

MAT220 第16讲：配对数据的假设检验

目录

1	总体均值假设检验的复习	2
1.1	检验的流程	2
2	什么是配对总体?	3
2.1	动机	3
2.2	配对总体的定义	3
2.3	组合配对总体	4
3	配对差值的抽样分布	4
3.1	配对数据的样本	5
3.2	$D = X - Y$ 的均值	5
3.3	$D = X - Y$ 的方差	6
3.4	差值的抽样分布	8
4	配对数据的假设检验	8
4.1	原假设	8
4.2	配对数据的检验流程	9
5	例：电影评分	10

在上一讲中，我们学习了如何检验关于单个总体参数的主张。我们考察了两类参数的检验：总体标准差 σ 已知或未知时的总体均值 μ ，以及二项分布的成功概率 p 。在本讲和下一讲中，我们将讨论如何检验两个总体是否具有相同的均值。事实上，这通常是一个相当复杂的问题，因此为了简化讨论，今天我们将假设两个总体以某种自然的方式相互配对。正如我们将看到的那样，我们可以利用这一假设构造适当的抽样分布，从而检验两个总体均值是否相同。

1 总体均值假设检验的复习

1.1 检验的流程

一般来说，假设检验的目标是提出一个主张，然后检验支持这一主张的证据。作为基准的主张称为原假设，记作 H_0 。随后，我们通过提出某种与其相反的备择假设来检验这一陈述，备择假设记作 H_1 。之后，我们抽取一个样本，并回答下面的问题。

