Introduction to Hypothesis Testing

Applied Statistics

Fall 2025

目录

Outline

1. 假设检验基础 (Basics about hypothesis testing)
    - 构建假设 (Constructing hypotheses)
    - 检验统计量 (Test Statistics)
    - 计算 P 值 (Calculating P-value)
    - 统计显著性 (Statistical Significance)
2. 错误类型与检验功效 (Types of errors and the power of the test)
3. 显著性检验的使用与滥用 (Use and abuse of tests of significance)

1 假设检验 (Hypothesis Testing)

- 假设检验是统计推断的第二种形式。
- 它评估数据中关于总体某个主张的证据。
- 该主张是关于总体参数的陈述，例如 μ：”总体均值是 45 吗？”
- 结论取决于数据与主张之间的不一致程度。

1.1 显著性检验的推理 (The Reasoning of Tests of Significance)

显著性检验示例

假设我们试图判断平均周薪是否等于 350。假设总体标准差为 100。假设我们抽取 25 名毕业生的样本，平均薪资为 200。基于这个样本数据，我们对该主张可以得出什么结论？

- 假设 $\mu = 350$，我们期望样本均值的抽样分布为：

$$\bar{x} \sim N(350, 100/\sqrt{25}) = N(350, 20)$$

- 观察到样本均值 $\bar{X} = 200$ 或更低的概率：

$$P(\bar{X} < 200) \text{ 小于} 0.0001$$

- 观察到的统计量如此不可能，以至于提供了令人信服的证据表明该主张不真实。
- 如果 $\bar{X}$ 接近 350（例如 355），则没有提供太多证据来推断总体均值不同于 350。

1.1.1 假设检验的推理步骤

- 假设检验从假设该主张为真开始。
- 基于该主张，我们可以推导出抽样分布的特征。
- 将数据与抽样分布进行比较：如果基于所述抽样分布，数据的证据非常不可能发生，那么我们得出结论，该主张可能不真实。

1.2 关键概念 (Critical Concepts)

- 两个假设：原假设 (null hypotheses) 和备择假设 (alternative hypotheses)。

- 两个可能的决策：
  - 得出结论：有足够证据反对原假设
  - 得出结论：没有足够证据反对原假设
- 两个可能的错误：
  - 第一类错误 α：拒绝一个真实的原假设。
  - 第二类错误 β：不拒绝一个错误的原假设。

审判类比

在审判中，陪审团或法官必须在两个假设之间做出决定。

- 原假设 H_0：被告无罪。
- 备择假设 H_1：被告有罪。

法官必须根据呈现的证据做出决定。在证据面前，任何人都可能看起来有点罪，所以证据必须超出合理怀疑才能推翻原假设。

用统计学的语言来说，定罪被告被称为拒绝原假设而支持备择假设。
也就是说，陪审团说有足够的证据得出结论被告有罪（即，有足够的证据支持备择假设）。

注意，陪审团并不是说被告是无辜的，只是说没有足够的证据支持备择假设。这就是为什么我们永远不说我们接受原假设。

有两种可能的错误：

- 第一类错误发生在我们拒绝一个真实的原假设时。即，陪审团给一个无辜的人定罪。
- 第二类错误发生在我们不拒绝一个错误的原假设时。即，一个有罪的被告被释放。

在司法系统中，第一类错误被视为更严重。我们尽量避免给无辜的人定罪。我们更愿意释放有罪的人。

假设 (Hypotheses)

- 统计检验所检验的主张称为原假设（H_0）。
- 该检验旨在评估**反对**原假设的证据的强度。

- 通常，原假设是" 无效应" 或" 无差异" 的陈述。
- 备择假设 H_a 是我们寻求证据支持的主张。

1.2.1 假设的形式

- 假设是关于总体**参数**的陈述。
- H_0 总是一个等式（$\theta = \theta_0$），其中 θ_0 是一个特定的数字。
  - 双尾检验 (Two-tailed hypothesis)：$H_0 : \theta = \theta_0$ 和 $H_a : \theta \neq \theta_0$
  - 右尾检验 (Right tail hypothesis)：$H_0 : \theta = \theta_0$ 和 $H_a : \theta > \theta_0$
  - 左尾检验 (Left tail hypothesis)：$H_0 : \theta = \theta_0$ 和 $H_a : \theta < \theta_0$
- 在薪资示例中，假设可以是 $H_0 : \mu = 350$ 和 $H_a : \mu < 350$

