MAT220 第13讲：更多关于估计的内容

目录

在第12讲中，我们引入了估计的概念，并用它在给定某个样本均值 $\bar{x}$ 的情况下猜测总体均值 μ 的值。在这样做时，我们有意假设真实总体标准差 σ somehow 已经知道。这个假设帮助我们简化了相关概念，在当时非常有用。然而，在现实世界中，如果我们不知道 μ，那么很可能也不知道 σ。在本讲中，我们将讨论在不假设 σ 已知的情况下如何进行估计。我们会在两个情形中这样做：像第12讲一样估计 μ，以及估计二项实验中的成功概率 p。

1 估计的复习

上一讲中，我们引入了置信区间的概念，用来估计真实总体均值 μ。主要思想是取一个样本均值 $\bar{x}$，并使用中心极限定理构造 $\bar{X}$ 的抽样分布

$$\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

这个分布把真实总体均值 μ 放在正中心，因此我们可以用它把 $\bar{x}$ 与 μ 联系起来。然后，我们在 $\bar{x}$ 周围构造一些小区间；根据构造，这些区间会在一定比例的情况下包含真实均值。

这种推定方式在现实世界中的一个常见用途出现在质量控制中。为了说明，假设有一家工厂生产大量螺丝。为了能装进它们被设计要装入的部件中，这些螺丝需要大约 20 mm 长，并且两边都有 0.1 mm 的容许误差。这意味着，例如，19.9 mm 的螺丝是可接受的，而长度为 19.89 mm 的螺丝则不可接受。

原则上，我们可以通过检查工厂生产的每一颗螺丝并测量其长度，来保证螺丝的高质量。这样，我们可以把所有这些数加起来，再除以螺丝总数，从而得到它们的真实平均长度。然而，实际中这会过于耗时，并且几乎不可能。因此，在这种情况下，我们可以改为取一些样本数据，并仅仅估计螺丝的真实平均长度。如果这个数落在 19.9 到 20.1 的范围内，那么这些螺丝就是可接受的。这就是质量控制的思想。

质量控制

质量控制使用样本数据来判断某个生产过程是否表现得可以接受。通常，检测每一个产品是不现实的，所以使用统计估计会更容易、更便宜。

1.1 复习：σ 已知时估计 μ

继续使用工厂的例子，假设长期研究告诉我们，总体标准差是已知的，其精确值为

$$\sigma = 1.2 \text{ mm}.$$

还假设质量控制部门取了一个由 $n = 50$ 颗螺丝组成的简单随机样本，并得到样本均值 $\bar{x} = 19.98$ mm。

现在我们要为真实平均长度 μ 构造一个 95% 置信区间。根据第12讲，σ 已知时 μ 的置信区间可以由误差范围 E 计算出来。如果把区间中心放在观测值 $\bar{x}$，则置信区间为

$$\left[\bar{x} - z_c\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_c\frac{\sigma}{\sqrt{n}}\right].$$

对于 95% 置信区间，我们从第12讲的表中知道 $z_c = 1.96$。因此，误差范围可以计算为

$$E = z_c\frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{1.2}{\sqrt{50}} \approx 0.333.$$