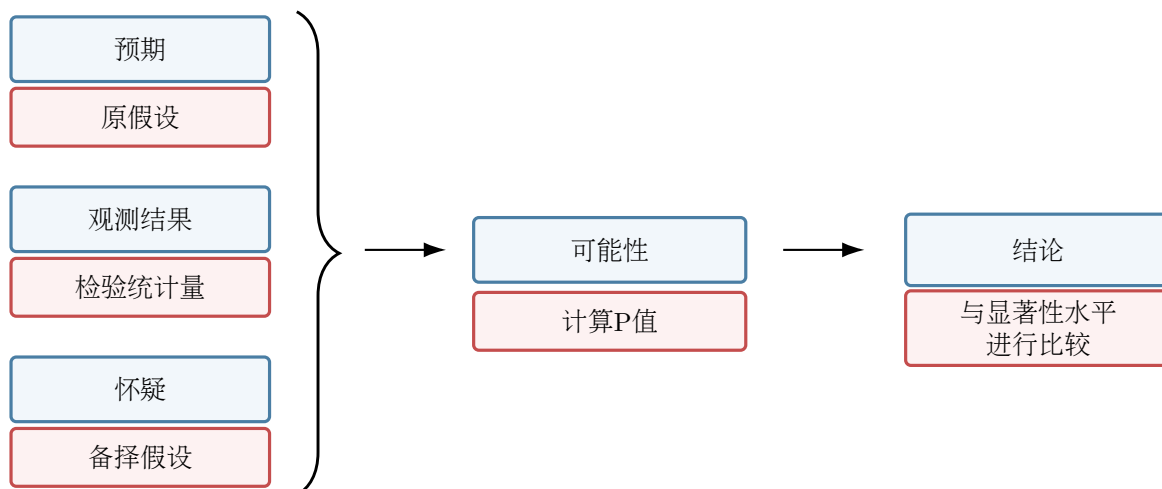
假设检验的核心问题

每一个假设检验都在回答同一个基本问题：

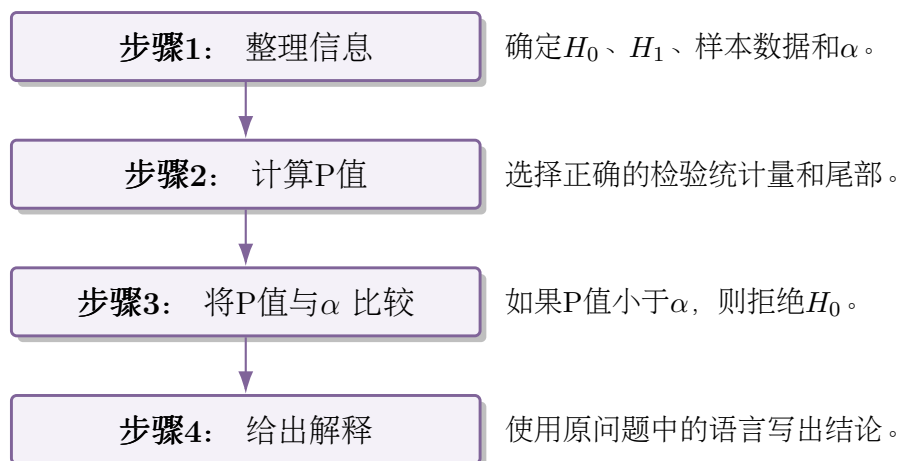
假设 H_0 为真，观测到的样本有多么异常？

具体的概率计算通常会根据问题的情境而改变。

一般来说，假设检验遵循相同的整体流程：



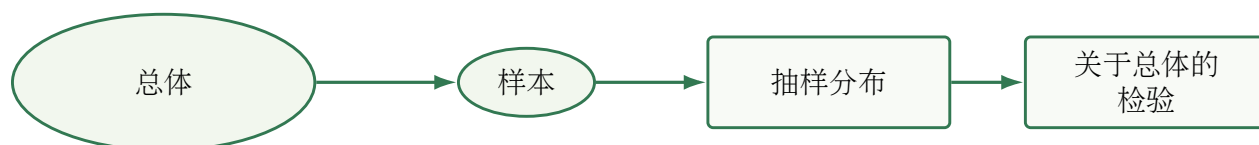
另一种说法是，假设检验的过程可以分为四个主要步骤：



2 什么是配对总体？

2.1 动机

到目前为止，我们的统计检验大多遵循以下模式：

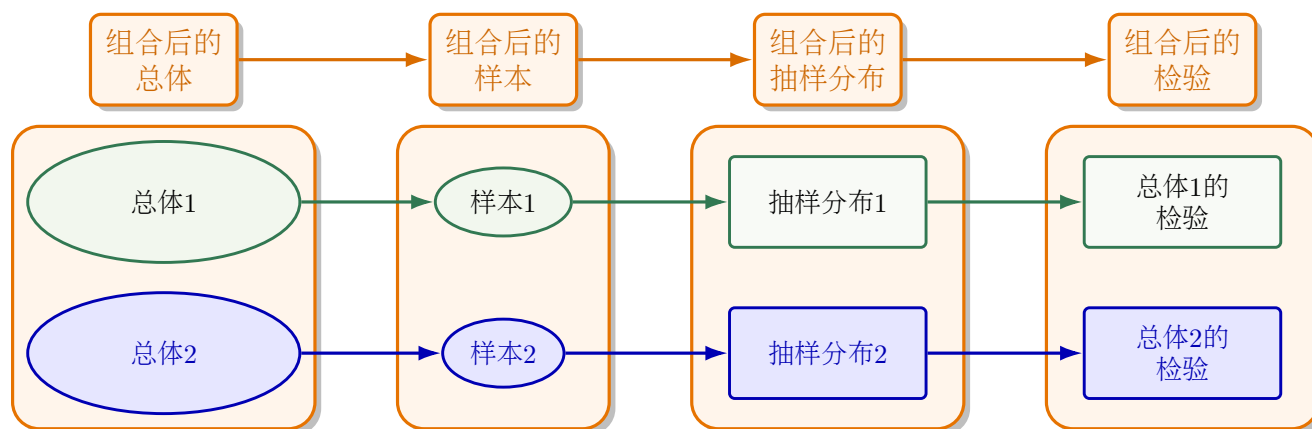


也就是说，我们对某个总体的未知总体参数提出一个主张，然后通过抽取一个随机样本并评估观测结果的可能性，检验这一主张是否合法。

现在，我们将研究如何使用假设检验比较两个不同的总体。我们要检验的主要主张是：

这两个总体的平均值是否相同？

用符号表示，如果第一个总体的均值为 μ_1 ，第二个总体的均值为 μ_2 ，那么我们要检验 $\mu_1 = \mu_2$ 是否成立。为此，我们将把两个总体组合成第三个总体，抽取一个样本，然后推导相应的抽样分布。用示意图表示，大致如下：