1.2.2 形成假设

- 假设取决于我们在进行统计显著性检验之前对问题的了解。
- 在查看数据之前形成假设（单侧或双侧，右尾或左尾），以防止 P 值操纵（P-Hacking）。

示例：香烟中的尼古丁含量

一个健康倡导组织检验某个品牌香烟的平均尼古丁含量是否大于广告值 1.4 毫克。这里，健康倡导组织怀疑香烟制造商出售的香烟尼古丁含量高于广告值，以便让消费者更上瘾，从而维持收入。因此，这是一个单侧检验：

$$H_0 : \mu = 1.4 \text{ mg} \quad \text{和} \quad H_a : \mu > 1.4 \text{ mg}$$

练习题：构建假设

构建以下情境的假设：

1. 经理想知道计算机的平均需求是否不同于 350 台。
2. 你想知道你有兴趣投资的共同基金的月回报率是否高于 1.1%。
3. 你想知道 CFA 一级考试的通过率是否低于 40%。

解答：

1. $H_0 : \mu = 350$，$H_a : \mu \neq 350$（双尾检验）
2. $H_0 : \mu = 1.1\%$，$H_a : \mu > 1.1\%$（右尾检验）
3. $H_0 : p = 0.4$，$H_a : p < 0.4$（左尾检验）

1.3 检验统计量 (Test Statistic)

- 显著性检验 (A test of significance) 基于估计参数的统计量。
- 如果原假设为真，我们期望统计量的值接近 H_0 中指定的参数值。
- 检验统计量总结样本证据：它衡量数据与原假设为真时期望的差异程度。

对于总体均值 μ 的检验，当 σ 已知时，使用 z 检验统计量：

$$z = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of the estimate}} = \frac{\text{估计值} - \text{假设值}}{\text{估计值的标准差}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

其中 μ_0 是原假设中假设的总体均值。

- 统计量的较大值表明数据与 H_0 不一致。

1.4 做出决策 (Making a Decision)

我们可以通过以下两种方法之一来决定检验统计量是否与原假设一致：

- 拒绝域方法 (The rejection region approach)（手动进行检验时使用）：将检验统计量与临界值 (critical value) 进行比较。
- P 值方法 (The p-value approach)（统计软件中使用）：计算观察到像检验统计量这样极端的结果的可能性，看概率是否足够小。

1.4.1 P 值 (P-Value)

- 原假设 H_0 陈述了我们试图反驳的主张。
- 衡量反对原假设证据强度的概率称为 P 值。
- P 值是在原假设 H_0 为真的条件下，观察到**至少与样本计算的检验统计量一样极端的检验统计量的概率**。
- 小的 P 值是反对 H_0 的证据，因为它们表明当 H_0 为真时，观察到的结果不太可能发生。
- P 值越小，反对 H_0 的证据越强。

示例：计算机需求

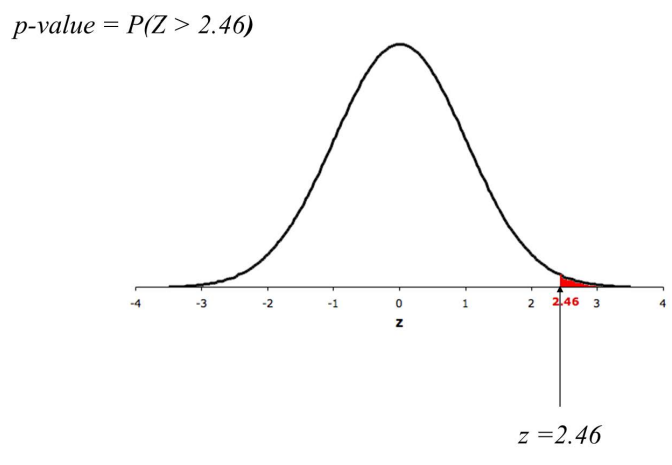
为了确定计算机月需求是否大于 170，我们收集了一个样本量为 400 的样本，均值为 178。已知总体标准差为 65。
问题：在原假设 ($H_0 : \mu = 170$) 为真的条件下，观察到至少与已观察到的样本均

