MAT220 第11讲：中心极限定理

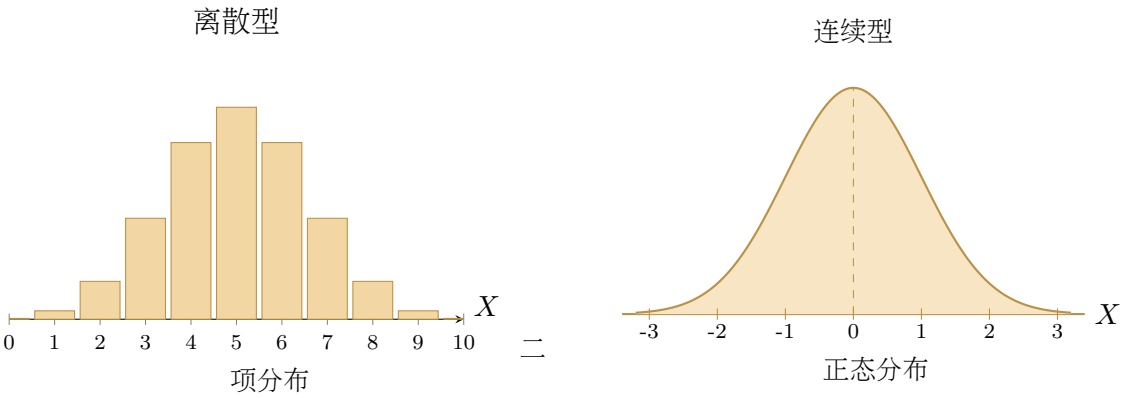
目录

在第10讲中，我们学习了正态分布以及 z 分数的使用方法。 在本讲中，我们解释为什么正态分布会如此频繁地出现在统计学中。核心思想是所谓的中心极限定理。它给出了一种方法：即使原来的系统一开始并不服从正态分布，我们仍然可以把来自许多系统的样本均值与一个正态分布联系起来。事实上，这个结果非常重要，所以我们会把整节课都用来讲它。首先，我们会仔细地陈述这个结果。然后，我们会用它来说明如何用正态曲线近似二项分布。

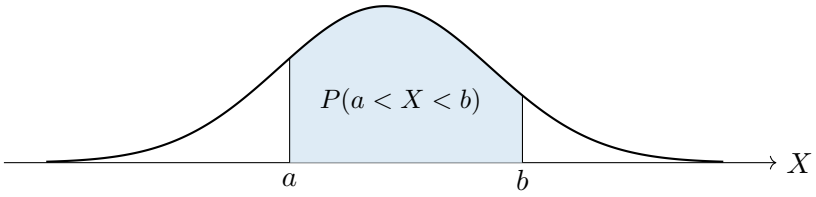
1 正态分布的简要复习

1.1 离散型随机变量与连续型随机变量

上一讲中，我们引入了正态分布。正态分布是由连续型随机变量构造出来的概率分布，而不是由离散型随机变量构造出来的。这与我们之前见过的二项分布看起来非常不同：



这里，根本区别在于 x 轴上可能出现的值。对于二项分布，随机变量数的是成功次数。这会使随机变量只取离散的值，并且 X 能取的每一个具体值都对应一个概率。相比之下，对于正态分布，随机变量 X 可以取任意实数值。因此，概率问题不再用柱子的高度来表达，而是用面积来表达：X 落在某个区域中的概率，由曲线下方对应区域的面积大小给出，如下图所示。



这里，随机变量 X 落在 a 与 b 之间区域内的概率，由蓝色阴影区域的面积来计算。

1.2 描述一个正态分布

一般来说，均值和标准差是描述数据集某些特征的两种不同方式。对于正态分布而言，只需要这两个数，就足以定义整个分布。这意味着我们可以用如下数学记号写出一个正态分布。

正态分布

正态分布是一种连续型概率分布，记作

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

这里，μ 是均值，σ 是标准差，σ^2 是方差。

均值和标准差足以完整描述一个正态分布。与数据集的情况一样，均值 μ 告诉我们分布的中心在哪里，标准差 σ 告诉我们分布有多分散。正态分布总是满足以下有用性质。

四个重要性质

正态分布具有以下性质。

1. 曲线呈钟形，最高点位于均值 μ 的正上方。

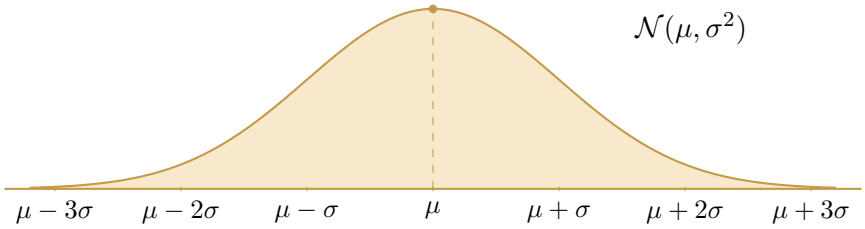
2. 曲线关于穿过 μ 的竖直线对称。

3. 两端的尾部趋近于水平轴，但永远不会碰到或穿过它。

4. 整条曲线下方的总面积为 1。

1.3 z 分数

上面列出的四个性质意味着，我们可以按照如下方式理解正态分布：



这里，图像是对称的，并且正好以 μ 为中心。由于 σ 与 μ 使用相同的物理单位，我们可以从图像中心出发，按照 σ 的增量向外“走”。这使我们能够用距离 μ 的远近来描述 X 的可能取值。例如，如果 X 取到一个离 μ 很远的值 x，那么 $x - \mu$ 可能是 3σ 或 -3σ；如果 X 取到一个相对接近 μ 的值，那么它可能离 μ 有 0.5σ。

