随机变量的定义与描述

公选课《可计量的社会》　何濯羽　2024年9月23日

目 录

CONCENTS

SECTION 01

随机变量的定义与种类

随机变量的定义

What is a random variable?

定义：随机变量指的是那些不能完全预测、可以重复取值的变量（variable）。随机变量为一场试验（experiment）中的每个事件（event）赋予了一个数。

符号：

我们通常用大写拉丁字母（例如 X）表示随机变量，用小写拉丁字母（例如 x）表示随机变量的取值。

由一个随机变量所有可能的取值组成的集合被称为该随机变量的“支撑集”（support set，英文中通常简称为 support）。我们通常用花体大写拉丁字母（例如 $\mathcal{X}$）表示支撑集。

简言之，X 是“对应关系”（或函数），$\mathcal{X}$ 是 X 的值域，x 是 $\mathcal{X}$ 中的某个元素。

例子：

我们可以用 X 表示成都东软学院任一学生的年级，用 x_1, x_2, x_3, x_4分别表示大一、大二、大三、大四这4个可能的年级。这样，随机变量 X 的支撑集为 $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$。

离散型随机变量

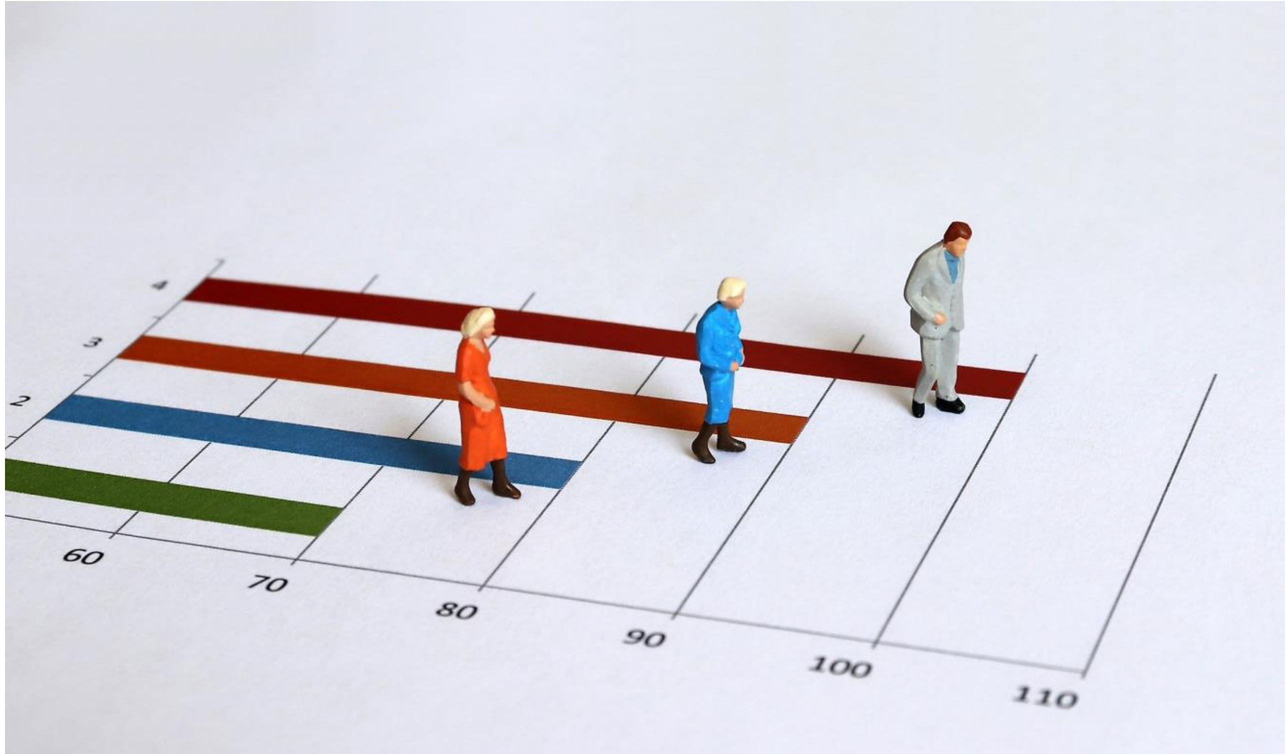
Discrete Random Variable

离散型随机变量指的是那些只能取得有限个（或可数无限个）数值的随机变量。

换言之，离散型随机变量的支撑集中含有有限个（或可数无限个）元素。

例： 我们用 A 表示成都市任一居民的年龄，

显然 A 可能的取值包含有限个正整数。



连续型随机变量

Continuous Random Variable

连续型随机变量指的是那些能取得（不可数）无限个数值的随机变量。
换言之，连续型随机变量的支撑集中含有（不可数）无限个元素。

例：如果我们只以整年记录，那么人的年龄是离散型变量。但是，如果我们以天数，甚至以时、分、秒来记录人的年龄，那么年龄可以理解为连续型变量。

个人观点：这个世界上也许并不存在真正意义上的连续型变量。物质可以无限细分吗？迄今为止，人们发现最小的基本粒子是夸克。能量可以无限细分吗？量子力学中能量跃迁理论的回答是：不可以。人类的感知与测量终究都是离散的。

完美描述随机变量

How to characterize a random variable?