事实上，将两个总体组合成第三个总体的过程相当复杂。因此，为了让问题更容易处理，我们将假设两个总体以某种有意义的自然方式相互配对。

2.2 配对总体的定义

在本讲剩余的部分中，我们只考虑配对总体。直观地说，如果一个总体中的数据点与另一个总体中的数据点之间存在自然的联系，那么这两个总体就是配对的。

配对总体

如果一个数据集中的每个观测值，在另一个数据集中都有唯一对应的观测值，那么我们称这两个总体是**配对的**。

在许多情况下，我们都可能需要研究配对数据。例如，以下情境中都可能出现配对数据：

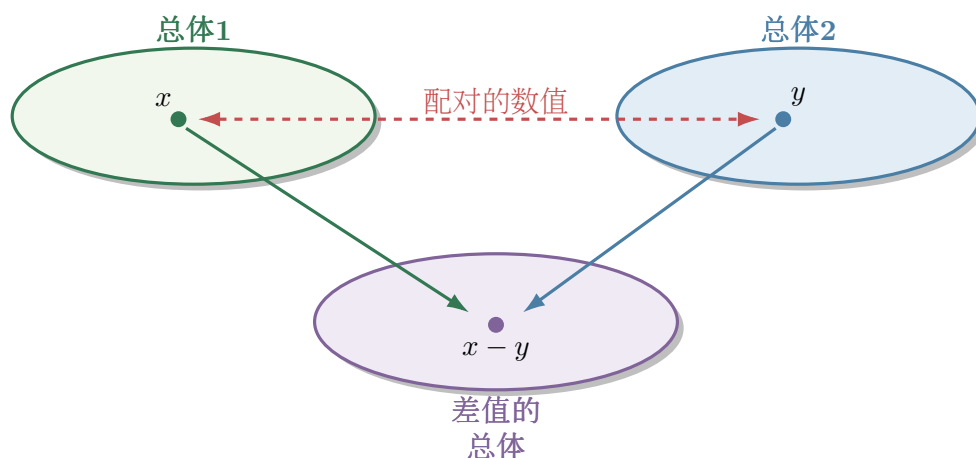
- 研究某种行为对同一组对象的影响，例如咖啡因对反应时间的影响。
- 研究多个测量设备的准确性，例如在同一个人身上使用的两种不同心率监测器的准确性。
- 研究自然匹配的对象，例如双胞胎，或者同一个人的左脚和右脚。

2.3 组合配对总体

假设两个配对总体分别由随机变量 X 和 Y 表示。由于两个总体是配对的，因此 X 中的任意观测值 x ，在 Y 中都有唯一对应的观测值 y 。这意味着我们可以考虑 $x - y$ ，它是一个有意义的数值，表示 X 的值与 Y 的值之差。为了简化记号，我们写作：

$$D = X - Y.$$

我们可以把它解释为一个新的随机变量，即 X 与 Y 的差。¹ 示意图如下：



3 配对差值的抽样分布

为了对差值 $X - Y$ 进行假设检验，我们需要描述 $\bar{D}$ 的抽样分布，其中 $\bar{D}$ 表示从总体 D 中抽取样本后得到的样本均值随机变量。至少，根据中心极限定理，当 n 足够大时， $\bar{D}$ 应当近似服从具有以下参数的正态分布：

$$\bar{D} \approx \mathcal{N}\left(\mu_D, \frac{\sigma_D^2}{n}\right).$$

¹因此使用字母“ D ” 表示英文单词“difference”。

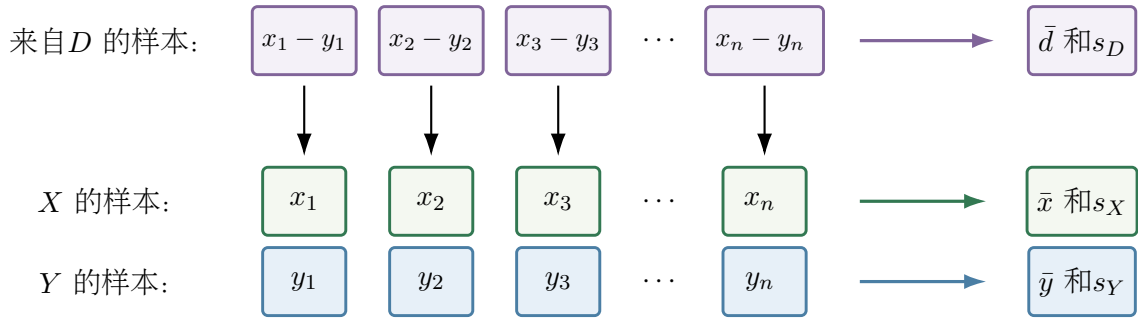
为了理解这一分布，我们需要理解随机变量 $D = X - Y$ 的均值 μ_D 和方差 σ_D^2 。