如果我们想在“距离 μ”的度量中去掉 σ，那么只需除以 σ，就会得到一个纯数。这个数称为 z 分数：

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

简而言之，一个值 x 的 z 分数表示 x 离 μ 有多远，以 σ 为单位来度量。z 越大，x 离 μ 越远。如果整体的 z 分数为负，那么 x 位于 μ 的左侧；如果 z 分数为正，那么 x 位于 μ 的右侧。

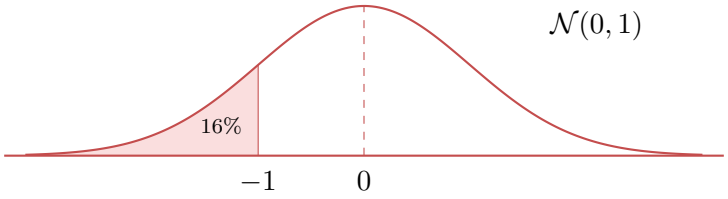
1.4 标准正态分布

在第10讲中，我们也花了一些时间讨论“标准”正态分布。这里的想法是：它只是正态分布中最简单的版本。它的均值等于 0，标准差等于 1。由于这个特殊分布非常好用，我们用特殊字母 Z 来表示它的随机变量。用数学记号写，标准正态分布可以表示为 $Z \sim \mathcal{N}(0,1)$。

方便的是，我们可以使用 z 表来近似标准正态分布的面积，也就是概率。这些表列出了常见 Z 值左侧的面积。例如，一个小型版本的表可能如下：

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

更详细的表可以在第10讲的附录中找到。从图像上看，表中的条目告诉我们给定值左侧的面积。例如，表中说 $P(Z < -1) \approx 0.16$, 这表示在 $\mathcal{N}(0,1)$ 中，-1 左侧的面积约占总面积的 16%：



像上面这样的表只给出给定值左侧的概率。其他概率可以利用正态分布的性质来计算。下面列出一些有用的事实。

概率的一般规则

对于 $Z \sim \mathcal{N}(0,1)$：

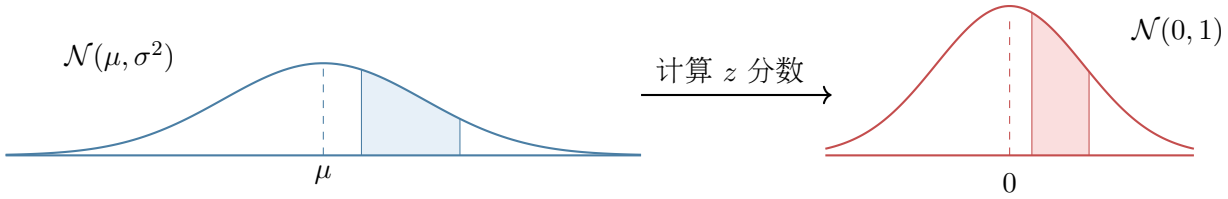
- $P(Z > z) = 1 - P(Z < z)$
- $P(z_1 < Z < z_2) = P(Z < z_2) - P(Z < z_1)$
- $P(Z > z) = P(Z < -z)$
- $P(Z = z) = 0$

1.5 计算其他概率

如果 $X \sim \mathcal{N}(\mu, \sigma^2)$ 是一般正态分布，那么我们可以通过计算 z 分数，把一个值 x 转换到标准正态分布中：

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

这个变换保持面积不变，因此 X 落在某个特定区域中的概率，可以先转换到标准正态分布，然后在那里进行计算：



例：身高

为了说明这一点，假设日本男性身高服从正态分布，均值为 171 cm，标准差为 5.6 cm。例如，我们可以通过 z 分数计算遇到身高超过 180 cm 的人的概率。这里我们要求的是 $P(X > 180)$，因此可以计算相应的 z 分数：