一个随机变量的取值总是在不确定的，那么我们如何能够完美地描述一个随机变量呢？

我们需要借助概率和函数。以下这些函数都可以完美地描述一个随机变量：

- ✓ 概率质量（或密度）函数
- ✓ 累积分布函数
- ✓ 分位数函数
- ✓ 矩量母函数

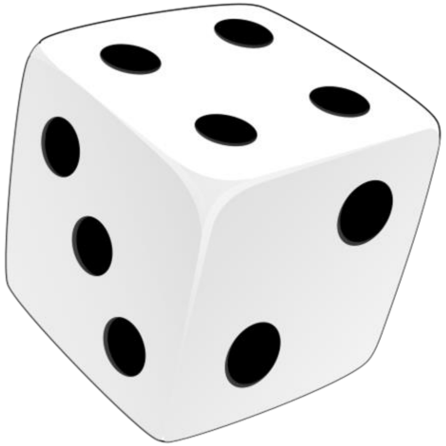
在本门课中，我们将粗略了解前两种函数。

SECTION 02

概率质量函数（PMF）

举个栗子

投掷一枚均匀的骰子。



概率质量函数

$x = 1 \rightarrow f_X(1) = \Pr(X = 1) = \frac{1}{6}$

$x = 3 \rightarrow f_X(3) = \Pr(X = 3) = \frac{1}{6}$

$x = 6 \rightarrow f_X(6) = \Pr(X = 6) = \frac{1}{6}$

PMF的定义

离散型随机变量 X 的概率质量函数（probability mass function，简记为 PMF）是

$$\boldsymbol{f_X(x)} = \begin{cases} \mathbf{Pr}(\boldsymbol{X} = \boldsymbol{x}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}} \\ \mathbf{0}, \boldsymbol{x} \notin \boldsymbol{\mathcal{X}} \end{cases}$$

简言之，概率质量函数的函数值是离散型随机变量在各个特定取值上的概率。因此，概率质量函数可以完美地描述离散型随机变量，即它可以提供一个离散型随机变量的完整信息。

如果 f_X 是离散型随机变量 X 的概率质量函数，那么我们可以简记为 $X \sim f_X$（读作“X 服从分布 f_X”）。当上下文语境足够清楚时，我们可以省略 f_X 的下标，即写作 f。

展示PMF

一般地，我们有 3 种方式向读者/观众展示一个概率质量函数。

1 函数解析式

例：如果随机变量 X 服从泊松分布 $Poiss(\lambda)$，那么它的 PMF 是 $f_X(x)=\frac{\lambda^x e^{-\lambda}}{x!}$，$x\in\mathbb{N}$。

2 概率分布列

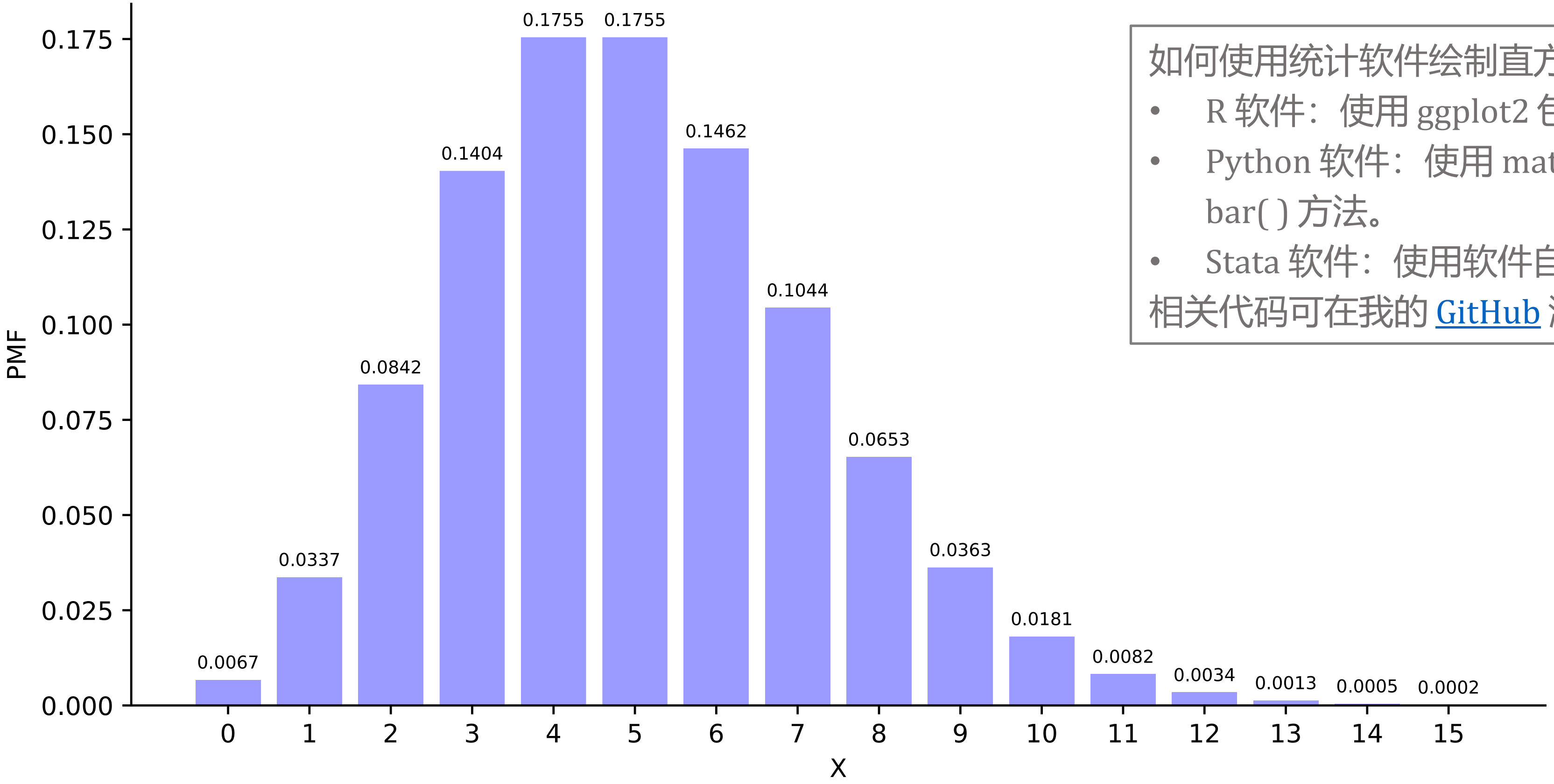
例：我们用 X 表示投掷一枚均匀硬币的结果，$X=1$ 代表正面朝上，$X=0$ 代表反面朝上，那么 X 的概率质量函数可以用下表（中国人教版高中教材称其为“概率分布列”）展示：

X 的取值	1	0
概率	0.5	0.5

展示PMF

3 概率分布直方图

例：如果随机变量 X 服从泊松分布 $Poiss(5)$，那么它的 PMF 是 $f_X(x) = \frac{5^x e^{-5}}{x!}$， $x \in \mathbb{N}$ 。这个 PMF 可以由下图所示的直方图展示出来：



如何使用统计软件绘制直方图?

