
第二章随机变量及其分布

2.2	连续型随机变量	1
2.2.1	正态分布	11
2.2.2	指数分布	16
2.2.3	均匀分布	21

2.2 连续型随机变量

离散随机变量只取有限个或可数无限个值，而连续型随机变量取不可数个值. 这就决定了不能用描述离散型随机变量的办法来刻画连续型随机变量.

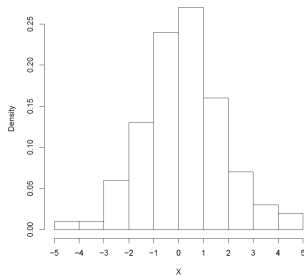
考虑一个例子. 假定步枪射手瞄准靶子在固定的位置进行一系列的射击. 令 X 是命中点与过靶心垂线的水平偏离值, 设 X 取值 $[-5\text{cm}, 5\text{cm}]$. X 是一个连续随机变量.

为了计算 X 落在某区间的概率, 将 $[-5, 5]$ 分为长为 1 厘米的小区间. 对于每个小区间, 以落在这个小区间的弹孔数除以弹孔总数得到落在这个区间的弹孔的相对频数. 设总弹孔数为 100. 我们得到下表:

区间	弹孔数	相对频数
$[-5, -4]$	1	0.01
$[-4, -3]$	1	0.01
$[-3, -2]$	6	0.06
$[-2, -1]$	13	0.13
$[-1, 0]$	24	0.24
$[0, 1]$	27	0.27
$[1, 2]$	16	0.16
$[2, 3]$	7	0.07
$[3, 4]$	3	0.03
$[4, 5]$	2	0.02

上表可以用下图来表示:

图 2.1: 弹孔位点分布图



我们注意每个矩形的底等于 1, 高为该矩形的区间所对应的相对频数, 所以面积为相对频数. 全部矩形的面积是 1. 对于 $[-5, 5]$ 的任一子区间, 我们可以根据上图估计弹孔落在该子区间的概率. 例如要估计 $0 < X \leq 2$ 的概率, 只要把区间中的两个矩形面积加起来, 结

果得到 0.43. 再譬如说要估计 $-0.25 < X \leq 1.5$ 中的概率, 我们应当计算该区间上的面积, 结果得到:

$$0.06 + 0.27 + 0.08 = 0.41.$$

如果第二批的 100 颗子弹射在靶子上, 我们就将获得另一个经验分布. 它与第一个经验分布多半是不同的, 尽管它们的外表可能相似. 如果把观察到的相对频数看作为某一“真”概率的估计, 则我们假定有一个函数, 它将给出任何区间中的精确概率. 这些概率由曲线下的面积给出. 由此我们得到如下定义:

X 称为连续型随机变量, 如果存在一个函数 f , 叫做 X 的概率密度函数, 它满足下面的条件:

1. 对所有的 $-\infty < x < +\infty$, 有 $f(x) \geq 0$;
2. $\int_{-\infty}^{+\infty} f(x)dx = 1$;
3. 对于任意的 $-\infty < a \leq b < +\infty$, 有 $P(a \leq X \leq b) = \int_a^b f(x)dx$.

Definition

注 1. 对于任意的 $-\infty < x < +\infty$, 有 $P(X = x) = \int_x^x f(u)du = 0$.

注 2. 如果 f 只取某有限区间 $[a, b]$ 的值, 令

$$\tilde{f}(x) = \begin{cases} f(x) & x \in [a, b], \\ 0 & \text{其它.} \end{cases}$$

则 \tilde{f} 是定义在 $(-\infty, +\infty)$ 上的密度函数, 且 $f(x)$ 和 $\tilde{f}(x)$ 给出相同的概率分布.

注 3. 假设有总共一个单位的质量连续地分布在 $a \leq x \leq b$ 上. 那么 $f(x)$ 表示在点 x 的质量密度且 $\int_c^d f(x)dx$ 表示在区间 $[c, d]$ 上的全部质量.

由于连续随机变量的概率是用积分给出的, 我们可以直接处理密度的积分而不是密度本身.

设 X 为一连续型随机变量. 则

$$F(x) = \int_{-\infty}^x f(u)du, \quad -\infty < x < +\infty \quad (2.1)$$

Definition

称为 X 的 (累积) 分布函数.