3.1 配对数据的样本

假设我们从 D 中抽取 n 个观测值，记为 $d_1, \dots, d_n$ 。由于每个观测值都是 $x - y$ 形式的差值，因此抽取 n 个观测值会得到：

$$d_1 = x_1 - y_1, \quad d_2 = x_2 - y_2, \quad \dots, \quad d_n = x_n - y_n.$$

将这些差值拆开后，可以恢复出大小为 n 的 X 样本和 Y 样本：



如果能够使用来自 x_i 和 y_i 的信息来描述样本统计量 $\bar{d}$ 和 s_D ，将会非常有帮助。至少，这可以帮助我们理解 X 、 Y 和 D 之间的均值与方差如何相互作用。

3.2 $D = X - Y$ 的均值

如果我们要计算样本 $d_1, \dots, d_n$ 的样本均值，可以利用每个 $d_i = x_i - y_i$ 的事实，将公式拆分为两个部分：

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i = \frac{1}{n} \sum_{i=1}^n (x_i - y_i) \stackrel{(*)}{=} \frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{n} \sum_{i=1}^n y_i = \bar{x} - \bar{y}.$$

在标记为 $(*)$ 的步骤中，我们使用以下性质将求和式拆开并重新排列：

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (x_i - y_i) &= \frac{1}{n} \left((x_1 - y_1) + (x_2 - y_2) + \dots + (x_n - y_n) \right) \\ &= \frac{1}{n} \left(x_1 + x_2 + \dots + x_n - y_1 - y_2 - \dots - y_n \right) \\ &= \frac{1}{n} \left((x_1 + x_2 + \dots + x_n) - (y_1 + y_2 + \dots + y_n) \right) \\ &= \frac{1}{n} (x_1 + x_2 + \dots + x_n) - \frac{1}{n} (y_1 + y_2 + \dots + y_n) = \frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{n} \sum_{i=1}^n y_i. \end{aligned}$$

差值的样本均值

对于配对数据,

$$\bar{d} = \bar{x} - \bar{y}.$$

换句话说, 差值的样本均值等于两个样本均值之差。

事实上, 在总体层面也会发生同样的现象。回忆一下, 随机变量的总体均值实际上由期望值给出。在离散型随机变量的情形中, 我们已经看到期望值的公式就是加权平均。对于连续型随机变量, 定义稍微复杂一些, 并涉及我们尚未学习的微积分。对我们而言, 重要的是以下性质:

$$E(X - Y) = E(X) - E(Y),$$

这一性质对所有具有有限期望值的随机变量都成立, 无论它们是连续型还是其他类型。利用这一性质, 可以使用 μ_X 和 μ_Y 表示总体均值 μ_D :

$$\mu_D = E(D) = E(X - Y) = E(X) - E(Y) = \mu_X - \mu_Y.$$

与样本统计量 $\bar{d}$ 相同, 差值的均值等于两个均值之差。

差值的总体均值

如果 $D = X - Y$, 那么

$$\mu_D = \mu_X - \mu_Y.$$

3.3 $D = X - Y$ 的方差

遗憾的是, 差值的方差不像均值那样容易处理。根据样本方差的公式,

$$s_D^2 = \frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n - 1}.$$

由上一节可知, $\bar{d} = \bar{x} - \bar{y}$ 。因此, 我们可以将求和式中的各项重写为:

$$(d_i - \bar{d})^2 = \left((x_i - y_i) - (\bar{x} - \bar{y}) \right)^2 = \left((x_i - \bar{x}) - (y_i - \bar{y}) \right)^2.$$

