

MAT220 第12讲：估计入门

目录

1	中心极限定理的复习	2
1.1	抽样分布	2
1.2	中心极限定理	2
2	估计的思想	3
2.1	σ 已知时的估计	3
2.2	样本均值周围的区间	4
2.3	置信水平的解释	5
3	构造置信区间	6
3.1	关于区间的提醒	6
3.2	抽样分布的 z 分数	7
3.3	求一个 50% 置信区间	7
3.4	求一个 95% 置信区间	8
4	置信区间的一般公式	9
4.1	临界值	9
4.2	一般置信区间	10
4.3	误差范围	11

在本讲中，我们将把正态分布的理论用于估计。基本问题很简单：如果我们不能测量整个总体，是否仍然可以对未知的总体参数，比如 μ ，作出合理猜测？正如我们将看到的，答案是可以。暂时我们会假设总体标准差 σ 已经知道。这样，我们就可以使用中心极限定理，把观测到的样本数据与真实总体均值联系起来。然后，利用正态分布的工具，我们会通过在样本均值 $\bar{x}$ 周围构造可接受的误差范围，来为 μ 的值作出估计。这种构造称为置信区间。

1 中心极限定理的复习

1.1 抽样分布

假设我们有某个总体分布，它的均值为 μ ，标准差为 σ 。如果我们取一个大小为 n 的简单随机样本，那么可以计算它的样本均值：

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

如果我们不断重复这个过程，就会得到许多可能的样本均值。所有这些样本均值的集合会形成一个新的分布，称为抽样分布。

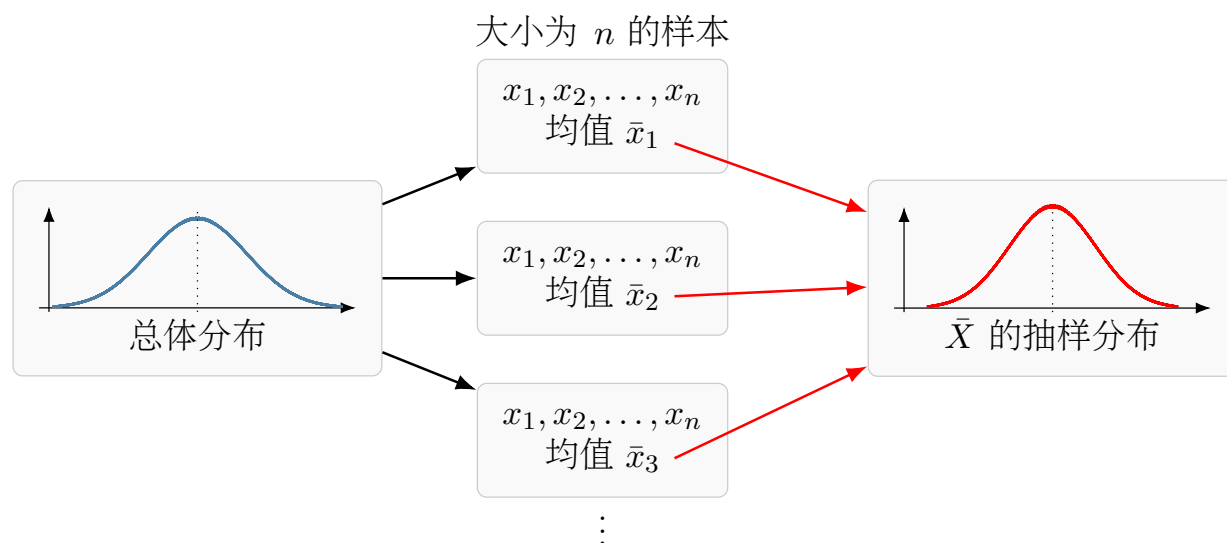
样本均值的抽样分布

给定一个概率分布，大小为 n 的样本对应的 $\bar{X}$ 的抽样分布，是由所有大小为 n 的简单随机样本所产生的全部样本均值形成的概率分布。

这里的重要点是，样本是随机抽取的。这意味着 $\bar{x}$ 的精确值也会是随机的。在我们取某个具体样本之前，我们不知道会得到哪些实际数值，因此也不知道 $\bar{x}$ 的精确值会是多少。所以，我们可以把样本均值提升为一个随机变量，记作 $\bar{X}$ 。它用来代表样本均值可能取到的所有值，而这些值取决于我们从总体中可能抽到的所有随机样本。

