

MAT220 第14讲：假设检验导论

目录

1 引言	2
1.1 统计检验的非正式结构	3
1.2 正态分布的简要回顾	3
2 假设检验	4
2.1 统计检验的一般结构	4
2.2 P值	4
2.3 显著性水平	5
2.4 不同类型的检验	6
2.5 第一类错误和第二类错误	7
3 示例检验：两枚硬币	8
3.1 可疑的硬币	8
3.2 修女们的硬币	9
3.3 检验的图形表示	9

在最近的课程中，我们学习了估计。估计使用样本数据来大致推测总体参数位于什么范围。现在，我们将通过进行主张的检验，继续讨论推断统计。我们不再估计参数，而是评估关于参数真实值的主张。这样的过程称为假设检验，它是推断统计中的一种强大工具。在接下来的四讲中，我们将详细讨论这种方法。今天，我们只介绍假设检验的整体结构。

1 引言

假设你正在散步，并遇到了两群人。其中一群人看起来危险而可疑，另一群人则是看起来值得信任的修女和孩子：



一个可疑的群体



一个值得信任的群体

你先走向那群看起来可疑的人。他们递给你一枚硬币，并提出下面的游戏：

不断抛掷这枚硬币。每次出现正面，我们就给你\$2。但是，每次出现反面，你必须给我们\$1。

起初，这个游戏似乎很慷慨，所以你决定参加。然而，在抛掷30次之后，你发现正面只出现了3次，而反面出现了27次。最终，你需要支付\$21，然后继续赶路。

你感到很可疑，于是决定进行一些计算。由于抛硬币是一个二项试验，你使用公式：

$$P(X = 3) = C_{30,3}(0.5)^3(0.5)^{27} \approx 3.78 \times 10^{-6},$$

这里假设硬币是公平的，并且 $p = q = 0.5$ 。根据上面的结果，你发现，在30次试验中只得到3次正面的概率极小。因此，你认为这枚硬币很可能有问题，并且 p 远低于50%。

接下来，你走向那群看起来值得信任的人，并开始与他们交谈。奇怪的是，第二群人也向你提出完全相同的游戏：

不断抛掷这枚硬币。每次出现正面，我们就给你\$2。但是，每次出现反面，你必须给我们\$1。

尽管你觉得这可能不是一个明智的决定，但你还是决定再抛掷30次硬币。这一次，正面出现了12次，反面出现了18次，因此你赢得了\$6。再次假设硬币是公平的，你计算：

$$P(X = 12) = C_{30,12}(0.5)^{12}(0.5)^{18} = C_{30,12}(0.5)^{30} \approx 0.081.$$

事实上，这个结果并不是特别不合理，因为在30次抛掷中恰好得到15次正面的概率也只有大约14%。仅根据这些信息，没有充分的理由认为第二枚硬币是不公平的。

1.1 统计检验的非正式结构

我们刚才已经看到了一个非常重要的统计过程——假设检验——背后的非正式思想。在两种情形中，我们都面对一个系统，并根据所进行的观察，对“我们使用的是一枚公平硬币”这一假设提出了质疑。在两种情形中，我们都进行了某种概率计算，并得出了两个不同的结论：

1. 对于可疑的硬币，我们认为随机观察到的结果极不可能发生。这使我们开始怀疑 $p = 0.5$ 这一预期，因此拒绝了硬币公平的想法；而
2. 对于修女们的硬币，随机观察到的结果与 $p = 0.5$ 相符，因此没有理由拒绝硬币是公平的这一主张。

