随机变量的定义与描述

高中统计学 第贰课　何濯羽 2024年2月12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?

口头语言的定义：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）。随机变量总是可以由概率分布函数刻画。随机变量为一场试验（experiment）中的每个事件（event）赋予了一个数。

这节课的核心知识!

数学语言的定义：随机变量是一个定义在样本空间 S 上的实值函数（real-valued function），即

$$X: S \to \mathbb{R}$$

换言之，对于样本空间内的每一个元素 $\omega \in S$，随机变量都赋予了它们一个（且仅有一个）实数。随机变量的值域自然是一个特定的实数集：$\{x \in \mathbb{R} | x = X(\omega), \omega \in S\}$。

随机变量的定义

What is a random variable?

注意：

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：

我们可以用 X 表示兰溪一中任一学生的年级，用 x_1, x_2, x_3 分别表示高一、高二、高三这三个可能的年级。这样，随机变量 X 的支撑集为 $\mathcal{X} = \{x_1, x_2, x_3\}$。

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。

借助指标函数（indicator function），我们可以轻松地构建离散型随机变量。

例如，我们用 W 表示兰溪市任一居民的月收入，然后我们可以构建一个离散型随机变量 X如下：

$$X = \mathbb{I}(W) = \begin{cases} 1, \text{当 } W \in [0, 5000] \\ 0, \text{当 } W \notin [0, 5000] \end{cases}$$

指标函数：满足特定条件时返回1，否则，返回0。

这表示，若某位居民的月收入在区间 $[0, 5000]$ 内，随机变量 X 取值为1，否则， X 取值为0。

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，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

在高中阶段，我们将重点学习前两种函数。

SECTION 02

概率质量函数（PMF）

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

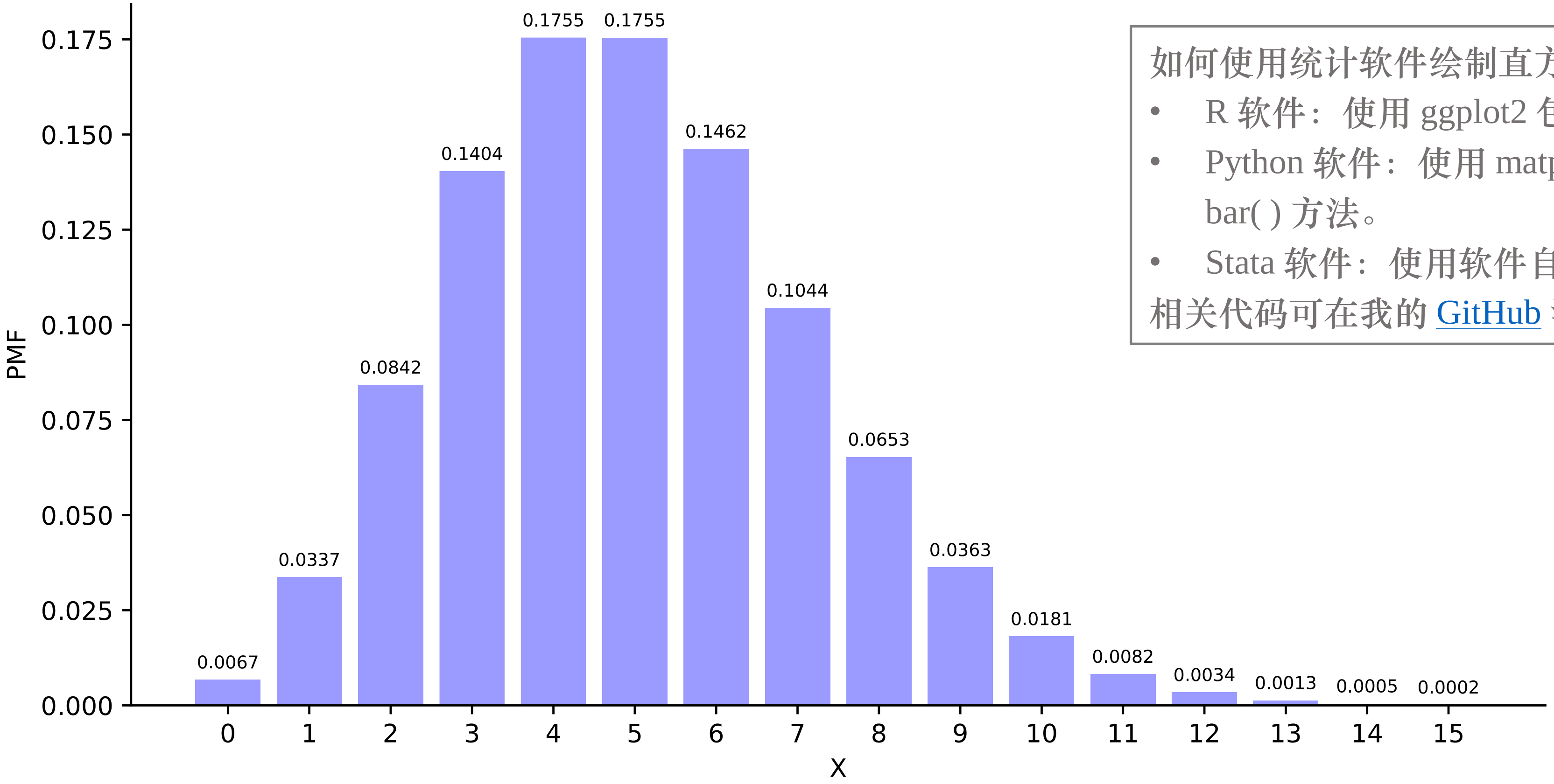
例：我们用 X 表示投掷一枚均匀硬币的结果，$X=1$ 代表正面朝上，$X=0$ 代表反面朝上，那么 X 的概率质量函数可以用下表（中国人教版高中教材称其为“概率分布列”）展示：