因此，以观测到的样本均值 $\bar{x} = 19.98$ 为中心的 95% 置信区间为

$$\text{CI} = [19.98 - 0.333,\ 19.98 + 0.333] = [19.647,\ 20.313].$$

解释

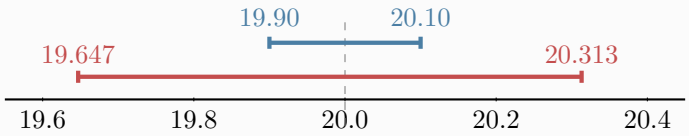
根据题目设定，平均长度的允许范围为

$$[19.90,\ 20.10].$$

但是我们得到的 95% 置信区间是

$$[19.647,\ 20.313].$$

这个区间比所要求的工程容许范围宽得多：

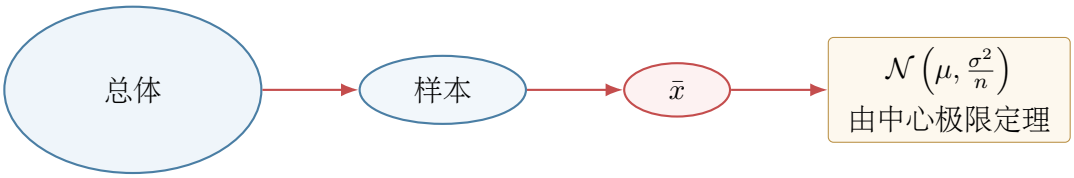


所以，这个样本并没有提供强有力的证据表明真实均值位于所要求的容许范围之内。不过，这不应该意味着我们自动关闭工厂。相反，我们必须接受：样本数据在两个方向上都不够有说服力。我们没有足够证据得出真实均值在期望的 20 mm ± 0.1 mm 之外，也没有足够证据得出真实均值在这些界限之内。

2 σ 未知时的估计

2.1 已知 σ 方法失效的位置

在前面的例子中，我们使用了以下工作流程：



然后，我们通过用

$$E = z_c \frac{\sigma}{\sqrt{n}}$$

计算出的误差范围来估计 μ。

然而，在许多现实情形中，我们根本不知道 σ。这意味着我们无法把这个数代入上面的公式，因此也就不能再使用这种方法估计 μ。相反，我们需要用某个已知的东西替代 σ，然后评估这种替代所带来的后果。

2.2 用 s 替代 σ

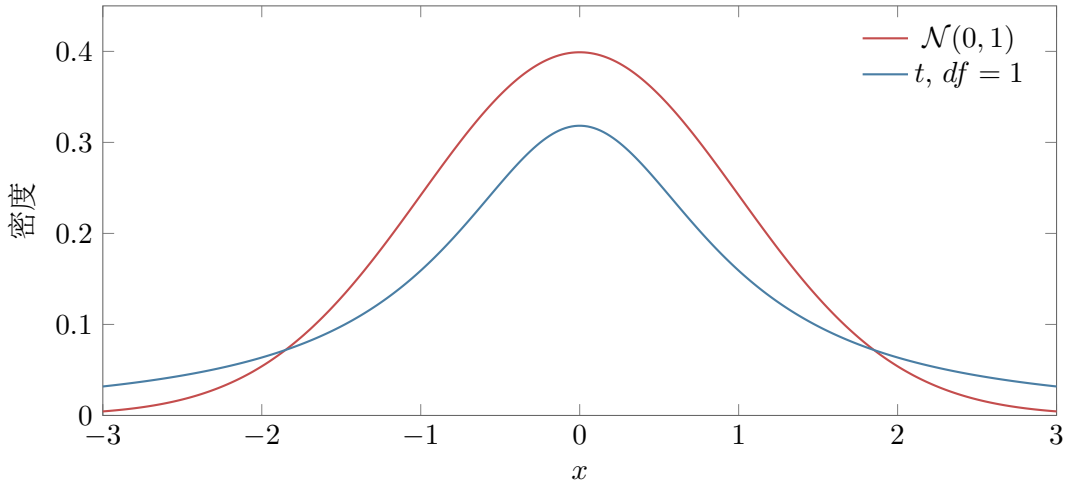
如果我们不知道真实总体标准差 σ 是什么，那么次好的选择似乎是用样本标准差替代它：

$$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}.$$

然而，如果我们做出这种替代，就会引入比之前更多的不确定性。因此，我们必须小心选择用于推导估计的概率分布。事实上，这时不再适合使用标准正态分布。相反，我们需要使用一种称为 Student 的 t 分布的修正版本。现在我们来具体看看这些 t 分布是什么样子。

