数据科学基础

Foundations of Data Science

3.1 数据分布与概率分布

陈振宇
南京大学智能软件工程实验室
www.iselab.cn

数据分布

给定一批数据，

我们如何来描述数据的分布？

频数分布与频率分布

- 频数是各个数据被观测到的次数。
- 频率是频数除以总次数。

例如，为了进一步改善节假日安排, 相关部门进行了一次关于黄金周过节方式的网络调查, 收到17452有效调查票。详细调查结果如右边表格。

过节方式	频数	频率
在家休息	7853	0.45
探亲访友	6632	0.38
外出度假	873	0.05
公司加班	873	0.05
其他	1221	0.07
合计	17452	1

累积频数与累积频率

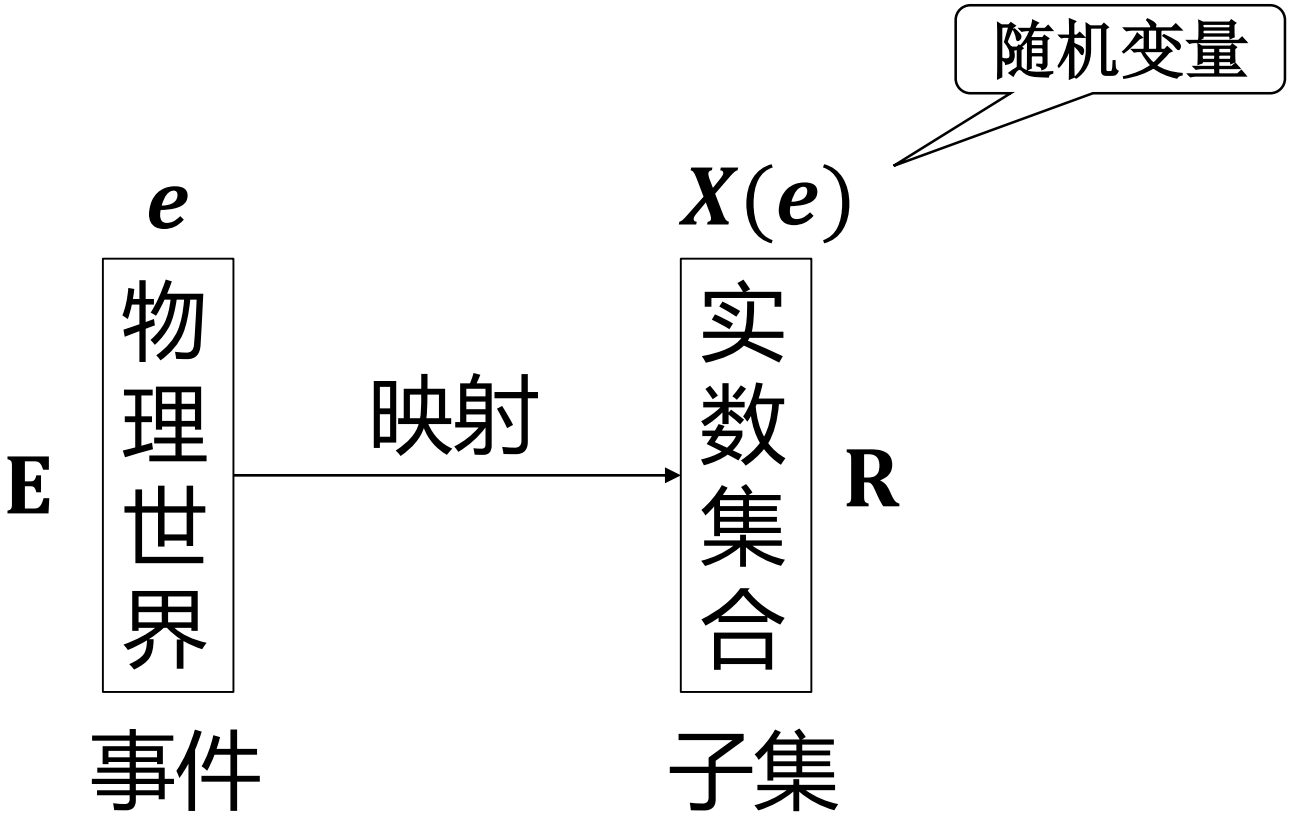
- 数据排序后（通常是降序）进行频数或者频率的累加。

过节方式	频数	频率	累积频数	累积频率
在家休息	7853	0.45	7853	0.45
探亲访友	6632	0.38	14485	0.83
其他	1221	0.07	15706	0.90
外出度假	873	0.05	16579	0.95
公司加班	873	0.05	17452	1.00
合计	17452	1		

