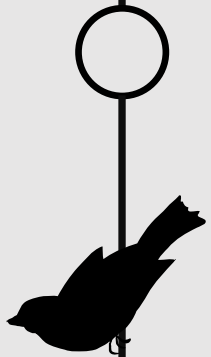
特殊的概率分布

高中统计学·第伍课

何濯羽 2024年2月14日



注：该幻灯片中的函数图像通过 R 软件或 TikZ 语言制作。需要代码的教师或学生可通过电子邮箱联系我（ianho0815@outlook.com），免费分享。

什么是概率分布？

概率分布（probability distribution）是指用于表述随机变量取值的概率规律。

若要全面了解随机变量，就必须知道随机变量的全部可能结果及各种可能结果发生的概率，即随机变量的概率分布。

不同随机变量的概率分布有着不同的表现形式。我们可以根据概率分布的“位置”和“形状”将一些相似的概率分布归类为同一个家族，称之为“分布族”（family of distributions）。

我们常用形似 $G(u)$ 的符号代表一个分布族，其中 G 是分布族的英文名称（或简称），而 u 被称为参数（parameter）。当我们知道某个概率分布是 G 时，参数 u 将帮助我们唯一确定分布 G 的全部特征——包括它累积分布函数、概率质量（或密度）函数、分位数、矩……

这节课的核心就是介绍概率分布宇宙中的五大家族。

注：人教版教材上只介绍了前 4 种概率分布，但我考虑到在第玖课中学生需要学习“卡方独立性检验”，故而加上了关于卡方分布的简介。

目录

壹　伯努利分布

伯努利试验

1713年，瑞士数学家雅各布·伯努利（Jakob I. Bernoulli）在他所著的《Ars Conjectandi》中提出了“伯努利试验”（Bernoulli trial）的概念：只有两种可能结果（“成功”和“失败”）的单次随机试验。

我们可以用 X 代表伯努利试验的结果，“成功”记作 1，“失败”记作 0。显然，X 是一个离散型随机变量。记 X 取值为 1 的概率为 p，X 取值为 0 的概率为 $q = 1 - p$，即

$$\Pr(X = 1) = p$$

$$\Pr(X = 0) = 1 - p$$

当随机变量 X 服从如上的概率分布时，我们称 X 服从“伯努利分布”（Bernoulli distribution），记作

$$X \sim Bernoulli(p)$$

其中，参数 p 代表伯努利试验的“成功”概率。

伯努利试验的例子

最常见的伯努利试验大概是“投掷硬币”吧。我们用 X 代表一枚均匀的硬币落地后的结果：“正面朝上”记作 1，“反面朝上”记作 0。那么，我们有

$$\Pr(X=1)=\Pr(X=0)=\frac{1}{2}$$

即使试验有 2 个以上可能的结果，我们也可以将它转化为伯努利试验。例如，我们可以用 X 代表一颗骰子落地后的结果：“朝上的点数大于 4”记作 1，“朝上的点数小于或等于 4”记作 0。那么，我们有

