数据科学基础知识

第一部分:概率论

ISEP第二年
2023-2024

基于 Nathalie Colin 和 Jean-Claude Guillerot 教授的课程

概率论

第三次会议（2023 年 10 月 13 日）：

第 4 章:实值随机变量的典型值。

第五章:特色功能

概率论

第 4 章:实随机变量的典型值。

- Ø4.1 期望值（平均值）
- Ø4.2 The Median（中间值）
- Ø4.3 众数（出现次数最多的值）
- Ø4.4 百分位数
- Ø4.5 随机变量的矩

实随机变量的典型值

介绍：

随机变量完全由以下任一定义：

Ø其累积分布函数(CDF)， FX (x) OR

Ø其概率密度函数， fX (x)

实际上，这些函数取决于一些通常未知的参数。然而，我们可以而且需要通过某些易于描述的参数来描述随机变量

措施：

Ø期望值（也称为平均值或平均值） Ø中位数（中间值） Ø众数:最可能的值Ø百分位数

Ø方差

期望值

定义（对于连续随机变量）：

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f_X(x) \cdot dx$$

E(X) :随机变量 X 的期望值

fX (x) : 随机变量 X 的概率密度

定义（对于离散随机变量）：

$$E(X) = \sum_i P_i \cdot x_i$$

Pi : 获得值xi的概率

评论：

Ø 如果概率密度函数绕点对称,= 且 ε ,则期望值等于该值：

( ) = 如果 ! ( ) = ! (2 − ∀ ∂

示例:正态分布N(m,σ2)围绕点 = Ø 对称 对于连续随机变量,如果满足以下条件,则期望值未定义：

积分不是有限的。

示例:柯西分布的期望值未定义

Ø 期望值与随机变量的单位相同
测量/尺寸。

示例:如果 X 表示以秒为单位的时间,则 E(X) 也是以[s]为单位的时间。

Ø 如果随机变量是离散X,则期望值可以等于不属于X域的值。

例:X:抛硬币后得到的反面数量,E(X)=0.5

根据相对频率解释预期值

考虑一个实验,其中有 N 个可能的结果 { $}$% ,⋯,# ,其中 N 个数值对应 = 1, ⋯ ,
通过应用程序 X($)

结果	频率 结果...	
w1	k1次k1次， x1=j(w1)	
w2	k2次k2次， x2=j(w2)	
……	……	……
wN	kN次试验	kN次， xN=j(wN)
可能的结果		

通过加权平均:

= + ⋯ + ## / # = / $ $ $

!! 值的相对频率 #, 这可以近似为概率
,,
当n变大并接近无穷大时,得到X = xi 。

$ = (= $) ≈ $ 所以 ≈ / # $ $ $ = ( )

随机变量的确定性函数的期望值

令随机变量 Y 为由 Y=g(X) 给出的随机变量 X 的变换。

我们要计算E(Y)

有两种可能：

1.首先计算fY(y)并推导出E(Y)（下一节）

2.更简单,例如,直接从 g(x) 得到连续的
随机变量：

$$(\quad)^{=}\quad(\;(\quad))=\&_{“\#}^{\$\#}\quad(\quad)\;_{!}\,(\quad)$$

中位数xe （“中间”值)(1/2)

中位数xe是将概率分布的上半部分与下半部分分开的值。形式上,中位数xe是满足以下条件的任何值：

$$P(X \leqslant 车辆) = P(X>车辆) =0.5 = FX (车辆)$$

警告:不要与预期值混淆!