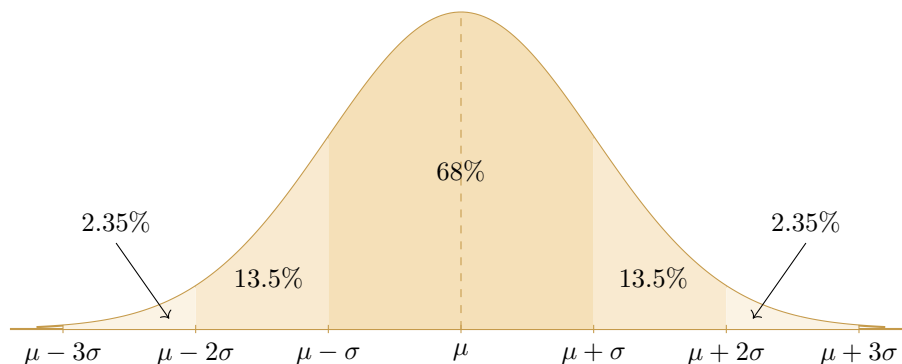
许多假设检验都遵循这种一般模式：我们先有一个预期，然后进行一次观察并产生怀疑，之后再 进行概率计算，以判断拒绝原始假设是否合理。这个概念流程可以概括如下：



在接下来的几讲中，我们将在不同情境下研究这一结构。现在，我们只描述其一般形式。为此，我们首先回顾正态分布的一些基本性质。

1.2 正态分布的简要回顾

回忆一下，正态分布由两个参数决定：总体均值 μ 和总体方差 σ^2 。当我们写作 $X \sim \mathcal{N}(\mu, \sigma^2)$ 时，分布的 x 轴表示随机变量 X 可能取得的值。由于这是一个连续随机变量， X 落在某一区域内的概率由图像下方的面积给出。因此，图像的形状表示 X 出现在附近各个位置的可能性。正态分布在中心附近达到最高点，然后迅速向两侧尾部下降，因此，随机选择一个 X 的值时，我们预计它很可能接近 μ 。事实上，根据第10讲的讨论， X 落在距离 μ 一个标准差以内的概率约为68%：

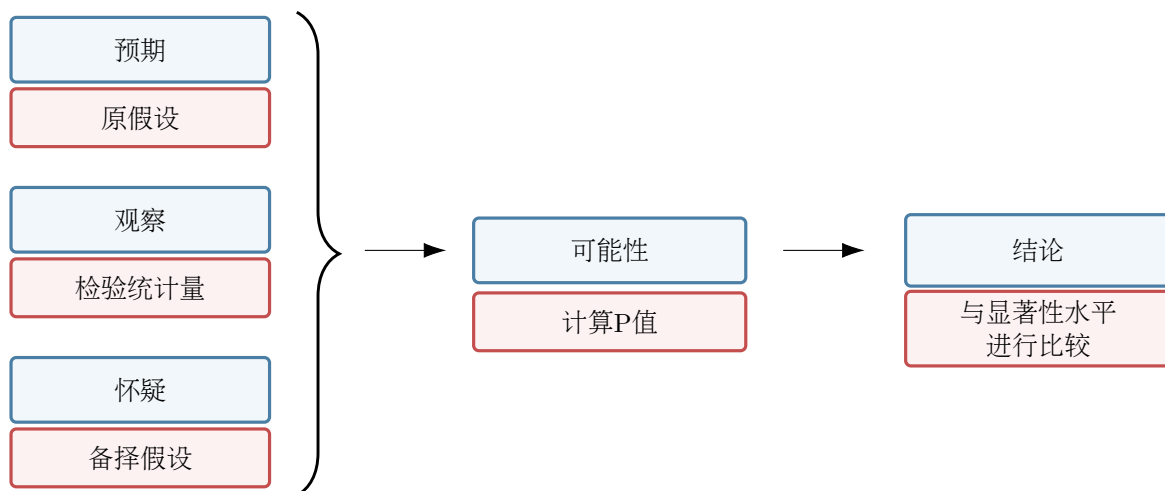


换句话说，如果我们不断随机选择 X 的值，那么大约有68%的时间，该值会落在区间 $[\mu - \sigma, \mu + \sigma]$ 内。相反，随机选择的 X 值远离中心的可能性非常低。这一观察是理解假设检验的基础。

2 假设检验

2.1 统计检验的一般结构

在引言中，我们描述了进行统计检验的一般流程。下面给出了相同的流程，并配上正确的专业术语。



现在，我们逐一解释这些术语。

重要术语

- 原假设 (H_0)： 我们想要检验的陈述。
- 备择假设 (H_1)： 反映我们怀疑的、与原假设相竞争的陈述。
- 检验统计量： 从样本数据中得到的数值信息。
- P值： 假设 H_0 为真时，得到与观察结果同样极端或更加极端的结果的概率。
- 显著性水平 (α)： 用于判断什么结果算作“过于不可能”的概率临界值。