2.3 Student 的 t 分布

Student 的 t 分布与标准正态分布有关。唯一主要的区别是，更多的面积分布在尾部。下面给出一个简单比较：



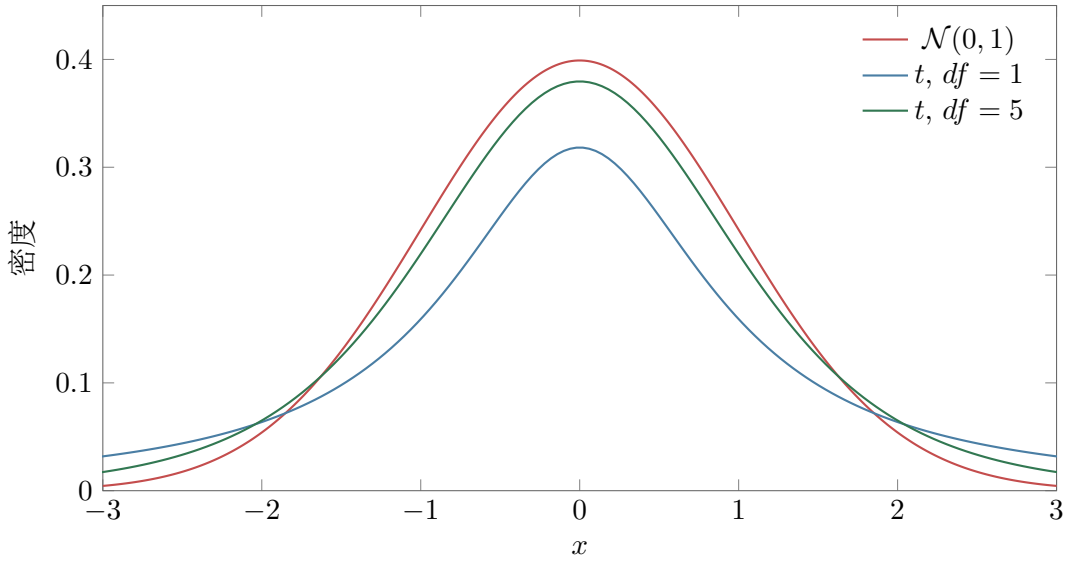
这里，标准正态分布用红色表示，t 分布用蓝色表示。注意，蓝色曲线仍然是对称的，并且仍然以 0 为中心。事实上，唯一明显的区别在尾部——它们更肥厚。[1] 这是有原因的：当 σ 未知时，我们有更多的不确定性，因此基于这个分布构造的置信区间应当比基于正态分布的对应区间更宽。

注意上图中，t 分布的标签里有一个额外的项 "df"。事实上，这是完整描述 t 分布所需要的一个额外输入。这个额外信息称为自由度，因此使用字母 df，并且它简单定义为

$$df = n - 1.$$

这里的想法是，这个数会随着样本大小的增加而增加。样本大小增加时，我们通常会预期样本标准差 s 受随机噪声的影响更小，因此它会成为真实总体参数 σ 的更好估计。如果 s 随着 n 增加而更接近 σ，那么我们就会预期 t 分布尾部中内置的不确定性会减少。当 n 非常大时，样本标准差 s 应当非常接近 σ，因此我们预期 t 分布会看起来非常类似于标准正态分布 $\mathcal{N}(0,1)$。事实上，情况正是如此：随着 df 增加，t 分布看起来越来越像标准正态分布：

[1] 对此的专业术语是峰度，它衡量一个分布的尾部有多"重"。



当 df 越来越大时，所得的 t 分布会越来越像正态分布。这给了我们另一种描述标准正态分布的方法：它是当 $df \to \infty$ 时标准 t 分布的极限。

Student 的 t 分布的性质

Student 的 t 分布具有以下性质。

1. 它们关于 0 对称。

2. 曲线下方的总面积为 1。

3. 它们的尾部比 $\mathcal{N}(0,1)$ 更大。

4. 它们需要一个额外输入：

$$df = n - 1.$$

5. 随着 df 变大，t 分布越来越接近 $\mathcal{N}(0,1)$。

2.4 t 的临界值

Student 的 t 统计量

假设原始总体服从均值为 μ 的正态分布。对于大小为 n 的样本，若样本均值为 $\bar{X}$，样本标准差为 S，[a] 则统计量

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

服从自由度为

$$df = n - 1$$

的 Student 的 t 分布。

[a] 这里我们把样本标准差本身也提升为一个随机变量。

这个表达式看起来几乎与 z 分数公式完全相同，只是我们把 σ 替换成了 s。一旦我们有了观测样

本，公式就变为：

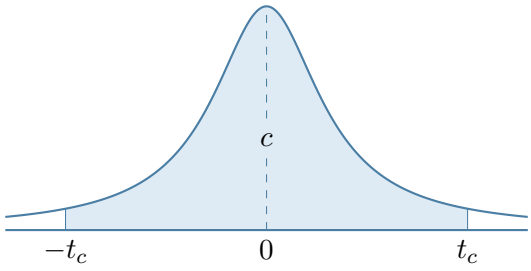
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad \xrightarrow{\text{未知 } \sigma} \quad t = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