值（即 $\bar{x} = 178$）一样极端的样本均值的概率是多少？

解答：

$$P(\bar{x} \geq 178) = P\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq \frac{178 - 170}{65/\sqrt{400}}\right) = P(Z \geq 2.46) = 0.0069$$

因此，检验统计量 $z = 2.46$，P 值为 0.0069。
这意味着，如果总体均值确实是 170，那么观察到样本均值至少为 178 的概率只有 0.69%。



1.5 统计显著性 (Statistical Significance)

- 基于 P 值的决策遵循的原则是：一个小概率事件" 不应该" 发生。那么，多小才算小呢？
- 将 P 值与**显著性水平 (significance level)** α 进行比较以做出决策。
- 决策规则：如果 P-值 $< \alpha$，拒绝原假设。
- 显著性水平 α 是决策阈值，由研究者确定。
- 常见的显著性水平：1%，5%，10%

练习题：解释 P 值

下表显示了三个假设检验的显著性水平 α 和 P 值。

		P 值
检验 1	0.05	0.10
检验 2	0.10	0.08
检验 3	0.10	0.05

拒绝 H_0 的证据最强的是：

1. 检验 1
2. 检验 2
3. 检验 3

解答：检验 3，因为其 P 值（0.05）小于 （0.10），且在三者中 P 值最小。

1.6 双尾检验中的 P 值

- ” 更极端” 情况的定义取决于假设。
- 在双尾检验中，更极端的情况可能在两端。
- 因此，双尾检验中的 P 值是两端概率的总和：

$$\text{P 值} = 2 \times P(Z \geq |z|)$$

其中 z 是检验统计量的实际值，$|z|$ 是其绝对值。

示例：糖饮料消费

含糖饮料的消费与体重增加和肥胖呈正相关。一份国家报告指出，青少年平均每天消费 298 卡路里，标准差为 435 卡路里。
你在你的大学调查了 100 名学生，发现平均消费为 262 卡路里。
问题：是否有证据表明你大学的平均每天卡路里消费与全国平均不同？假设显著性水平为 5%。
解答：

- 假设：$H_0 : \mu = 298$，$H_a : \mu \neq 298$（双尾检验）
- 检验统计量：

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{262 - 298}{435/\sqrt{100}} = \frac{-36}{43.5} \approx -0.828$$

- P 值：$2 \times P(Z \geq |-0.828|) = 2 \times P(Z \geq 0.828) \approx 2 \times 0.204 = 0.408$
- 由于 P 值 $= 0.408 > \alpha = 0.05$，我们无法拒绝原假设。
- 结论：没有足够的证据表明你大学的平均每天糖饮料消费与全国平均不同。

1.7 假设检验结论总结

单尾检验（左尾）	双尾检验	单尾检验（右尾）
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$
P 值 $= P(Z \leq z)$	P 值 $= 2 \times P(Z \geq \lvert z \rvert)$	P 值 $= P(Z \geq z)$

表 1: 不同假设下 P 值的计算区域

1.8 假设检验的结论 (Conclusions of a Hypothesis)

- 我们将做出两个决策之一：拒绝 H_0 或无法拒绝 H_0。
- 如果我们拒绝原假设，我们得出结论：有足够的证据推断备择假设为真。
- 如果我们不拒绝原假设，我们得出结论：没有足够的统计证据推断备择假设为真。
- 我们从不”接受”（或证明）原假设。

1.8.1 比较两种结论

- 当我们做出拒绝原假设而支持备择假设的决定时，这是支持备择假设的更强情况。
- 如果我们未能拒绝原假设，这只是缺乏证明相反的证据。
- 在单尾检验的情况下，如果我们交换原假设和备择假设，随后未能拒绝新的原假设，这种”未能拒绝”对主张的支持较弱。
- 因此，我们总是把我们要证明的主张作为备择假设。