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。

PMF的性质

任一概率质量函数必定满足以下两个性质：

- 值域：$0 \le f_X(x) \le 1$；
- 总和：$\sum_x f_X(x) = 1$。

我们利用概率三大公理（第壹课的内容）可以很轻易地证明这两个性质必定成立。

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

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

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。

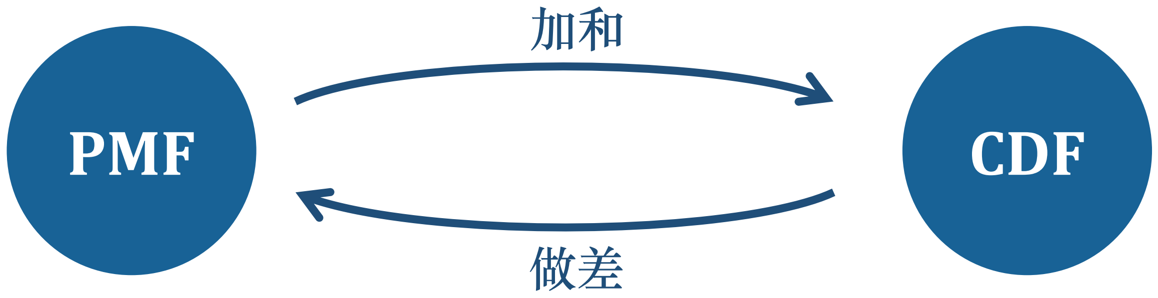
CDF 和 PMF 的关系

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：

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

反之，X 的概率质量函数也可以由它的累积分布函数求得：

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

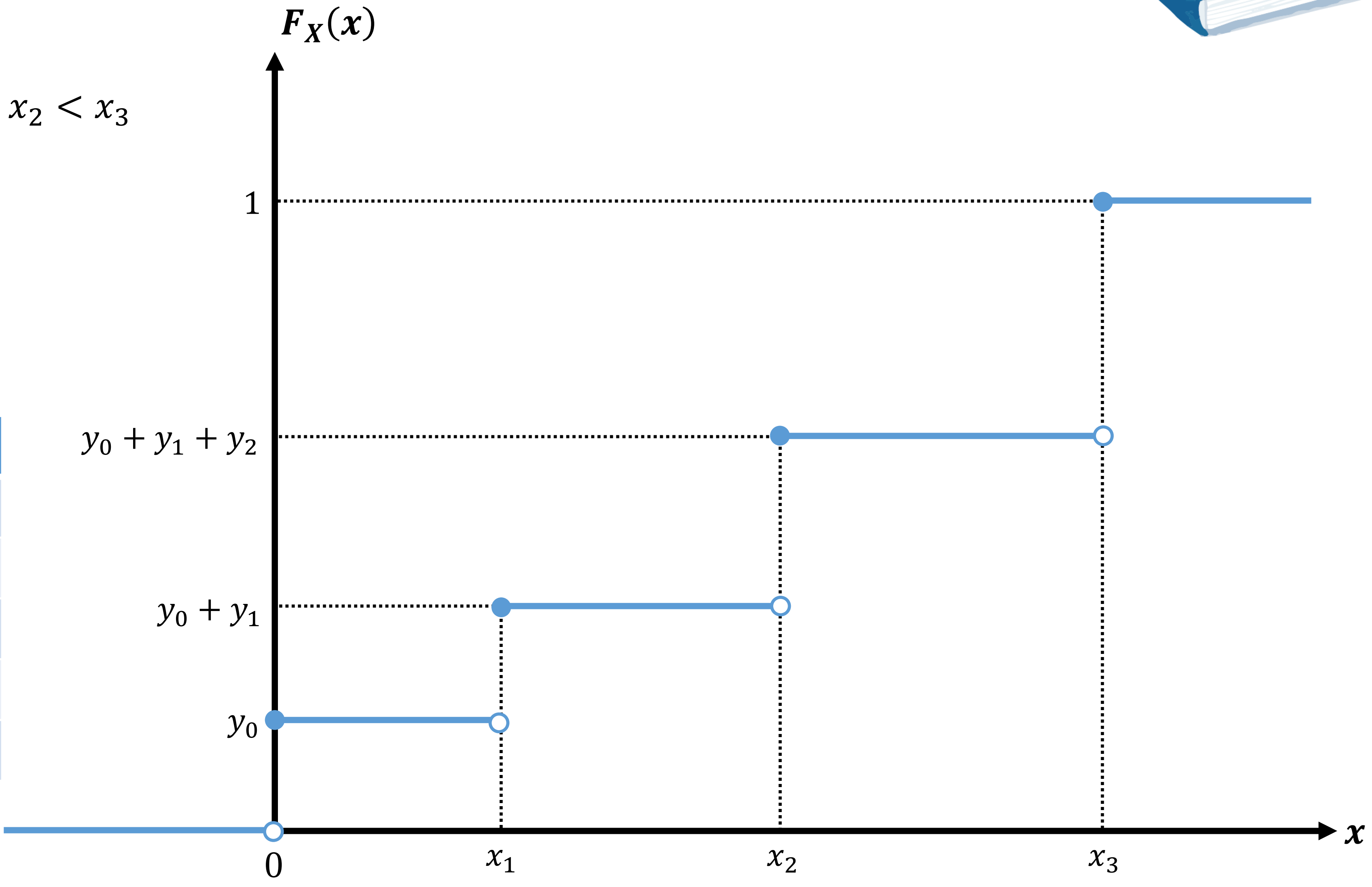