1.2 中心极限定理

一般来说，中心极限定理是一个数学结果，它告诉我们 $\bar{X}$ 的抽样分布是什么样子。如果原始总体是正态的，那么 $\bar{X}$ 的抽样分布就是精确的正态分布：



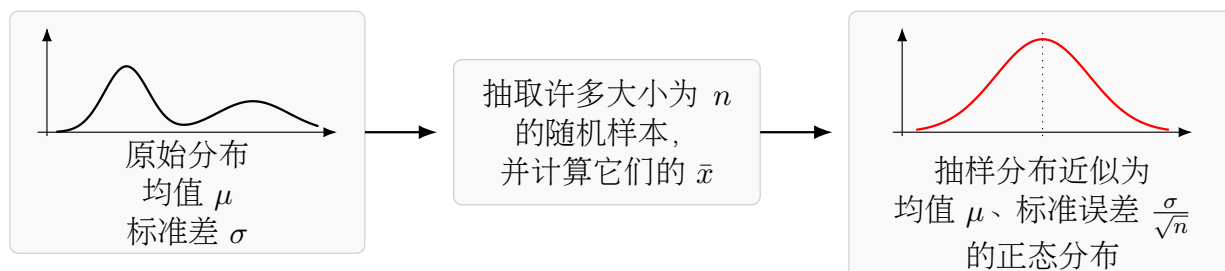
此外，如果原始总体不是正态的，那么中心极限定理说明，只要 n 足够大，得到的 $\bar{X}$ 的抽样分布也会近似正态。形式上，中心极限定理的一种表述如下。

中心极限定理

如果总体的均值为 μ ，标准差为 σ ，那么对于足够大的 n ， $\bar{X}$ 的抽样分布满足：

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

这意味着抽样分布以真实总体均值 μ 为中心，但它的分散程度小于原始总体的分散程度。样本大小 n 越大，样本均值的分散程度就越小。



在中心极限定理的表述中，抽样分布的方差等于特定的量 $\frac{\sigma^2}{n}$ 。它的平方根 $\frac{\sigma}{\sqrt{n}}$ 是抽样分布的标准差。为了避免把它与原始总体的标准差 σ 混淆，我们使用不同的术语。

标准误差

$\bar{X}$ 的抽样分布的标准差是

$$\frac{\sigma}{\sqrt{n}}.$$

这称为**标准误差**。

2 估计的思想

2.1 σ 已知时的估计

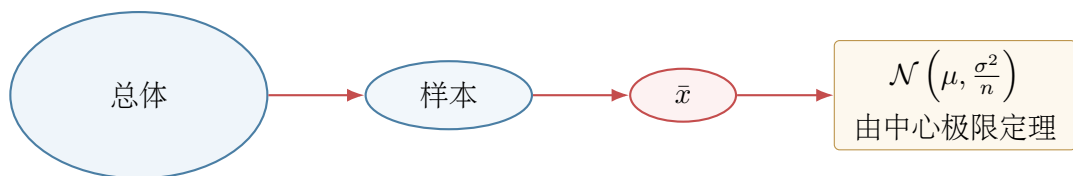
在真实的统计问题中，我们通常不知道真实总体均值 μ 。知道总体的真实平均数这样的信息非常有用，而且它当然是我们常常想知道的参数。然而，在实际操作中，我们不可能走进现实世界，把总体中的每一个对象都测量一遍，从而计算真实均值。