- R 软件：使用 ggplot2 包。
- Python 软件：使用 matplotlib.pyplot 包中的 bar( ) 方法。
- Stata 软件：使用软件自带的 histogram 指令。

相关代码可在我的 GitHub 浏览。

SECTION
03

累积分布函数（CDF）

定义

随机变量 X 的累积分布函数（cumulative distribution function，简记为 CDF）是

$$\boldsymbol{F_X(x) = \Pr(X \leq x), \qquad x} \in \mathbb{R}$$

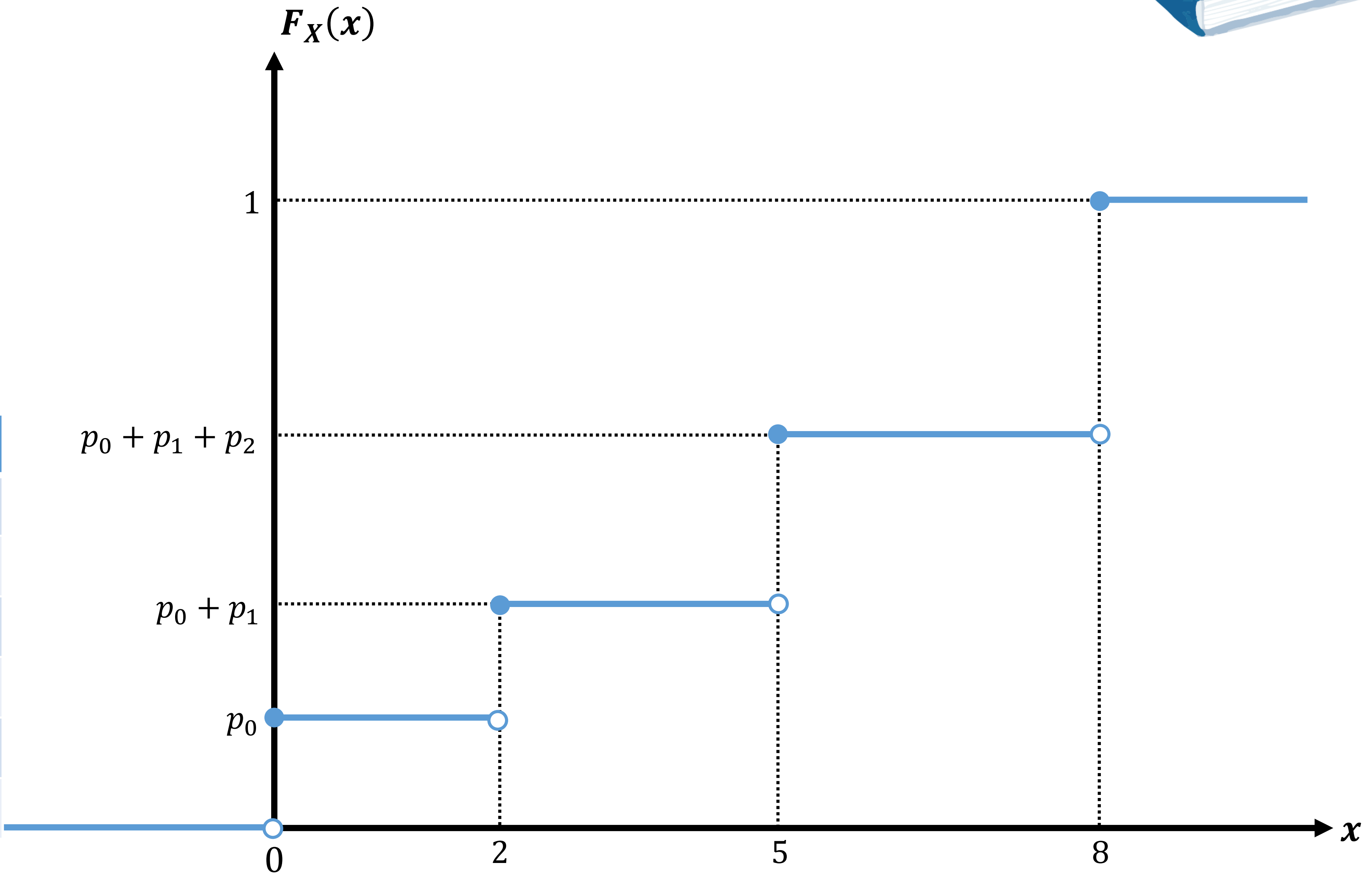
简言之，累积分布函数在 x 上的函数值是随机变量的取值小于或等于 x 的概率。

相似地，如果 F_X 是随机变量 X 的累积分布函数，那么我们可以简记为 $X \sim F_X$（读作“X 服从分布 F_X”）。当上下文语境足够清楚时，我们可以省略 F_X 的下标，即写作 F。

