

MAT120: 论证 第15讲讲义

正态分布

上一讲中，我们开始学习统计学，了解了如何描述数据，并引入了两个核心概念：集中趋势和变异程度。这一讲中，我们研究最著名的连续型概率分布：正态分布。我们将学习什么是正态曲线，什么是标准正态分布，以及 z -score 如何让我们在不同的正态分布之间来回转换。我们还会学习，为什么对于连续变量而言，概率是通过曲线下的面积来衡量，而不是通过柱子的高度来衡量。下一讲中，我们将使用这些工具开始学习假设检验。

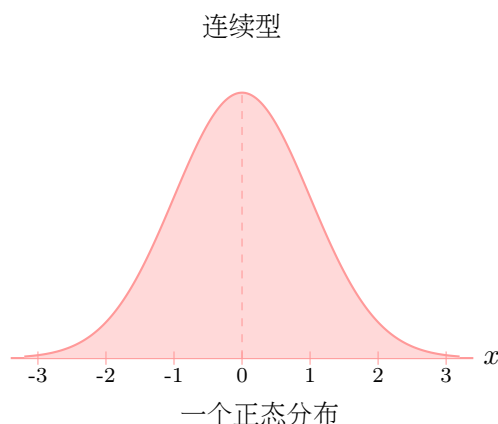
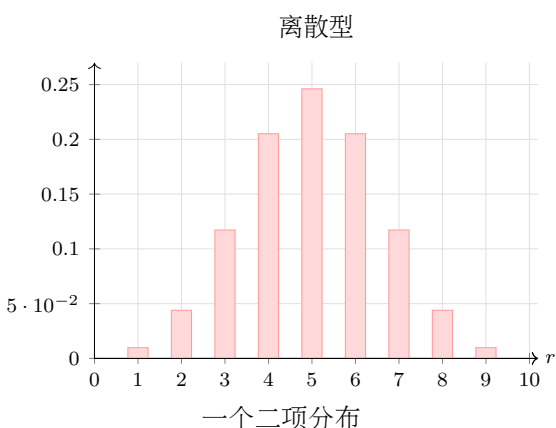
1 什么是正态分布？

1.1 离散型与连续型随机变量

随机变量是指其取值由随机性决定的变量。在当前阶段，有两类重要的随机变量需要记住：

- **离散型随机变量**：可以取分离的数值，例如 $0, 1, 2, 3, \dots$ 。
- **连续型随机变量**：可以在某个区间内取任意实数值。

在第13讲中，我们研究了一种由离散型随机变量构造出来的概率分布，也就是二项分布。今天我们要转向一种由连续型随机变量构造出来的概率分布。



连续型概率分布比我们目前见过的离散型分布具有更多的结构。这意味着有更多内容需要讨论，因此它们也可能更令人困惑。所以，我们的策略是利用极限，把连续情形和我们已经熟悉的离散情形联系起来，从而学习连续情形。

1.2 一个关于极限的比喻

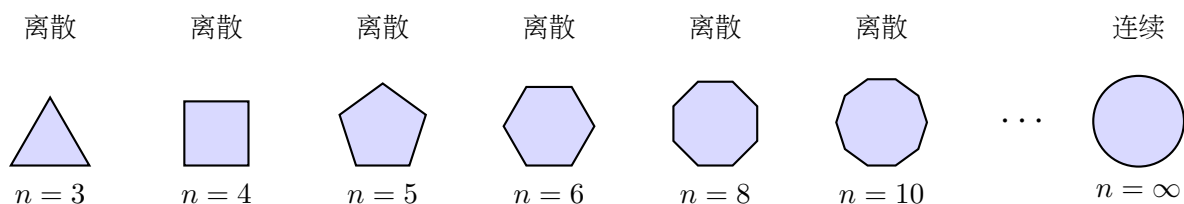
假设我们有一堆木头，它们的横截面具有不同的形状：



对于第一根木头来说，三角形的横截面会让它滚动起来非常别扭。如果我们换到下一根横截面为正方形的木头，就会发现它稍微容易滚一点。接着我们会注意到，五边形横截面的木头更容易滚，而六边形横截面的木头又更容易一些。我们看到，随着横截面的边数增加，木头会变得越来越