$$\Pr(X=1)=\frac{1}{3},\Pr(X=0)=\frac{2}{3}$$

伯努利分布

根据在第贰课所学的知识，我们知道若想充分了解一个离散型随机变量的概率分布，我们需要知道它的累积分布函数（CDF）或者概率质量函数（PMF）。

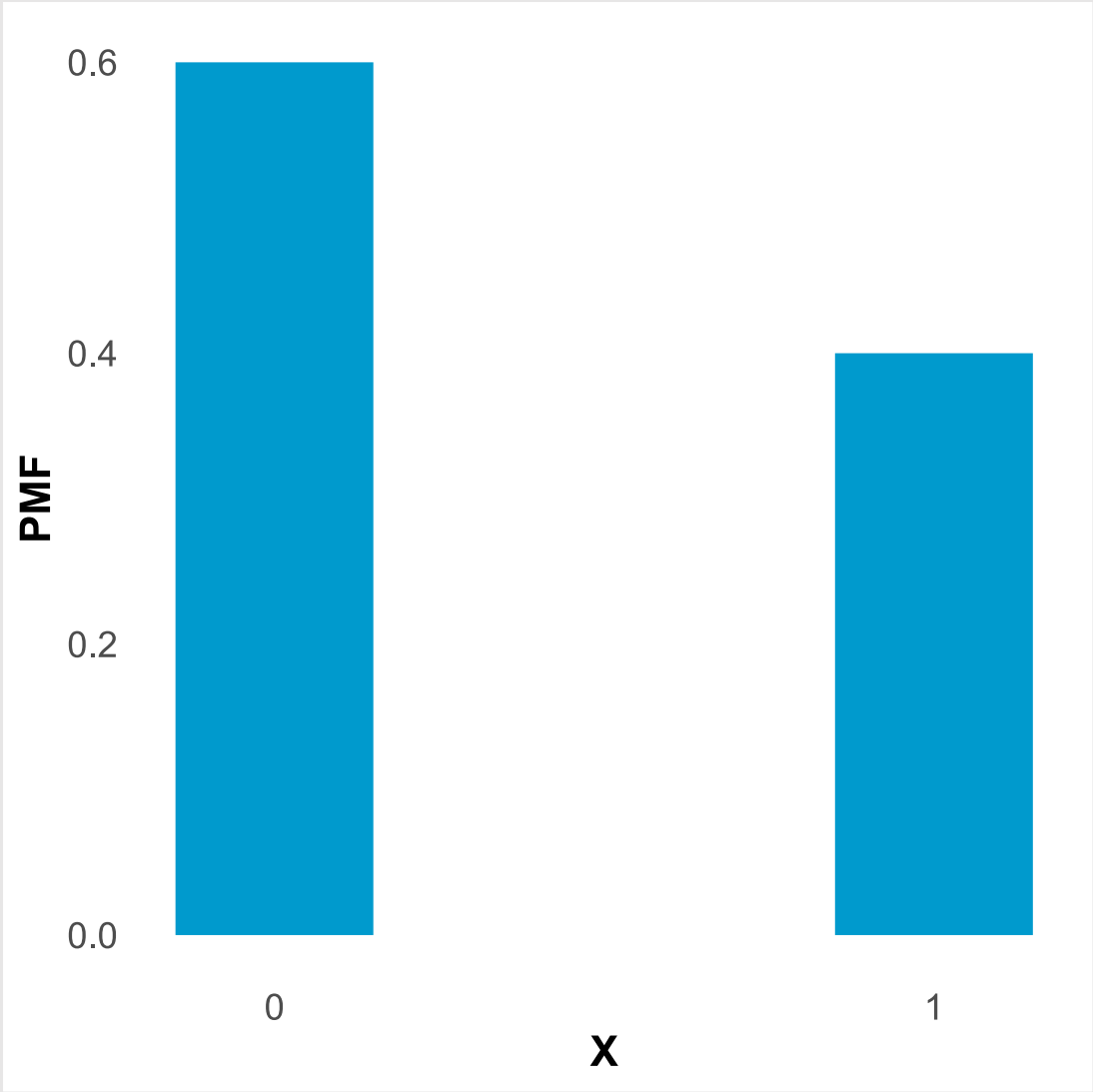
当 $X \sim Bernoulli(p)$ 时，X 的概率质量函数是

$$f_X(x) = \Pr(X = x) = \begin{cases} p\ (x = 1) \\ 1 - p\ (x = 0) \end{cases}$$

X 的累积分布函数是

$$F_X(x) = \sum_{k \le x} \Pr(X = k) = \begin{cases} 0\ (k < 0) \\ p\ (0 \le k < 1) \\ 1\ (k \ge 1) \end{cases}$$

图：$Bernoulli(0.4)$ 的 PMF



伯努利分布的矩

请计算：当 $X \sim Bernoulli(p)$ 时，X 的数学期望是什么？方差呢？

$$E(X) = 1 \cdot \Pr(X = 1) + 0 \cdot \Pr(X = 0) = 1 \cdot p + 0 \cdot (1 - p) = p$$

$$D(X) = (1 - p)^2 \cdot \Pr(X = 1) + (0 - p)^2 \cdot \Pr(X = 0) = (1 - p)^2 p + (-p)^2 (1 - p) = p(1 - p)$$

更复杂的计算将告诉我们 X 的偏度和峰度：

$$Skew(X) = \frac{1 - 2p}{\sqrt{p(1 - p)}} \qquad Kurt(X) = \frac{1 - 6p(1 - p)}{p(1 - p)}$$

我们可以看出：当我们知道 X 服从伯努利分布时，只需再确定参数 p（即伯努利试验“成功”的概率），我们就可以唯一确定该伯努利分布——即我们可以知道该分布的 CDF、PMF、所有阶矩等。

贰 二项分布

二项分布

二项分布（binomial distribution）是 n 个相互独立的相同的伯努利试验中成功次数的离散型概率分布，其中每次试验的成功概率为 p，失败概率为 $q = 1 - p$。当 $n = 1$ 时，二项分布就是伯努利分布。

进行 n 次相互独立的相同的伯努利试验后，成功次数为 x 的概率是多少?

回答这个问题时，我们需要利用第壹课中学习的组合数（combination）的概念。

从 n 个试验结果中抽取 x 个结果合为一组，共有多少不同的组合?

$$\Pr(X = x) = \boxed{C_n^x} \cdot \boxed{p^x} \cdot \boxed{(1-p)^{n-x}}$$

抽到的 x 个结果均为成功的概率

未抽到的 $n - x$ 个结果均为失败的概率

例

假设我们进行了3次相互独立的相同的伯努利试验……

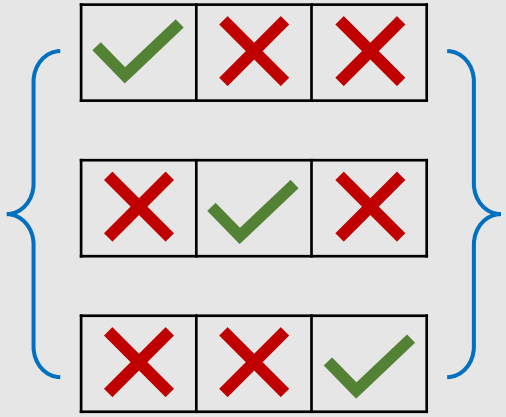
3次结果均为成功的情况

$$\Pr(X=3)=C_3^3\cdot p^3\cdot(1-p)^0=p^3$$

3次结果均为失败的情况

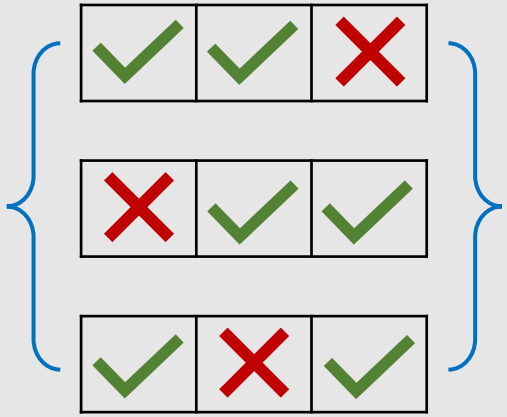
$$\Pr(X=0)=C_3^0\cdot p^0\cdot(1-p)^3=(1-p)^3$$