遗憾的是, 展开括号后会出现一个额外的交叉项:

$$\begin{aligned} (d_i - \bar{d})^2 &= \left((x_i - \bar{x}) - (y_i - \bar{y}) \right)^2 \\ &= (x_i - \bar{x})^2 - 2(x_i - \bar{x})(y_i - \bar{y}) + (y_i - \bar{y})^2. \end{aligned}$$

这意味着, 如果我们尝试分解 s_D^2 的公式并将其重新排列成较小的部分, 将得到:

$$\begin{aligned}
s_D^2 &= \frac{1}{n-1} \sum_{i=1}^n \left((x_i - \bar{x})^2 - 2(x_i - \bar{x})(y_i - \bar{y}) + (y_i - \bar{y})^2 \right) \\
&= \underbrace{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}_{s_X^2} + \underbrace{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}_{s_Y^2} - \frac{2}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).
\end{aligned}$$

虽然代数运算中出现了方差 s_X^2 和 s_Y^2 ，但同时也出现了一个额外的量。这个额外的量称为样本协方差，通常记作 s_{XY} 或 $\text{cov}(x, y)$ 。

差值的样本方差

对于配对数据，样本方差可以写为

$$s_D^2 = s_X^2 + s_Y^2 - 2 \text{cov}(x, y).$$

首先计算各个差值：

$$d_i = x_i - y_i,$$

然后根据差值列表计算样本标准差 s_D 。

在总体层面也会出现同样的问题。一般来说，

$$\text{Var}(X - Y) \neq \text{Var}(X) - \text{Var}(Y).$$

这可以使用随机变量方差 Var 的定义进行验证。再次利用期望值 E 对和与差具有良好性质这一事实，可以得到：

$$\begin{aligned}
\text{Var}(X - Y) &= E \left[((X - Y) - E(X - Y))^2 \right] \\
&= E \left[((X - Y) - (E(X) - E(Y)))^2 \right] \\
&= E \left[((X - E(X)) - (Y - E(Y)))^2 \right] \\
&= E \left[(X - E(X))^2 \right] + E \left[(Y - E(Y))^2 \right] - 2E \left[(X - E(X))(Y - E(Y)) \right] \\
&= \text{Var}(X) + \text{Var}(Y) - 2 \text{Cov}(X, Y),
\end{aligned}$$

这与样本方差的结果完全对应。

粗略地说，总体协方差反映了 X 和 Y 之间的线性关系。我们将在第18讲中深入研究这一概念。现在，我们将关于总体方差 σ_D^2 的讨论总结如下。

差值的总体方差

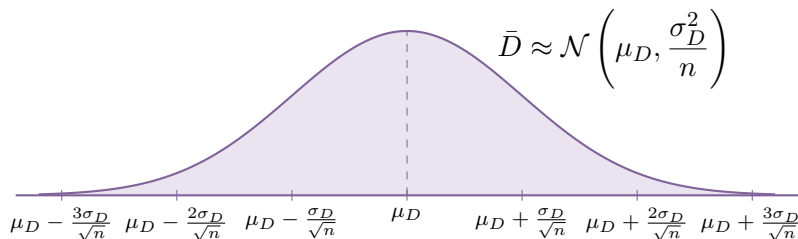
对于配对随机变量，总体方差可以写为

$$\sigma_D^2 = \sigma_X^2 + \sigma_Y^2 - 2 \text{Cov}(X, Y).$$

特别地，一般来说并不成立 $\sigma_D^2 = \sigma_X^2 + \sigma_Y^2$ 。

3.4 差值的抽样分布

如果我们随机抽取 n 个相互独立的配对，并计算它们的差值，那么中心极限定理告诉我们，其抽样分布应当近似服从正态分布：



利用上一节的结果，现在可以看到，如果 X 的均值为 μ_X 、方差为 σ_X^2 ，而 Y 的均值为 μ_Y 、方差为 σ_Y^2 ，那么差值 $D = X - Y$ 的参数为 $\mu_D = \mu_X - \mu_Y$ 和 $\sigma_D^2 = \sigma_X^2 + \sigma_Y^2 - 2\text{Cov}(X, Y)$ 。