回忆一下，当 σ 已知时，我们使用标准正态分布中的临界值 z_c。从几何上看，这对应于标准正态分布中面积恰好为 c 的区域。然后，我们使用 z 分数公式把这个区域映回抽样分布 $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$，从而构造置信区间来估计 μ。

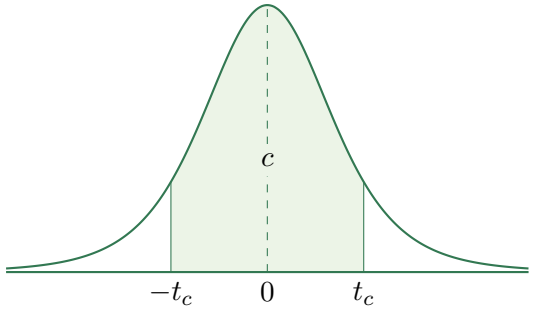
相比之下，当 σ 未知时，我们改用 Student 的 t 分布中的临界值。类比临界值 z_c，我们把 t 分布中的临界值写作 t_c。从几何上看，这些临界值发挥完全相同的作用：它们给出 t 分布中一个区域的端点，该区域以原点为中心，并围住恰好为 c 的面积。由于 t 分布依赖自由度 df，所以 t_c 的精确值依赖两条信息：

1. 置信水平 c，
2. 自由度 $df = n - 1$。

例如，考虑前面画出的 $df = 1$ 和 $df = 5$ 的 t 分布。这里，由于两条曲线形状不同，临界值 t_c 也会改变：



$df = 1$ 时的面积 c



$df = 5$ 时的面积 c

可以看到，当 df 较小时，t_c 的临界值会更大。常见的 t_c 值表写在下面。注意，它比对应的 z_c 表复杂得多：

df	80%	90%	95%	98%	99%	99.5%
1	3.078	6.314	12.706	31.821	63.657	127.321
2	1.886	2.920	4.303	6.965	9.925	14.089
3	1.638	2.353	3.182	4.541	5.841	7.453
4	1.533	2.132	2.776	3.747	4.604	5.598
5	1.476	2.015	2.571	3.365	4.032	4.773
6	1.440	1.943	2.447	3.143	3.707	4.317
7	1.415	1.895	2.365	2.998	3.499	4.029
8	1.397	1.860	2.306	2.896	3.355	3.833
9	1.383	1.833	2.262	2.821	3.250	3.690
10	1.372	1.812	2.228	2.764	3.169	3.581
12	1.356	1.782	2.179	2.681	3.055	3.428
15	1.341	1.753	2.131	2.602	2.947	3.286
20	1.325	1.725	2.086	2.528	2.845	3.153
25	1.316	1.708	2.060	2.485	2.787	3.078
30	1.310	1.697	2.042	2.457	2.750	3.030
40	1.303	1.684	2.021	2.423	2.704	2.971
60	1.296	1.671	2.000	2.390	2.660	2.915
120	1.289	1.658	1.980	2.358	2.617	2.860
∞	1.282	1.645	1.960	2.326	2.576	2.807

这里，每一列对应置信水平 c 的一个常见值。每一行根据某个常见自由度给出 t_c 的值。最下面一行表示 df 任意变大的情况。在这种情况下，t_c 的值会越来越接近 z_c 的值，并最终在极限中趋近于它。

2.5 σ 未知时的置信区间

既然我们可以使用临界值 t_c，就可以按照通常推导误差范围的方法，描述观测值 $\bar{x}$ 周围的置信区间。在这种情况下，我们必须使用 t 统计量：

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

而不是通常的 z 分数公式。使用临界值 $\pm t_c$ 并求解，可以得到误差范围：

$$E = t_c \frac{s}{\sqrt{n}}.$$

注意它与 σ 已知的情况非常相似——这里我们只是把 σ 替换成了 s，并且使用 t 分布的临界值，而不是通常的 z_c。观测到的样本均值 $\bar{x}$ 周围的置信区间可以描述如下。

σ 未知时 μ 的置信区间

如果总体标准差 σ 未知，我们使用样本标准差 s。μ 的置信区间变为：

$$[\bar{x} - E, \bar{x} + E] = \left[\bar{x} - t_c \frac{s}{\sqrt{n}},\ \bar{x} + t_c \frac{s}{\sqrt{n}}\right],$$