PMF → CDF

X 的取值	概率 $\Pr(X = x)$
0	y_0
x_1	y_1
x_2	y_2
x_3	y_3

$0 < x_1 < x_2 < x_3$

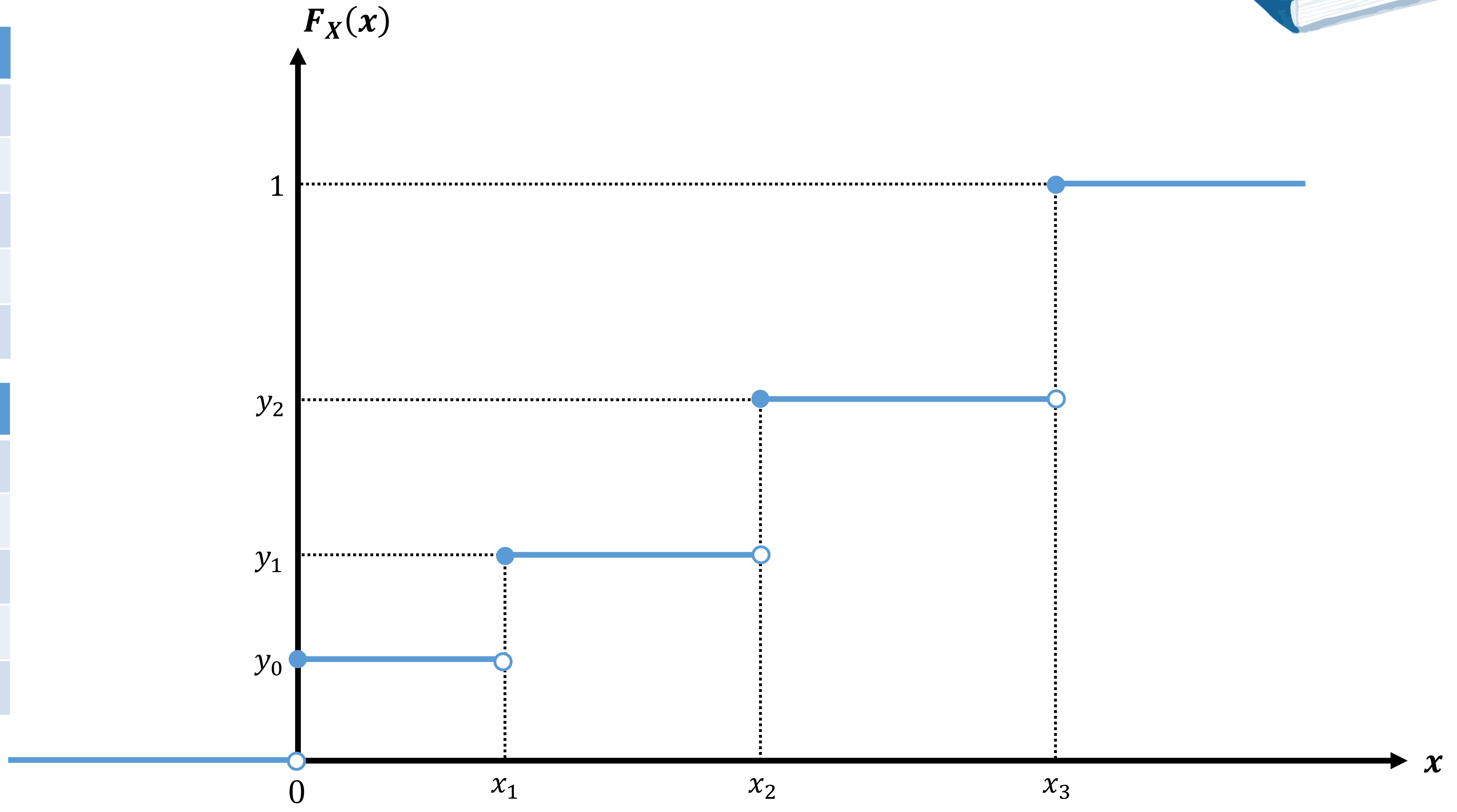
X 的取值	累积概率 $\Pr(X \le x)$
$x < 0$	0
$0 \le x < x_1$	y_0
$x_1 \le x < x_2$	$y_0 + y_1$
$x_2 \le x < x_3$	$y_0 + y_1 + y_2$
$x \ge x_3$	$y_0 + y_1 + y_2 + y_3 = 1$



