连续型随机变量 1

续本达

清华大学 工程物理系

2023-10-07 清华

复习

随机变量

定义 (随机变量)

设随机试验的样本空间为 $S = \{e\}$.
$X = X(e)$ 是定义在样本空间 S 上的实单值函数，称 $X = X(e)$ 为 随机变量 (random variable, 略写为 r.v.)。

定义 (分布函数)

设 X 为随机变量, x 是任意实数, 称函数

$$F(x) = P(X \le x), \quad -\infty < x < +\infty$$

为 X 的分布函数，也称为累积分布函数 (cumulative distribution function).

常见的离散型随机变量分布律

伯努利试验 0-1 分布、二项分布、泊松分布、超几何、几何分布

随机变量

定义 (随机变量)

设随机试验的样本空间为 $S = \{e\}$.
$X = X(e)$ 是定义在样本空间 S 上的实单值函数，称 $X = X(e)$ 为 随机变量 (random variable, 略写为 r.v.)。

定义 (分布函数)

设 X 为随机变量, x 是任意实数, 称函数

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

为 X 的分布函数，也称为累积分布函数 (cumulative distribution function).

常见的离散型随机变量分布律

伯努利试验 0-1 分布、二项分布、泊松分布、超几何、几何分布

随机变量

定义 (随机变量)

设随机试验的样本空间为 $S = \{e\}$.
$X = X(e)$ 是定义在样本空间 S 上的实单值函数，称 $X = X(e)$ 为 随机变量 (random variable, 略写为 r.v.)。

定义 (分布函数)

设 X 为随机变量, x 是任意实数, 称函数

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

为 X 的分布函数，也称为累积分布函数 (cumulative distribution function).

常见的离散型随机变量分布律

伯努利试验 0-1 分布、二项分布、泊松分布、超几何、几何分布

连续型随机变量

定义

设 X 是随机变量, 若存在一个非负可积函数 $f(x)$，使得

$$F(x)=\int_{-\infty}^{x} f(t)\mathrm{d}t, -\infty<x<+\infty$$

其中 $F(x)$ 是它的分布函数，亦称累积分布函数，则称 X 是连续型随机变量。称 $f(x)$ 为它的 概率密度函数 ，简称概率密度或密度函数。

$$\begin{aligned} F'(x_0) &= \lim_{\Delta x\to 0^+} \frac{F(x_0+\Delta x)-F(x_0)}{\Delta x} \\ &= \lim_{\Delta x\to 0^+} \frac{P(x_0<X\le x_0+\Delta x)}{\Delta x} = f(x_0) \\ \Longrightarrow f(x_0)\Delta x &\approx P(x_0<X\le x_0+\Delta x) \end{aligned}$$

定义

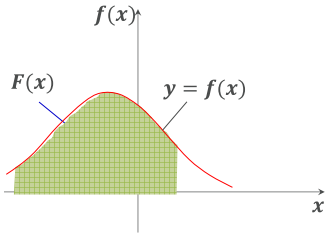
设 X 是随机变量, 若存在一个非负可积函数 $f(x)$，使得

$$F(x)=\int_{-\infty}^{x} f(t)\mathrm{d}t, -\infty<x<+\infty$$

其中 $F(x)$ 是它的分布函数，亦称累积分布函数，则称 X 是连续型随机变量。称 $f(x)$ 为它的 概率密度函数 ，简称概率密度或密度函数。

$$\begin{aligned} F'(x_0) &= \lim_{\Delta x\to 0^+} \frac{F(x_0+\Delta x)-F(x_0)}{\Delta x} \\ &= \lim_{\Delta x\to 0^+} \frac{P(x_0<X\le x_0+\Delta x)}{\Delta x} = f(x_0) \\ \Longrightarrow f(x_0)\Delta x &\approx P(x_0<X\le x_0+\Delta x) \end{aligned}$$