其中 $df = n - 1$。

已知和未知两种情况的关键区别总结如下。

	σ 已知	σ 未知
分数	$z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$	$t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$
临界值	z_c	t_c
误差范围	$E = z_c \dfrac{\sigma}{\sqrt{n}}$	$E = t_c \dfrac{s}{\sqrt{n}}$
置信区间	$[\bar{x} - E, \bar{x} + E]$	$[\bar{x} - E, \bar{x} + E]$

2.6 例题：σ 未知时的螺丝

回到第1节中的工厂例子。现在假设工厂不知道真实总体标准差 σ。 相反，一个由 50 颗螺丝组成的样本给出：

$$n = 50, \qquad \bar{x} = 19.98, \qquad s = 1.1 \text{ mm}.$$

如果我们现在想为真实平均长度 μ 构造一个 95% 置信区间，那么必须使用 s 而不是 σ，并且必须使用 t 统计量而不是 z 分数。由于 $n = 50$，自由度为 $df = n - 1 = 49$。从 t 表中，对于 $df = 49$ 的 95% 置信区间，临界值约为

$$t_c \approx 2.010.$$

使用 $s = 1.1$ 以及上面的数值，我们计算误差范围为

$$E = t_c \frac{s}{\sqrt{n}} = 2.010 \cdot \frac{1.1}{\sqrt{50}} \approx 0.313.$$

因此，以观测到的 $\bar{x}$ 为中心的 95% 置信区间为

$$\text{CI} = [19.98 - 0.313,\ 19.98 + 0.313] = [19.667,\ 20.293].$$

这里，我们得到与前一个例子相同的解释。

3 估计二项比例

3.1 定义样本比例

假设我们观察到二项实验的一部分。我们恰好观察了 n 次试验，并数出 r 次成功。由于样本中的随机变化，这还不足以推断真实的成功概率 p。然而，我们仍然可以用这些信息来估计 p 的值。和上一节以及第12讲一样，我们会围绕样本数据构造一个置信区间，从而给出估计值周围的适当误差范围。我们的第一个目标是得到一个有意义的样本统计量。

样本比例

如果在 n 次二项试验中观察到 r 次成功，那么样本比例定义为：

$$\hat{p} = \frac{r}{n}.$$

由此，我们也定义 $\hat{q} = 1 - \hat{p}$。

量 $\hat{p}$ 将作为真实成功概率 p 的点估计。

3.2 定义抽样分布

在第11讲讨论二项分布的正态近似时，我们提到，通常用来数二项实验中成功次数的二项随机变量 X 可以分解成若干 Y_i 的和，其中 Y_i 是为实验中的每一次单独试验定义的随机变量：

$$Y_i = \begin{cases} 1 & \text{如果第 } i \text{ 次试验成功,} \\ 0 & \text{如果第 } i \text{ 次试验失败.} \end{cases}$$

用符号表示，我们在第11讲中看到：

$$X = Y_1 + Y_2 + \cdots + Y_n$$

因此，把 X 除以 n 会得到一个像样本均值一样起作用的量。现在我们把这个思想与样本比例结合起来：如果定义

$$\hat{P} = \frac{X}{n} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n},$$

就可以引入一个记作 $\hat{P}$ 的随机变量。 事实上，这正是我们在第11讲中做过但没有命名的事情。这样写时，可以看到 $\hat{P}$ 是一个随机变量，它的观测值 $\hat{p}$ 像从大小为 n 的样本中取得的样本均值一样起作用——上面的观测值 r 是 n 个单独随机变量 Y_i 的结果之和，然后我们除以 n 得到均值。

我们在第11讲中看到，这些单独随机变量 Y_i 中的每一个都满足

$$E(Y_i) = p \qquad \text{并且} \qquad \mathrm{Var}(Y_i) = pq.$$

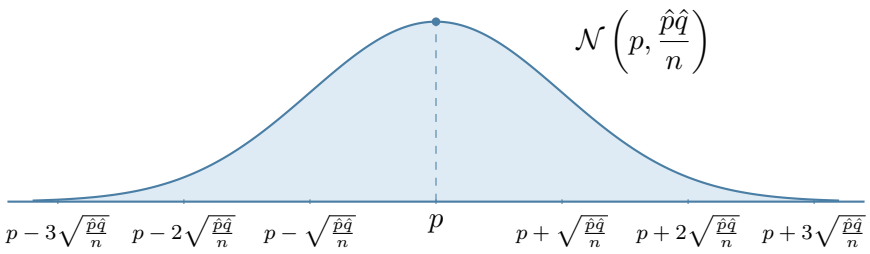
中心极限定理告诉我们，所有样本均值的集合近似服从正态分布，其中心是形成样本的原始变量的均值，而新的方差是原始变量的方差除以样本大小 n。在这里，中心极限定理告诉我们，$\hat{P}$ 近似服从参数为