在本讲中，我们将假设：

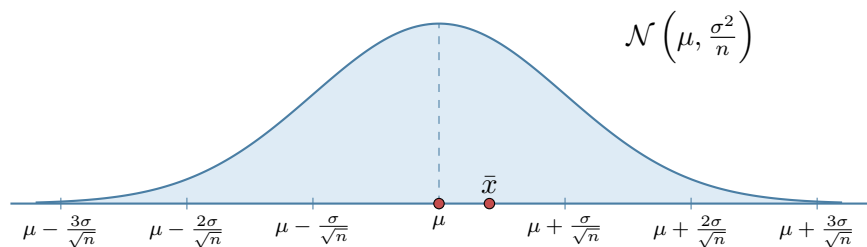
- 总体均值 μ 未知，但是
- 总体标准差 σ 恰好是已知的。

这显然是不现实的情形。然而，从理论上说，这是估计 μ 的最简单设置。下一讲中，我们会放松这个假设，并在 σ 也未知的现实情形中估计 μ 。

假设 σ 已知的好处是，我们可以使用中心极限定理，把样本均值与真实总体均值联系起来：



从图像上看，如果我们取一个简单随机样本，那么中心极限定理告诉我们，观测到的 $\bar{x}$ 值会位于这个正态分布中的某个位置：

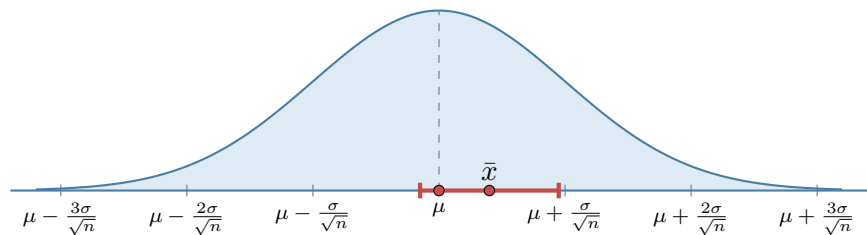


假设 σ 已知，使我们能够利用抽样分布 $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ 的结构。例如，我们不知道 μ 在哪里，但我们知道 $\bar{X}$ 约有 68% 的概率落在 μ 的一个标准误差之内。为什么？因为 $\bar{X}$ 的随机性被编码在分布的形状中。第10讲中的 z 分数计算告诉我们，分布面积的 68% 位于中心的一个标准误差之内。

重要的是，即使我们不知道 μ 实际上在哪里，这些想法仍然成立。根据中心极限定理， $\bar{X}$ 的随机性服从一个以真实总体均值为中心的正态分布。正如我们马上会看到的，即使我们不知道 μ 是什么，也至少可以把它与观测值 $\bar{x}$ 联系起来。

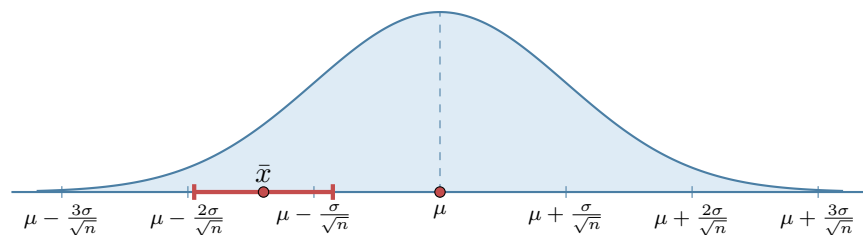
2.2 样本均值周围的区间

如果我们随机选取一个样本，那么所得均值 $\bar{x}$ 不一定会与 μ 完全相同。事实上，由于随机噪声，我们通常会预期每一个样本均值都与真实均值略有不同。不过，至少我们总是可以定义一种意义，使得 $\bar{x}$ 与 μ 足够接近。既然我们不知道 μ 是什么，与其说“这就是 μ 的精确位置”，不如说“ μ 可能在这里，在某个可接受的误差范围内”。在数学上，这个“可接受的误差范围”会表现为一个区间：



在上图中，我们随机选到的 $\bar{x}$ 值周围有一个固定长度的区间。虽然 μ 与 $\bar{x}$ 不同，但它仍然位于 $\bar{x}$ 的这个邻域中。

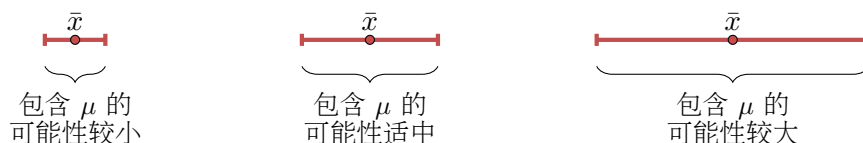
当然，有时样本在统计上是不太可能出现的，这会导致所得的 $\bar{x}$ 值远离中心：