$$z = \frac{180 - 171}{5.6} \approx 1.61.$$

这个变换保持概率不变，因此我们问题的答案可以通过计算下面的概率得到：

$$P(X > 180) \approx P(Z > 1.61),$$

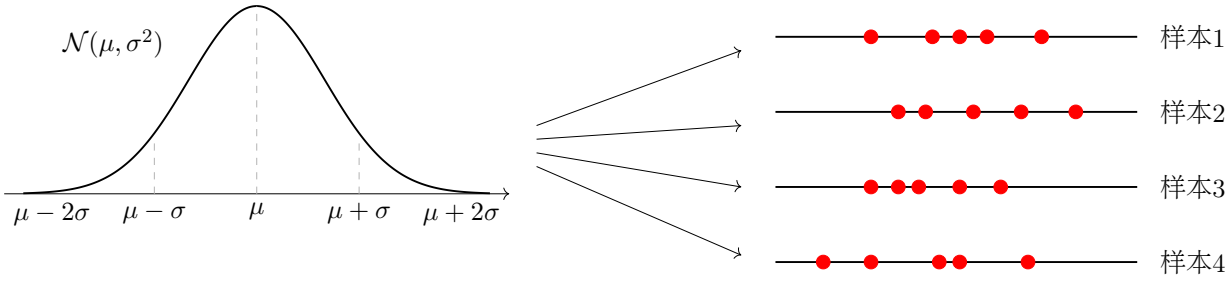
结果约为 5.4%。

2 抽样分布

2.1 对随机变量进行抽样

现在考虑一般正态分布 $X \sim \mathcal{N}(\mu, \sigma^2)$。正如我们之前提到的，$X$ 落在任何给定区域中的概率，可以由该区域中曲线下方的面积决定。因此，我们会预期 X 更可能落在对应面积较大的区域中。

现在，假设我们从这个分布中抽取一个随机值的样本。为了简单起见，我们随机选取 5 个值。使用第1讲的术语，这称为大小为 5 的简单随机样本。一般来说，这 5 个值可以落在 x 轴上的任何位置。因此，如果我们反复抽取大小为 5 的简单随机样本，它们可能看起来相当不同：



对于每一个单独样本，我们总是可以计算样本均值：

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}.$$

现在，关键观察是：由于我们的样本由随机值组成，样本均值 $\bar{x}$ 的具体值也会是随机的。因此，我们可以把样本均值本身解释为一个随机变量，并把这个随机变量记作 $\bar{X}$。接下来，我们将探究随机变量 $\bar{X}$ 的分布是什么样的。

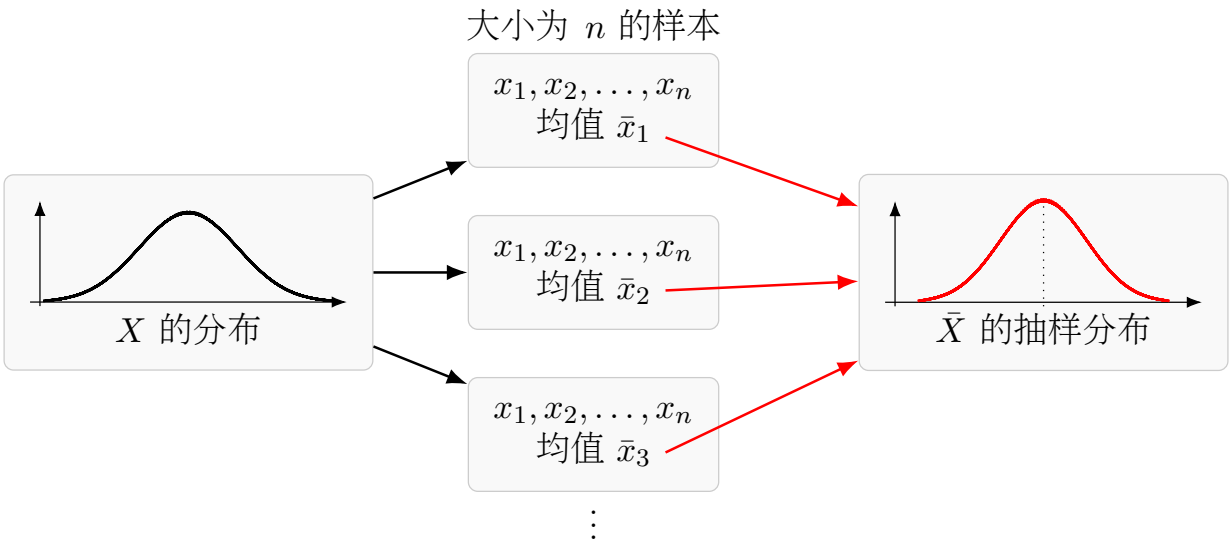
2.2 样本均值的抽样分布

想象从一个总体中抽取所有可能的大小为 n 的简单随机样本。我们可以计算每个样本对应的样本均值，并把所有这些结果收集在一起，形成一个新的概率分布。这个新对象称为抽样分布。

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。

我们可以把抽样分布的构造想象成如下过程：



换句话说，我们从一个概率分布开始，取所有大小为 n 的简单随机样本，计算所有对应的样本均值 $\bar{x}$，然后把它们收集在一起。由于这些样本一开始就是随机抽取的，样本均值 $\bar{x}$ 可以被看作随机变量 $\bar{X}$ 的取值。因此，所有可能的 $\bar{x}$ 的集合具有概率分布的结构。

大小为 n 的样本应当被看作随机变量 $X_1, \ldots, X_n$，它们每一个都与 X 具有相同的分布，并且

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}.$$

3 中心极限定理

正如我们可能预期的，不同的概率分布通常会有不同的抽样分布。中心极限定理告诉我们，根据原始分布的形状，这些抽样分布会呈现什么样子。现在我们给出两个相关结果：一个针对正态分布，另一个针对更一般的概率分布。