2.2 P值

根据上面的定义，假设检验中的P值并不只是计算某个结果发生的概率。相反，P值是在假设原假设为真的前提下，计算得到当前样本数据或更加极端结果的概率。在第1节中，我们对两枚硬币进行了二项概率计算，并看到了这种思想的原型。事实上，这些概率计算还不是这些假设检验的P值。为了得到P值，我们需要进行更加详细的计算。

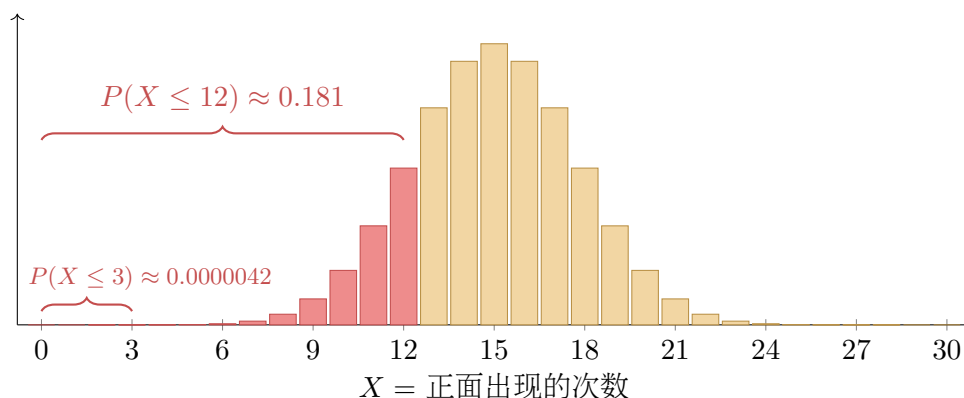
对于可疑的硬币，实际观察结果为 $X = 3$ ，也就是正面出现3次。计算这个观察结果的P值时，我们假设硬币是公平的，即 $p = 0.5$ ，然后计算正面出现3次或更少的概率：

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \approx 0.0000042.$$

对于第二枚硬币，我们观察到 $X = 12$ 。同样，P值通过计算正面出现12次或更少的概率得到：

$$P(X \leq 12) = P(X = 0) + P(X = 1) + \dots + P(X = 12) \approx 0.181.$$

从图形上看，我们是在计算由 $n = 30$ 和 $p = 0.5$ 构成的二项分布左尾中的概率：



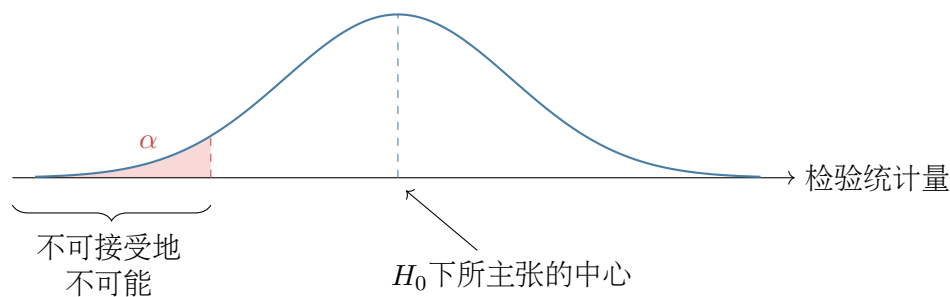
2.3 显著性水平

假设检验首先假设原假设实际上是正确的，然后研究这一假设所产生的结果。正如第1.2节所述，在大多数情况下，我们预计随机选取的检验统计量会落在其概率分布中心附近，尤其是当它服从正态分布时。如果我们观察到一个远在分布尾部的结果，那么这个结果在原始模型下出现的概率就非常低。