1次成功2次失败的情况



$$\Pr(X=1)=C_3^1\cdot p^1\cdot(1-p)^2=3p(1-p)^2$$

2次成功1次失败的情况



$$\Pr(X=2)=C_3^2\cdot p^2\cdot(1-p)^1=3p^2(1-p)$$

二项分布的PMF

当离散型随机变量 X 代表 n 个独立的相同伯努利试验中结果为“成功”的次数时，我们称 X 服从二项分布，记作

$$X \sim B(n,p)$$

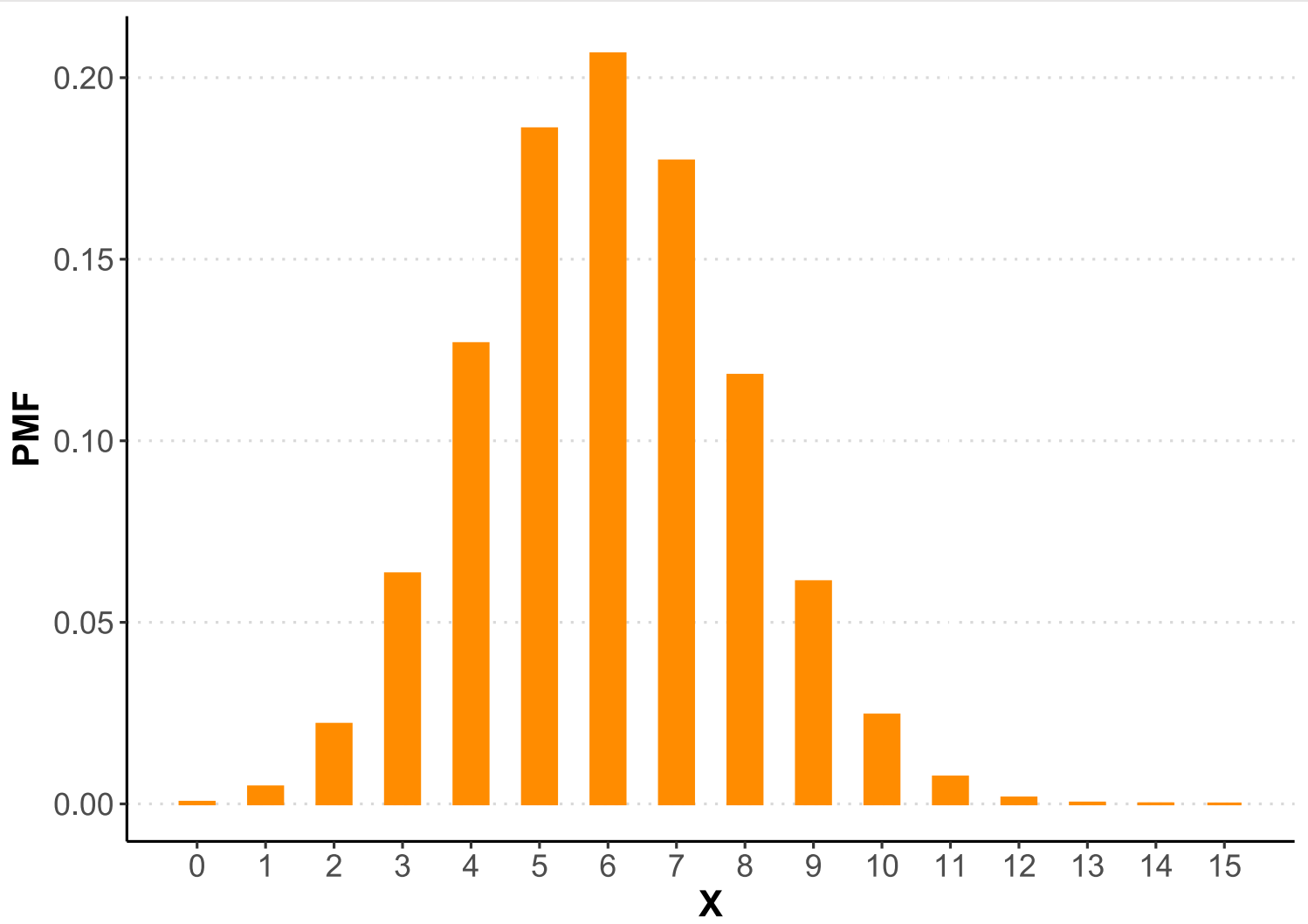
与伯努利分布不同的是，唯一确定一个二项分布需要两个参数：伯努利试验的重复次数 $n \in \mathbb{N}$ 和单个伯努利试验的成功概率 $p \in [0,1]$。

当 $X \sim B(n,p)$ 时，它的概率质量函数是

$$f_X(x;n,p) = \Pr(X = x) = \mathrm{C}_n^x \cdot p^x \cdot (1-p)^{n-x}$$

二项分布的累积分布函数较为复杂，需要使用“不完全贝塔函数”（incomplete beta function）表示，故在此不做展示。

二项分布的图示



图：$B(15, 0.4)$ 的 PMF

二项分布的 PMF 具有特殊的对称性：

$$f_X(x; n, p) = f_X(n - x; n, 1 - p)$$

利用组合数的对称性即可证明该等式。

二项分布的矩

请计算：当 $X \sim B(n,p)$ 时，X 的数学期望是什么？方差呢？

答：我们可以用 $X_1, X_2, \ldots, X_n$ 分别表示第 1、2、……、n 次的伯努利试验的结果。那么，我们有

$$X = X_1 + X_2 + \cdots + X_n$$

于是，

$$E(X) = E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n) = np$$

$$D(X) = D(X_1 + X_2 + \cdots + X_n) = D(X_1) + D(X_2) + \cdots + D(X_n) = np(1-p)$$

这两个等号成立分别需要满足什么条件？

我们可以看出：当我们知道 X 服从二项分布时，只需再确定参数 n 和 p，我们就可以唯一确定该二项分布——即我们可以知道该分布的 CDF、PMF、所有阶矩等。

叁

超几何分布

超几何分布

超几何分布（hypergeometric distribution）也是一种离散型概率分布，描述了从有限个物品中抽出 n 个物品（抽出不放回），成功抽出 k 次指定种类的物品的概率。

例：假设我们共有 N 个样品，并且已知其中有 K 个不合格样品。我们从这 N 个样品中抽出 n 个（抽出不放回）。请问，在抽出的 n 个样品中，有 k 个不合格样品的概率是多少?

这种情况可以总结为如下所示的表格：

	已抽出	未抽出	总数
不合格品	k	$K-k$	K
合格品	$n-k$	$N-K-n+k$	$N-K$
总数	n	$N-n$	N

这种表格被称为“列联表”（contingency table），是将数据按两个属性分类后所列出的频数表。它相当于频数分布表，不同于我们在第肆课中提到的联合分布列（相当于概率分布表）。

推导概率

	已抽出	未抽出	总数
不合格品	k	$K-k$	K
合格品	$n-k$	$N-K-n+k$	$N-K$
总数	n	$N-n$	N

为了计算在抽出的 n 个样品中出现 k 个不合格样品的概率，我们需要回答以下问题：

1) 从 N 个样品中抽出 n 个样品总共有多少种不同的抽取方式？ C_N^n
2) 在 n 个抽出的样品中出现 k 个不合格样品，相当于我们从 K 个不合格品中抽出了 k 个，并从 $N-K$ 个合格品中抽出了 $n-k$ 个。请问总共有几种不同的抽取方式？ $C_K^k C_{N-K}^{n-k}$
3) 根据古典概型，抽出的 n 个样品中出现 k 个不合格样品的概率是多少？