在本课程中，我们不会证明这些定理。相反，我们主要关注中心极限定理在估计和假设检验等技术中的应用。所以，我们的目标只是陈述并理解这个结果。

3.1 版本1：原始总体为正态分布时的抽样分布

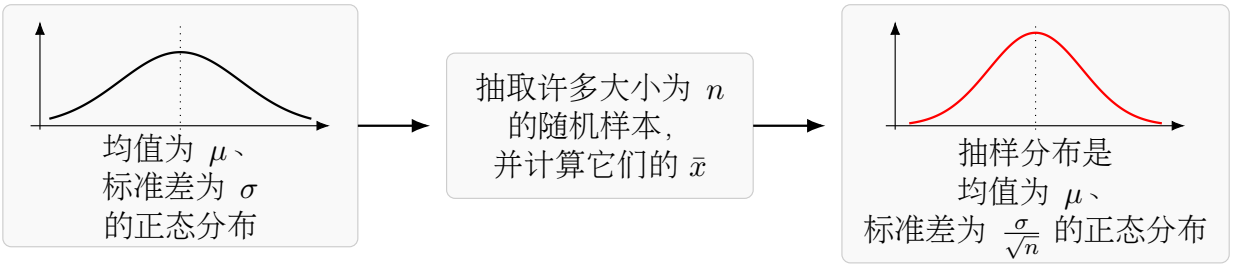
首先假设原始概率分布是正态分布，即 $X \sim \mathcal{N}(\mu, \sigma^2)$。第一个结果把 $\bar{X}$ 的抽样分布与参数 μ 和 σ 联系起来。

中心极限定理：版本1

设 X 是满足 $X \sim \mathcal{N}(\mu, \sigma^2)$ 的随机变量。 那么：

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

这个结果说明：当我们从均值为 μ、方差为 σ^2 的正态分布开始时，所有样本均值的集合会排列成另一个正态分布。它具有相同的中心 μ，但方差变成了 $\frac{\sigma^2}{n}$：



粗略地说，如果我们从 $\mathcal{N}(\mu, \sigma^2)$ 中抽取一组值，那么样本均值 $\bar{x}$ 会把这些单个值中的随机噪声平均掉，并返回一个更靠近中心的数。因此，平均而言，我们会预期这些样本均值更接近分布的真实中心。这个想法部分解释了为什么抽样分布的方差 $\frac{\sigma^2}{n}$ 比原始分布的方差更小。正如公式所示，n 的值越大，抽样分布就越集中。这是因为较大的样本对可能扭曲平均数的单个离群值不那么敏感。

3.2 版本2：原始总体不一定为正态分布时

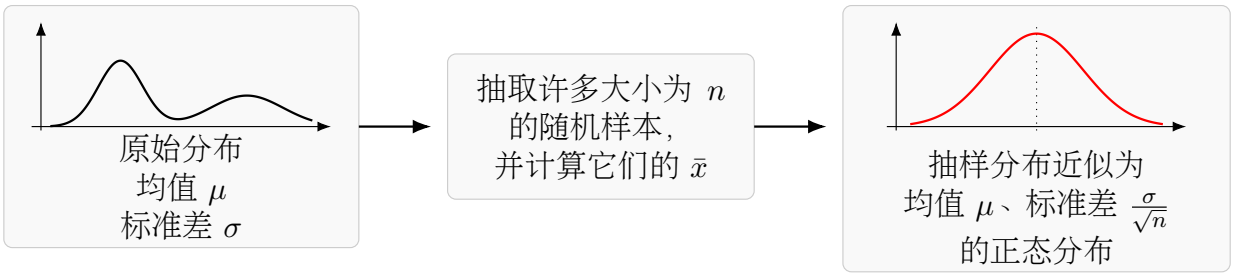
第一个结果说明，如果我们从正态分布构造 $\bar{X}$ 的抽样分布，那么结果仍然是正态分布。令人惊讶的是，中心极限定理还有一个强得多的版本：

中心极限定理

假设 X 具有某个概率分布，并且该分布具有有限均值 μ 和有限方差 σ^2。对于足够大的样本大小 n，通常取 $n \geq 30$，样本均值 $\bar{X}$ 近似服从正态分布，其参数为：

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

换句话说，中心极限定理说明，$\bar{X}$ 的抽样分布看起来会近似正态，均值为 μ，方差为 $\frac{\sigma^2}{n}$。



在这两个结果中，抽样分布的标准差都有相同的一般形式。事实上，这个形式有一个特殊名称：它叫做标准误差。

标准误差

$\bar{X}$ 的抽样分布的标准差是

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

这称为**标准误差**。

标准误差衡量样本均值在不同样本之间变化的程度。随着样本大小 n 增加，标准误差减小，这意味着所得的抽样分布会更紧密地集中在中心 μ 附近。如前所述，这与我们的直觉相符：较大的样本对离群值不那么敏感，也就是说，它们给出更稳定的平均数。

3.3 为什么中心极限定理重要

在现实世界中，我们通常无法直接获得总体参数。例如，如果我们要研究东京居民的平均身高，那么我们不可能出去测量城市中数百万人的身高。相反，在现实中，我们只能从总体中抽取样本并计算一些样本统计量。