注 4. $F(x)$ 表示的是随机变量的数值小于或等于 x 的概率, 即

$$F(x) = P(X \leq x) \quad -\infty < x < +\infty. \quad (2.2)$$

由式 (2.2) 定义的 F 为 X 的 (累积) 分布函数的一般定义. 它适用于任意的随机变量. 设 X 为一离散型随机变量, 它以概率 $\{p_1, \dots, p_n, \dots\}$ 取值 $\{a_1, \dots, a_n, \dots\}$. 则

$$F(x) = \sum_{a_i \leq x} p_i.$$

分布函数 F 具有下列性质:

(1) F 是非减的函数;

对任何 $x_1 < x_2$ 都有, $F(x_2) - F(x_1) = P(x_1 < X \leq x_2) \geq 0$

(2) $0 \leq F(x) \leq 1, x \in R$, 且 $\lim_{x \rightarrow -\infty} F(x) = 0; \lim_{x \rightarrow +\infty} F(x) = 1$.

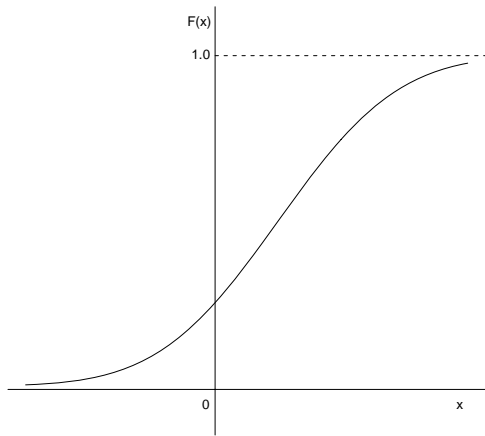
(3) $F(x)$ 右连续;

$$F(x) = P(X \leq x) = \sum_{i: a_i \leq x} P(X = a_i) = \sum_{i: a_i \leq x} p_i$$

对于连续随机变量, 如果 $F(x)$ 在点 x 的导数存在, 则

$$f(x) = F'(x).$$

连续随机变量的分布函数的图象如下图所示.



下面我们介绍常见的连续型分布. 它们包括正态分布, 指数分布和均匀分布.

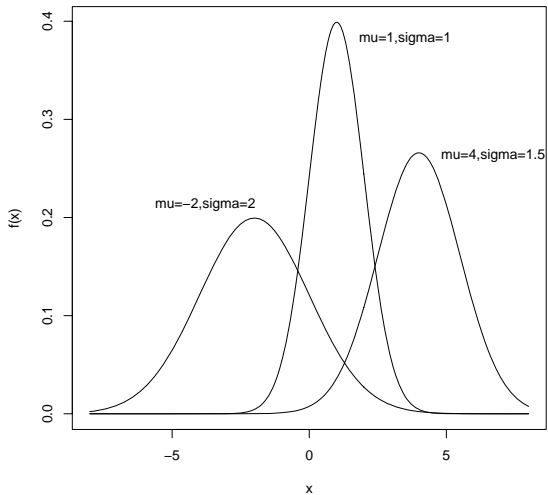
2.2.1 正态分布

如果一个随机变量 X 具有概率密度函数

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < +\infty \quad (2.3)$$

其中 $-\infty < \mu < +\infty$, $\sigma^2 > 0$, 则称 X 为一正态随机变量, 记为 $X \sim N(\mu, \sigma^2)$. 以 (2.3) 为密度的分布称为参数为 μ 和 σ^2 的正态分布.

具有参数 $\mu = 0$, $\sigma = 1$ 的正态分布称为标准正态分布. 用 $\Phi(x)$ 和 $\phi(x)$ 表示标准正态分布 $N(0, 1)$ 的分布函数和密度函数.



in R

```
x <- seq(-4, 4, length = 401)
plot(x, dnorm(x), type = 'l') # N(0, 1)
N(1, 1.52):
lines(x, dnorm(x, mean = 1, sd = 1.5), lty = 'dashed')
```

↑Code

↓Code

从上图可以看出, 正态分布的密度函数是以 $x = \mu$ 为对称轴的对称函数. μ 称为位置参数. 密度函数在 $x = \mu$ 处达到最大值, 在 $(-\infty, \mu)$ 和 $(\mu, +\infty)$ 内严格单调. 同时我们看到, σ 的大小决定了密度函数的陡峭程度. 通常称 σ 为正态分布的形状参数.

以 $F(x)$ 记正态分布 $N(\mu, \sigma^2)$ 的概率分布函数, 则恒有 $F(x) = \Phi(\frac{x-\mu}{\sigma})$. 所以任一正态分布的概率分布函数都可通过标准正态分布的分布函数计算出来.