PMF → CDF

X 的取值	概率 $\Pr(X=x)$
0	p_0
2	p_1
5	p_2
8	p_3

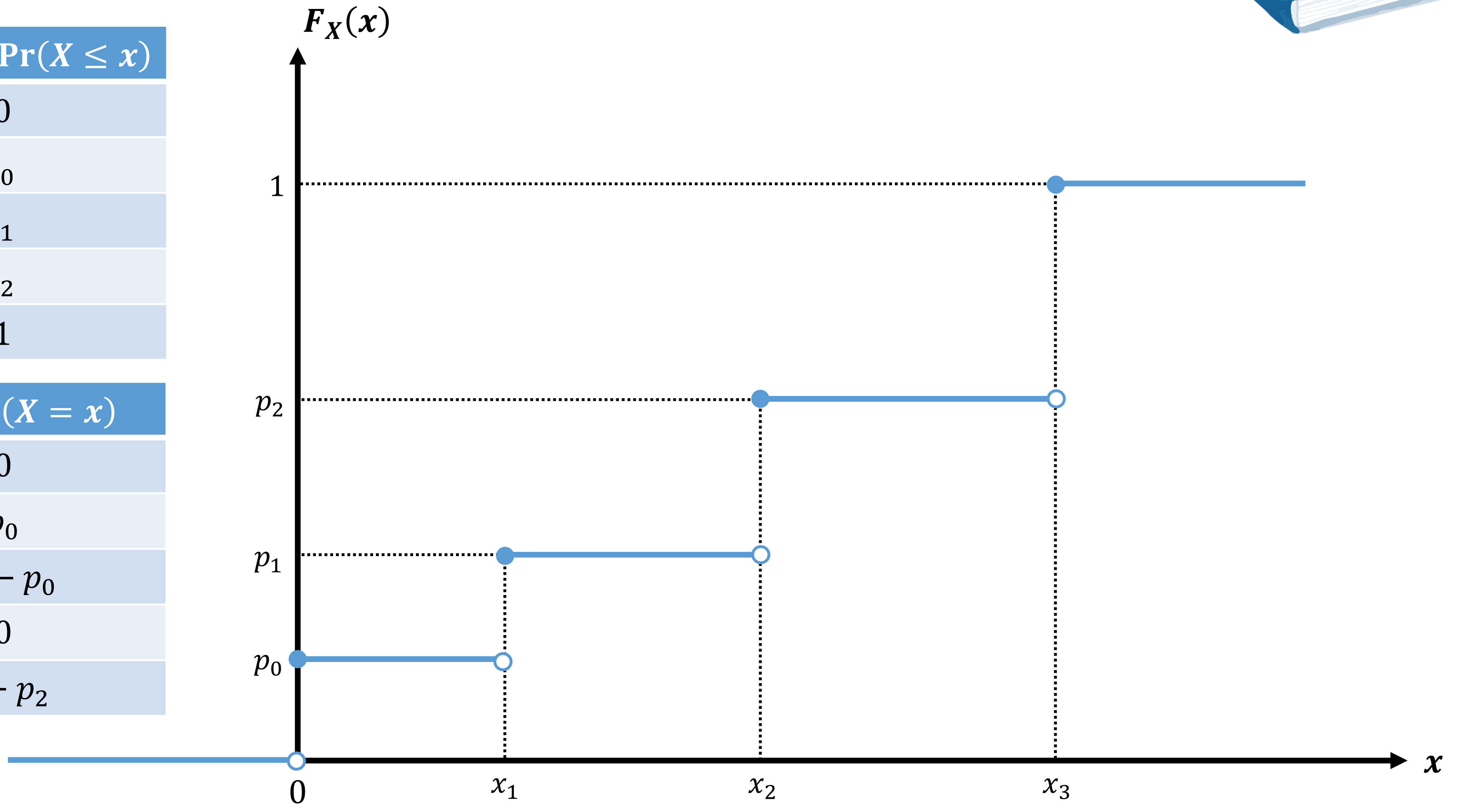
	累积概率 $\Pr(X \le x)$
-1	0
0	p_0
1	p_0
4	$p_0 + p_1$
5	$p_0 + p_1 + p_2$
9	$p_0 + p_1 + p_2 + p_3 = 1$



CDF → PMF

	累积概率 $\Pr(X \le x)$
$x < 0$	0
$0 \le x < x_1$	p_0
$x_1 \le x < x_2$	p_1
$x_2 \le x < x_3$	p_2
$x \ge x_3$	1

x 的取值	概率 $\Pr(X = x)$
$x = -1$	0
$x = 0$	p_0
$x = x_1$	$p_1 - p_0$
$x \in (0, x_1)$	0
$x = x_3$	$1 - p_2$



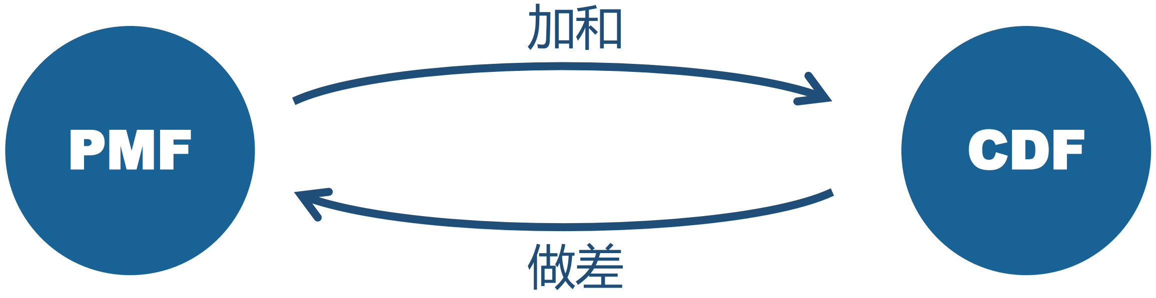
CDF 和 PMF 的关系

当 X 为离散型随机变量时，X 的累积分布函数可以由它的概率质量函数求得：

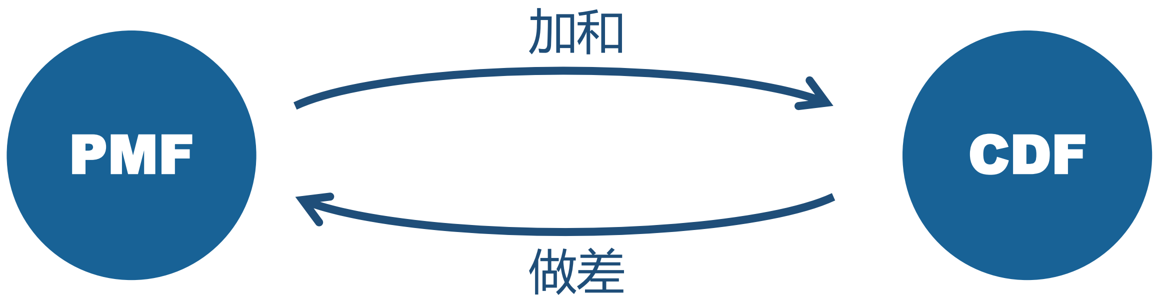
$$F_X(x) = \Pr(X \le x) = \sum_{k \le x} \Pr(X = k) = \sum_{k \le x} f_X(k)$$