由于随机性的本质，这必然会发生。事实上，给定 $\bar{X}$ 的某个随机值，我们最多只能说类似这样的话：

我们以 $c\%$ 的置信度认为，这个区间包含 μ 。

这最终是关于不确定性的陈述：我们并不知道 μ 在哪里，而只能说，这种特定长度的区间会在某个比例的情况下捕捉到它。当然，如果我们把区间做得更长，那么它击中 μ 的可能性会更大；如果我们把区间做得更短，那么它包含 μ 的可能性会更小：



因此，我们的置信程度似乎与区间长度直接相关：如果我们想非常确定随机得到的 $\bar{x}$ 周围的区间包含 μ ，那么应该把区间做得很大。相反，如果区间很小，那么我们就不太确定它会包含 μ 。这给出了置信区间的思想。

置信区间

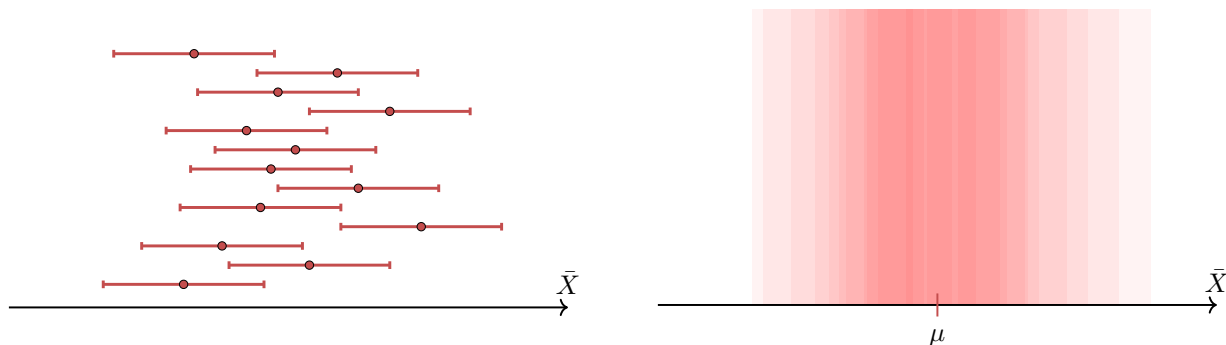
置信区间是对未知总体参数的区间估计。置信水平描述的是，在重复抽样中，该方法捕捉到真实参数的频率。

我们经常用字母 c 表示置信水平。这里的想法是， c 可以写成百分比，并且不同的置信水平对应不同的置信区间。如果我们想更有信心，那么就增大 c ，这会增加所得区间的长度。

2.3 置信水平的解释

为了说明，假设我们已经构造了一个置信水平为 $c = 95\%$ 的置信区间。根据构造，这意味着该方法产生的随机区间会在 95% 的情况下包含真实总体均值 μ 。换句话说，如果我们取 100 个简单随机样本，并围绕它们的样本均值计算 95% 置信区间，那么平均来说，我们会预期其中 95 个区间包含 μ ，而 5 个不包含。

下图有助于说明置信水平的长期含义：如果我们不断重复抽样，那么所得的大多数置信区间都会包含 μ ：



在右图中，我们把左侧的多个置信区间叠加在一起。可以看到，许多区间在真实总体均值 μ 附近重叠。

在讨论置信区间的构造之前，我们必须先澄清一个常见误解。

常见误解

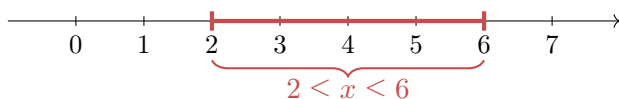
假设我们围绕某个观测到的 $\bar{x}$ 构造了一个 95% 置信区间。这里的“95%”并不表示该区间包含 μ 的概率是 95%。相反，它指的是整个方法的长期成功率。

说“这个区间有 95% 的概率包含 μ ”这类话，会暗示 μ 的值在以某种方式随机变化。这是不正确的： μ 是未知的，但它的值是完全固定的。对于围绕某个选定 $\bar{x}$ 的任意区间，均值 μ 要么在这个区间中，要么不在。相反，在分布中随机移动的是 $\bar{x}$ 的值，这用记号 $\bar{X}$ 表示。置信区间会随着 $\bar{x}$ 的值移动，而置信水平指的是这种移动。