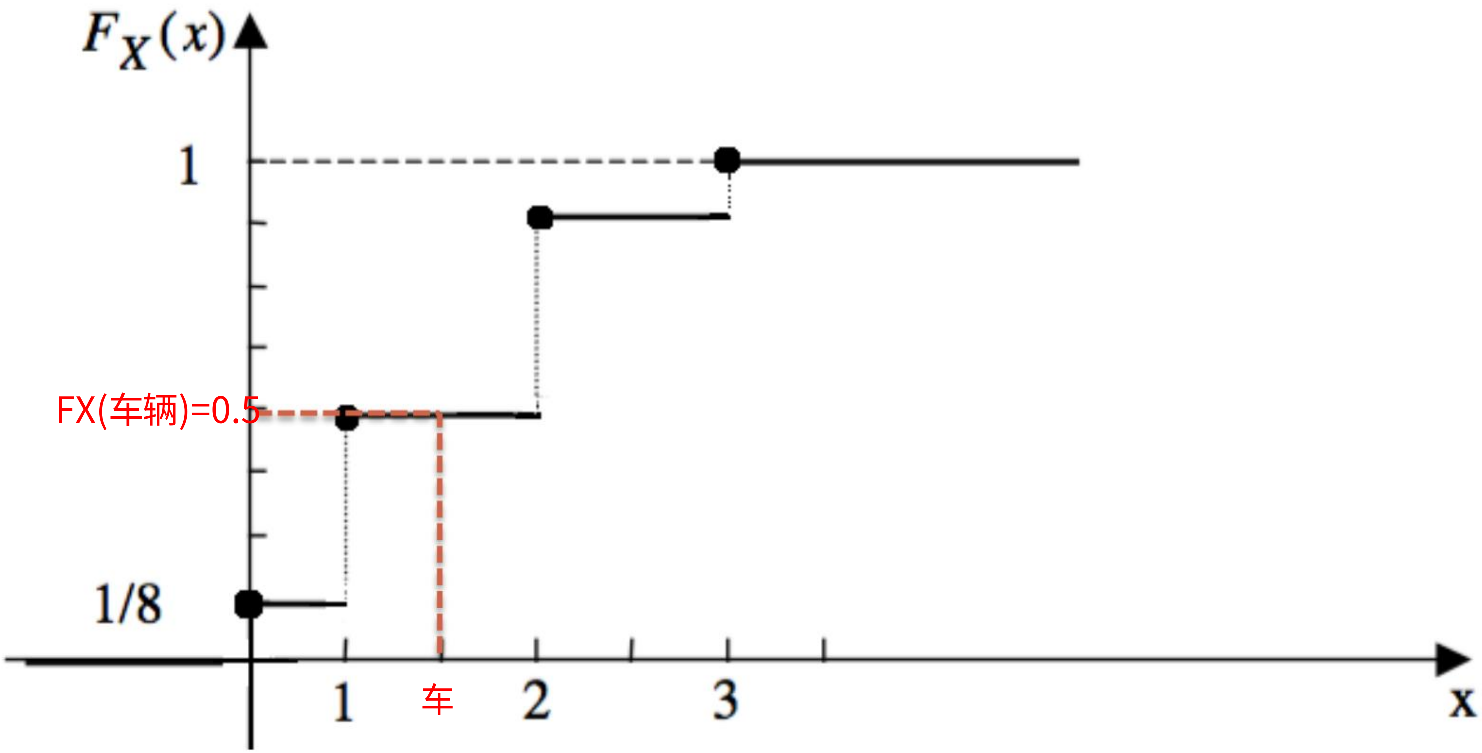
备注:如果概率密度函数是对称的 E(X) = xe

与预期值（均值)相比,中位数的一个有趣特性是,它不会因一小部分极大或极小的值而产生偏差,因此可以更好地表示“典型”值来表征分布。

中位数xe （“中间”值)(2/2)