练习题：计算 P 值

对原假设 $H_0 : \mu = \mu_0$ 的检验给出检验统计量 $z = 1.89$。

1. 如果备择假设是 $H_a : \mu > \mu_0$，P 值是多少？
2. 如果备择假设是 $H_a : \mu < \mu_0$，P 值是多少？
3. 如果备择假设是 $H_a : \mu \neq \mu_0$，P 值是多少？

解答：

1. 单尾（右尾）：$P = P(Z > 1.89) \approx 0.0294$
2. 单尾（左尾）：$P = P(Z < 1.89) \approx 0.9706$（观察到的结果不在极端方向）
3. 双尾：$P = 2 \times P(Z > \lvert 1.89 \rvert) = 2 \times P(Z > 1.89) \approx 2 \times 0.0294 = 0.0588$

1.9 单尾还是双尾检验？

- 注意，双尾假设的 P 值是相应单尾假设 P 值的两倍。
- 因此，如果双尾假设被拒绝，相应的单尾假设也保证被拒绝。
- 双尾检验是更保守的检验，也是大多数情况下的默认检验。

1.9.1 注意” 搜寻显著性”

- **” 警惕搜寻显著性 (Beware of searching for significance)”**：先检查数据，然后决定进行数据指示方向的单侧检验以寻找” 显著” 结果是错误的。
- 假设总是基于问题或理论；不应纯粹由数据驱动。
- 单尾检验不太保守，仅在有强有力的先验理论证明时才应使用。

练习题：环境保护

一家回收公司的财务分析师计算出，如果每个家庭每周平均纸张回收量超过 2.0 磅，该公司将盈利。
从一个大型社区抽取了 148 个家庭的随机样本，记录了每个家庭每周丢弃用于回收的纸张重量。
样本均值为 2.180 磅，标准差为 0.981 磅。
这些数据是否提供了足够的证据，使分析师得出结论：回收厂将盈利？
解答：

- 假设：$H_0 : \mu = 2.0$，$H_a : \mu > 2.0$（右尾检验，因为只有超过 2.0 磅才盈利）
- 检验统计量（由于 n 较大，使用 z 检验）：

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{2.180 - 2.0}{0.981/\sqrt{148}} \approx \frac{0.180}{0.0806} \approx 2.23$$

- P 值：$P(Z > 2.23) \approx 0.0129$
- 在显著性水平 $\alpha = 0.05$ 下，由于 P 值 $= 0.0129 < 0.05$，我们拒绝原假设。
- 结论：有足够的证据表明每个家庭每周平均纸张回收量超过 2.0 磅，因此回收厂将盈利。

2 错误类型与检验功效 (Types of Errors and the Power of the Test)

2.1 错误类型 (Types of Errors)

在假设检验中，存在两种可能的错误类型：

- **第一类错误 (Type I error)**：拒绝了一个真实的原假设（即拒绝了一个真的 H_0）。
- **第二类错误 (Type II error)**：没有拒绝一个错误的原假设（即没有拒绝一个假的 H_0）。

每种错误都有相应的概率：

$$P(\text{第一类错误}) = \alpha$$

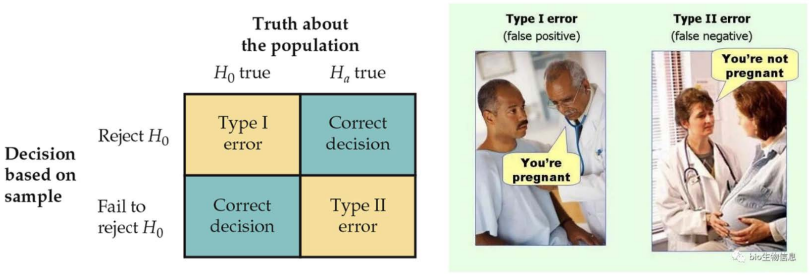
$$P(\text{第二类错误}) = \beta$$

实际上，这就是显著性水平 α（假阳性 (False Positive) 的概率）和 β（假阴性 (False Negative) 的概率）的定义。

2.2 错误类型的决策矩阵 (Decision Matrix for Error Types)