但是，这里存在一个问题：在实际应用中，我们几乎从来不知道额外协方差项的数值。这意味着，即使我们恰好知道方差 σ_X^2 和 σ_Y^2 ，这些信息仍然不足以计算 σ_D^2 。因此，我们用样本标准差 s_D 替代 σ_D ，并使用学生 t 分布。

$$z = \frac{\bar{d} - \mu_D}{\sigma_D / \sqrt{n}} \quad \xrightarrow{\text{用 } s_D \text{ 替代 } \sigma_D} \quad t = \frac{\bar{d} - \mu_D}{s_D / \sqrt{n}}$$

σ_D 未知 学生 t 分布, $df = n - 1$

在进行假设检验时，后一个式子就是相应的检验统计量。

4 配对数据的假设检验

4.1 原假设

通常，在检验配对数据时，我们要回答的问题是：

这两个总体的平均值是否相同？

如果 μ_X 是总体1的均值， μ_Y 是总体2的均值，那么在数学上可以将这一问题重写为：

$$\mu_X = \mu_Y \quad \text{或者等价地, } \mu_X - \mu_Y = 0.$$

由于 $\mu_D = \mu_X - \mu_Y$ ，因此原假设为

$$H_0 : \mu_X - \mu_Y = 0.$$

每当我们检验两个均值是否相等时，这始终是原假设的正确形式。

备择假设取决于我们所怀疑的方向：

$$H_1 : \begin{cases} \mu_X - \mu_Y < 0, & \text{左尾检验,} \\ \mu_X - \mu_Y > 0, & \text{右尾检验,} \\ \mu_X - \mu_Y \neq 0, & \text{双尾检验.} \end{cases}$$

事实上，选择正确的备择假设可能具有一定难度。

选择正确的备择假设

如果你怀疑总体1的平均值小于总体2的平均值，那么：

$$\mu_X < \mu_Y \implies \mu_X - \mu_Y < 0$$

因此，备择假设应写为 $H_1 : \mu_X - \mu_Y < 0$ 。

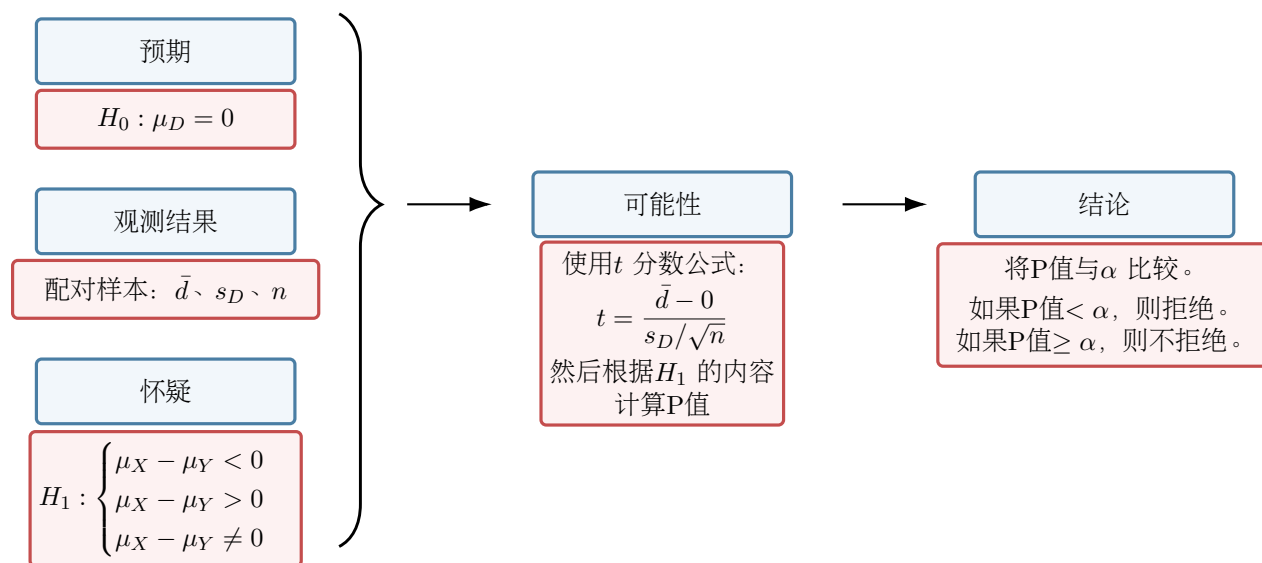
相反，如果你怀疑总体1的平均值大于总体2的平均值，那么：

$$\mu_X > \mu_Y \implies \mu_X - \mu_Y > 0$$

因此，备择假设应写为 $H_1 : \mu_X - \mu_Y > 0$ 。

4.2 配对数据的检验流程

根据第3.4节的讨论，我们通常无法得到两个总体的协方差，因此将 σ_D^2 视为未知。这意味着，在原假设成立时，检验统计量服从学生 t 分布，而不是正态分布。综合以上内容，配对数据假设检验的完整流程如下：