若我们用 X 表示 n 个样品中不合格样品的个数，则 $\Pr(X=k)=\dfrac{C_K^k C_{N-K}^{n-k}}{C_N^n}$

超几何分布的PMF

当离散型随机变量 X 代表从 N 个物品中抽取（不放回）的 n 个物品中具有特定属性的物品个数时，我们称 X 服从超几何分布，记作

$$X \sim Hypergeometric(N, K, n)$$

其中，K 代表已知的 N 个物品中具有特定属性的物品个数。

当 $X \sim Hypergeometric(N, K, n)$ 时，它的概率质量函数是

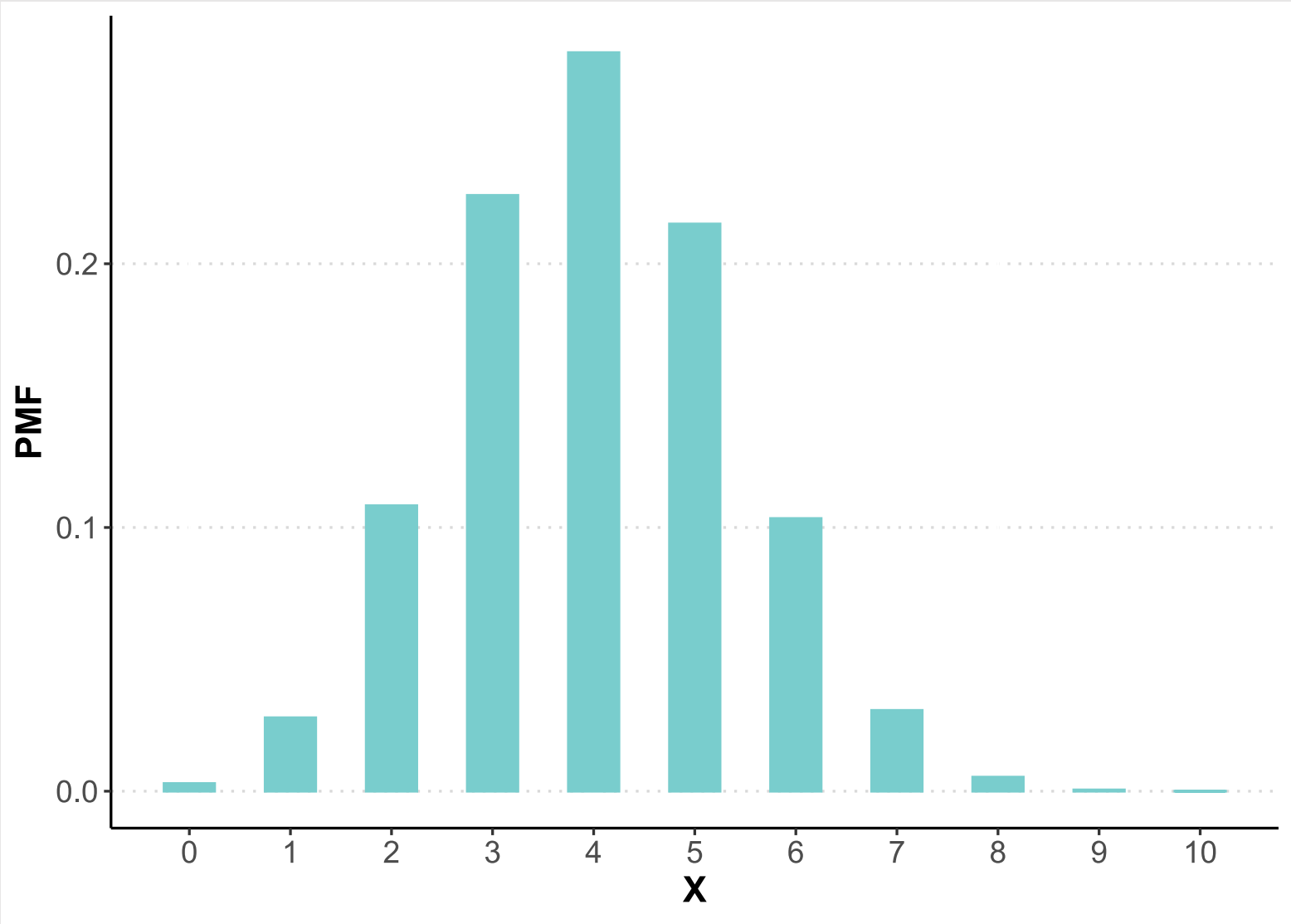
$$f_X(x; N, K, n) = \Pr(X = x) = \frac{\mathrm{C}_K^k \mathrm{C}_{N-K}^{n-k}}{\mathrm{C}_N^n}$$

其中，$x \in \mathbb{N}$ 且 $\max\{0, n + K - N\} \leq x \leq \min\{n, K\}$。

超几何分布的累积分布函数较为复杂，需要使用“广义超几何函数”（generalized hypergeometric function）表示，故在此不做展示。

超几何分布的图示

图：$Hypergeometric(50, 10, 20)$ 的 PMF



超几何分布的 PMF 也具有特殊的对称性：

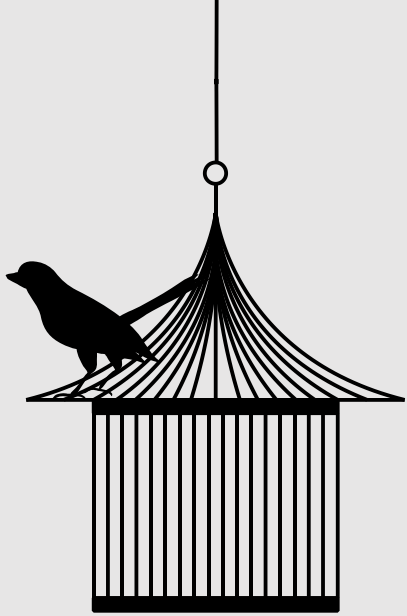
$$f_X(x; N, K, n) = f_X(n - x; N, N - K, n)$$

$$f_X(x; N, K, n) = f_X(K - x; N, K, N - n)$$

$$f_X(x; N, K, n) = f_X(x; N, n, K)$$

肆　正态分布

正态分布



正态分布（normal distribution），又名高斯分布（Gaussian distribution），是统计学中最常用的连续型概率分布。当连续型随机变量 X 服从数学期望为 μ、方差为 σ^2 时，我们记作

$$X \sim N(\mu, \sigma^2)$$

其中，μ 被称为位置参数（location parameter），决定了正态分布的中心位置；σ^2 被称为尺度参数（scale parameter），决定了正态分布的散布幅度。当我们已知某随机变量服从正态分布时，只需要再确定 μ 和 σ^2 的值，我们就可以唯一确定该正态分布（以及它的所有信息）。