设 $X \sim N(\mu, \sigma^2)$, 则 $\frac{X-\mu}{\sigma} \sim N(0, 1)$.

↑Example

↓Example

证: 由

$$\begin{aligned} P\left(\frac{X-\mu}{\sigma} \leq x\right) &= P(X \leq \sigma x + \mu) \\ &= \int_{-\infty}^{\sigma x + \mu} \frac{1}{\sqrt{2\pi}\sigma} e^{-(t-\mu)^2/(2\sigma^2)} dt \\ &\stackrel{t=\sigma z + \mu}{=} \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \end{aligned}$$

求数 k 使得对于正态分布的变量有 $P(\mu - k\sigma < x < \mu + k\sigma) = 0.95$.

↑Example

↓Example

解:

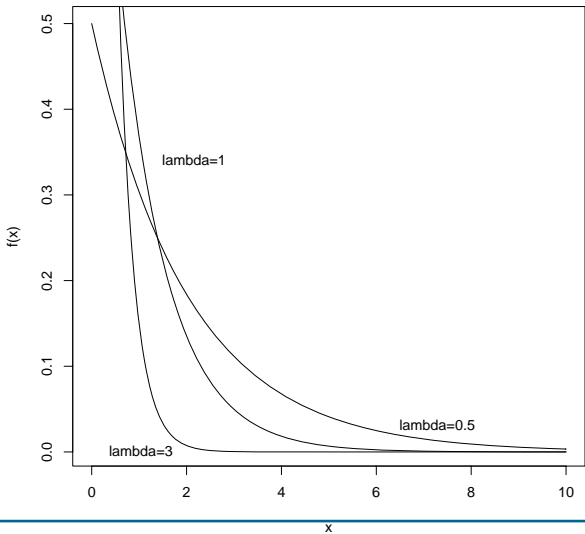
2.2.2 指数分布

若随机变量 X 具有概率密度函数

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0, \\ 0 & x \leq 0, \end{cases} \quad (2.5)$$

其中 $\lambda > 0$ 为常数, 则称 X 服从参数为 λ 的指数分布.
指数分布的分布函数为

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x > 0, \\ 0 & x \leq 0. \end{cases} \quad (2.6)$$



in R

dexp, rexp, pexp, qexp

↑Code

↓Code

从图 (2.5) 可以看出, 参数 λ 愈大, 密度函数下降得愈快.

指数分布经常用于作为各种“寿命”的分布的近似. 令 X 表示某元件的寿命. 我们引进 X 的失效率函数如下:

$$h(x) = \lim_{\Delta x \rightarrow 0} \frac{P(x \leq X \leq x + \Delta x | X > x)}{\Delta x}.$$

失效率表示了元件在时刻 x 尚能正常工作, 在时刻 x 以后, 单位时间内发生失效的概率. 则如果

$$h(x) \equiv \lambda \quad (\text{常数}), \quad 0 < x < +\infty,$$

X 服从指数分布. 即指数分布描述了无老化时的寿命分布.

设 X 表示某种电子元件的寿命, $F(x)$ 为其分布函数。若假设元件无老化, 即元件在时刻 x 正常工作的条件下, 其失效率保持为某个常数 λ , 与 x 无关。试证明 X 服从指数分布。

↑Example

↓Example

解:

从而结论得证。

指数分布的最重要的特点是“无记忆性”。即若 X 服从指数分布，则对任意的 $s, t > 0$ 有

$$P(X > s + t \mid X > s) = P(X > t). \quad (2.7)$$

即寿命是无老化的。可以证明，指数分布是唯一具有性质 (2.7) 的连续型分布。

2.2.3 均匀分布

设 $-\infty < a < b < +\infty$, 如果随机变量 X 具有密度函数

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & \text{其它,} \end{cases} \quad (2.8)$$

则称该随机变量为区间 $[a, b]$ 上的均匀分布, 记作 $U[a, b]$. 如此定义的 $f(x)$ 显然是一个概率密度函数. 容易算出其相应的分布函数为

$$F(x) = \begin{cases} 0, & x \leq a, \\ \frac{x-a}{b-a}, & a < x \leq b, \\ 1, & x > b. \end{cases}$$

in R

dunif, runif, punif, qunif

↑Code

↓Code

在计算时因四舍五入而产生的误差可以用均匀分布来描述.