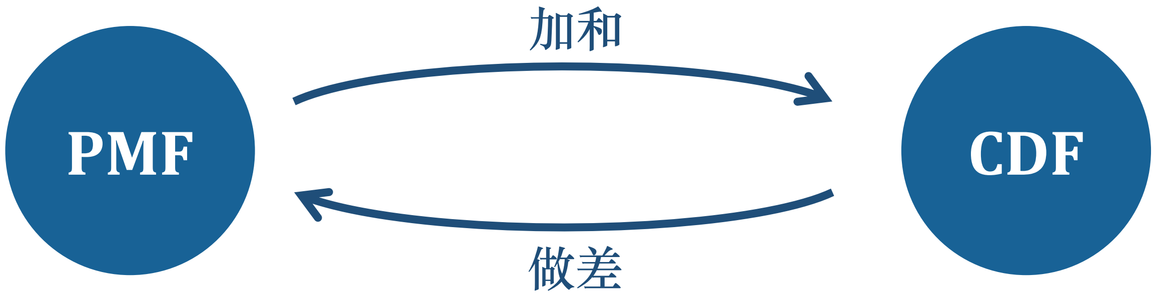
CDF → PMF

X 的取值	累积概率
$x < 0$	0
$0 \le x < x_1$	y_0
$x_1 \le x < x_2$	y_1
$x_2 \le x < x_3$	y_2
$x \ge x_3$	1

X 的取值	概率
$x = -1$	0
$x = 0$	y_0
$x = x_1$	$y_1 - y_0$
$x \in (0, x_1)$	0
$x = x_3$	$1 - y_2$



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。

CDF的性质

任一累积分布函数必定满足以下四个性质：

- $\lim_{x\to-\infty} F_X(x)=0$， $\lim_{x\to+\infty} F_X(x)=1$；
- $F_X(x)$ 是一个非单调递减（non-decreasing）函数；
- $F_X(x)$ 是一个右连续（right-continuous）函数，即 $\lim_{x\to x_0^+} F_X(x)=F_X(x_0)$；
- $F_X(x)$ 在每一点上均有左极限，即 $\lim_{x\to x_0^-} F_X(x)$ 对于任意 x_0 存在。