当然，由于随机性的本质，偶然得到一个异常极端的结果始终是可能的。真正需要考虑的问题是：

我们的运气需要差到什么程度，才应该开始怀疑原假设有问题？

显著性水平 α 通过给出一个判断“运气过差”的临界值来回答这个问题：



当我们得到样本数据后，P值计算的是：假设数据的真实中心由 H_0 给出时，得到这个样本或更加极端样本的概率。如果这个概率低得无法接受，即小于 α ，那么我们就认为 H_0 所主张的总体参数值很可能是错误的。最终的判断规则可以总结如下：

判断规则

- 如果计算得到的P值小于或等于 α ，那么观察结果在 H_0 下过于不可能，因此我们**拒绝** H_0 。
- 如果计算得到的P值大于 α ，那么观察结果还没有不可能到足以拒绝 H_0 的程度，因此我们**不拒绝** H_0 。

在实际应用中，我们会在检验开始时选择显著性水平。常见的取值为

$$\alpha = 0.01, \quad \alpha = 0.02, \quad \alpha = 0.05.$$

需要特别注意的是，“不拒绝 H_0 ”并不意味着我们已经证明 H_0 为真。它只意味着样本没有提供足够的证据来反对 H_0 。

2.4 不同类型的检验

备择假设 H_1 总是一个与原假设相反的陈述。事实上，有多种方式可以反对一个“等于”的陈述。通常，原假设会声称总体参数的真实值等于某个特定值。这个主张可能以三种方式出错：

- 真实值可能小于所主张的值；
- 真实值可能大于所主张的值；
- 真实值可能只是与所主张的值不同。

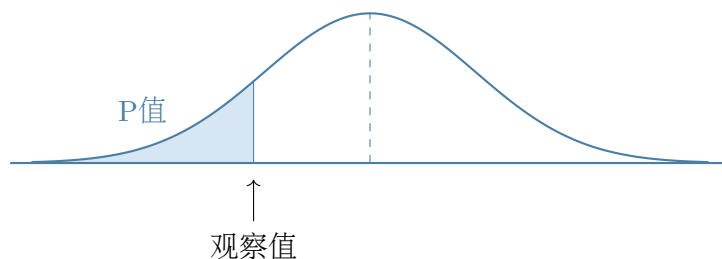
这三种检验分别称为左尾检验、右尾检验和双尾检验。下面我们将说明，备择假设如何决定给定检验中P值的正确计算方法。