越容易滚动。当然，完美的圆形木头会是最容易滚动的。

如果我们有很多很多根这样的木头，并不断增加横截面的边数，那么就会得到一系列大致像下面这样排列的木头：

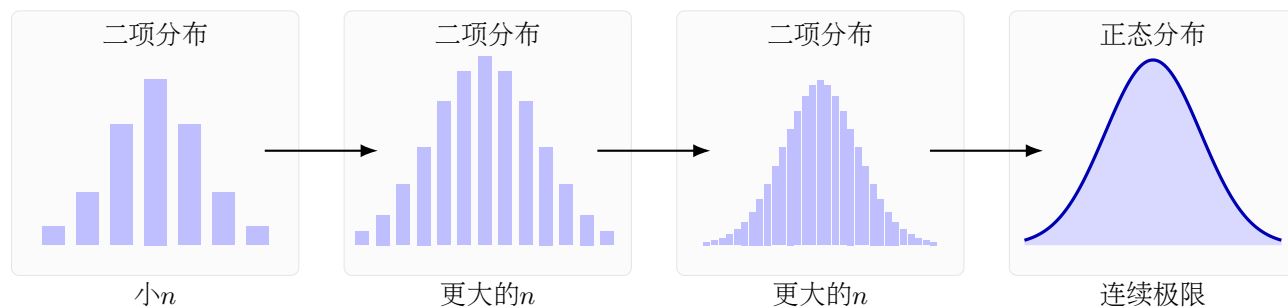


这个过程的极限就是一个圆，我们可以把它看成是一种具有无穷多条边的形状。如果把边数理解为一个离散量，那么我们会发现：在不断增加边数的过程中，这些多边形在极限下逼近了一个连续形状。

1.3 极限过程

我们关于滚木头的类比，其实有一个更深层的思想，也可以应用到随机变量上：我们可以通过考察一系列二项分布，并不断增加试验次数，从而让柱子的数量不断增加，以此来“逼近”一个正态分布。换句话说：一条连续曲线可以看作许多离散柱子的极限。

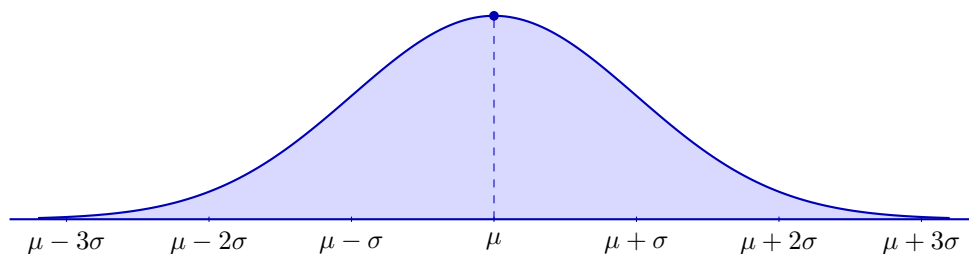
回忆一下，在二项分布中，图上的柱子总数是 $n + 1$ ，因为可能的取值从 0 到 n 。当 n 增加时，二项分布中的柱子会变得更细、更密。整个分布看起来会越来越平滑，并最终开始像一条曲线：



这条极限下呈钟形的曲线，就叫作正态分布。

1.4 正态分布与正态曲线

正态密度函数的图像叫作正态曲线。由于它的形状，它也常常被称为钟形曲线。



正态分布是统计学中最重要的对象之一。在实际中，我们会在人类特征、物理测量、工业数据、测量误差以及许多其他情境中看到正态分布。

一个正态分布有几个重要性质。

正态分布的重要性质

1. 曲线呈钟形，最高点位于均值 μ 的正上方。
2. 曲线关于经过 μ 的竖直线对称。
3. 两侧尾部会无限接近水平轴，但永远不会碰到或穿过它。
4. 整条曲线下方的总面积为1。

1.5 数学描述

既然正态分布是由一条图像表示的，我们自然可以问：与之对应的函数是什么？这个函数有一个专门的名字，叫作正态密度函数。虽然我们并不需要直接使用这个函数，但它写作：

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

尽管这个函数显然有些复杂，但我们还是应该注意几个重要事实。首先，这个函数只有一个变量 x 。这个 x 就是连续型随机变量，它构成了图像的横轴。接着我们看到，除了 x 之外，函数 $f(x)$ 还需要两个重要参数：