练习：尝试证明第 2 个性质。

注：第 1、3、4 点已超出高中数学范围，可选择忽略，也可尝试用函数图象理解。

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

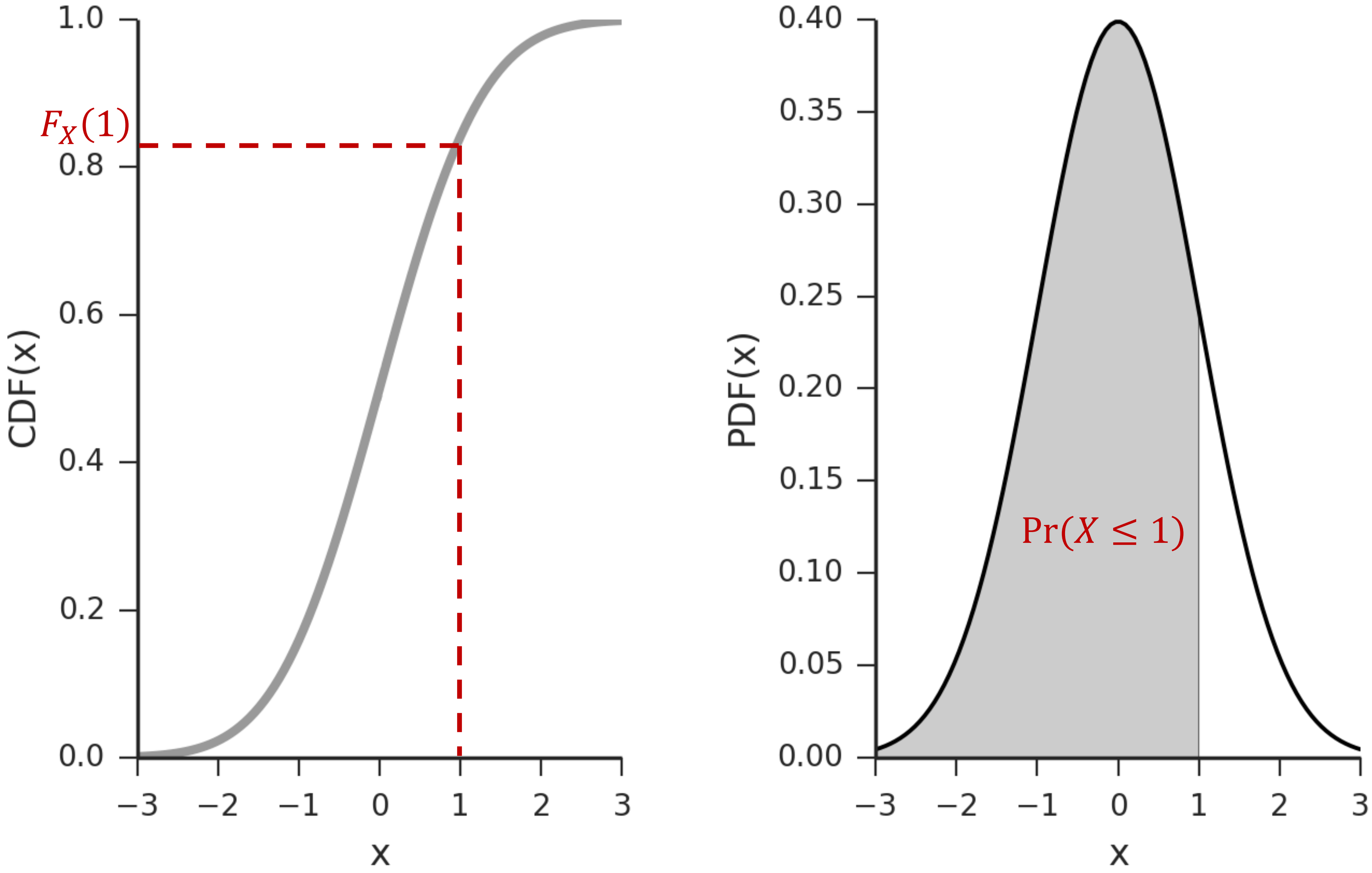
$$\boldsymbol{F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt, \qquad x \in \mathbb{R}}$$