3 构造置信区间

3.1 关于区间的提醒

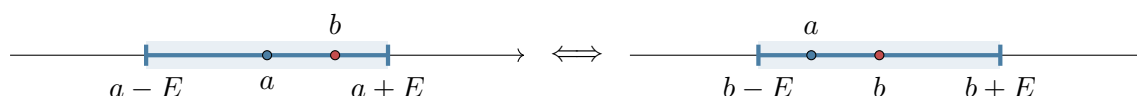
一般来说，一个连通区间完全由它的端点描述：我们只需写出区间从哪里开始、到哪里结束，然后就默认区间本身包含中间的一切。例如，区间 $[2, 6]$ 表示所有满足 $2 \leq x \leq 6$ 的数 x ：



现在假设 a 和 b 是实数轴上的两个点。说“ b 距离 a 在 E 以内”，等价于说“ a 距离 b 在 E 以内”。用符号写为：

$$a - E \leq b \leq a + E \quad \Longleftrightarrow \quad b - E \leq a \leq b + E.$$

从几何上看，如果 b 位于以 a 为中心、长度为 $2E$ 的区间内，那么 a 也位于以 b 为中心、长度为 $2E$ 的区间内：



在构造置信区间时，我们的目标是指定每个区间的两个端点。我们要寻找位于观测值 $\bar{x}$ 两侧的两个端点 $\bar{x}_1$ 和 $\bar{x}_2$ ，使得 $\bar{x}$ 正好处在中心。这样，用符号表示就是：

$$\mu \text{ 位于以 } \bar{x} \text{ 为中心的区间中} \iff \bar{x} - E \leq \mu \leq \bar{x} + E.$$

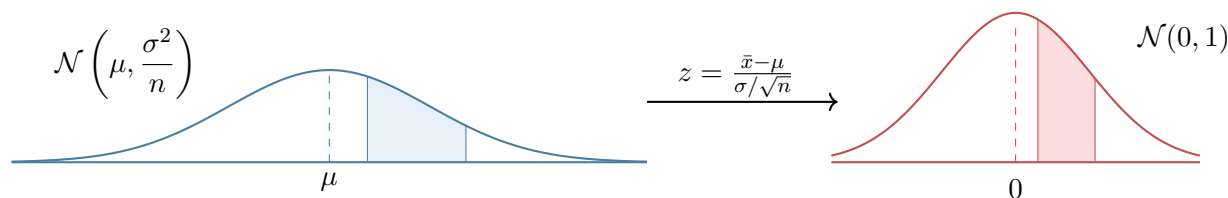
这个观察是置信区间背后的关键思想：如果 $\bar{X}$ 接近 μ ，那么 μ 也接近 $\bar{X}$ 。

3.2 抽样分布的 z 分数

根据中心极限定理，我们知道，对于足够大的 n ， $\bar{X}$ 的抽样分布可以近似为 $\bar{X} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ 。由于这是一个正态分布，我们可以通过构造 z 分数，把它与标准正态分布 $\mathcal{N}(0, 1)$ 联系起来。 z 分数公式测量一个值与中心之间的距离，然后除以该分布的标准差。对于抽样分布，中心是 μ ，而分布的尺度由标准误差 $\frac{\sigma}{\sqrt{n}}$ 衡量。因此，我们必须使用的 z 分数公式是：