X 的概率密度函数是

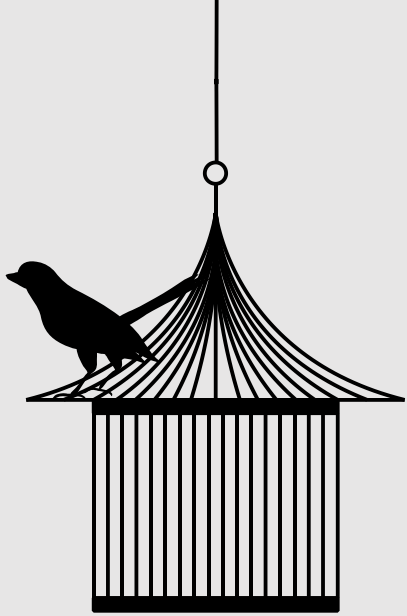
$$f_X(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (x \in \mathbb{R})$$

在 Excel（2007以上版本）中，正态分布的 PDF 和 CDF 的函数值均可运用 `NORM.DIST( )` 函数求得。

X 的累积分布函数是

$$F_X(x;\mu,\sigma^2) = \int_{-\infty}^{x} f_X(t)\,dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\,dt \quad (x \in \mathbb{R})$$

标准正态分布



我们把数学期望为 0、方差为 1 的正态分布称为“**标准正态分布**”（standard normal distribution），记作 $N(0,1)$。标准正态分布的**概率密度函数**是

$$\phi(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\quad(z\in\mathbb{R})$$

标准正态分布的**累积分布函数**是

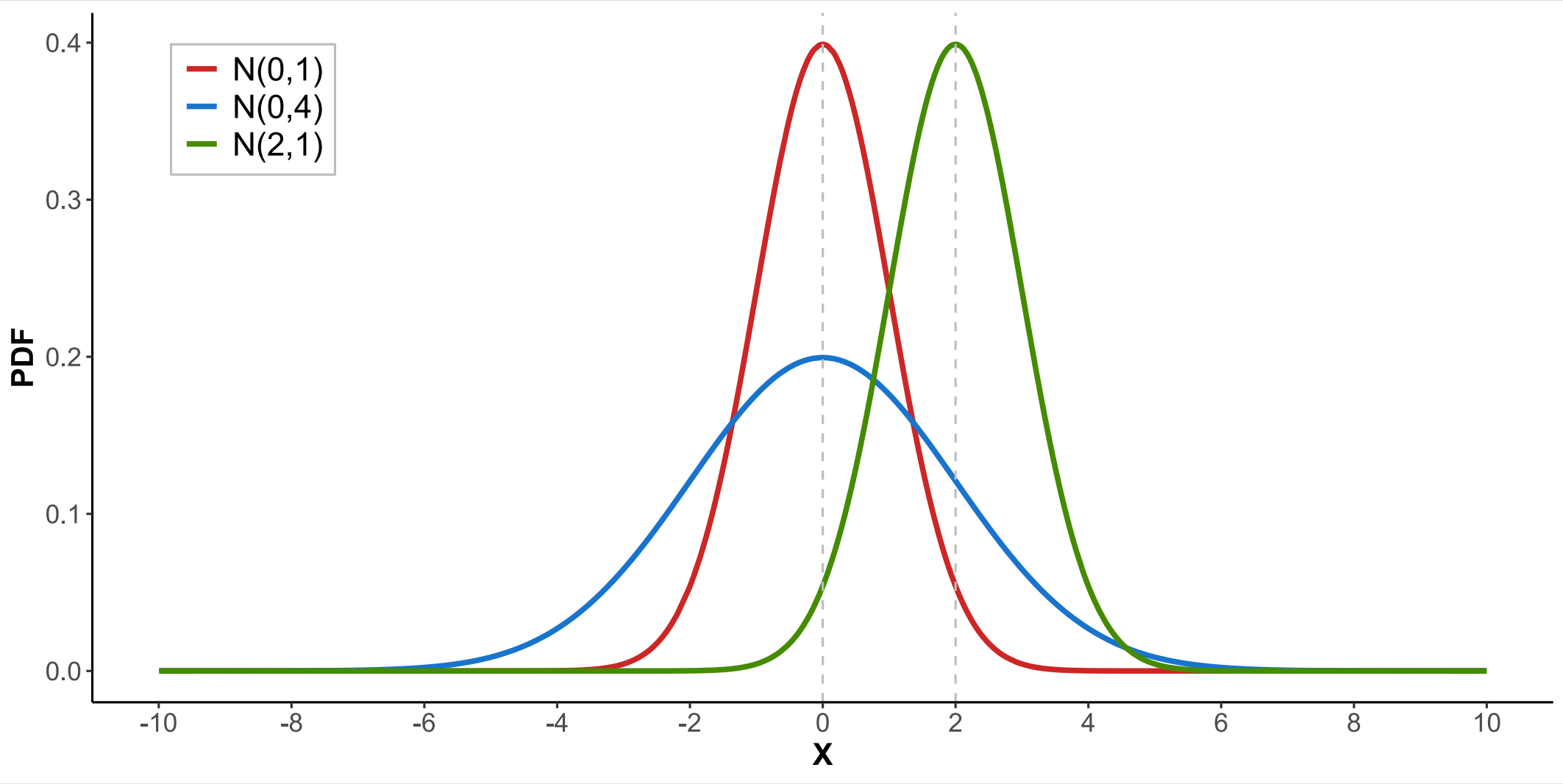
$$\Phi(z)=\int_{-\infty}^{z}\phi(t)\,dt=\int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}\,dt\quad(z\in\mathbb{R})$$

z	$\Pr(Z\le z)$	$\Pr(Z>z)$	$\Pr(\lvert Z\rvert>z)$	$\Pr(\lvert Z\rvert\le z)$
0.00	0.5000	0.5000	1.0000	0.0000
1.00	0.8413	0.1587	0.3173	0.6827
1.65	0.9505	0.0495	0.0989	0.9011
1.96	0.9750	0.0250	0.0500	0.9500
2.00	0.9772	0.0228	0.0455	0.9545
2.33	0.9901	0.0099	0.0198	0.9802
2.58	0.9951	0.0049	0.0099	0.9901
3.00	0.9987	0.0013	0.0027	0.9973