离散随机变量的中值xe示例。 xe可以使用 CDF 图检索：

X:抛硬币3次后反面的数量



对于此示例， xe可以是区间 1≤ xe <2 中的任何值 按照约定

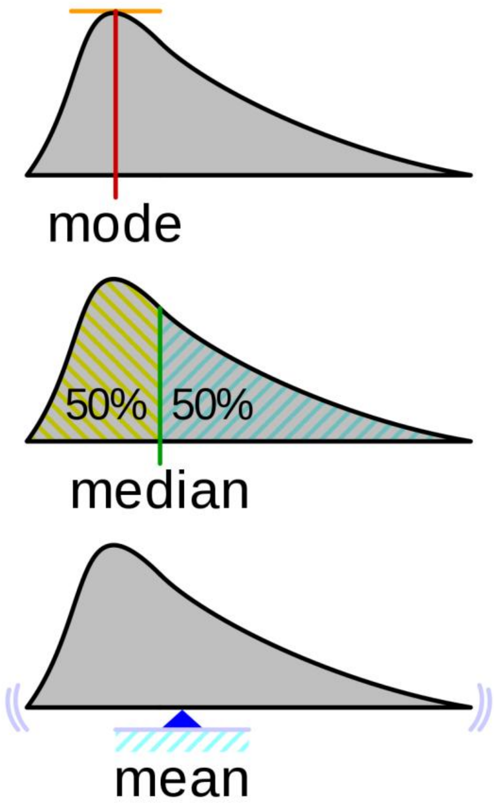
xe=1.5

众数：（最常见的值）

模式是xmode 的值,其中
概率质量函数取其
最大值。

Ø对于连续cas,它是
密度曲线中最大值的横坐标。

Ø对于离散情况,为xi值
哪个概率最大。



任意数据的众数、中位数和
平均值的比较
概率密度函数。

来源:维基百科

Pth百分位数或 P% 百分位数

百分位 P%,也称为 centile,是不被超过的概率为 P/100 的xp值：

$$F(x_p) = \int_{-\infty}^{x_p} f_X(x) \cdot dx = \frac{P}{100}$$

四分位数:百分位数 25 % ，50% 和 75%

随机变量的矩:定义

o 矩:期望值的概括

随机变量的n阶矩是函数g (X) = Xn的期望值,记为mn

对于连续随机变量

$$\boxed{m_n = \int_{-\infty}^{+\infty} x^n \cdot f_X(x) \cdot dx}$$

对于离散随机变量

$$\boxed{m_n = \sum_i x_i^n . P(X = x_i)}$$

尤其: m0 = 1 且m1 = E(X)

第二中心矩也称为“方差”

二阶矩： m2

E(X2) （在物理学中很重要）

$$m_2 = \int_{-\infty}^{+\infty} x^2 f_X(x) dx$$ 警告： m2不能为负数

方差 :第二中心矩 Var(X) 或σ2它允许测量随机变量所取值与预期值 x 的分散程度。它是分散度的度量。

$$Var(X) = \sigma_X^2 = E(X - m_1)^2$$

对于连续随机变量

$$\sigma_X^2 = \int_{-\infty}^{+\infty} (x - m_1)^2 \cdot f_X(x) \cdot dx$$

对于离散随机变量

$$\sigma_X^2 = \sum_i (x_i - m_1)^2 \cdot P_i$$

重要特性

期望值：

- 常数 b 的期望值等于该常数E(b) = b
- 线性：期望值是线性运算符。对于任意两个随机变量 X 和 Y，对于任意两个实数 a 和 b：

$$E\ (aX + bY) = aE(X) + b\ E(Y)$$

- 产品：一般来说，E(XY) ≠ E(X) E(Y)。然而，如果 X 和 Y 独立，则等式成立。

方差：

$$\sigma_X^2 = Var(X) = E(X^2) - E^2(X) = m_2 - m_1^2$$

如果c是常数：

( ) = ” ( )

和

(± =) ( )

标准差

$$\sigma_X = \sqrt{E(X^2) - E^2(X)} > 0$$

锻炼

给定随机变量X的期望值m1和标准差

χ 计算Y的期望值和方差。

$$Y = \frac{X - m_1}{\sigma_X}$$

比奈梅-切比雪夫定理 (1/2)：

（也称为不平等）

令X为具有有限期望值 m 的随机变量,并且
有限非零方差σ2。那么对于任意实数 我 > 0,

$$P\{|X - m_1| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2} \qquad \varepsilon > \sigma \geq 0$$

解释:随机变量的方差越小,其偏离均值超过的概率就越小

e 单位。

Ø该定理强调了期望值和方差在随机变量描述中的作用。

比奈梅-切比雪夫定理 (2/2)：

另一个版本：

$$\varepsilon = n\sigma \quad \Rightarrow \quad P\left(\left\{\left|X - m_1\right| \geq n\sigma\right\}\right) \leq \frac{1}{n^2} \qquad (n > 1)$$