概率分布

我们需要从特定数据到一般规律的抽象，才能奠定数据科学的基础。

随机变量



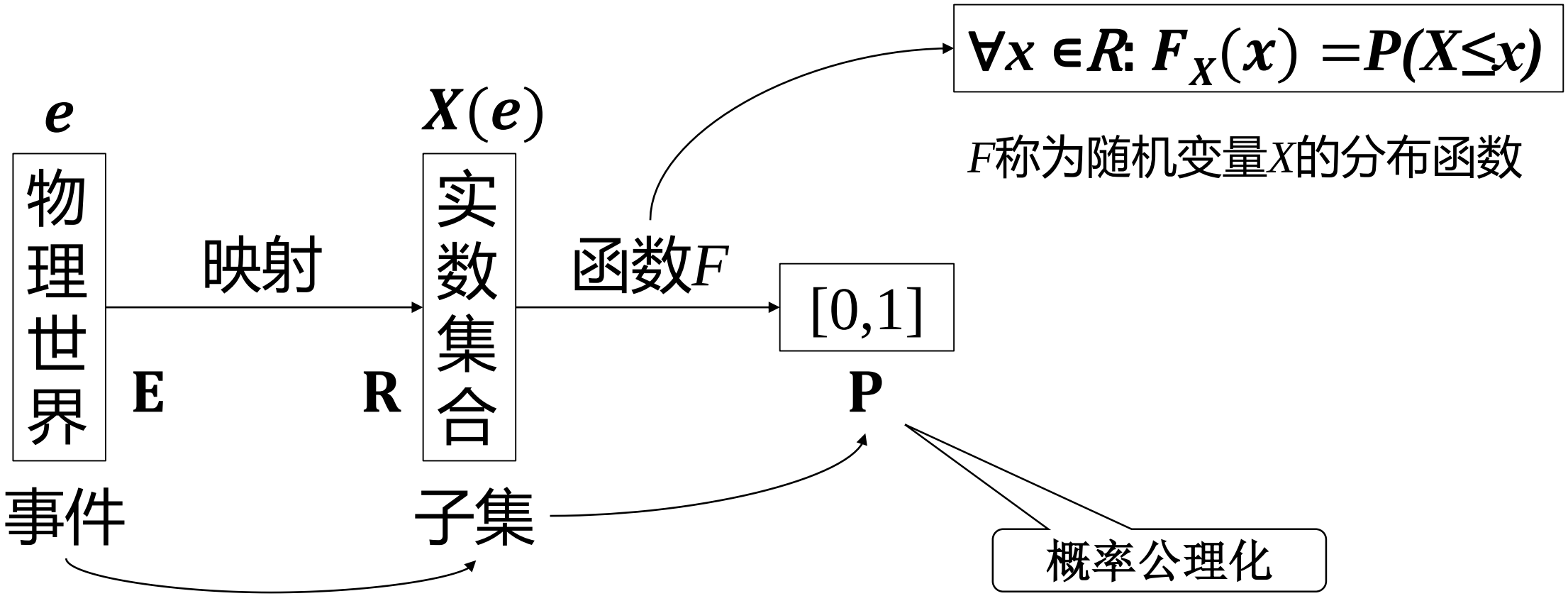
随机变量通常使用大写字母X,Y,Z等表示，并将e省略。

随机变量示例

将一枚硬币抛掷3次，我们感兴趣的是三次投掷中，出现H的总次数,而对H, T出现的次序不关心。以X记三次投掷中出现H的总次数,那么对于样本空间中的每一个样本点e, X都有一个值与之对应：

e	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
X	3	2	2	2	1	1	1	0
Y	1	2	3	4	5	6	7	8
Z	2	3	3	3	0	0	0	1

概率分布


$\forall x \in R: F_X(x) = P(X \leq x)$
F称为随机变量X的分布函数
e
$X(e)$
物理世界
实数集合
映射
函数F
[0,1]
E
R
P
事件
子集
概率公理化

分布函数

- 归一性：$0 \le F(x) \le 1, \forall x \in R$，且

$$F(-\infty) = \lim_{x \to -\infty} F(x) = 0, \mathrm{F}(+\infty) = \lim_{x \to +\infty} F(x) = 1$$

- 单调不减性：

若$x_1 \le x_2$，则有$F(x_1) \le F(x_2)$；

- 右连续性：对任意$x_0 \in R$，有

$$F(x_0) = \lim_{x \to x_0^+} F(x)$$

离散随机变量

【离散型随机变量】

一个随机变量X的可能取值为有限个或可列无穷多个，则称X为离散型随机变量。

离散随机变量-概率分布律

【概率分布律】

X是一个离散型随机变量，其一切可能值为$\{x_1, x_2, \cdots, x_n, \cdots\}$,

X的概率分布律是其取各值时的概率

$$P(X = x_i) = p_i, i = 1,2, \cdots, n, \cdots$$

其中$P(x_i) \geq 0, (i = 1,2, \cdots), \sum_{i=1}^{\infty} p_i = 1$。

概率分布律

从物理世界引入离散型随机变量完成实数映射

e	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
X	3	2	2	2	1	1	1	0

从随机变量值到概率值构建概率分布律

X	3	2	1	0
P	1/8	3/8	3/8	1/8

概率分布律

离散随机变量-分布函数

根据离散随机变量的概率分布律, 可以通过下式求得分布函数:

$$F(X) = P(X \le x) = \sum_{x \le i} P\{X = x_i\} = x_i = \sum_{x_i \le x} p_i$$

X	3	2	1	0
P	1/8	3/8	3/8	1/8

互相转换

$$F(x) = \begin{cases} 0: & x < 0 \\ \frac{1}{8}: & 0 \le x < 1 \\ \frac{1}{2}: & 1 \le x < 2 \\ \frac{7}{8}: & 2 \le x < 3 \\ 1: & x \ge 3 \end{cases}$$