在 Excel（2007以上版本）中，标准正态分布的 PDF 和 CDF 的函数值均可运用 `NORM.S.DIST( )` 函数求得。

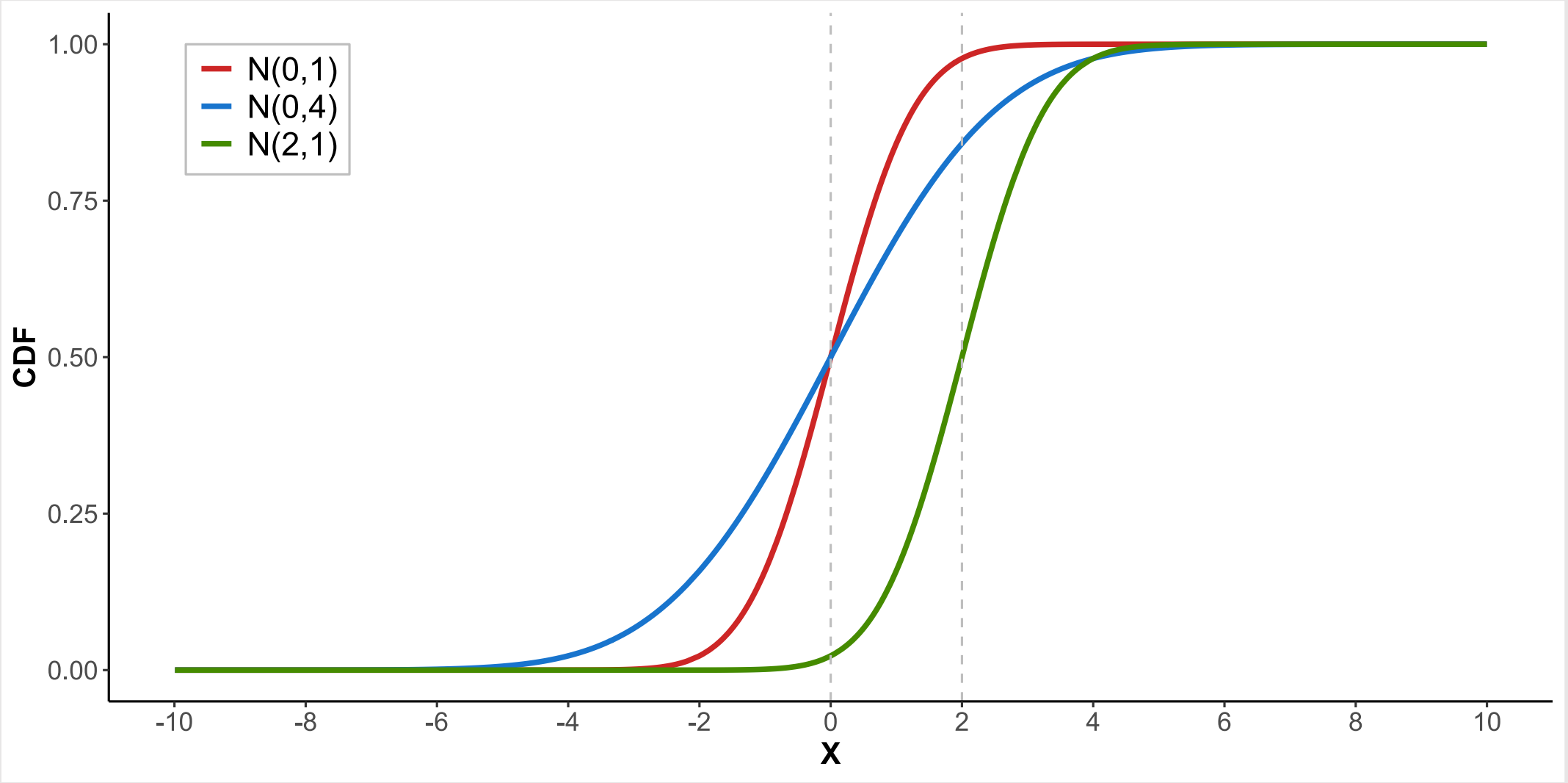
正态分布的图示

正态分布的概率密度函数图象被称为“钟形曲线”（bell curve），因为它形似于寺庙里的大钟。图象关于正态分布的数学期望呈轴对称，即 $\phi(\mu - x) = \phi(\mu + x)$。

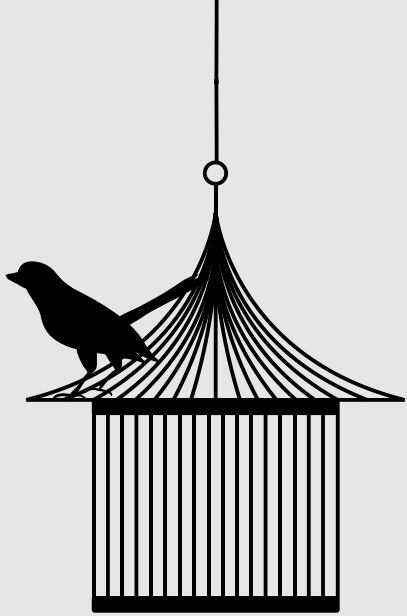


正态分布的图示

正态分布的累积分布函数图象是一个 S 型曲线——先平坦，后骤升，最后恢复平坦。斜率最大的点对应正态分布的数学期望。



正态分布的特点



定理 1： 如果 $X \sim N(\mu, \sigma^2)$，那么 $\frac{X-\mu}{\sigma} \sim N(0,1)$。

这就是“随机变量减去其数学期望，然后除以其标准差”这一操作被称为“标准化”（standardization）的原因。

推论： 如果 $X \sim N(\mu, \sigma^2)$，那么

$$\Pr(a \leq X \leq b) = \Pr\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

定理 2： 如果 $X_1 \sim N(\mu_1, \sigma_1^2)$ 与 $X_2 \sim N(\mu_2, \sigma_2^2)$ 相互独立，那么 $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$。