关于总体的真实情况 (Truth about the population)		**基于样本的决策 (Decision based on sample)**
H_0 为真	H_a 为真	
第一类错误 (Type I error)	正确决策 (Correct decision)	拒绝 H_0 (Reject H_0)
		假阳性 (False positive)
正确决策 (Correct decision)	第二类错误 (Type II error)	不拒绝 H_0 (Fail to reject H_0)
		假阴性 (False negative)

表 2: 假设检验中的错误类型



类比示例：怀孕测试 (Analogy: Pregnancy Test)

- **假阳性 (False positive)**：测试显示怀孕，但实际上并未怀孕。（对应第一类错误）
- **假阴性 (False negative)**：测试显示未怀孕，但实际上已怀孕。（对应第二类错误）

2.3 第二类错误的概率 (Probability of a Type II Error)

2.3.1 定义与计算思路

第二类错误发生在未能拒绝一个错误的原假设时。

回顾计算机需求的例子，我们如果检验统计量落在 $z > 1.65$ 的范围内，则拒绝 H_0 而支持 $H_1 : \mu > 170$。这等价于样本均值大于 175.34：

$$\bar{x} > 175.34.$$

当真实均值大于 170，但我们得到的样本均值 $\bar{x}$ 小于 175.34（不在拒绝域内）时，就会发生第二类错误：

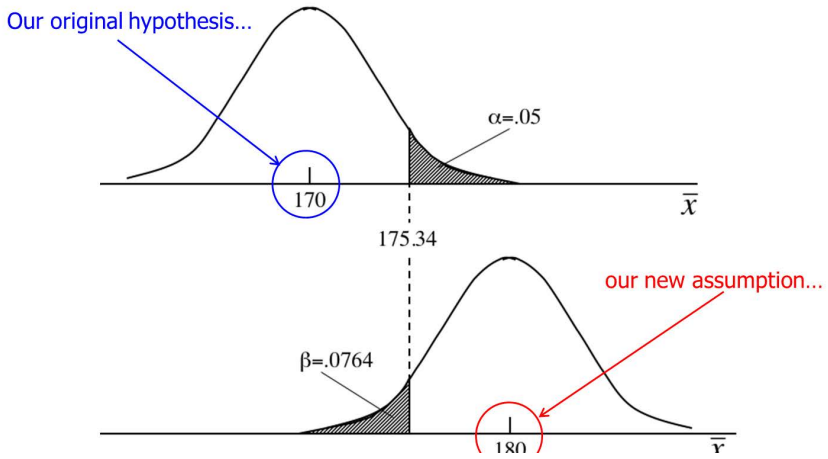
$$\beta = P(\bar{x} < 175.34 \mid \mu > 170)$$

2.3.2 计算难点

尽管我们可以在假设检验中定义第二类错误，但 β 很难计算，因为它要求 H_a 下有一个具体的真实值。

计算涉及在以 H_0 为中心的分布和以 H_a 下真实参数为中心的分布下找到临界值。

示例： 如果平均需求是 \$180，那么第二类错误的概率是



2.4 α 变化对 β 的影响 (Effects on β of Changing α)

虽然我们通常不知道 β 的实际值，但我们知道对于给定的检验，β 受 α 和 n 的影响。

- 降低显著性水平 α 会增加 β 的值，反之亦然。
- 将临界值线 (critical value line) 向右移动（以减小 α）将意味着在下方曲线下 β 的面积更大。

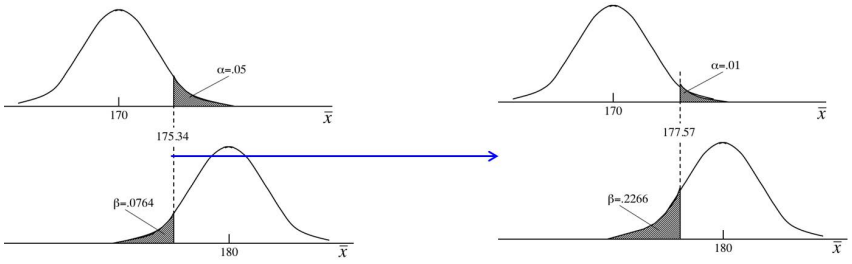


图 1: α 与 β 的权衡关系

2.5 样本量 n 变化对 β 的影响 (Effects on β of Changing n)