提醒：积分（integration）的概念已超出高中数学范围。

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，

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

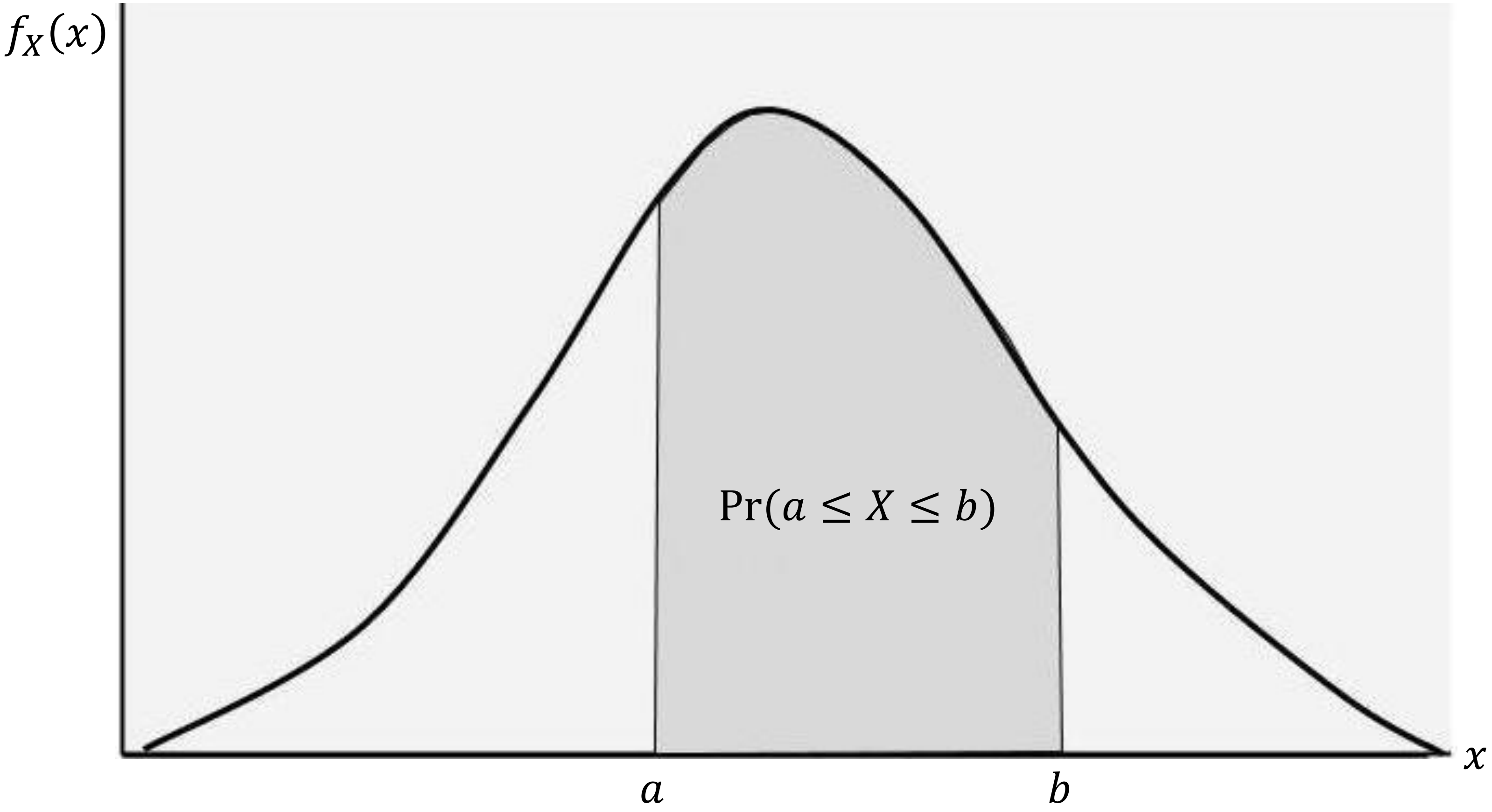
CDF与PDF的关系



1.0
$F_X(1)$
0.8
0.6
0.4
0.2
0.0
CDF(x)
−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)
$\Pr(X \leq 1)$
−3
−2
−1
0
1
2
3
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。



PDF的性质

任一概率密度函数必定满足以下两个性质：

- 值域：$f_X(x) \geq 0$；
- 总和：$\int_{-\infty}^{+\infty} f_X(x)\, dx = 1$。

注意：与 PMF 不同的是，PDF 的函数值可能大于 1。关于这一点的解释是：PMF 的函数值对应的是概率（probability），而 PDF 的函数值对应的是概率密度（probability density）——这是两个不同的概念。

SECTION 05

分位数简介

分位数的定义

分位数（quantile），亦称分位点，是指将一个随机变量的概率分布分为几个等份的数值点。

例如，随机变量 X 的四分位数是把 X 的概率分布分为 4 个等份的 3 个数值点。

我们最常用的分位数是“百分位数”（percentile）——将随机变量的概率分布分为 100 个等份。

大多数分位数都可以转换为百分位数的形式，例如：X 的第 1 四分位数相当于 X 的第 25 百分位数。

现实中，我们可能无法知晓随机变量的概率分布。这时，我们可以对随机变量进行多次观测，然后将所有观察值从小到大排序后，找到相应的分位数。通过这种手段找到的分位数被称为样本分位数（sample quantile）。

