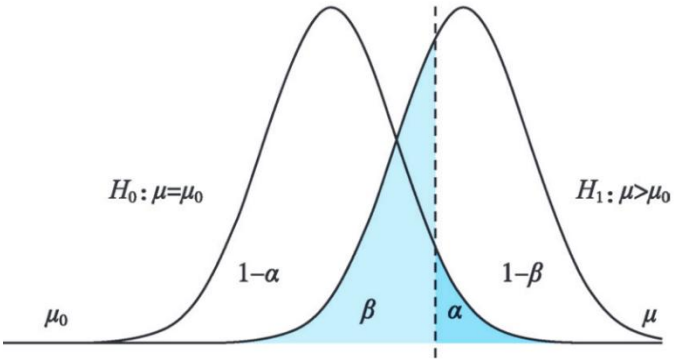

图 6-4　I类错误和II类错误示意图

(李康,贺佳主编, 2024, p. 65) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

1️⃣ I 类错误（Type I Error，α）

定义：

> **H_0 实际成立，却被你拒绝了**

通俗说法：

- **把“没差别”错判成“有差别”**
- **错杀好人**

数学符号：

$$P(\text{拒绝}H_0 \mid H_0\text{为真}) = \alpha$$

📌 关键点：

- α **不是算出来的**
- α 是你**事先设定并承诺承担的风险上限**

2️⃣ II 类错误（Type II Error，β）

定义：

> **H_0 实际不成立，却没被你拒绝**

通俗说法：

- **真的有差别，但你没看出来**
- **放过坏人**

数学符号：

$$P(\text{不拒绝}H_0 \mid H_0\text{为假}) = \beta$$

📌 关键点：

- β **通常算不准**
- 因为真实总体分布在 H_0 不成立时往往未知

三、假设检验在“偏向谁”（非常重要）

🎯 核心结论

> **经典假设检验，优先控制 I 类错误（α）**

为什么？

- 理论基础：**小概率原理**
- 逻辑立场：

> “如果在 H_0 成立下，这种结果几乎不可能发生，
> 那我才拒绝 H_0。”

所以：

- α 是“硬性控制”

- β 往往是“被动承受”

四、用单侧 t 检验建立直觉

设：

$$H_0: \mu = \mu_0 H_1: \mu > \mu_0$$

情形 1：I 类错误（α）

- H_0 **真的成立**
- 但由于抽样误差：

$$t \geq t_{\alpha,\nu}$$

- 你拒绝了 H_0，得出“$\mu > \mu_0$”

👉 **I 类错误，概率 ≤ α**

情形 2：II 类错误（β）

- H_0 **实际上不成立（$\mu > \mu_0$）**
- 但样本偏小：

$$t < t_{\alpha,\nu}$$

- 你没拒绝 H_0

👉 **II 类错误，概率 = β**

五、α 和 β 的此消彼长关系（重点）

⚠️ 一个残酷事实

在样本量固定时：

- α↓ → β↑
- α↑ → β↓

也就是说：

- 你越谨慎不“错杀”（α 小）
- 就越容易“漏判”（β 大）

📌 这不是计算问题，是**统计学的物理极限。**

六、那有没有“两全其美”的办法?

✅ 唯一正道：**增加样本量**

- 样本量↑
  → 抽样分布变窄
  → 拒绝域与接受域分离更清晰

结果是：

- α↓
- β↓

📌 **重要直觉**：

严谨不是靠“调 α”，
而是靠“多收数据”。

七、检验效能（Power）

检验效能 = 在真实存在差异时，你能把它检验出来的能力。

1️⃣ 正式定义

检验效能（power）：

$$\text{Power} = 1 - \beta$$

含义是：**当 H_0 实际不成立时，按既定显著性水平 α，正确拒绝 H_0 的概率**

也就是说：

- β：漏判概率
- **1 − β：抓住真相的概率**

📌 你原文里已经把 β 讲得非常清楚了，**power 只是换了一个“主动视角”。**

2️⃣ 为什么“效能”必须单独拎出来说？

因为：

- β 是**被动承受的失败**
- **power 是方法本身的能力指标**

换句话说：

> **α 是你的态度，power 是你的实力**

教材为什么要强调它？因为在实际研究中，大家真正关心的是：

- “我这个实验**有没有能力**发现真实差异？”
- “是不是我方法不行 / 样本不够，而不是‘没差别’？”