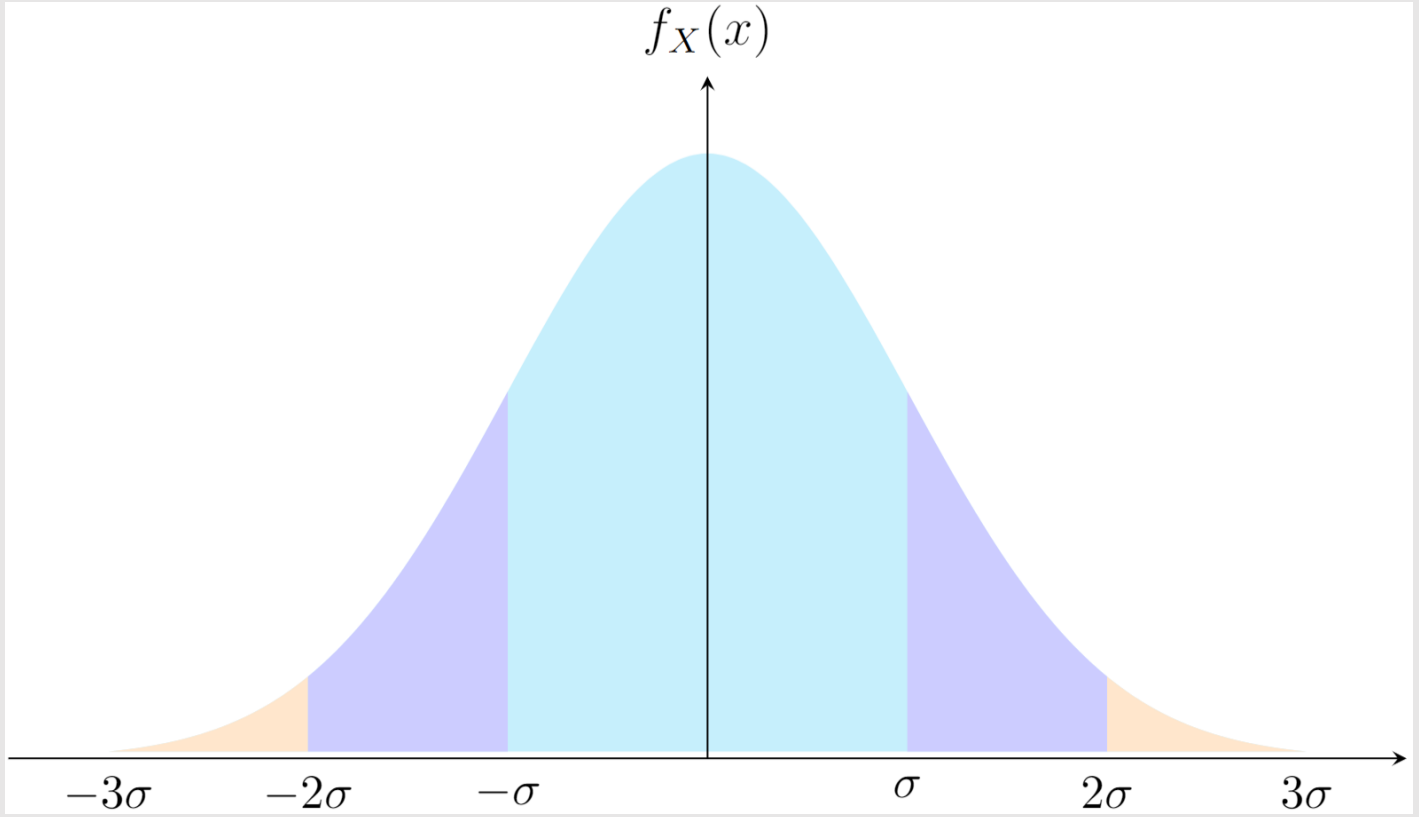
该定理可以推广至 n 个正态随机变量相加的情况。

正态分布的特点

当 $X \sim N(\mu, \sigma^2)$ 时，我们有

$$\Pr(|X-\mu|<\sigma) \approx 0.68$$
$$\Pr(|X-\mu|<2\sigma) \approx 0.95$$
$$\Pr(|X-\mu|<3\sigma) \approx 0.997$$

这三个正态分布的性质被频繁地应用于假设检验（将在第玖课进行介绍）。



伍

卡方分布

卡方分布

当 $Z \sim N(0,1)$ 时，随机变量 Z 的平方也服从一个特别的概率分布：卡方分布（chi-squared distribution）。

$$Z \sim N(0,1) \implies Z^2 \sim \chi^2(1)$$

如果 $Z_1, Z_2, \ldots, Z_n$ 为 n 个相互独立的标准正态随机变量，那么

$$Z_1^2 + Z_2^2 + \cdots + Z_n^2 = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$$