分位数的名称

中文	英文
二分位数（亦称中位数）	median
三分位数	tertile / tercile
四分位数	quartile
五分位数	quintile
六分位数	sextile
七分位数	septile
八分位数	octile
十分位数	decile
十二分位数	duodecile
十六分位数	hexadecile
二十分位数	vigintile
百分位数	percentile
千分位数	millile

这些英文单词的前缀均源自拉丁语中代表相应数字的单词。了解这些前缀对扩充代数、几何、统计领域的英语词汇量有所帮助。

样本分位数的计算

假设下列数是我们对随机变量 X 进行 16 次观测后记录下的观测值。

25	30	47	31
26	2	1	26
30	50	46	30
45	29	21	3

排序

1	2	3	21
25	26	26	29
30	30	30	31
45	46	47	50

请问，第 1 样本四分位数是多少?

解:

将所有观测值从小到大排序后，我们发现第 4 个数和第 5 个数分别为 21 和 25。这意味着区间 $[21,25)$ 内的任何一个实数都可以当作“第 1 样本四分位数”。人教版教材建议我们把第 4 个数和第 5 个数的平均数，即 $\frac{21+25}{2}=23$，当作第一个样本四分位数。

样本分位数的计算

计算**第 p 样本百分位数**的步骤：

1) 将所有观测值按从小到大排序；
2) 计算 $i = n \times p\%$；
3) 当 i 是整数时，第 p 样本百分位数为第 i 项和第 $(i+1)$ 项观测值的平均数；当 i 不是整数而大于 i 的比邻整数为 j 时，第 p 样本百分位数为第 j 项观测值。

例：第 1 样本四分位数相当于第 25 样本百分位数。所以，

1	2	3	21
25	26	26	29
30	30	30	31
45	46	47	50

$$i = n \times p\% = 16 \times 25\% = 4 \in \mathbb{Z}$$

$$\Rightarrow Q(0.25) = \frac{21+25}{2} = 23$$

样本分位数的计算

注意：

在上一页的例题中，将第 4 个数和第 5 个数的平均数当作第 1 样本四分位数只是人教版教材的观点。事实上，任何一个处于区间 [21,25) 内的实数都可被称为第 1 样本四分位数。

不同的学者、书籍、软件对于上述情况中分位数的计算持有不同的观点，没有统一的标准。

因此，尤其是在使用统计软件计算样本分位数时，请务必阅读相应软件的官方说明书，了解其计算分位数时所用的算法。例如，

- 在 R 软件的 stats 包（3.6.2 版本）中，计算样本分位数的 `quantile( )` 函数提供了 9 种不同的算法来应对上述情况，其中“type = 2”对应人教版教材中介绍的平均取值法。
- 在 Microsoft Excel（2007 以上版本）中，可以计算样本百分位数的函数有两个：`PERCENTILE.INC( )` 和 `PERCENTILE.EXC( )`，但是这两个函数的算法均不是人教版教材中的平均取值法（Excel 的统计函数主要是供商务人士使用的，所以函数的算法通常只满足金融学和会计学的要求）。

分位数函数的定义

随机变量 X 的第 $100\tau\%$ 分位数函数（quantile function）是

$$\boldsymbol{Q_X(\tau) = F_X^{-1}(\tau), \qquad \tau \in (0,1)}$$

简言之，分位数函数是累积分布函数的“特殊化”反函数。

“特殊化”三字不可忽略！累积分布函数是非单调递减函数，这意味着它并不总是拥有反函数。若要确保上述定义的分位数函数对于任意 X 恒存在，必须对反函数进行“特殊化”——人教版教材中将第 i 项和第 $(i+1)$ 项的平均数当作相应的分位数就是一种“特殊化”。

我们已知 CDF 可以完美地描述（离散/连续型）随机变量，此刻我们发现了分位数函数与 CDF 的等价关系，所以分位数函数也是可以完美描述（离散/连续型）随机变量的一个工具。

What’s next?

下节课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）。

感谢在座各位的聆听！

何濯羽　　　　2024年2月12日