- 通过增加样本量 n，我们可以在不增加第一类错误概率的前提下，降低第二类错误的概率 β。
- 增加样本量 n 会减少抽样分布的变异性，从而提高统计推断 (statistical inference) 的有效性。

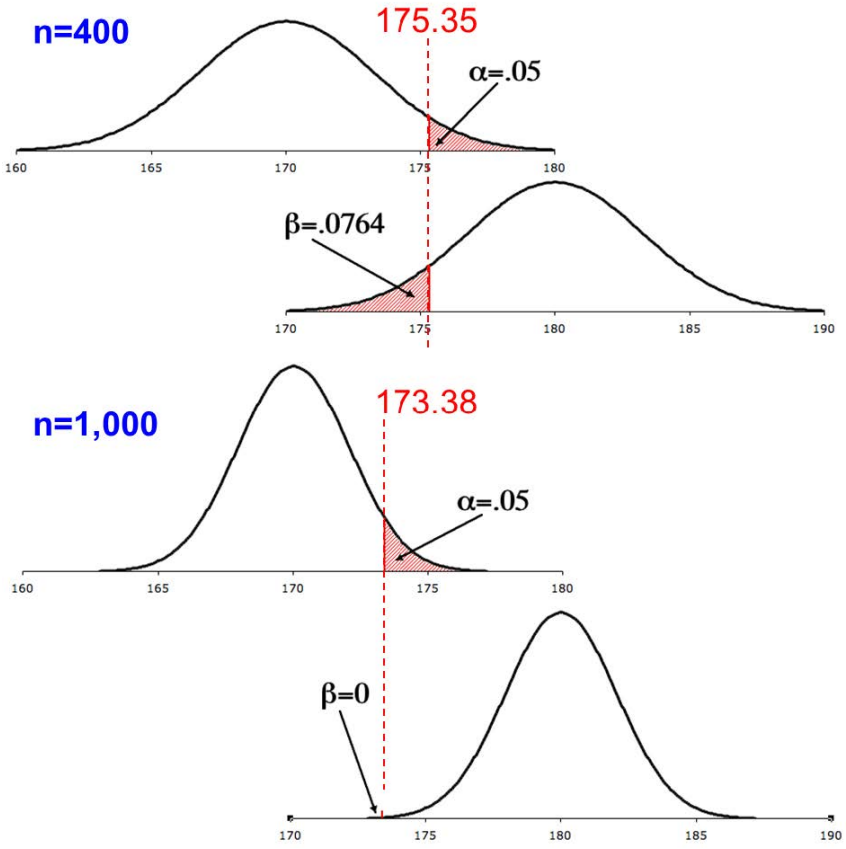


图 2: 比较 $n = 400$ 和 $n = 1000$ 时的 β

示例：比较 $n = 400$ 和 $n = 1000$ 时的 β 通常，更大的样本量会带来更小的 β，即检验更有可能检测到真实存在的效应。

2.6 检验的功效 (Power of a Test)