其中，参数 n 对应的是独立的标准正态随机变量的个数，被称为“自由度”（degree of freedom）。

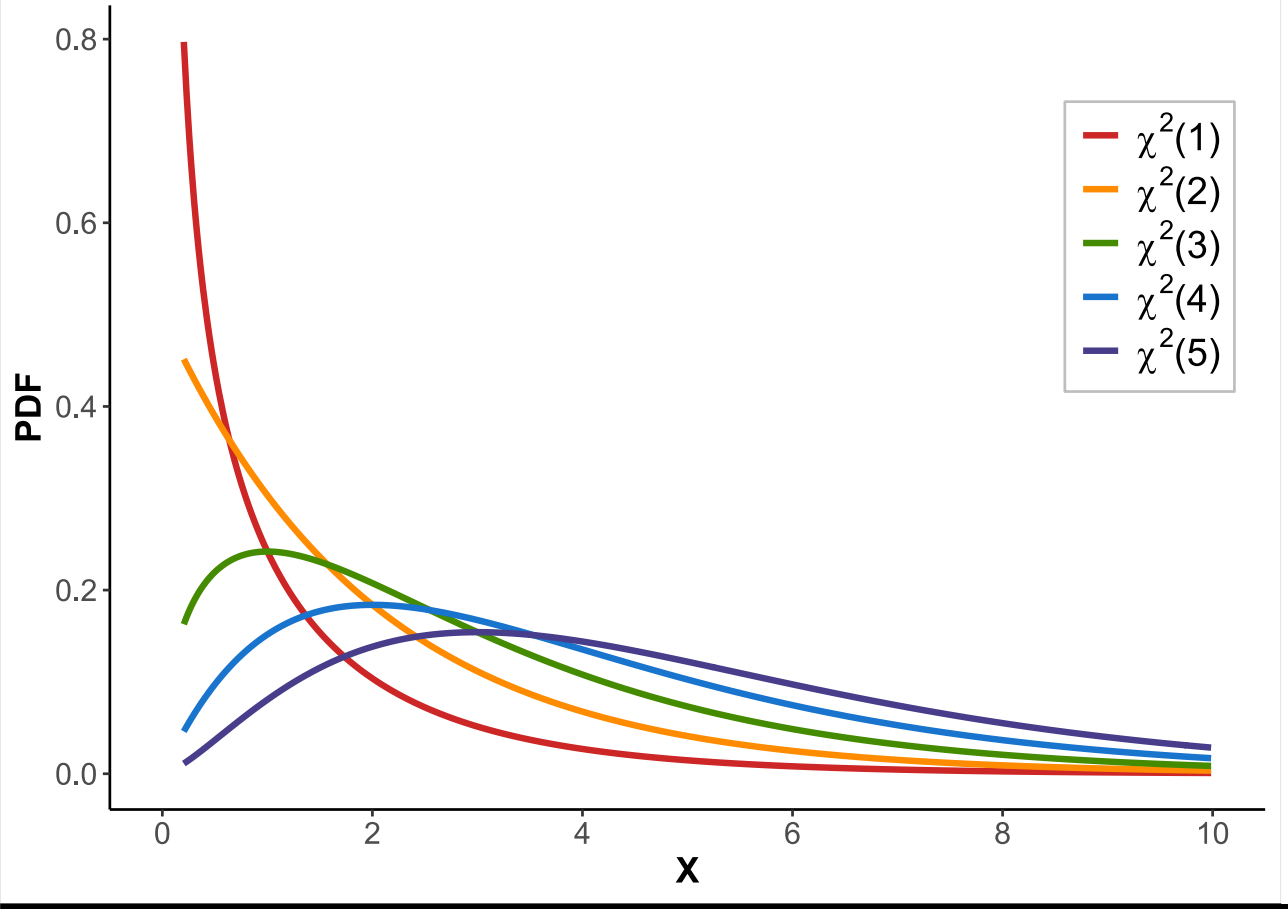
自由度是统计学中非常重要的概念，但不在高考范围内，故在此不做更多介绍。

卡方分布的概率密度函数和累积分布函数的解析式较为复杂，需要使用伽马函数（Gamma function）表示，故在此不做展示。在 Excel（2007以上版本）中，卡方分布的 PDF 和 CDF 均可运用 CHISQ.DIST( ) 函数求得。

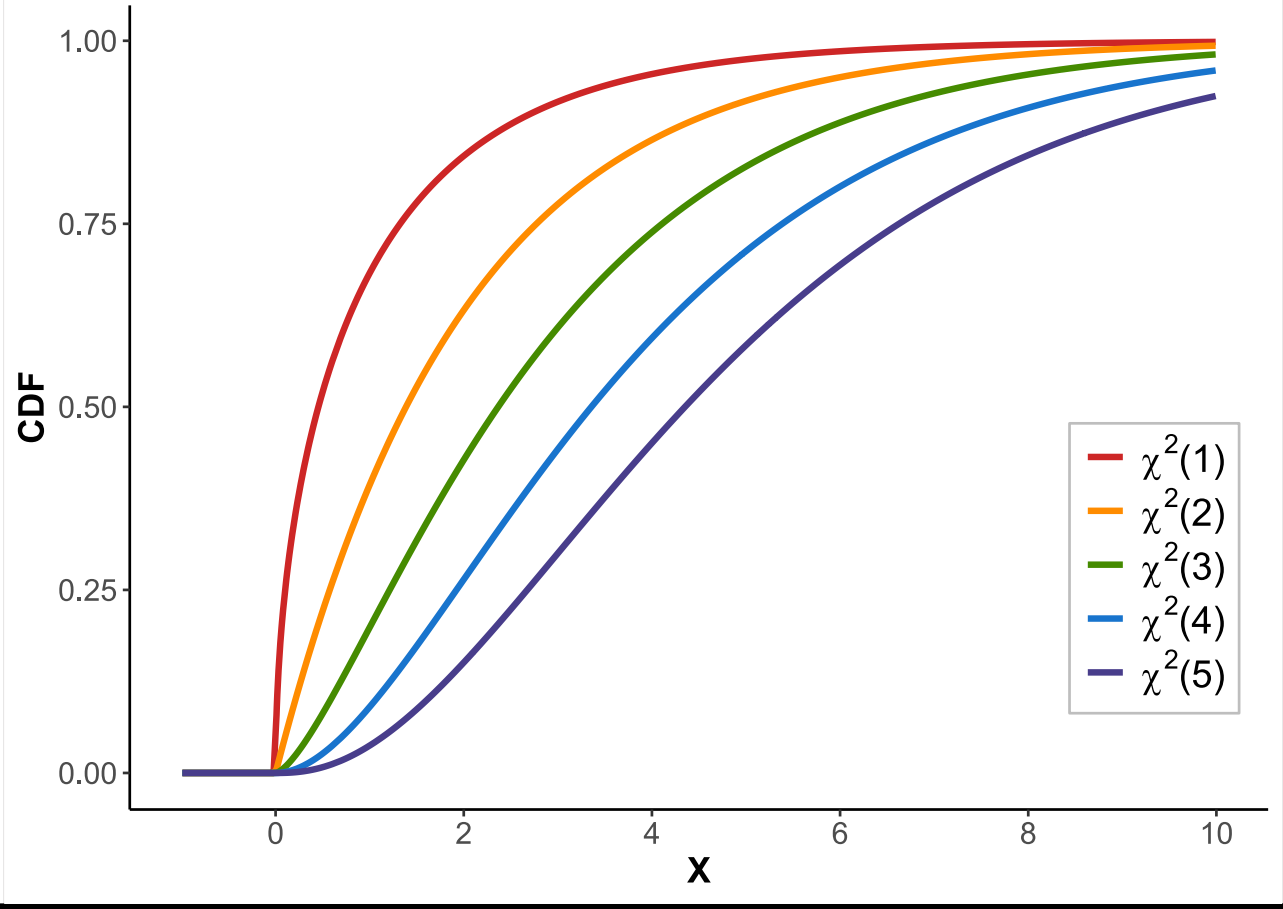
卡方分布的数学期望为 n，方差为 $2n$。

卡方分布的图示

图：卡方分布的 PDF



图：卡方分布的 CDF




感谢观看！
何濯羽 2024年2月14日