反之，X 的概率质量函数也可以由它的累积分布函数求得：

$$f_X(x) = \Pr(X = x) = \Pr(X \le x) - \Pr(X < x) = F_X(x) - \lim_{k \to x} F_X(k)$$



CDF 和 PMF 的关系



通过上述讨论和计算，我们发现，CDF 和 PMF 传达的信息是等价的。

我们已知 PMF 提供了离散型随机变量的完整信息，因此，CDF 也可以提供离散型随机变量的完整信息。

SECTION
04

概率密度函数（PDF）

PMF的失效

在前两节中，我们一直关注离散型随机变量的描述。那么，如何完美地描述一个连续型随机变量呢？连续型随机变量是否有概率质量函数（PMF）呢？

概率质量函数不能描述连续型随机变量！原因是：对于任一连续型随机变量 X，

$$\Pr(X = x) = 0, \qquad \forall x \in \mathcal{X}$$

即连续型随机变量的 PMF 在支撑集的任意一点上的函数值恒为 0。

这一结论的证明需要利用高等数学中“极限”的概念：

$$\Pr(X = x) = \lim_{\epsilon \to 0} \Pr(x \le X \le x + \epsilon) = \lim_{\epsilon \to 0} [F_X(x + \epsilon) - F_X(x)] = 0$$

PDF的定义

我们已知 PMF 无法描述连续型随机变量，但是 CDF 仍然可以完美地描述连续型随机变量。
也许，我们可以通过连续型随机变量的 CDF 逆推出类似 PMF 的函数。

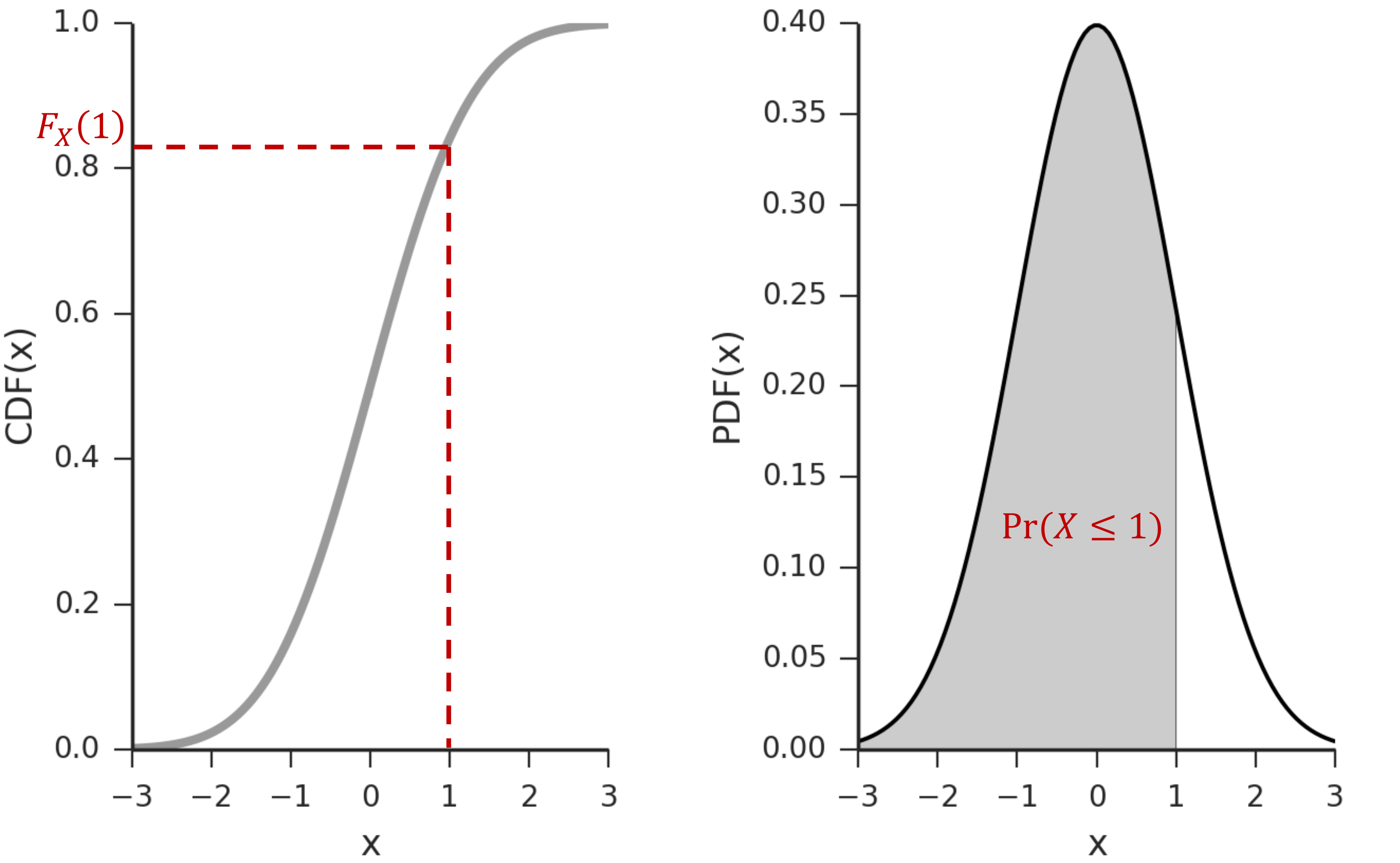
连续型随机变量 X 的概率密度函数（probability density function，简记为 PDF）是

$$\boldsymbol{F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt, \qquad x \in \mathbb{R}}$$