2.4.1 左尾检验

左尾检验研究真实值是否小于原假设所主张的值。形式上，检验通常从下面的陈述开始：

$$H_0 : \text{真实值} = \text{所主张的值}, \quad H_1 : \text{真实值} < \text{所主张的值}.$$

顾名思义，在左尾检验中，我们通过取观察值左侧的面积来计算P值：

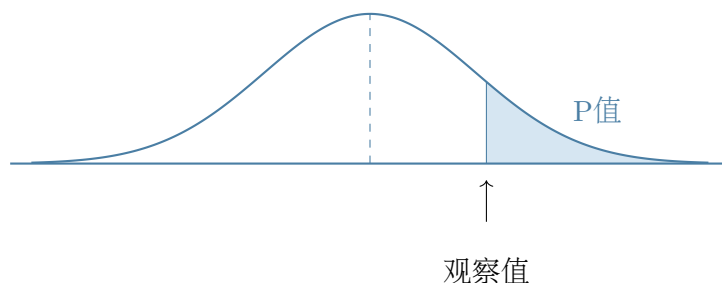


2.4.2 右尾检验

在右尾检验中，备择假设研究真实值是否大于原假设所主张的值：

$$H_0 : \text{真实值} = \text{所主张的值}, \quad H_1 : \text{真实值} > \text{所主张的值}.$$

在这种情况下，P值是观察到的样本统计量右侧的面积：

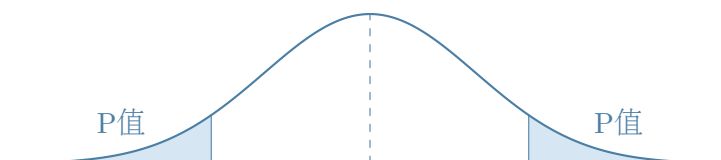


2.4.3 双尾检验

在双尾检验中，我们研究真实值是否与原假设所主张的值不同：

$$H_0: \text{真实值} = \text{所主张的值}, \quad H_1: \text{真实值} \neq \text{所主张的值}.$$

回忆一下，如果两个数 a 和 b 不同，那么必有 $a < b$ 或 $b < a$ 。因此，双尾检验通过同时合并两个尾部的概率来计算P值：



2.5 第一类错误和第二类错误

假设检验本身始终包含一定程度的不确定性。如果得到一个低于显著性水平的P值，这并不能证明原假设为假。它只意味着样本提供了反对原假设的证据。同样，如果我们不拒绝原假设，这也不能证明原假设确实为真。它只意味着没有足够的证据来拒绝原假设。

由于存在这种固有的不确定性，我们始终有可能意外得出错误的结论。在进行假设检验时，我们主要可能犯两种错误。它们称为第一类错误和第二类错误。

第一类错误和第二类错误

- 当 H_0 实际上为真，但我们拒绝了 H_0 时，就发生了**第一类错误**。
- 当 H_0 实际上为假，但我们没有拒绝 H_0 时，就发生了**第二类错误**。

如果我们在第一次抽样时碰巧得到一个统计上极不可能的样本，就可能发生第一类错误。例如，抛掷一枚硬币30次，并得到27次反面和3次正面，从严格意义上说并非不可能。然而，这一结果极其不可能，因此我们不得不拒绝硬币公平这一主张。但是，在这样做时，我们无法保证自己并不是恰好生活在一个极不可能的结果发生了的世界中。在这种情况下拒绝原假设，可能就是第一类错误的一个例子。

第二类错误发生在我们得到与原假设相符的样本数据时。在这种情况下，即使原假设实际上为假，我们也没有充分的理由拒绝它。如果样本观察得到的P值高于 α ，就可能发生这种情况。为了说明这一点，回忆一下，修女们的硬币在30次试验中得到了12次正面。我们认为这个观察结果与公平硬币相当一致，因此没有理由拒绝硬币公平这一主张。但是，如果这枚硬币真实的正面概率实际上是 $P(H) = 0.4$ 呢？我们的样本数据不能排除公平硬币的可能性，即使公平这一主张实际上可能是错误的。

假设检验可能产生四种结果：

	不拒绝 H_0	拒绝 H_0
H_0 为真	正确判断	第一类错误
H_0 为假	第二类错误	正确判断

显著性水平 α 控制犯第一类错误的概率。例如，使用 $\alpha = 0.05$ 意味着，在重复使用该检验时，它被设计为大约有5%的时间会拒绝一个真实的原假设。

3 示例检验：两枚硬币

现在，我们将说明如何把前面的抛硬币例子表示为正式的假设检验。为了方便后续讨论，我们将使用随机变量 $\hat{P}$ ，并用正态近似来描述这些试验。我们在之前的课程中估计真实二项概率 p 时已经见过这个随机变量。下面我们只给出与样本试验对应的P值，而把P值的实际计算推迟到下一讲。现在只需要回忆，随机变量 $\hat{P}$ 近似服从参数为

$$\hat{P} \sim \mathcal{N}\left(p, \frac{pq}{n}\right)$$

的正态分布。

重要的是，真实的二项概率 p 位于这个分布的中心。因此，我们可以把原假设理解为关于这个正态分布中心的主张。观察到的样本数据最终会表现为这个分布 x 轴上的点。换句话说，上面的正态分布告诉我们应该如何组织和衡量我们的“运气”。