- μ 是分布的均值。
- σ 是分布的标准差。

从上面的函数可以看出，除了像 π 这样的常数之外， μ 和 σ 是函数中仅有的参数信息。因此，每一个正态分布都被 μ 和 σ 这两个数完全刻画。这意味着，不同的 μ 和 σ 取值会对应不同的正态曲线。正因为如此，我们通常把一个正态分布记作：

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

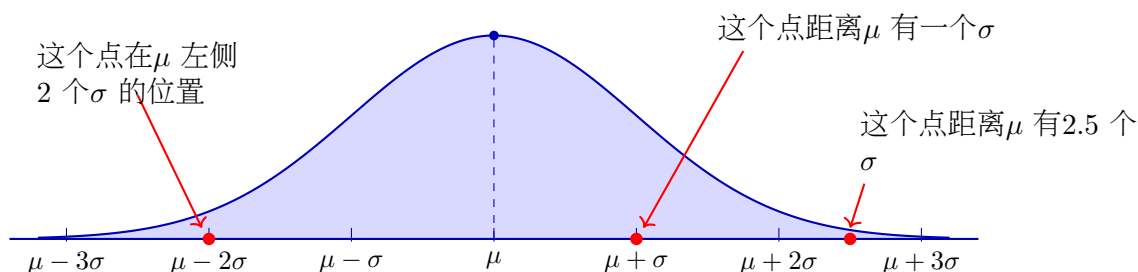
这句话读作：“连续型随机变量 X 服从一个均值为 μ 、方差为 σ^2 的正态分布。”

这是正态分布一个相当特别的性质。总体均值 μ 告诉我们分布的中心在哪里，而总体标准差 σ 则衡量数据有多分散。通常来说，它们只是帮助描述一个数据集的两个数字。然而，在正态分布的情形中， μ 和 σ 这两个数完全描述了整个分布。因此，符号 μ 和 σ 就是我们掌握整条曲线所需的全部信息。

2 标准正态分布

2.1 z-score

根据前面的讨论，我们看到，任何正态分布 $\mathcal{N}(\mu, \sigma^2)$ 的 x 轴都可以用 μ 和 σ 来描述。如果我们随机选取随机变量 X 的一个值，它就会落在这个 x 轴上的某个位置。由于这个 x 轴以 μ 为中心，并以 σ 为单位刻度，我们就可以把一个数值的位置表示成它离中心有多远。例如：



z -score，也叫作标准分数，就是把这些信息变成一个纯粹的数字，它告诉我们：一个观测值距离均值有多少个标准差。

z -score 的定义

如果 x 是一个均值为 μ 、标准差为 σ 的正态分布中的数值，那么它的 z -score 定义为

$$z = \frac{x - \mu}{\sigma}.$$

这个公式可以逐部分理解：

- $x - \mu$ 衡量的是 x 距离均值有多远；
- 再除以 σ ，就把这个距离转换成“多少个标准差”；
- 最终结果可以是正的，也可以是负的，这分别对应均值右侧和左侧。

例如，在上面的图中：

- 位于 $\mu - 2\sigma$ 的点，其 z -score 是-2；
- 位于 $\mu + \sigma$ 的点，其 z -score 是1；
- 位于 $\mu + 2.5\sigma$ 的点，其 z -score 是2.5。

有时我们知道 z -score，想反推出原始数值 x 。从公式

$$z = \frac{x - \mu}{\sigma},$$

出发，整理可得：

$$x = \mu + z\sigma.$$

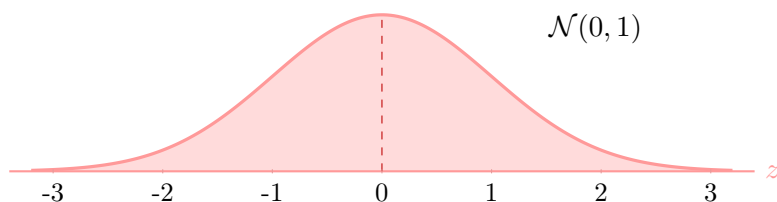
所以， z -score 与原始分数包含的是同样的信息，只是写法不同而已。

2.2 标准正态分布

标准正态分布就是均值为0、标准差为1 的正态分布。我们通常把它写作

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

注意，这里我们用特殊字母 Z 来表示标准正态随机变量，而不是前面使用的 X 。作为一个约定，我们会把标准正态分布 $\mathcal{N}(0, 1)$ 画成红色，这样就能在视觉上把它和一般的正态分布区分开来，后者我们会画成蓝色。标准正态分布如下图所示。