解释:该定理保证不超过1/n2的分布值与均值相差n 个或更多标准差（或者等效地,超过 1 — 1/n2的分布值与均值相差小于n 个标准差）。

示例:如果 n=10,对于任何随机变量 X,其偏离均值超过 10σ 的概率小于 1%。

概括：

非中心矩：

$$m_n = \int_{-\infty}^{+\infty} x^n \cdot f_X(x) \cdot dx$$

中心时刻：

$$\mu_n = E\left(\left(X - m_1\right)^n\right) = \int_{-\infty}^{+\infty} \left(x - m_1\right)^n \cdot f_X(x) \cdot dx$$

任意顺序的时刻之间的一些属性：

（使用算子 E(.) 的线性证明）

$$C_n^r = \frac{n!}{r! \cdot (n - r)!}$$

$$\mu_n = \sum_{r=0}^{n} C_n^r \cdot (-1)^r \cdot m_1^r \cdot m_{n-r}$$

$$m_n = \sum_{r=0}^{n} C_n^r \cdot m_1^r \cdot \mu_{n-r}$$

概率论

第五章 :特征功能 实数随机变量

- 定义
- 相关定理

特征函数的定义 characteristic function

实数随机变量 X 的特征函数是 t 的复值函数,定义为:

$$\begin{cases} R & \to & C \\ t & \mapsto & \varphi_X(t) = E\{e^{jtX}\} \end{cases}$$

其中 t ∈ 是确定性参数, j2 = –1 是虚数单位。

对于密度为fX(x)的连续随机变量 X:

$$\varphi_X(t) = \int_{-\infty}^{+\infty} e^{jtx} f_X(x) dx$$

对于离散随机变量 X:

$$\varphi_X(t) = \sum_i P_i \cdot e^{j \cdot t \cdot x_i}$$

和 $P_i = P(X = x_i)$

备注: Ø

特征函数是概率密度函数的傅里叶变换。 Ø 特征函数完全定义了概率分布。

示例 1:令 X 为参数p具有伯努利分布的随机变量 X。和

$$X = \begin{cases} 1 & P\{X = 1\} = p \\ 0 \quad \text{和} & P\{X = 0\} = q \end{cases} \quad \text{et} \quad p + q = 1$$

$$\varphi_X(t) = \sum_{i=0}^{1} P_i e^{jtx_i} = qe^{jt0} + pe^{jt1}$$

$$\varphi_X(t) = pe^{jt} + q$$

示例 2:设 X 为指数随机变量:

$$f_X(x) = \alpha e^{-\alpha x}, \ \alpha > 0, \ x \geq 0$$

$$\varphi_X(t) = \int_0^{+\infty} \alpha \, e^{-\alpha x} e^{jtx} dx$$

$$\varphi_X(t) = \varphi_{\exp}(t) = \frac{\alpha}{\alpha - jt}$$

力矩生成函数：

如果随机变量X的矩达到n 阶,则特征函数φX在整条实线上连续可导 n 次,我们可以用它来求矩

$$\boxed{\left.\frac{d^n\varphi_X(t)}{dt^n}\right|_{t=0} = j^n m_n}$$

示例:E(X) 和 Var(X) 的计算

二项分布：

$$P\{X = k\} = C_n^k p^k q^{n-k} \qquad 0 \le X \le n$$

其中p + q = 1

特征函数为:

$$\varphi_X(t) = \left(pe^{jt} + q\right)^n$$

一阶和二阶导数为:

$$\varphi_X'(t) = n\left(pe^{jt} + q\right)^{n-1}.jpe^{jt}$$

$$\varphi_X''(t) = jnp\left[(n-1)\left(pe^{jt} + q\right)^{n-2}.jpe^{2jt} + \left(pe^{jt} + q\right)^{n-1} je^{jt}\right]$$

在点 t = 0

$$\boxed{\varphi_X'(0) = jnp = jm_1 \Rightarrow m_1 = np}$$

$$\varphi_X''(0) = j^2\left[n^2p^2 + npq\right] = j^2 m_2 \qquad m_2 = n^2p^2 + npq$$

$$\boxed{Var\{X\} = \sigma_X^2 = m_2 - m_1^2 = npq}$$

定理： 反演公式：

傅立叶反演定理

如果特征函数φX可积,则FX绝对连续,因此X具有概率密度函数。

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \varphi_X(t) e^{-jxt} dt$$

累积分布函数和特征函数之间存在一一对应的关系。