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}.$$

这会把 $\bar{X}$ 的抽样分布转换为标准正态分布，并允许我们使用 z 表计算抽样分布中的面积：

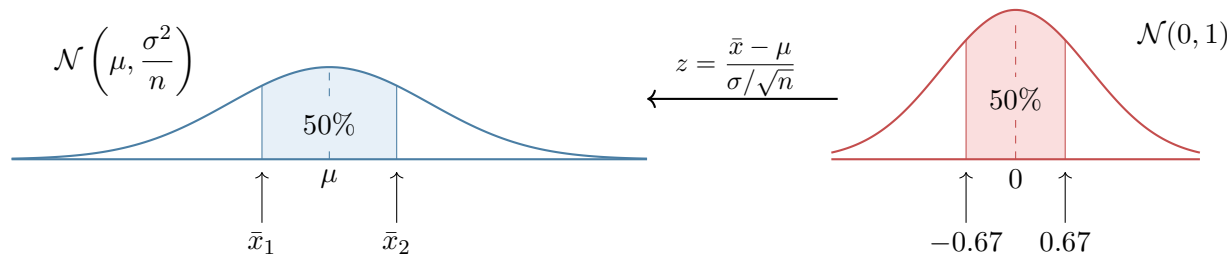


3.3 求一个 50% 置信区间

假设我们希望 $\bar{X}$ 以 50% 的概率接近 μ 。我们可以先在标准正态分布中寻找以 0 为中心、面积为 50% 的区域，从而描述 $\bar{X}$ 的值周围的置信区间。然后，我们可以使用 z 分数公式并反向计算，找出同一面积在抽样分布 $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ 中对应的边界。根据 z 表，可以看到，使两侧尾部面积相等并满足

$$P(z_1 < Z < z_2) = 0.5$$

的 z_1 和 z_2 的值正好是 $z_1 = -0.67$ 和 $z_2 = 0.67$ 。现在，我们可以利用这两个值，通过 z 分数公式反向求出抽样分布中的对应值：



这些值会围住 μ 周围 50% 的面积，也就是说， $\bar{X}$ 有 50% 的概率落在该区域中。

我们可以通过把 z 取为 ± 0.67 并进行代数求解，找到 $\bar{x}_1$ 和 $\bar{x}_2$ 的值。使用 $z = 0.67$ ，有：

$$0.67 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \implies \bar{x} = \mu + \frac{0.67\sigma}{\sqrt{n}}.$$

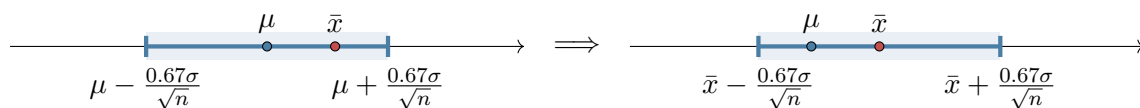
类似地，使用 $z = -0.67$ 得到：

$$-0.67 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \implies \bar{x} = \mu - \frac{0.67\sigma}{\sqrt{n}}.$$

现在，这两个端点可以用来描述一个围绕 μ 的区间，使得 $\bar{X}$ 有 50% 的概率落在其中。用符号表示，根据构造，该区间满足：

$$P\left(\mu - \frac{0.67\sigma}{\sqrt{n}} < \bar{X} < \mu + \frac{0.67\sigma}{\sqrt{n}}\right) = 0.5.$$

这里我们使用第3.1节中的观察：如果 $\bar{X}$ 在 50% 的情况下接近 μ ，那么这也意味着 μ 以概率 0.5 接近 $\bar{X}$ ：



因此，如果保持区间长度不变，并把它改为以 $\bar{X}$ 为中心，那么根据构造，这个区间会在 50% 的情况下包含 μ 。换句话说，我们刚刚围绕 $\bar{X}$ 构造了一个 50% 置信区间。写作：

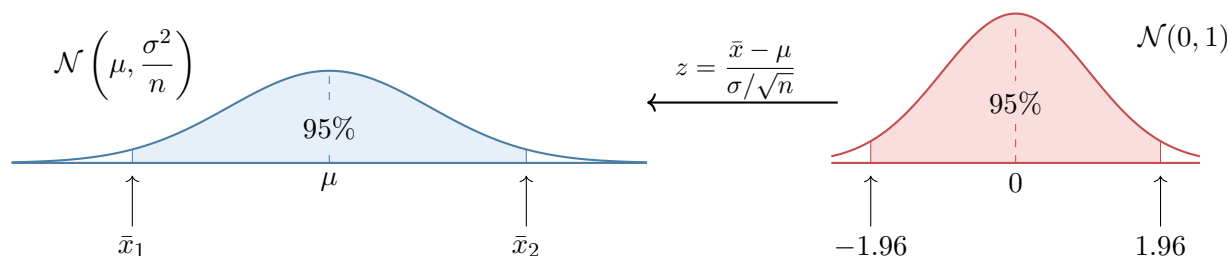
$$\left[\bar{X} - \frac{0.67\sigma}{\sqrt{n}}, \bar{X} + \frac{0.67\sigma}{\sqrt{n}}\right].$$