2.6.1 定义

检验的功效 (Power of a test)：$1-\beta$，表示当原假设为假时，正确拒绝原假设的概率。

功效是当效应真实存在时，正确检测到该效应的概率。

2.6.2 比较检验的功效

给定相同的备择假设、样本量和显著性水平，如果一个检验比另一个检验有更高的功效，则称第一个检验更有效力，是更优选的检验。

提高检验功效的方法：

- 增加样本量 n
- 增大显著性水平 α（但会增加第一类错误风险）
- 增大效应量（effect size），即 H_0 与 H_a 的参数值差异更大。

练习题：检验功效

问题 1：正确拒绝原假设的概率是：

A. p 值。

B. 检验的功效。

C. 显著性水平。

问题 2：假设检验的功效是：

A. 等价于显著性水平。

B. 不犯第二类错误的概率。

C. 不随小样本量的增加而改变。

解答：

1. B. 检验的功效定义为 $1-\beta$，即当 H_0 为假时正确拒绝它的概率。
2. B. 功效是 $1-\beta$，即不犯第二类错误的概率。选项 C 错误，因为增加样本量通常会提高功效。

练习题：Calculating β

对于假设检验：$H_0 : \mu = 22$; $H_1 : \mu < 22$; $n = 220$; $\sigma = 6$; $\alpha = 0.10$
计算当实际均值为 21 时的第二类错误概率。

1. **阶段 1：确定拒绝域**临界值计算：$\bar{x} < 21.48$（具体计算过程略）。
2. **阶段 2：计算第二类错误概率**

$$\beta = P(\bar{x} > 21.48 \mid \mu = 21)$$

在 $\mu = 21$ 的正态分布下，计算 $\bar{x} > 21.48$ 的概率。

3 显著性检验的使用与滥用 (Use and Abuse of Tests of Significance)

3.1 选择显著性水平 α (Choosing a Significance Level)

选择 α 时通常考虑的因素：

- 当原假设实际上为真时，拒绝原假设的后果是什么？
- **示例：** H_0：被告无罪；H_a：被告有罪。第一类错误（冤枉好人）的后果非常严重。
- 你正在进行一项初步研究吗？如果是，你可能想要一个更大的 α，这样你就不太可能错过一个有趣的结果。

没有免费的午餐： α 越小，当原假设实际上为假时，未能拒绝原假设的机会就越高（即 β 越大）。

3.1.1 常见做法与误解

- 报告 P 值，并且每当 $P \leq 0.05$ 时将结果描述为“统计显著”，这是一种常见做法。
- 然而，“显著”和“不显著”之间没有明显的界限，只有随着 P 值减小而逐渐增强的证据。
- P 值没有明显的分界点：4.9% 和 5.1% 之间没有实际区别。

3.2 统计显著性不代表什么 (What Statistical Significance Does Not Mean)

- 统计显著性仅说明观察到的效应是否可能仅仅是由于随机抽样造成的偶然。

- 统计显著性可能没有实际重要性：统计显著性并不告诉你效应的大小，只告诉你存在一个效应。
- 一个效应可能太小而不相关。并且，在样本量足够大的情况下，即使是最微小的效应也能达到显著性。

3.3 不要忽略不显著性 (Don't Ignore Lack of Significance)

- 未能发现结果的统计显著性意味着“不拒绝原假设”。这与实际“支持”原假设非常不同。
- 例如，样本量可能太小，无法克服总体中的巨大变异性。
- 缺乏证据不是零效应的证据：考虑围绕“零”的宽置信区间。
- 检查置信区间以说明参数可能值的范围，即使 H_0 未被拒绝。

3.4 警惕搜寻显著性 (Beware of Searching for Significance)

- 假设应基于理论或至少基于不同的数据集。
- 当同时检验大量原假设时，即使所有原假设都为真，也可以预期其中几个检验会显示显著性。
- 上述两点并不意味着探索性数据分析是坏事。探索性分析常常导致有趣的发现。然而，如果手头的数据暗示了一个有趣的理论，那么就用一个新的数据集来检验那个理论！