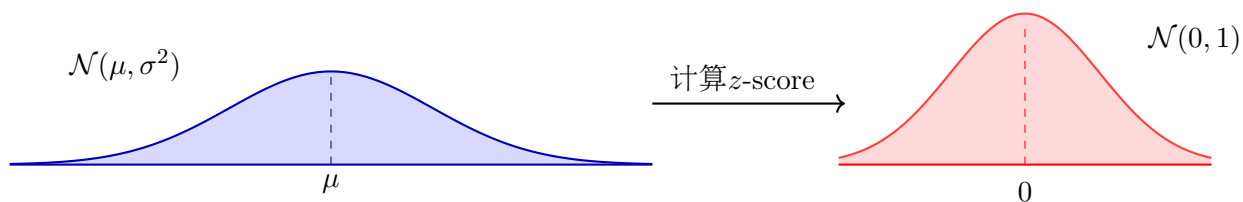
注意这里的横轴：它以0 为中心，并以1 为步长向左右延伸。

2.3 为什么标准正态分布很有用

任何一个正态分布都可以通过把它的观测值转换成 z -score，而变成标准正态分布。这就是为什么标准正态分布如此重要：我们不需要为每一条可能的正态曲线都去学习一张不同的面积表，而是只需要使用一张通用的参考曲线。

核心思想

每一个正态分布都可以被翻译到标准正态的世界里。一旦做到了这一点，概率就可以通过同一张标准正态面积表来读取。



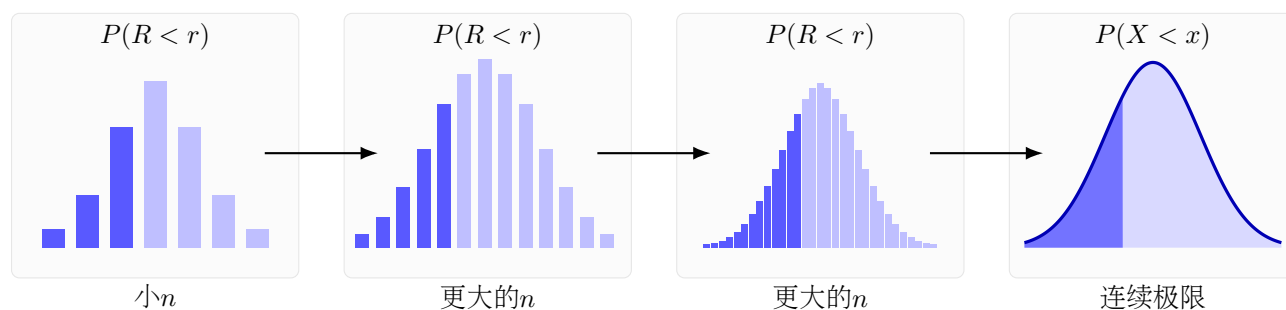
3 面积与概率

3.1 从柱高到面积

在处理二项分布时，每一根柱子的高度就给出了离散型随机变量取到那个值的精确概率。例如，在 $n = 10$ 且 $p = 0.5$ 的二项分布中：

- 取到恰好 $r = 5$ 的概率，就是 $r = 5$ 处那根柱子的高度，大约是0.24；
- 取到 $r < 5$ 的概率，则是把 $r = 0, 1, 2, 3, 4$ 这些柱子的高度加起来，结果大约是0.38。