我们可以用概率分布律或分布函数来描述离散型随机变量。

离散随机变量示例

设一汽车在开往目的地的道路上需经过四组信号灯,每组信号灯以p的概率允许或禁止汽车通过。以X表示汽车首次停下时,它已通过的信号灯的组数(信号灯工作独立),求X的分布律。

解：以p表示每组信号灯禁止汽车通过的概率, 易知X的分布律为

X	0	1	2	3	4
p_k	p	$(1-p)p$	$(1-p)^2p$	$(1-p)^3p$	$(1-p)^4$

我们假设$p=\frac{1}{2}$，则代入得

X	0	1	2	3	4
p_k	0.5	0.25	0.125	0.0625	0.0625

试写出相应的分布函数和画出相应的分布函数图。

连续随机变量

- 一个靶子是半径为2米的圆盘, 设击中靶上任意同心圆盘上的点的概率与该圆盘的面积成正比, 并设射击都能击中靶, 以X表示弹着点于圆心的距离。 试求随机变量X的分布函数。

连续随机变量

- 解: 若$X<0$, 则$\{X\le x\}$是不可能事件, 于是 $F(x)=P\{X\le x\}=0$. 若$0\le x\le 2$, 由题意, $P\{0\le X\le x\}=kx^2$, k是某一常数, 为确定k的值, 取$x=2$, 有$P\{0\le X\le 2\}=2^{2k}$, 但已知$P\{0\le X\le 2\}=1$, 故得$k=\frac{1}{4}$, 即$P\{0\le X\le x\}=\frac{x^2}{4}$, 于是 $F(x)=P\{X\le x\}=P\{X<0\}+P\{0\le X\le x\}=\frac{x^2}{4}$
- 若$X>2$, 由题意, 有$F(x)=P\{X\le x\}=1$. 综合上述, 即得X的分布函数为

$$F(x)=\begin{cases}0, x<0\\ \frac{x^2}{4}, 0\le x<2\\ 1, x\ge 2\end{cases}$$

- X的分布函数也可以表示为$F(x)=\int_{-\infty}^{x}f(t)dt$, 其中$f(t)=\frac{t}{2}, 0<t<2;\ f(t)=0$, 其他.

连续随机变量

【连续随机变量】对于随机变量X，其分布函数为$F(x)$，如存在非负可积函数$f(x)$，使得对于任意实数x，有$F(x)=\int_{-\infty}^{x} f(t)dt$，则称X为连续型随机变量，$f(x)$称为$X$的概率密度函数。

概率密度函数性质

1. $f(x) \geq 0$
2. $\int_{-\infty}^{+\infty} f(x)dx = 1$
3. 对于任意实数$a, b(a < b)$，

   都有$P(a < X \leq b) = \int_a^b f(x)dx = F(b) - F(a)$
4. F若$f(x)$在点x处连续, 则有$F'(x) = f(x)$, 即分布函数$F(x)$是概率密度的一个原函数
5. 对于连续性随机变量X, X取任一指定实数值a的概率均为0, 即$P\{X = a\} = 0$

连续随机变量示例

设随机变量X具有概率密度

$$f(x)=\begin{cases} kx, 0 \le x < 3 \\ 2-\dfrac{x}{2}, 3 \le x < 4 \\ 0, \text{其他} \end{cases}$$

(1) 确定常数k; (2) 求X的分布函数$F(x)$; (3) 求$P\{1 < X \le \frac{7}{2}\}$.

连续随机变量示例

解: (1) 由$\int_{-\infty}^{+\infty} f(x)dx = 1$得$\int_0^3 kxdx + \int_3^4 (2 - x/2)dx = 1$, 解得$k = \frac{1}{6}$.

(2) 所以分布函数为$F(x) = \begin{cases} 0, x < 0 \\ \int_0^x \frac{t}{6} dt, 0 \le x < 3 \\ \int_0^3 \frac{t}{6} dt + \int_3^x 2 - \frac{t}{2} dt, 3 \le x < 4 \\ 1, x \ge 4 \end{cases}$

(3) $P\{1 < X \le \frac{7}{2}\} = F(\frac{7}{2}) - F(1) = \frac{41}{48}$.

更多详情，请咨询课程助教。