当连续型随机变量 X 的累积分布函数 $F_X(x)$ 在 $x \in S \subseteq \mathbb{R}$ 上可导（differentiable）时，

$$f_X(x) = \underbrace{\boxed{\frac{d}{dx}F_X(x)}}_{\text{导数}}, \qquad x \in S$$

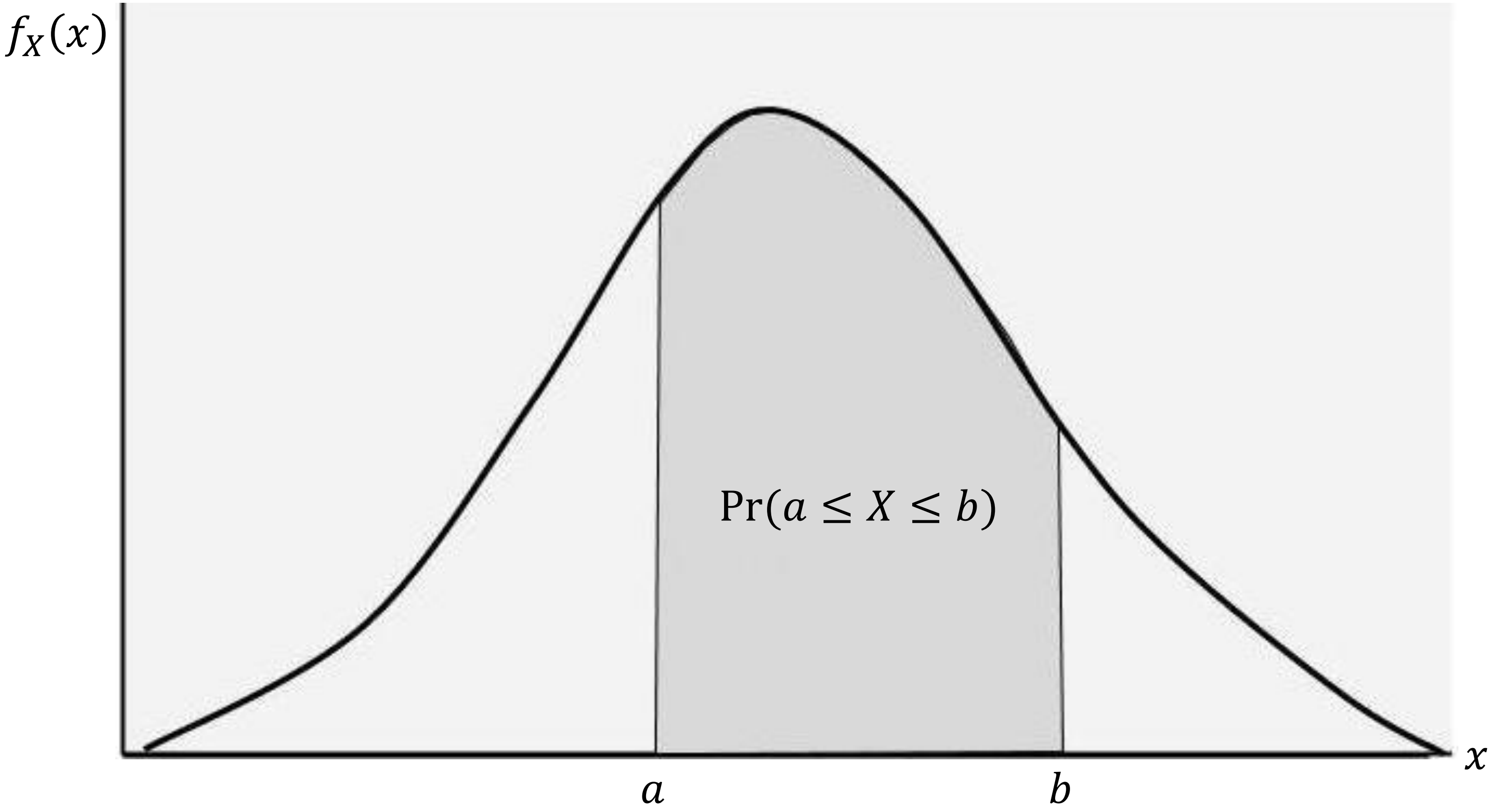
CDF与PDF的关系



1.0
0.8
0.6
0.4
0.2
0.0
CDF(x)
$F_X(1)$
−3
−2
−1
0
1
2
3
x
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
PDF(x)
$\Pr(X \leq 1)$
x

PDF的图象

PDF 的函数曲线与 x 轴之间的面积对应的是随机变量的取值落在特定范围内的概率。



What's next?

接下来，我们将介绍那些不能完美描述随机变量，但是可以快速总结随机变量特征的工具。

他们拥有一个共同的名字——矩（moment）。

SECTION
05

矩的定义与计算

原始矩的定义

随机变量 X 的 m 阶原始矩（m-th raw moment，简称 m 阶矩）是

$$E(X^m)$$

其中，$E(\cdot)$ 是期望算子（expectation operator），代表针对 X^m 的概率分布的特定算法，而非 X^m 的函数。

当 $E(|X^m|) = \infty$ 时，我们称“随机变量 X 的 m 阶原始矩**不存在**”。

矩的存在

在很多统计学教材和论文中，我们都可以看见 $E(|X^m|) < \infty$ 形式的假设或条件，这一假设或条件通常就是为了确保随机变量 m 阶矩的存在。

Theorem 6.1 Weak Law of Large Numbers (WLLN)
If $Y_i \in \mathbb{R}^k$ are i.i.d. and $\boxed{\mathbb{E}\|Y\| < \infty}$, then as $n \to \infty$,

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \underset{p}{\longrightarrow} \mathbb{E}[Y].$$

Theorem 6.3 Multivariate Lindeberg-Lévy Central Limit Theorem (CLT). If $Y_i \in \mathbb{R}^k$ are i.i.d. and $\boxed{\mathbb{E}\|Y\|^2 < \infty}$ then as $n \to \infty$

$$\sqrt{n}\left(\overline{Y} - \mu\right) \underset{d}{\longrightarrow} \mathrm{N}(0, \boldsymbol{V})$$

where $\mu = \mathbb{E}[Y]$ and $\boldsymbol{V} = \mathbb{E}\left[(Y-\mu)(Y-\mu)'\right]$.