当 n 越来越大时，这种“把许多细柱相加”的过程，就会越来越像一个连续面积：



正如图所展示的那样，对于连续型随机变量，我们把概率理解为曲线下的面积，而不是柱高（或者柱高之和）。

离散概率与连续概率

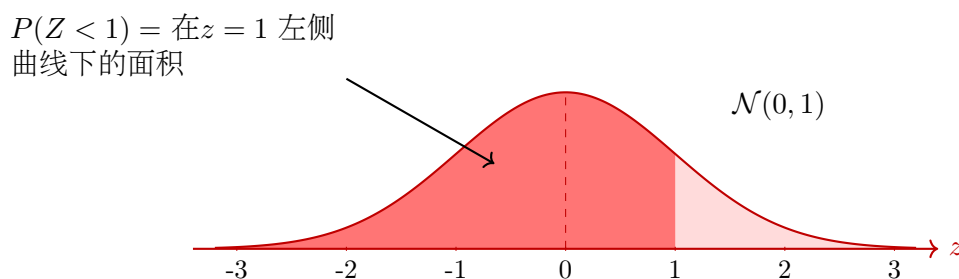
- 在离散情形中，概率是通过把各个离散结果的概率相加得到的。
- 在连续情形中，概率是通过测量曲线下的面积得到的。

3.2 如何找到这些面积

再来看标准正态分布 $\mathcal{N}(0, 1)$ 。根据上面的讨论，当我们写

$$P(Z < z),$$

时，我们的意思是：标准正态曲线在点 z 左侧下方的面积。例如，概率 $P(Z < 1)$ 就由下面阴影部分的面积给出：



如果你学过一点微积分，那么这些面积可以通过积分求出。然而，这通常比较困难，所以在实践中，我们往往会改用 z 表。

我们可以把 z 表理解为：它列出了标准正态分布中一些常见区域所对应的答案。在本课程中，我们主要使用下面这张表：

z	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0
$P(Z < z)$	0.02	0.07	0.16	0.31	0.50	0.69	0.84	0.93	0.98

这里，第二行给出了特定 z 值左侧的面积。例如，在上面的图中，这张表告诉我们，阴影面积是0.84，也就是说，曲线下总面积中有84% 位于 $z = 1$ 的左侧。用概率语言来说，这表示随机变量 Z 取值小于1 的概率大约是84%。

3.3 理解面积

z 表列出的是标准正态随机变量 Z 在标准正态曲线下的面积。我们把这张表解释为对概率

$$P(Z < z),$$

的近似回答，也就是 z 左侧的面积。利用正态分布的基本性质，我们还可以用这张表来描述其他区域的面积。

如何求其他面积

- z 左侧的面积：直接读取 $P(Z < z)$ 。
- z 右侧的面积：用 $P(Z > z) = 1 - P(Z < z)$ 。
- 两个值之间的面积：用 $P(a < Z < b) = P(Z < b) - P(Z < a)$ 。
- 负值：利用对称性，当 $a > 0$ 时，

$$P(Z < -a) = 1 - P(Z < a).$$

对于连续型随机变量，任何一个单独精确点的概率都等于零。例如，

$$P(Z = 0.5) = 0.$$

这并不意味着这个点是不可能的。它只是意味着，单独一个点的宽度为零，因此面积也为零。

3.4 计算其他正态分布中的概率

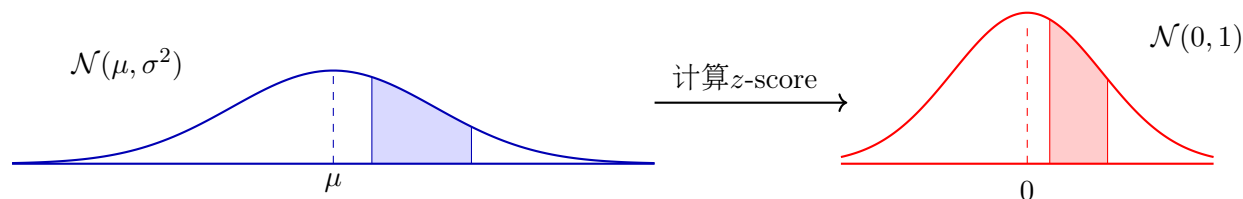
在第2 节中我们提到过，我们可以使用 z -score 公式

$$z = \frac{x - \mu}{\sigma}$$

把任意一个正态分布 $\mathcal{N}(\mu, \sigma^2)$ 转换成标准正态分布 $\mathcal{N}(0, 1)$ 。这里有用的一点在于：这个变换过程会保持面积比例不变，因此，我们可以借助 z -score 和上面的 z 表，来确定 $\mathcal{N}(\mu, \sigma^2)$ 中的概率。对于随机变量 $X \sim \mathcal{N}(\mu, \sigma^2)$ ，我们有：

$$P(X < x) = P\left(Z < \frac{x - \mu}{\sigma}\right).$$

换句话说，为了求 $P(X < x)$ 的值，我们只需要先计算点 x 的 z -score，然后到 z 表中查找对应的面积。这个思路对其他各种面积也同样成立。例如，像 $P(x_1 < X < x_2)$ 这样的概率，也可以通过分别求出 x_1 和 x_2 的 z -score 来确定：