3️⃣ 一个你已经暗示、但没点破的关键事实

在第六节其实已经说了这一点：

> 增加样本量 → α↓，β↓

等价地说：

> **样本量↑ → power↑**

📌 所以可以直接补一句极重要的总结：

> **power 不是靠“放宽 α”换来的，而是靠设计与样本量撑起来的。**

4️⃣ 不同检验方法，power 本身就不同（你提到但没展开）

你引用的教材那段话，其实在说一个**高级但非常重要的点**：

> 在相同 α、相同真实差异 Δ ≠ 0 的情况下
> **不同统计方法的检验效能是不同的**

直觉解释：

- 方法 1 的抽样分布更集中 / 噪声更小
- → 在同样阈值下，更容易落入拒绝域
- → **power 更高，样本更省**

这也是为什么：

- 配对 t 检验 > 两独立样本 t 检验
- 方差分析合理建模 > 乱做多次 t 检验

📌 **这是“方法论优劣”的统计学标准，而不只是显不显著。**

5️⃣ 教材那个 0.80 的例子，应该这样理解

> 若 $1 - \beta = 0.80$
> 表示在 H_0 不成立的前提下
> 100 次实验中，理论上有 80 次能拒绝 H_0

真正的含义是：

- 你不是“保证发现差异”
- 而是**承认：我还有 20% 的概率会看不见真相**

📌 所以 power 本质上是：**“我是否有资格得出‘没发现差异’这句话”**

八、这一节你必须形成的终极认知

1️⃣ 假设检验 = 风险决策

不是：

- “判断谁对谁错”

而是：

- **在有限证据下，选择哪种错误更可接受**

2️⃣ α 是态度，不是结果

- α = 0.05
  ≠ “结果可信度 95%”
- α = 0.05
  = “我接受最多 5% 的错杀风险”

3️⃣ “不显著” ≠ “没差别”

- 很多真实差异
- 只是被 β 吞掉了

第四节 假设检验与区间估计的关系

假设检验回答“要不要动手”，区间估计回答“动多少才算数”。

一、它们在干什么（角色分工）

1️⃣ 假设检验（Hypothesis Test）

👉 本质是一个**是 / 否的决策工具**

- 问题形式：**“这个差别存在吗？”**
- 输出结果：
  - 拒绝 / 不拒绝 H_0
  - 或一个 **P 值**
- 特点：
  - 操作简单
  - 决策明确
  - **但信息很粗**

📌 你可以把它理解为：**“红灯 / 绿灯系统”**

2️⃣ **区间估计（Confidence Interval）**

👉 本质是一个**量级 + 不确定性的描述工具**

- 问题形式：**“这个差别大概有多大？可能落在哪？”**
- 输出结果：
  - 一个区间（如 (a, b)）
- 特点：
  - 有方向
  - 有大小
  - 能讨论 **实际意义**

📌 更像：**“带误差条的仪表盘”**

二、为什么它们对同一数据能给出“同样的结论”

关键等价关系

在同一 α 水平下：

- 若 $1-\alpha$ **置信区间不包含 0**
  👉 等价于 **拒绝 H_0：差值 = 0**

换句话说：

- **区间不跨 0** → 差别“站得住”
- **区间跨 0** → 差别“不稳”

📌 所以：

- 假设检验 ≈ **“有没有差别”**
- 区间估计 ≈ **“差别站在哪个范围”**

三、P 值为什么“好用但危险”

教材这一段，其实在提醒你一个**非常现实的问题。**

P 值的优点

- 给非统计专业者一个：

  简单、可执行的判断规则

- 适合回答：

  “我现在要不要拒绝 H_0？”

P 值的致命问题（样本量一大就暴露）

样本量足够大时，任何微小差别都能让 P 值变得很小。

结果是：

- **统计学意义** ✓
- **实际意义** ❌

📌 这时：

- 假设检验会说：
  👉 “有显著差异”
- 但你不知道：
  👉 “这差异有没有用？”

四、区间估计为什么更“像科学判断”

区间估计能多回答一层问题

不仅是：“有没有差别？”

而是：**“这个差别在数量级上，值不值得在乎？”**

例如：

- 血清甘油三酯下降：
  - 统计上显著 ✓
  - 但区间显示下降极小
    👉 **可能没有临床价值**

📌 所以教材那句话的真实含义是：

> **假设检验只判断“存在性”，**
> **区间估计才能判断“重要性”。**

五、图 6-6 想告诉你的真正信息（不用看图也能懂）

> **有统计学意义 ≠ 有实际意义**

- 研究 1、2、3：
  - 都显著
- 但只有研究 1：
  - 区间大小达到 **专业上“有用”的阈值**

📌 这就是为什么：

真正严肃的结论，必须“检验 + 区间”一起报。

六、给你一个“以后永远不会乱”的终极对照表

你关心的问题	用谁
有没有差别	假设检验
差别大不大	区间估计
要不要拒绝 H_0	假设检验
差别有没有实际价值	区间估计
给决策者一个明确结论	假设检验
给专业判断足够信息	区间估计

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