数学期望

离散型随机变量 X 的数学期望（expectation 或 expected value，简称“期望”）是 X 的一阶原始矩 μ_X，即

$$\boldsymbol{\mu_X = E(X) = \sum_{x \in \mathcal{X}} x \cdot f_X(x)}$$

其中，$f_X(\cdot)$ 是 X 的概率质量函数（PMF），$\mathcal{X}$ 是 X 的支撑集。

例： 随机变量 Y 的概率分布列如下所示。Y 的数学期望是多少?

Y 的取值	-1	0	1
概率	0.5	0.2	0.3

$$E(Y) = \sum_{y \in \mathcal{Y}} y \cdot f_Y(y) = -1 \times 0.5 + 0 \times 0.2 + 1 \times 0.3 = -0.2$$

数学期望与平均数

离散型随机变量的数学期望是试验中所有可能的结果乘以其相应的概率的总和。

注意：数学期望不等于我们平时常说常用的平均数！

我们日常所说的平均数在统计学中被称为“简单算术平均数”（ordinary arithmetic mean）：

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

其中，X_i 代表我们观测到的数值，n 代表观测的次数。显然，这和数学期望的定义式不同！

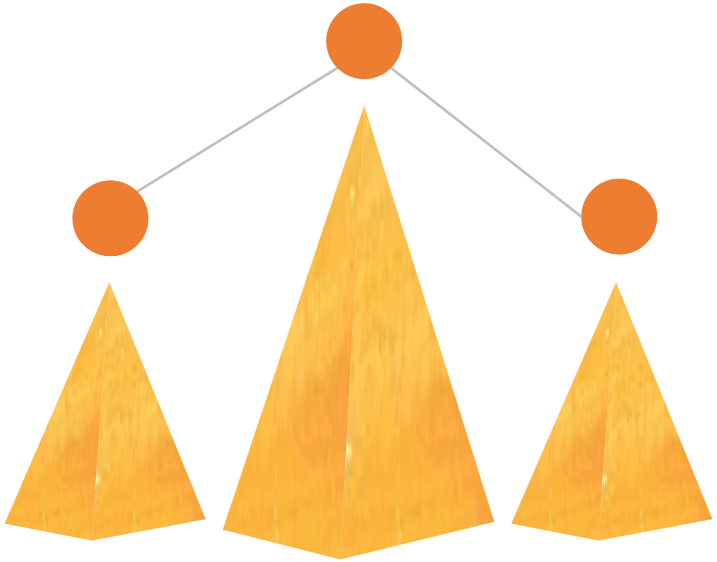
实际上，数学期望可以理解为随机变量 X 所有可能取值的加权算数平均数（weighted arithmetic mean），权重就是各个可能取值相应的概率。

$$\bar{X} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i} \Rightarrow E(X) = \frac{\sum_{x\in\mathcal{X}} x \cdot \Pr(X = x)}{\sum_{x\in\mathcal{X}} \Pr(X = x)} = \sum_{x\in\mathcal{X}} x \cdot \Pr(X = x) = \sum_{x\in\mathcal{X}} x \cdot f_X(x)$$

数学期望的含义

数学期望反映的是一个随机变量取值的中心趋势（central tendency）——它回答了“随机变量的取值集中趋向于何值？”

数学期望并非唯一可以反映中心趋势的统计工具；中位数（median）、众数（mode）等都可以反映中心趋势。



数学期望的局限

Expectation is not everything!（数学期望不意味着一切！）

数学期望，如同接下来我们将要学习的二阶矩、三阶矩等，只反映了随机变量所服从的概率分布的一个特征。数学期望并不能提供给我们关于一个概率分布的全部信息。

例一

X 的取值	-1	0	1
概率	0.3	0.4	0.3

Y 的取值	-2	0	2
概率	0.3	0.4	0.3

显然，X 和 Y 服从的是两个不同的概率分布，但是它们的数学期望是相等的。

$$E(X) = 0 = E(Y)$$

数学期望的局限

Expectation is not everything!（数学期望不意味着一切！）

数学期望，如同接下来我们将要学习的二阶矩、三阶矩等，只反映了随机变量所服从的概率分布的一个特征。数学期望并不能提供给我们关于一个概率分布的全部信息。

例二

X 的取值	0	1	10000
概率	0.4	0.59	0.01

当 X 的可能取值中存在极端值时，即使它出现的概率很低，X 的数学期望仍会收到巨大影响：

$$E(X) = 100.59$$

中心矩和方差

随机变量 X 的 m 阶中心矩（m-th central moment）是

$$E\{[X-E(X)]^m\}$$

离散型随机变量 X 的方差（variance）是 X 的二阶中心矩 σ_X，即

$$\boldsymbol{\sigma_X^2 = D(X) = \sum_{x \in X} (x-\mu_X)^2 \cdot f_X(x)}$$

中国大陆高中、本科教材一般使用 $D(X)$ 代表方差（来源于英文单词 deviation，意为“离差”），而西方教材一般使用 $Var(X)$ 或 $var(X)$。