中心极限定理提供了一种数学方式，把样本统计量与真实总体均值联系起来。虽然我们可能永远无法理解总体分布的真实结构，但至少可以预期，如果样本足够大，通常 $n \geq 30$，那么所有样本均值的分布近似为正态分布。特别地，抽样分布的中心仍然是真实的总体均值。因此，中心极限定理给了我们额外的表达能力：我们现在可以讨论样本均值 $\bar{x}$ 与总体均值 μ 之间的理论关系。由样本统计量推断总体参数的思想，正是推断统计学的核心。在接下来的几讲中，我们会详细研究这个思想：我们会看到一些根据样本数据实际估计 μ 的方法。

3.4 例题：豚鼠体重

假设一个大型豚鼠总体的均值为 950g，标准差为 150g。假设该总体并不服从正态分布。现在，假设我们随机抽取 30 只豚鼠，并计算样本均值 $\bar{x}$。我们要回答下面的问题：

> 30 只豚鼠的平均体重在 900g 到 1000g 之间的概率是多少?

由于 $n = 30$，我们可以使用中心极限定理来近似 $\bar{X}$ 的抽样分布：

$$\bar{X} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) = \mathcal{N}\left(950, \frac{150^2}{30}\right).$$

等价地，标准误差为

$$\sigma_{\bar{X}} = \frac{150}{\sqrt{30}} \approx 27.39.$$

所以我们也可以写成

$$\bar{X} \approx \mathcal{N}(950, 27.39^2).$$

为了回答上面的问题，我们要求的是概率 $P(900 \leq \bar{X} \leq 1000)$。为此，我们需要先把两个端点都转换为 z 分数：

$$z_1 = \frac{900 - 950}{150/\sqrt{30}} \approx -1.83,$$

以及

$$z_2 = \frac{1000 - 950}{150/\sqrt{30}} \approx 1.83.$$

因此，

$$P(900 \leq \bar{X} \leq 1000) \approx P(-1.83 \leq Z \leq 1.83).$$

使用标准正态表，我们有：

$$P(Z < 1.83) = 0.96638 \qquad \text{且} \qquad P(Z < -1.83) = 0.03362,$$

由此可得：

$$P(900 \leq \bar{X} \leq 1000) \approx P(-1.83 \leq Z \leq 1.83) = 0.96638 - 0.03362 = 0.93276.$$

所以，随机抽取 30 只豚鼠，并使其平均体重在 900g 与 1000g 之间的概率约为 93%。

3.4.1 为什么这有用

如果我们被迫直接计算这个概率，那么会遇到非常麻烦的计算。首先，如果我们试图列出从 1000 只豚鼠中抽取 30 只豚鼠的所有可能样本，可能样本的数量为

$$\binom{1000}{30} \approx 2.43 \times 10^{57}.$$

若要用古典概率的黄金公式直接计算精确概率，我们需要：

1. 列出所有大小为 30 的可能样本；
2. 计算每一个样本的样本均值；
3. 数出有多少样本均值位于 900 与 1000 之间；
4. 然后除以 $C_{1000,30}$，也就是可能样本的总数。

当然，这比简单地使用中心极限定理来近似解要困难得多。

4 二项分布的正态近似

4.1 大型二项实验的问题

回忆第8讲，二项实验是具有以下性质的概率系统：

1. 同一个试验被重复多次；
2. 每次试验永远只有两个可能结果，我们称之为成功和失败；
3. 每次试验具有相同的概率；
4. 各次试验相互独立。

二项分布由两个主要数字指定：n 表示试验次数，p 表示成功概率。二项实验中的核心问题，是求在 n 次尝试中得到某个特定成功次数，或某个成功次数范围的精确概率。为了用数学方式正确地写出这一点，我们定义了一个离散型随机变量 X，它的作用是数出 n 次试验中的成功次数。随后我们看到，在 n 次试验中成功 r 次的概率由公式给出：

$$P(X = r) = C_{n,r}p^r q^{n-r},$$

其中 $q = 1 - p$。

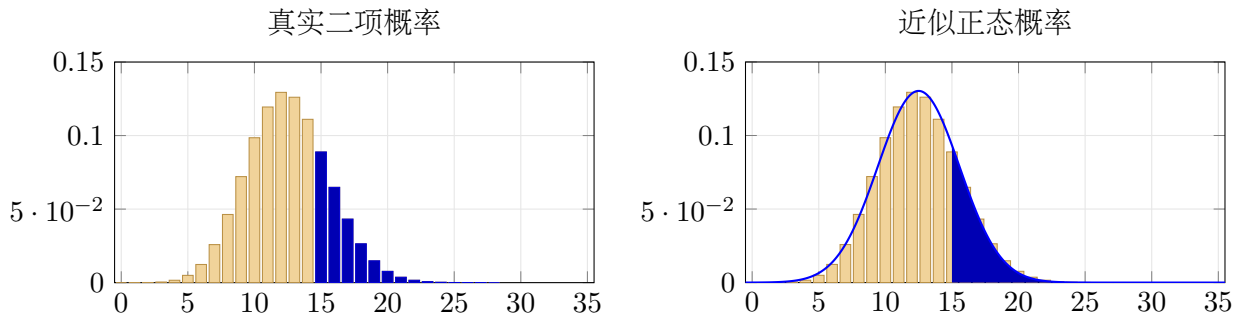
现在想象一个相当大的二项实验，比如 $n = 50$ 且 $p = 0.25$。如果我们被要求计算像 $P(X \geq 15)$ 这样的复杂概率，那么这需要我们计算许多个别概率并把它们相加：

$$\begin{aligned} P(X \geq 15) &= P(X = 15) + P(X = 16) + \cdots + P(X = 49) + P(X = 50) \\ &= C_{50,15}p^{15}q^{35} + C_{50,16}p^{16}q^{34} + \cdots + C_{50,49}p^{49}q^1 + C_{50,50}p^{50}q^0, \end{aligned}$$