这正是 z -score 的真正威力：它让我们可以利用标准正态表来处理任意一个正态分布。

例题

设 $X \sim \mathcal{N}(10, 2^2)$ 。求一个随机选出的数值落在11 和14 之间的概率。用符号写，就是求 $P(11 \leq X \leq 14)$ 。

解答

第1 步：把 x 值转换成 z 值。

$$z_1 = \frac{11 - 10}{2} = 0.5, \quad z_2 = \frac{14 - 10}{2} = 2.0.$$

第2 步：从表中读取左尾面积。

$$P(Z < 0.5) = 0.69, \quad P(Z < 2.0) = 0.98.$$

第3 步：相减得到区间概率。

$$P(11 \leq X \leq 14) = P(Z < 2.0) - P(Z < 0.5) = 0.98 - 0.69 = 0.29.$$

所以这个概率是

$$0.29.$$

也就是说，从这个正态模型中随机选取一个数值，它落在11 到14 之间的概率大约是29%。

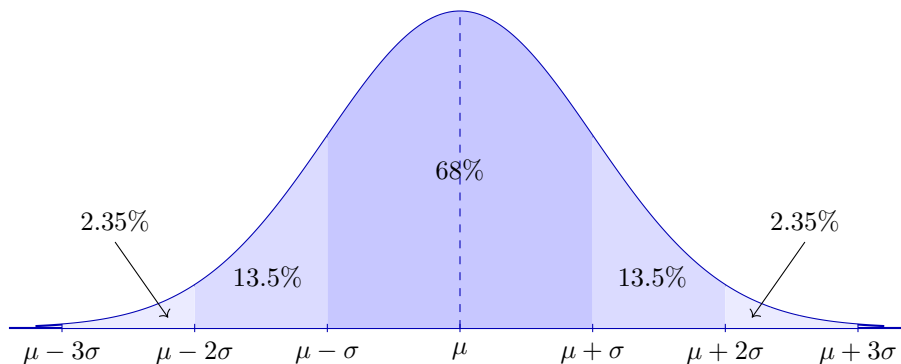
3.5 68-95-99.7 规则

对于正态分布来说，标准差为我们提供了一个非常有用的几何刻画。

经验法则

对于一个正态分布：

- 大约有68% 的数据落在均值左右1 个标准差以内；
- 大约有95% 的数据落在均值左右2 个标准差以内；
- 大约有99.7% 的数据落在均值左右3 个标准差以内。



剩下的0.3% 位于两端极端尾部，每一边大约是0.15%。

3.6 例题：标准正态曲线下的面积

利用上面的简化 z 表，再结合对称性，求出下面各个面积。

练习

- (a) 求 $z = -1.00$ 左侧的面积。
- (b) 求 $z = 2$ 左侧的面积。
- (c) 不用 z 表，求 $z = 0$ 左侧的面积。
- (d) 不用 z 表，求整条曲线下方的总面积。
- (e) 求 $z = 1$ 右侧的面积。
- (f) 求区间 $-1 < Z < 2$ 的面积。

解答

- (a) 由表可知，

$$P(Z < -1.00) = 0.16.$$

- (b) 由表可知，

$$P(Z < 2) = 0.98.$$

- (c) 由于对称性，总面积的一半位于0 左侧，因此

$$P(Z < 0) = 0.50.$$

- (d) 整条曲线下方的总面积始终等于1。

- (e)

$$P(Z > 1) = 1 - P(Z < 1) = 1 - 0.84 = 0.16.$$

- (f)

$$P(-1 < Z < 2) = P(Z < 2) - P(Z < -1).$$

由表可得，

$$P(-1 < Z < 2) = 0.98 - 0.16 = 0.82.$$