5 例：电影评分

假设奥康奈尔博士想检验公众是否认为 *Twilight* 和 *The Twilight Saga: New Moon* 同样优秀。定义

$\mu_X = \textit{Twilight}$ 的十分制平均评分,

$\mu_Y = \textit{The Twilight Saga: New Moon}$ 的十分制平均评分.

在亲自看过两部电影之后, 他怀疑人们平均而言可能认为 *Twilight* 比 *The Twilight Saga: New Moon* 更好。为了检验这一主张, 他随机抽取了30人, 并要求每个人为两部电影评分。

对于每个人, 他计算两部电影评分的差值 $d_i = x_i - y_i$, 其中 x_i 是对 *Twilight* 的评分, y_i 是对 *The Twilight Saga: New Moon* 的评分。根据配对数据, 他得到:

$$\bar{d} = 0.6 \quad \text{并且} \quad s_D = 1.5.$$

我们将使用显著性水平 $\alpha = 0.03$ 进行检验。

步骤1：提取信息

根据检验的设定, 重要的数值为:

$$\mu_{D,0} = 0, \quad n = 30, \quad \bar{d} = 0.6, \quad s_D = 1.5, \quad \alpha = 0.03.$$

原假设表示两部电影具有相同的平均评分:

$$H_0 : \mu_X - \mu_Y = 0.$$

由于我们定义了 $D = X - Y$, 因此认为 *Twilight* 的评分高于 *The Twilight Saga: New Moon* 的怀疑可以写为:

$$H_1 : \mu_X - \mu_Y > 0.$$

因此, 这是一个右尾检验。

步骤2：计算P值

由于我们不知道 σ_D , 因此不能使用 z 分数。相反, 我们使用以下检验统计量:

$$t = \frac{\bar{d} - 0}{s_D / \sqrt{n}} = \frac{0.6 - 0}{1.5 / \sqrt{30}} \approx 2.19.$$

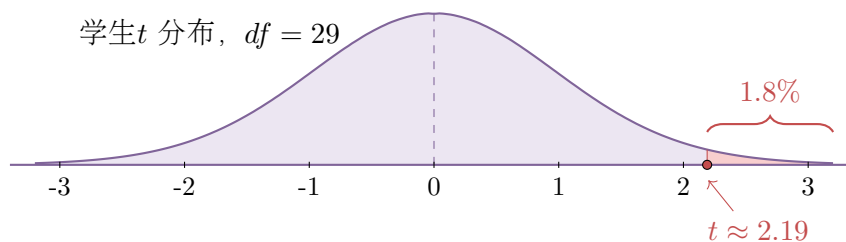
由于 $n = 30$, 自由度为

$$df = n - 1 = 29.$$

因此，P值为

$$P\text{值} = P(T > 2.19) \approx P(T > 2.2) = 1 - P(T < 2.2) \approx 1 - 0.98204 = 0.01796.$$

我们通常使用的表格只提供到小数点后一位的数值，因此在上述近似中使用了 $t = 2.2$ 。从图形上看，P值就是下图阴影部分的面积：



换句话说，如果两部电影的真实平均评分确实相同，那么从30人的随机样本中得到如此大或更大的 t 统计量的概率约为1.8%。

步骤3：将P值与 α 比较

在这个例子中，我们需要在显著性水平 $\alpha = 0.03$ 下进行检验。P值约为0.018，小于0.03。因此，我们拒绝原假设。

步骤4：给出解释

解释

在3% 的显著性水平下，样本提供了有说服力的证据，表明公众平均而言对 *Twilight* 的评分高于对 *The Twilight Saga: New Moon* 的评分。

这里，我们没有分别对任何一部电影的平均评分作出单独的主张。相反，我们检验的是两部电影平均评分之间的差值。