或者也许使用补事件：

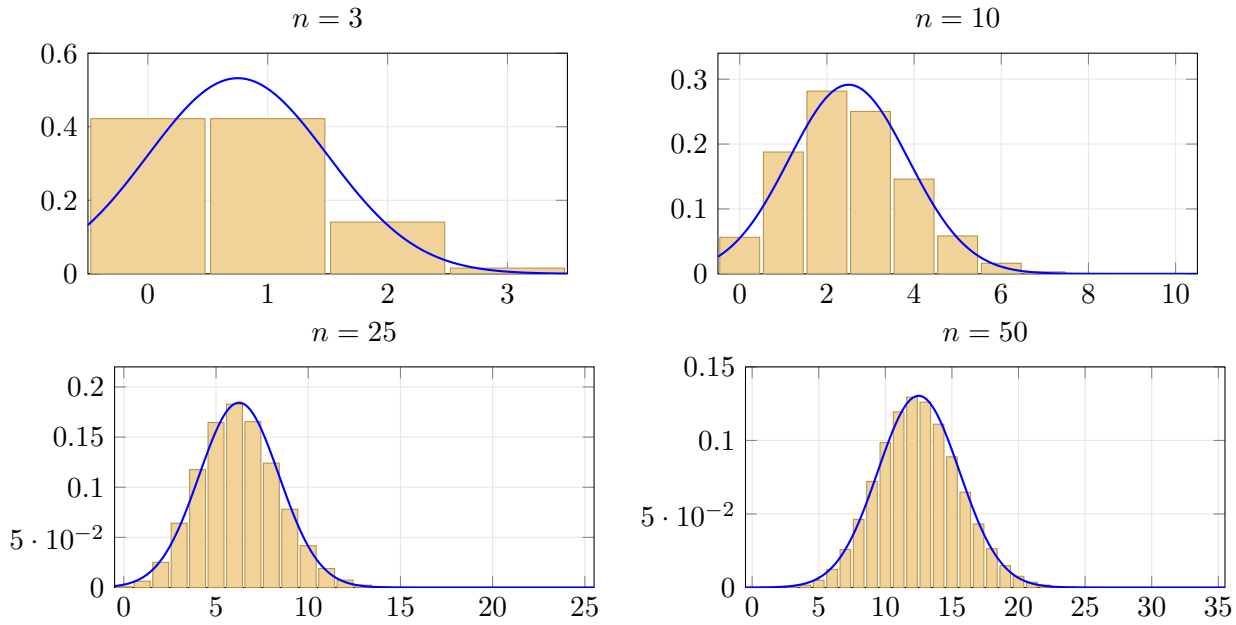
$$\begin{aligned} P(X \geq 15) &= 1 - P(X < 15) \\ &= 1 - (P(X = 0) + P(X = 1) + \cdots + P(X = 13) + P(X = 14)) \\ &= 1 - P(X = 0) - P(X = 1) - \ldots - P(X = 13) - P(X = 14) \\ &= 1 - \left(C_{50,0}p^0 q^{50}\right) - \left(C_{50,1}p^1 q^{49}\right) - \cdots - \left(C_{50,13}p^{13}q^{37}\right) - \left(C_{50,14}p^{14}q^{36}\right) \end{aligned}$$

无论哪种方法，我们都会遇到庞大而困难的计算。相反，在这种情况下，直接用正态曲线近似该分布，并使用 z 分数得到一个粗略答案，可能会更容易。从图像上看，真实概率对应于从 15 到 50 所有柱子高度的总和；而正态近似则是在一条大致匹配该分布的正态曲线下，计算 15 右侧的面积：



当然，如果我们考虑一个更大的二项实验，比如 $n = 1000$，这一点会更加清楚。突然之间，计算像 $P(X \geq 400)$ 这样的概率，会涉及对公式 $C_{n,r}p^r q^{n-r}$ 进行 601 次不同计算，然后再进行一个相当大的求和。

在第10讲中，我们通过观察二项分布在适当缩放后如何开始类似正态曲线，来引入正态分布。从视觉上看，随着 n 越来越大，柱子的数量不断增加，直到图像开始类似一条连续曲线。这意味着随着 n 增加，二项分布会越来越像正态曲线。因此，用 z 分数近似概率会变得越来越合适。下面的图像中，我们把正态曲线拟合到较小的二项分布上。注意，随着 n 变大，正态曲线越来越好地拟合二项分布：