方差的计算

X 的概率分布列如下所示。它的方差是多少？

X 的取值	0	1	2
概率	0.3	0.2	0.5

X 的数学期望为 $E(X) = 0 \times 0.3 + 1 \times 0.2 + 2 \times 0.5 = 1.2$。接下来，我们可以计算 X 各个取值与数学期望的差，然后再计算各个差的平方数。

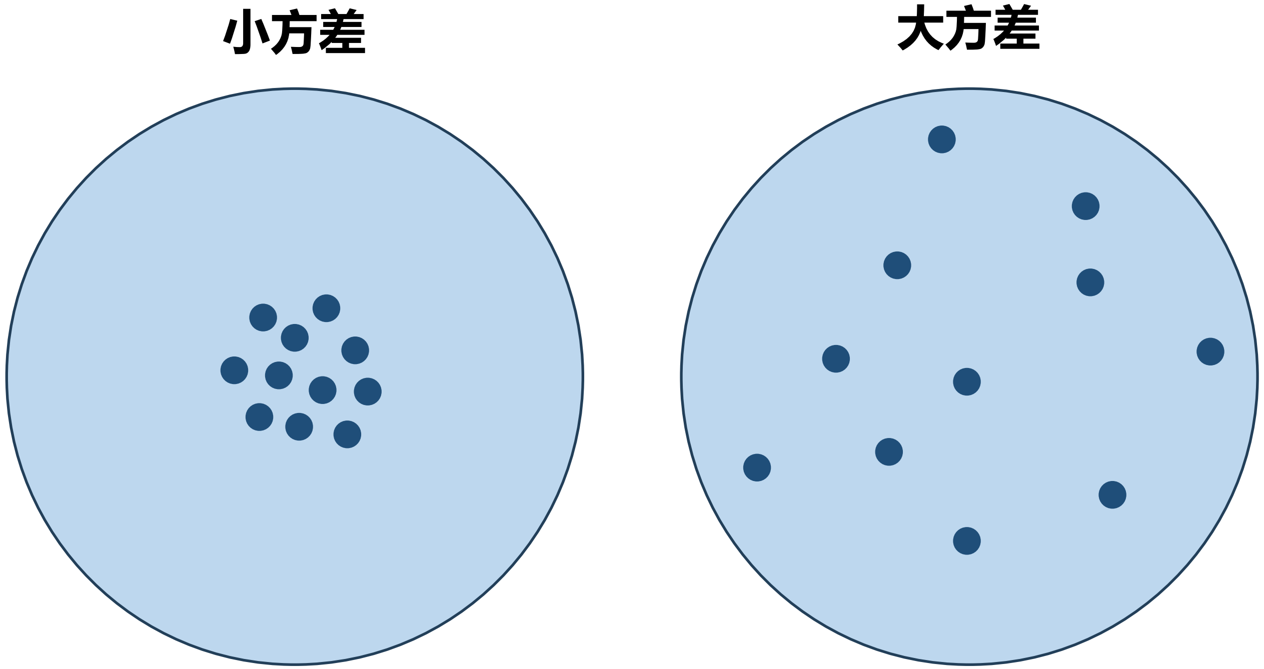
X 的取值	0	1	2
$E(X)$		1.2	
做差	-1.2	-0.2	0.8
平方	1.44	0.04	0.64
概率	0.3	0.2	0.5

最后，将这些平方差乘以各自相应的概率，加总，我们就得到了 X 的方差：

$$D(X) = 1.44 \times 0.3 + 0.04 \times 0.2 + 0.64 \times 0.5 = 0.76$$

方差的含义

方差反映的是一个随机变量取值的离散程度（dispersion）——它回答了“随机变量的取值有多么分散？”



标准差的定义

“方差”的单位是古怪的。例如，随机变量 X 代表一所学校内男生的身高，单位为米（m），那么 X 的方差 $D(X)$ 的单位就是平方米（m²）。这个方差的值到底有什么现实含义呢?

为了解决这个问题，统计学家们选择对方差进行开方，并将这种统计工具称为“标准差”（standard deviation）。对于离散型随机变量 X，它的标准差为

$$\sigma_X = SD(X) = \sqrt{D(X)} = \sqrt{\sum_{x \in \mathcal{X}} (x - \mu_X)^2 \cdot f_X(x)}$$

显然，标准差的单位和随机变量的单位是保持一致的。

方差为什么方？

使用一阶中心矩来衡量离散程度的问题在于：等概率的“正距离”和“负距离”会相互抵消。

例子

X 的取值	-1000	1000
概率	0.5	0.5

X 的数学期望为 $E(X)=0$，一阶中心矩也为 0，尽管 X 的离散程度看起来非常大。

为了解决这个正负相互抵消的问题，我们可以在一阶中心矩中加入“绝对值”：

$$\sum_{x\in\mathcal{X}}|x-\mu_X|\cdot f_X(x)$$

这个统计工具被称为“平均绝对离差”（mean absolute deviation），记为 $MAD(X)$。

使用平均绝对离差来衡量离散程度并没有问题。但是，统计学家们更喜欢使用“平方”而非“绝对值”。

$$\sum_{x\in\mathcal{X}}(x-\mu_X)^2\cdot f_X(x)$$

这不是为了符合二阶中心矩的定义，而是因为考虑到计算机语言的特点：现实中，大多数统计研究或应用都需要统计软件的帮助（笔算再快也快不过 Excel）。在计算机语言中，平方运算是一个如同加减乘除般简单的任务。然而，绝对值运算是一个条件语句（conditional statement）：当输入值大于或等于 0 时，返回原数值；当输入值小于 0 时，返回原数值的相反数。编写过软件代码的同学应该知道，条件语句的代码行数往往会超过普通运算代码的行数，这也意味着在计算机眼中，绝对值运算是一个比平方运算更加繁琐的任务。

使用“平方”的另一个优点是：“平方”惩罚了极端值。

当随机变量的取值极大或者极小时，这些极端值与数学期望的距离会在平方之后被放大，使得它们对方差的增大有着更显著的作用。

感谢各位的聆听！

公选课《可计量的社会》　何濯羽　2024年9月23日