Summary

- **假设检验基础 (Basics of Hypothesis Testing)：**
  - 假设检验是统计推断的第二种形式，用于评估关于总体参数的主张。
  - 涉及两个假设：原假设 H_0（通常是无效应或无差异）和备择假设 H_a（我们寻求证据支持的主张）。
  - 检验统计量衡量数据与原假设之间的差异程度。
  - P 值是在原假设为真的条件下，观察到至少与样本结果一样极端的结果的概率。
  - 小的 P 值提供反对原假设的证据。
- **决策与错误 (Decisions and Errors)：**
  - 基于 P 值与显著性水平 α 的比较做出决策：如果 P 值 $< \alpha$，拒绝 H_0。
  - 第一类错误（α）：拒绝一个真实的原假设。
  - 第二类错误（β）：不拒绝一个错误的原假设。
  - 在大多数情况下，我们更关心控制第一类错误。
- **假设形式 (Forms of Hypotheses)：**
  - 单尾检验（左尾或右尾）：当有方向性的先验知识时使用。
  - 双尾检验：当只关心差异而不关心方向时使用。
  - 假设应在查看数据之前根据研究问题或理论确定，以避免 P 值操纵。
- **计算与解释 (Calculation and Interpretation)：**
  - 对于总体均值 μ 的检验（σ 已知），检验统计量为 $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$。
  - P 值的计算取决于备择假设的方向。
  - 结论应表述为" 拒绝 H_0" 或" 无法拒绝 H_0"，而不是" 接受 H_0"。
- **错误类型 (Types of Errors)**:
  - 第一类错误 (Type I error)：拒绝真的 H_0，概率为 α（显著性水平）。
  - 第二类错误 (Type II error)：不拒绝假的 H_0，概率为 β。
- **检验功效 (Power of a Test)**:
  - 定义：$1 - \beta$，即当 H_0 为假时正确拒绝它的概率。
  - 影响因素：α、样本量 n、效应大小。

  - 提高功效：增加 n，增大 α（权衡），或研究更大的效应。

- **选择显著性水平 (Choosing Significance Level α)**:
  - 考虑犯第一类错误的后果。
  - 初步研究可使用较大 α。
  - 常用 $\alpha = 0.05$，但应结合背景和 P 值解释。
- **正确解释显著性 (Correct Interpretation of Significance)**:
  - 统计显著性仅表明效应不太可能仅由随机性导致。
  - 不意味着效应具有实际重要性或规模大。
  - 样本量大时，微小效应也可能显著。
- **不显著的结果 (Non-Significant Results)**:
  - “未能拒绝 H_0”不等于“接受 H_0”或证明没有效应。
  - 可能由于样本量不足或变异性大。
  - 应结合置信区间评估效应可能范围。
- **避免滥用 (Avoiding Abuse)**:
  - 假设应基于理论，而非数据窥探。
  - 多重检验会自然产生一些“显著”结果。
  - 探索性发现需在独立数据上验证。
- **实际应用 (Practical Applications)：**
  - 假设检验广泛应用于科学、医学、商业等领域，用于检验理论、评估效应等。
  - 正确解释 P 值和统计显著性至关重要。
  - 避免滥用显著性检验，如数据窥探、P 值操纵等。
- **关键公式 (Key Formulas)：**
  - 检验统计量（z 检验）：$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$
  - P 值：
    * 右尾检验：$P(Z \geq z)$
    * 左尾检验：$P(Z \leq z)$
    * 双尾检验：$2 \times P(Z \geq |z|)$
    * $P(\text{Type I error}) = \alpha$

- $P(\text{Type II error}) = \beta$
- $\text{Power} = 1 - \beta$

本章核心要点 (Core Takeaways)

- **理解假设检验的逻辑：**从原假设为真出发，计算观察到样本结果的概率，如果概率很小，则拒绝原假设。
- **正确构建假设：**原假设通常包含等式，备择假设反映研究主张。
- **计算和解释 P 值：**P 值越小，反对原假设的证据越强。
- **做出决策：**比较 P 值与显著性水平 α。
- **理解错误类型：**第一类错误是弃真（α，假阳性），第二类错误是取伪（β，假阴性）。
- **功效：** $1 - \beta$，检测真实效应的能力。受 α、n 和效应量影响。
- **选择 α：**需权衡两类错误的后果。常用 0.05，但非绝对标准。
- **谨慎使用单尾检验：**仅在理论或先验知识支持方向性预测时使用。
- **避免常见误解：**不拒绝原假设并不意味着原假设为真，只是证据不足。
- **解释结果：**统计显著 $\neq$ 实际重要；不显著 $\neq$ 没有效应。
- **避免滥用：**预先设定假设，避免数据窥探（p-hacking），用新数据验证探索性发现。