接下来我们会看到，这些正态曲线如何通过应用中心极限定理得到。

4.2 把中心极限定理应用于二项分布

假设我们有一个固定 n、p 和 q 的二项实验。令 X 为数成功次数的随机变量。从某种角度看，我们可以把每一次单独试验理解为一个独立的概率系统，并且每次试验都有自己的随机变量。如果定义：

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

那么 X 可以写成如下和：

$$X = Y_1 + Y_2 + \cdots + Y_{n-1} + Y_n.$$

这个和中的每一项，成功时贡献 1，失败时贡献 0。因此，和 X 会数出 n 次试验后的总成功次数。现在注意：如果我们把 X 除以 n，就会得到一个很像样本均值的东西：

$$\frac{X}{n} = \frac{Y_1 + Y_2 + \cdots + Y_{n-1} + Y_n}{n}.$$

中心极限定理说明，随机变量 $Y_1, \ldots, Y_n$ 的样本均值，会近似排列成正态分布 $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$，其中 μ 和 σ^2 是单次试验变量 Y_i 的均值和方差。在这个例子中，原始变量是某一个 Y_i，也就是实验中单次试验对应的变量。处理二项实验时，我们知道均值的对应物是期望值。对于单个变量 Y_i，有：

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

期望值[1]为：

$$\mathrm{E}(Y_i) = P(Y_i = 0) \cdot 0 + P(Y_i = 1) \cdot 1 = q \cdot 0 + p \cdot 1 = p.$$

此外，方差为：

$$\begin{aligned} \mathrm{Var}(Y_i) &= \mathrm{E}\left((Y_i - \mathrm{E}(Y_i))^2\right) \\ &= \mathrm{E}\left((Y_i - p)^2\right) \\ &= P(Y_i = 0)(0 - p)^2 + P(Y_i = 1)(1 - p)^2 \\ &= q(-p)^2 + p(q)^2 \\ &= qp^2 + pq^2 \\ &= pq(p + q) \\ &= pq. \end{aligned}$$

总结一下：变量 $\frac{X}{n}$ 表现得像二项实验中每次单独试验结果的某种样本均值。每一次试验都具有相同的期望值 p，以及相同的方差 pq。应用中心极限定理可知，$\frac{X}{n}$ 近似为正态分布，具体为：

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

我们可以利用这个信息，为原始变量 X 构造一个新的近似。关于正态分布的一些复杂代数[2]给出以下事实：

$$\frac{X}{n} \approx \mathcal{N}\left(p, \frac{pq}{n}\right) \qquad \Longrightarrow \qquad X \approx \mathcal{N}\left(np, npq\right).$$

因此，我们得到了 X 的正态近似。

[1]回忆一下，对于离散型随机变量，随机变量的期望值大致为

$$\mathrm{E}(X) = \sum \text{概率} \times \text{结果}$$

。

[2]如果必须写出来：我们可以把 $\frac{x}{n}$ 代入定义正态分布的概率密度函数中，得到

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

这几乎看起来像 $\mathcal{N}(n\mu, n^2\sigma^2)$ 的定义，只是前面的系数不一样。如果我们在表达式前额外加上 $\frac{1}{n}$，那么 $\frac{1}{n} f\left(\frac{x}{n}\right)$ 就会与 $\mathcal{N}(n\mu, n^2\sigma^2)$ 的定义完全匹配。