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

的正态分布。

这似乎正好给了我们所需要的东西：我们刚刚确认了样本比例有一个近似正态分布，并且真实均值位于中心。从这里开始，看起来我们可以构造一些置信区间了。然而，有一个根本问题：我们正在试图估计 p，即真实成功概率。如果我们不知道 p 的精确值，那么我们也不知道 $\frac{pq}{n}$ 的值！换句话说，如果我们继续下去，想构造某种误差范围，那么在已经不知道 p 的情况下，我们到底该如何计算它?

在这里，我们必须接受：$\hat{P}$ 的真实抽样分布无法满足我们的需要。相反，我们只能做次好的事情，也就是用点估计替换方差项。实际中，我们用观测值 $\hat{p}$ 和 $\hat{q}$ 替换 p 和 q，得到以下近似：



这将成为我们的估计设定。

$\hat{P}$ 的抽样分布

对于二项实验，样本比例近似服从正态分布：

$$\hat{P} \approx \mathcal{N}\left(p, \frac{\hat{p}\hat{q}}{n}\right).$$

当

$$n\hat{p} > 5 \qquad \text{并且} \qquad n\hat{q} > 5$$

时，这个近似是合理的。

3.3 p 的置信区间

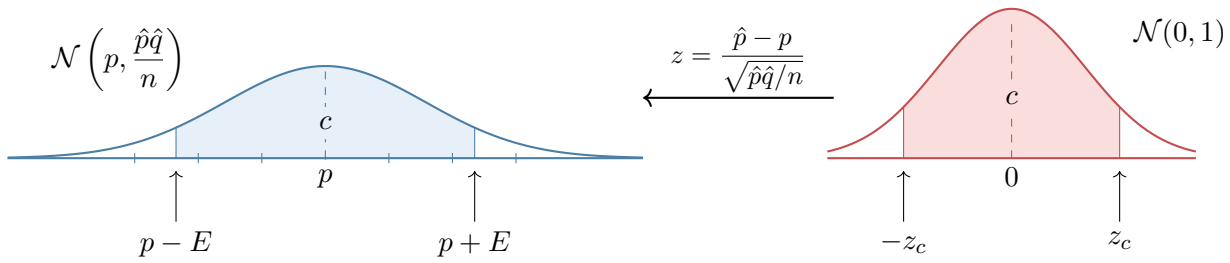
现在我们像之前一样构造置信区间。估计标准误差为

$$\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

因此，z 分数的公式变为：

$$z = \frac{\hat{p} - p}{\sqrt{\hat{p}\hat{q}/n}}.$$

为了推导置信水平为 c 的置信区间，我们在 $\mathcal{N}(0,1)$ 中找到临界值 z_c，使得 $P(-z_c < Z < z_c) = c$。然后反向使用 z 分数公式，把这个区域映回抽样分布：



误差范围 E 由 z_c 乘以标准误差给出。在这个情形中：

$$E = z_c\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

因此，观测值 $\hat{p}$ 周围的置信区间可以定义如下。

二项比例的置信区间

给定 n 次试验中的 r 次成功，

$$\hat{p} = \frac{r}{n}, \qquad \hat{q} = 1 - \hat{p}.$$

真实成功概率 p 的置信区间为

$$\left[\hat{p} - z_c\sqrt{\frac{\hat{p}\hat{q}}{n}},\ \hat{p} + z_c\sqrt{\frac{\hat{p}\hat{q}}{n}}\right].$$