在观测到某个具体样本均值 $\bar{x}$ 之后，我们把围绕它的置信区间写作：

$$\left[\bar{x} - \frac{0.67\sigma}{\sqrt{n}}, \bar{x} + \frac{0.67\sigma}{\sqrt{n}}\right].$$

3.4 求一个 95% 置信区间

现在假设我们想构造一个 95% 置信区间。同样的逻辑仍然成立，只不过这一次，我们要寻找在 $\mathcal{N}(0, 1)$ 中以 0 为中心、面积为 95% 的 z 值。再次使用 z 表，这些值恰好是 ± 1.96 。现在我们使用 z 分数公式，找到抽样分布中对应端点的值：



现在使用 $z = \pm 1.96$ 而不是 ± 0.67 ，可得：

$$1.96 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \implies \bar{x} = \mu + \frac{1.96\sigma}{\sqrt{n}}$$

$$-1.96 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \implies \bar{x} = \mu - \frac{1.96\sigma}{\sqrt{n}}$$

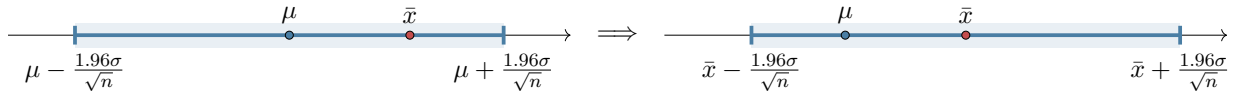
由与之前相同的逻辑， $\bar{x}_1$ 和 $\bar{x}_2$ 之间的区域会在 95% 的情况下包含 $\bar{X}$ ：

$$P\left(\mu - \frac{1.96\sigma}{\sqrt{n}} < \bar{X} < \mu + \frac{1.96\sigma}{\sqrt{n}}\right) = 0.95.$$

因此，如果我们取同一个区间，并改为以 $\bar{X}$ 为中心，那么 μ 也会在 95% 的情况下包含在这个新区间中：

$$P\left(\bar{X} - \frac{1.96\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right) = 0.95.$$

这一次，从几何上看，我们围绕 $\bar{x}$ 构造了一个更宽的区域：¹



用符号表示，围绕 $\bar{X}$ 的 95% 置信区间可以写成：

$$\left[\bar{X} - \frac{1.96\sigma}{\sqrt{n}}, \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right],$$

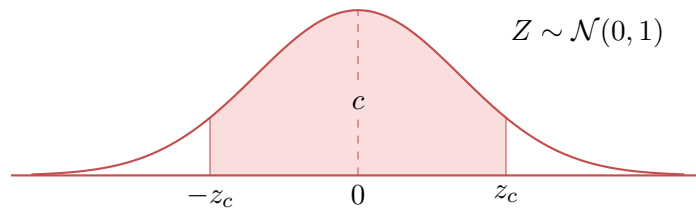
或者，如果已经观测到了 $\bar{x}$ 的值，则写成：

$$\left[\bar{x} - \frac{1.96\sigma}{\sqrt{n}}, \bar{x} + \frac{1.96\sigma}{\sqrt{n}}\right].$$

4 置信区间的一般公式

4.1 临界值

现在假设我们想构造某个置信水平 c 的置信区间。沿用上一节的思路，我们首先需要寻找标准正态分布中与置信水平 c 对应的面积：



¹这幅图并非按比例绘制。事实上，真实区间应当大约是上一节所画区间的 3 倍。

一般来说, c 可以取不同的值。因此, $\pm z_c$ 的精确值会根据你想让面积变得多大或多小而改变。这个数 z_c 有一个特殊名称: 它称为置信水平 c 的临界值。