二项分布的正态近似

如果 X 按照试验次数 n 和成功概率 p 服从二项分布，并且使用粗略规则 $np > 5$ 和 $nq > 5$，那么 X 可以用如下正态分布近似：

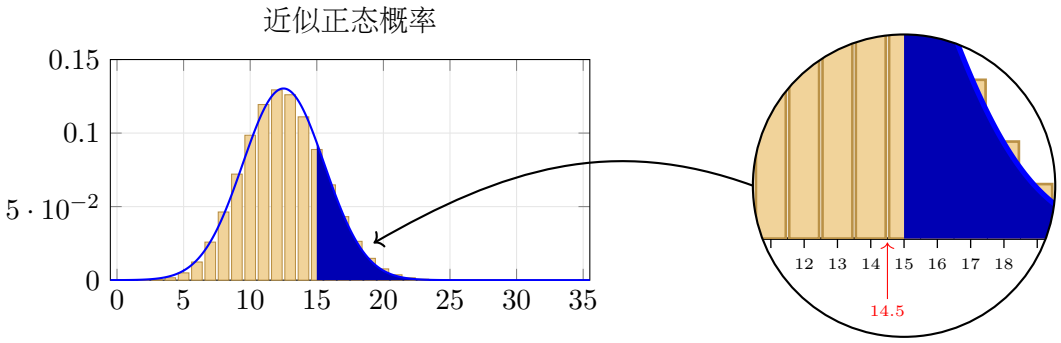
$$X \approx \mathcal{N}(np, npq).$$

换一种写法，我们说 $X \approx Y$，其中 $Y \sim \mathcal{N}(np, npq)$。

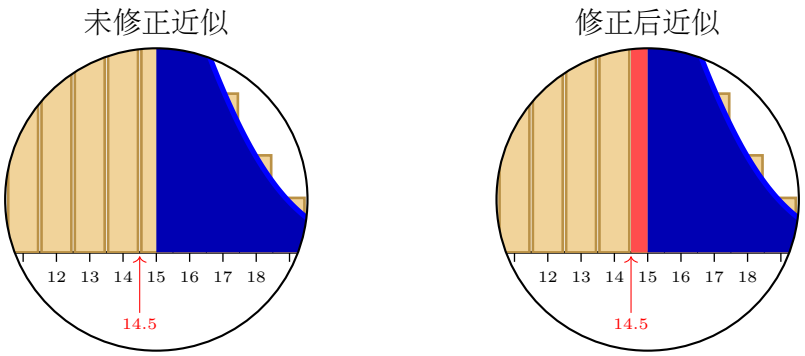
注意，条件 $np > 5$ 和 $nq > 5$ 是为了确保原始二项分布足够大，并且不太偏斜，从而使近似有意义。例如，注意第4.1节中的图：$n = 3, p = 0.25$ 的近似并不真正贴合原始分布。

4.3 连续性修正

根据上一节，当二项分布足够大并且不太偏斜时，我们可以用正态分布 $\mathcal{N}(np, npq)$ 来近似它。后者使我们能够计算 z 分数，并通过面积求出概率。虽然这可能很有用，但在用平滑面积近似柱状图时，存在一个真正的几何问题。我们再次考虑第4.1节中 $n = 50$、$p = 0.25$ 的例子。如果把图放大，我们会注意到，近似的连续面积正好把 $X = 15$ 对应的柱子切成两半：



因此，蓝色区域对应的面积会小于真实概率。我们可以通过让面积近似从柱子的起点开始计数，而不是从柱子中间开始计数，来修正这一点：



一般来说，这会比未修正的版本给出更接近的近似。 这个调整称为连续性修正。

连续性修正的规则

当用正态随机变量 Y 近似二项随机变量 X 时，将每个截断点移动 0.5，使正态面积完整覆盖二项分布的柱子。

二项概率		正态近似
$P(X \geq r)$	变为	$P(Y \geq r - 0.5)$
$P(X > r)$	变为	$P(Y > r + 0.5)$
$P(X \leq r)$	变为	$P(Y \leq r + 0.5)$
$P(X < r)$	变为	$P(Y < r - 0.5)$
$P(a \leq X \leq b)$	变为	$P(a - 0.5 \leq Y \leq b + 0.5)$

例如，如果我们想近似 $P(15 \leq X)$，那么应计算 $P(14.5 \leq Y)$；如果想近似 $P(X \leq 12)$，那么应计算 $P(Y \leq 12.5)$。

5 例题：复古服装

现在我们来完整地看一个二项近似的例子。假设我去下北泽，想买一些很酷的复古服装。根据过去的经验，我知道在任意一家店里找到有意思、值得购买的衣服的概率大约是 25%。我们假设各家店相互独立，并且这个概率在不同店之间保持不变。我走进了 25 家不同的店。我们尝试求出下面问题的近似答案：

> 25 家店中，有 8 家或更多店有好衣服的概率是多少?

假设我们可以把这建模为一个二项实验，定义随机变量

$$X = \text{有好衣服的店的数量.}$$

根据假设，变量 X 服从二项分布，其中 $n = 25$，$p = 0.25$，且 $q = 1 - 0.25 = 0.75$。

5.1 方法1：精确的二项计算

为了以后参考，我们首先精确计算这个结果。题目要求的是 $P(X \geq 8)$。可以直接计算为：

$$\begin{aligned} P(X \geq 8) &= P(X = 8) + P(X = 9) + \cdots + P(X = 25) \\ &= C_{25,8}(0.25)^8(0.75)^{17} + C_{25,9}(0.25)^9(0.75)^{16} + \cdots + C_{25,25}(0.25)^{25}(0.75)^0. \end{aligned}$$

使用计算器可得：

$$P(X \geq 8) \approx 0.27349.$$

5.2 方法2：正态近似

为了使用正态近似，我们首先必须检查近似条件：

$$np = 25(0.25) = 6.25 \qquad \text{且} \qquad nq = 25(0.75) = 18.75.$$

两者都大于 5，所以正态近似是合理的。由于 $n = 25$，$p = 0.25$，$q = 0.75$，近似用的正态分布为：

$$X \approx \mathcal{N}(np, npq) = \mathcal{N}(6.25, 4.6875).$$

如果不使用连续性修正，我们通过计算 z 分数来求近似概率：

$$z = \frac{8 - 6.25}{\sqrt{4.6875}} \approx 0.81.$$

因此：

$$P(X \geq 8) \approx P(Z > 0.81) = 1 - P(Z < 0.81) \approx 1 - 0.79103 = 0.20897.$$

可以看到，这明显小于精确答案 0.27349。然而，如果我们在计算概率之前进行连续性修正，就会把 $P(X \geq 8)$ 改为 $P(Y \geq 7.5)$。把这个新值代入 z 分数公式，得到：

$$z = \frac{7.5 - 6.25}{2.17} \approx 0.58 \qquad \Longrightarrow \qquad P(Y \geq 7.5) \approx P(Z \geq 0.58) \approx 0.28096.$$

注意，这明显更接近真实值 0.27349。