3.4 方法

要估计二项成功概率 p：

步骤 1: 计算

$$\hat{p} = \frac{r}{n}, \qquad \hat{q} = 1 - \hat{p}.$$

步骤 2: 检查正态近似是否合理：

$$n\hat{p} > 5 \qquad \text{并且} \qquad n\hat{q} > 5.$$

步骤 3: 找到所选置信水平对应的临界值 z_c。

步骤 4: 计算误差范围：

$$E = z_c\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

步骤 5: 写出置信区间：

$$[\hat{p} - E, \hat{p} + E].$$

4 例题：篮球球员考察

假设某篮球队只想雇用真实罚球命中率至少为 70% 的球员。用符号表示，球队想要证据说明 $p \geq 0.70$。

你去考察一名球员，并观察他投罚球。你不能观察他一整天，所以你取他的表现作为样本，并决定计算一个 95% 置信区间来估计他的真实投篮成功率 p。

4.1 例1：95次出手命中65次

假设这名球员投了 $n = 95$ 次，命中其中 $r = 65$ 次。那么我们对成功概率的点估计为

$$\hat{p} = \frac{65}{95} \approx 0.684,$$

因此 $\hat{q} = 1 - \hat{p} \approx 0.316$。在继续之前，我们应先确认近似条件。我们有：

$$n\hat{p} = 95(0.684) \approx 65, \qquad \text{并且} \qquad n\hat{q} = 95(0.316) \approx 30.$$

两者都大于 5，所以正态近似是合理的。对于 95% 置信区间，有 $z_c = 1.96$。因此，误差范围为：

$$E = z_c\sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96\sqrt{\frac{(0.684)(0.316)}{95}} \approx 0.093.$$

因此，围绕观测值 $\hat{p} = 0.684$ 的 95% 置信区间为：

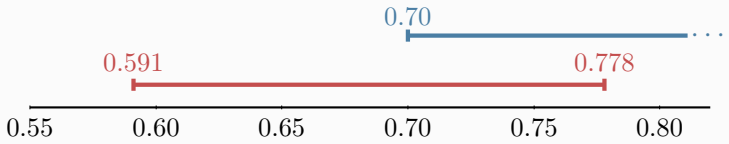
$$\text{CI} \approx [0.684 - 0.093,\ 0.684 + 0.093] = [0.591,\ 0.778].$$

解释

根据题目设定，我们寻找的目标阈值是 $p = 0.70$。然而，我们的置信区间是

$$[0.591,\ 0.778].$$

这个区间跨过了阈值：



因此，我们遇到的情况与第1.1节类似——我们不能决定性地说这名球员的真实投篮命中率高于或低于 70%。简单地说，这个样本在任一方向上都没有提供足够有说服力的证据。

4.2 例2：1000次出手命中773次

发现第一次观察没有结论之后，你回去第二次观察这名球员，这一次观察得更久。假设现在你看到这名球员投了 $n = 1000$ 次，命中其中 $r = 773$ 次。现在，我们对成功概率的点估计为

$$\hat{p} = \frac{773}{1000} = 0.773,$$

因此 $\hat{q} = 1 - \hat{p} = 0.227$。同样，我们应先再次检查近似条件是否成立。我们有：

$$n\hat{p} = 1000(0.773) = 773, \qquad \text{并且} \qquad n\hat{q} = 1000(0.227) = 227.$$

两者都大于 5，所以正态近似是合理的。再次使用临界值 $z_c = 1.96$，我们计算误差范围：

$$E = z_c\sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96\sqrt{\frac{(0.773)(0.227)}{1000}} \approx 0.026.$$

于是，围绕观测值 $\hat{p} = 0.773$ 的 95% 置信区间为：

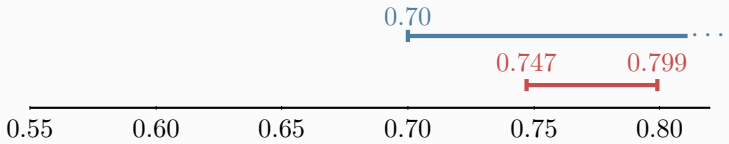
$$\text{CI} = [0.773 - 0.026,\ 0.773 + 0.026] = [0.747,\ 0.799].$$

解释

在这个情况下，我们的置信区间为

$$[0.747,\ 0.799].$$

整个区间现在都明显高于目标阈值 0.70：



由于这种方法大约在 95% 的情况下会捕捉真实的 p 值，我们有令人信服的证据表明真实的 p 值高于 0.70。因此，这个样本提供了令人信服的证据，表明这名球员达到了球队的标准。