临界值

对于置信水平 c , 临界值 z_c 是满足

$$P(-z_c < Z < z_c) = c,$$

的正数, 其中

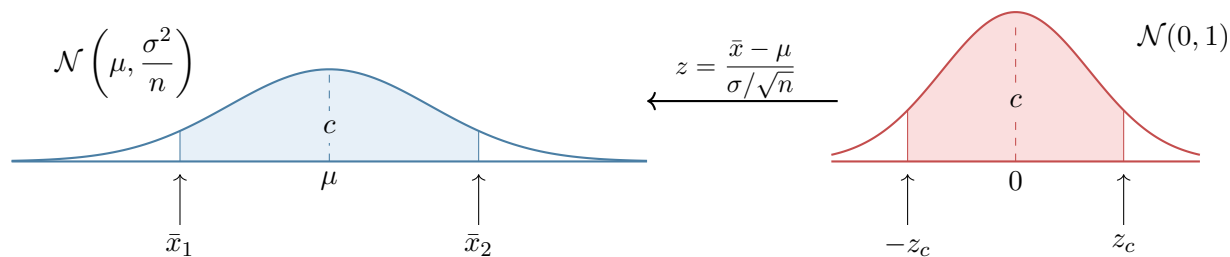
$$Z \sim \mathcal{N}(0, 1).$$

z_c 的精确值通常可以从 z 表或计算器中找到。然而, 在实际中, 我们经常计算常见置信水平的置信区间, 比如 90% 或 99%。为了方便, 下面的表列出了一些常见的临界值。

置信水平 c	临界值 z_c
0.70, 或 70%	1.04
0.75, 或 75%	1.15
0.80, 或 80%	1.28
0.85, 或 85%	1.44
0.90, 或 90%	1.645
0.95, 或 95%	1.96
0.98, 或 98%	2.33
0.99, 或 99%	2.58

4.2 一般置信区间

现在假设我们想构造某个置信水平 c 的置信区间。沿用上一节的思路, 我们首先需要寻找标准正态分布的中央面积 c 。一旦找到与 c 对应的临界值, 我们就可以求出 $\bar{X}$ 的对应端点:



解方程

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

来求 $\bar{x}_1$ 和 $\bar{x}_2$, 可得抽样分布中的端点为:

$$\mu - \frac{z_c \sigma}{\sqrt{n}} \quad \text{和} \quad \mu + \frac{z_c \sigma}{\sqrt{n}}.$$

因此，根据构造，我们找到了一个以 μ 为中心、以概率 c 包含 $\bar{X}$ 的区间。用符号表示：

$$P\left(\mu - \frac{z_c\sigma}{\sqrt{n}} < \bar{X} < \mu + \frac{z_c\sigma}{\sqrt{n}}\right) = c.$$

利用“接近”这一关系的对称性，我们可以把这个区间拿来，并围绕随机变量 $\bar{X}$ 构造一个随机区间，因此这个区间也会以相同的概率包含 μ ：

$$P\left(\bar{X} - \frac{z_c\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{z_c\sigma}{\sqrt{n}}\right) = c.$$

这就给出了置信区间的一般公式。

σ 已知时 μ 的置信区间

给定样本均值 $\bar{x}$ 、已知总体标准差 σ 、样本大小 n 以及置信水平 c ， μ 的置信区间为：

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

4.3 误差范围

注意，量

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

出现在置信区间的一般公式中。这个距离 E 告诉我们估计中可接受的误差界。因此， E 称为**误差范围**。使用这个术语，置信区间可以简单写成如下形式。

误差范围

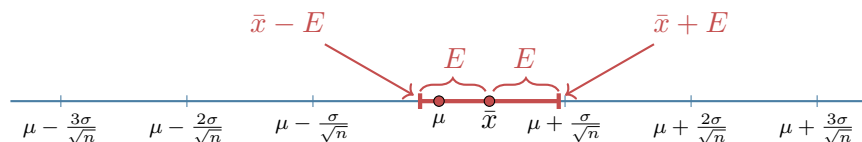
对于 σ 已知时 μ 的置信区间，误差范围为

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

置信区间为

$$[\bar{x} - E, \bar{x} + E].$$

从几何上看，长度 E 是置信区间总长度的一半：



因此，要在 σ 已知的情况下求 μ 的置信区间，我们按照以下步骤：

步骤 1： 使用置信水平 c 找到临界值 z_c 。

步骤 2: 计算误差范围:

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

步骤 3: 写出区间:

$$[\bar{x} - E, \bar{x} + E].$$