在绘制观察数据之前，我们先说明如何正确表述这些检验。

3.1 可疑的硬币

我们先回到第一枚硬币。

步骤1：陈述假设。

我们首先假设硬币是公平的，但是怀疑出现正面的概率实际上低于0.5。因此，原假设和备择假设分别为

$$H_0 : p = 0.5, \quad \text{以及} \quad H_1 : p < 0.5.$$

根据备择假设可以看出，这是一个左尾检验。

步骤2：记录检验统计量。

根据样本数据，在 $n = 30$ 次抛硬币中，我们观察到 $r = 3$ 次正面。因此，样本比例为

$$\hat{p} = \frac{r}{n} = \frac{3}{30} = 0.1.$$

步骤3：计算P值。

通常，我们会在这里直接计算P值。但是，现在我们只给出根据正态近似得到的计算结果：

$$\hat{P} \sim \mathcal{N}\left(0.5, \frac{(0.5)(0.5)}{30}\right).$$

因此，

$$P(\hat{P} \leq 0.1) \approx 0.0000134.$$

注意，由于这是左尾检验，我们计算的是观察到的 $\hat{p}$ 左侧的概率。

步骤4：与显著性水平比较。

使用显著性水平 $\alpha = 0.05$ ，显然有

$$0.0000059 < 0.05.$$

因此，我们必须拒绝原假设 H_0 。

解释

在30次抛硬币中只出现3次正面的样本数据，为反对硬币公平这一假设提供了极强的证据。相反，这些数据支持正面概率小于0.5的怀疑。

3.2 修女们的硬币

现在，我们对第二枚硬币进行相同的检验。

步骤1：陈述假设。

我们再次检验这枚硬币是否不利于出现正面：

$$H_0 : p = 0.5, \quad \text{以及} \quad H_1 : p < 0.5.$$

步骤2：记录检验统计量。

在这种情况下，我们在 $n = 30$ 次抛掷中观察到 $r = 12$ 次正面，因此

$$\hat{p} = \frac{r}{n} = \frac{12}{30} = 0.4.$$

步骤3：计算P值。

在原假设下进行连续性校正后，P值近似为

$$P(X \leq 12) \approx P(Z \leq -0.913) \approx 0.181.$$

步骤4：与显著性水平比较。

再次使用显著性水平 $\alpha = 0.05$ ，可以看到

$$0.181 > 0.05.$$

因此，我们不拒绝 H_0 。

解释

在30次抛硬币中出现12次正面的样本，没有提供足够强的证据来说明真实的正面概率低于0.5。这并不能证明硬币是公平的，只表示这个观察结果与公平硬币相当一致。

3.3 检验的图形表示

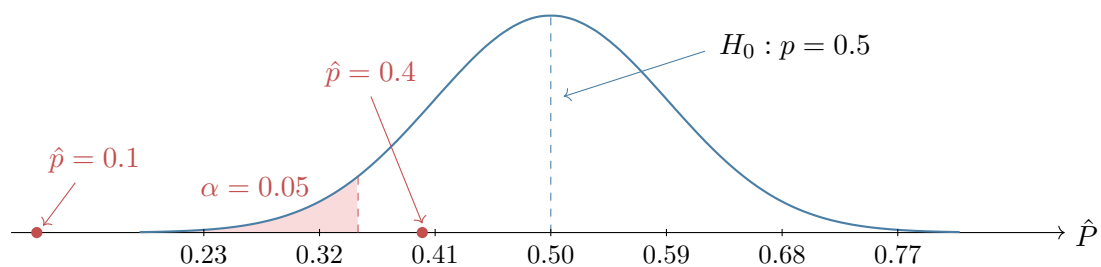
在原假设下，观察到的样本比例的抽样分布近似为

$$\hat{P} \sim \mathcal{N}\left(0.5, \frac{(0.5)(0.5)}{30}\right).$$

这个分布的标准差为

$$\sqrt{\frac{(0.5)(0.5)}{30}} \approx 0.091.$$

可疑的硬币产生了 $\hat{p} = 0.1$ ，它位于左尾非常远的位置。修女们的硬币产生了 $\hat{p} = 0.4$ ，它低于中心，但并不是特别极端。



如图所示，与修女们的硬币对应的左尾P值面积大于临界值 $\alpha = 0.05$ ，因此我们没有充分的理由拒绝这枚硬币是公平的这一假设。另一方面，对于可疑的硬币，观察到的样本比例 $\hat{p} = 0.1$ 位于分布左尾非常远的位置，这意味着在第一次收集样本数据时得到这种结果的可能性极低。0.1左侧的面积远小于0.05，因此我们拒绝了可疑硬币是公平的这一假设。