概率密度函数



检验一个函数能否作为连续性随机变量的概率密度函数：

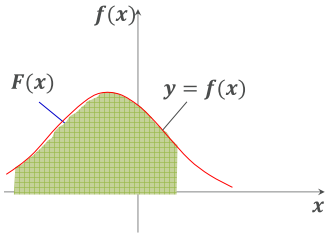
- 在 $f(x)$ 的连续点处，$f(x)=F'(x)$
  $f(x)$ 描述了 X 在 x 附近单位长度的区间内取值的概率。
- 对于任意实数 $a,b(a\le b)$

$$
\begin{aligned}
&P(a<X\le b)=P(a\le X\le b)\\
=&P(a<X<b)=P(a\le X<b)\\
=&\int_a^b f(x)\mathrm{d}x=F(b)-F(a)
\end{aligned}
$$

$$f(x)\ge 0$$

$$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x=1$$

概率密度函数



检验一个函数能否作为连续性随机变量的概率密度函数：

- 在 $f(x)$ 的连续点处，$f(x)=F'(x)$
  $f(x)$ 描述了 X 在 x 附近单位长度的区间内取值的概率。
- 对于任意实数 $a,b(a\le b)$

$$
\begin{aligned}
&P(a<X\le b)=P(a\le X\le b)\\
=&P(a<X<b)=P(a\le X<b)\\
=&\int_a^b f(x)\mathrm{d}x=F(b)-F(a)
\end{aligned}
$$

$$
\begin{aligned}
f(x)&\ge 0\\
\int_{-\infty}^{+\infty}f(x)\mathrm{d}x&=1
\end{aligned}
$$

连续型随机变量取任一常数的概率为零

对于连续型随机变量 X，$P(X=a)=0$，其中 a 是随机变量 X 的一个可能的取值。

$$\{X=a\} \subset (a-\Delta x < X \leq a), \Delta x > 0$$

$$\Longrightarrow 0 \leq P(X=a) \leq P(a-\Delta x < X \leq a) = \int_{a-\Delta x}^{a} f(x)\mathrm{d}x$$

$$\Longrightarrow 0 \leq P(X=a) \leq \lim_{\Delta x \to 0^+} \int_{a-\Delta x}^{a} f(x)\mathrm{d}x = 0$$

$$\Longrightarrow P(X=a) = 0$$

连续型随机变量取任一常数的概率为零

对于连续型随机变量 X，$P(X=a)=0$，其中 a 是随机变量 X 的一个可能的取值。

$$\{X=a\} \subset (a-\Delta x < X \leq a), \Delta x > 0$$

$$\Longrightarrow 0 \leq P(X=a) \leq P(a-\Delta x < X \leq a) = \int_{a-\Delta x}^{a} f(x)\mathrm{d}x$$

$$\Longrightarrow 0 \leq P(X=a) \leq \lim_{\Delta x \to 0^+} \int_{a-\Delta x}^{a} f(x)\mathrm{d}x = 0$$

$$\Longrightarrow P(X=a) = 0$$

连续型随机变量取任一常数的概率为零

对于连续型随机变量 X，$P(X=a)=0$，其中 a 是随机变量 X 的一个可能的取值。

$$\{X=a\} \subset (a-\Delta x < X \leq a), \Delta x > 0$$

$$\Longrightarrow 0 \leq P(X=a) \leq P(a-\Delta x < X \leq a) = \int_{a-\Delta x}^{a} f(x)\mathrm{d}x$$

$$\Longrightarrow 0 \leq P(X=a) \leq \lim_{\Delta x \to 0^+} \int_{a-\Delta x}^{a} f(x)\mathrm{d}x = 0$$

$$\Longrightarrow P(X=a) = 0$$

常见的连续性随机变量分布

(1) 均匀分布 $f(x;a,b)=\dfrac{1}{b-a}, \quad a<x<b$

(2) 指数分布 $f(x;\lambda)=\lambda e^{-\lambda x}, \quad x\geq 0$

(3) 正态分布 $f(x;\mu,\sigma^2)=\dfrac{1}{\sqrt{2\pi}\sigma}e^{\frac{(x-\mu)^2}{2\sigma^2}}, -\infty<x<\infty$

(4) 伽玛分布 $f(x;\alpha,\lambda)=\dfrac{\lambda^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}, \quad x\geq 0$

(5) 卡方分布 $f(x;n)=\dfrac{x^{n/2-1}}{2^{n/2}\Gamma(n/2)}e^{-x/2}, \quad x\geq 0$

(6) 贝塔分布 $f(x;a,b)=\dfrac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}, 0<x<1$

(7) 柯西分布 $f(x)=\dfrac{1}{\pi}\dfrac{1}{1+x^2} \quad -\infty<x<\infty$

(8) 朗道分布 $f(x;\beta)=\frac{1}{\xi}\phi(\lambda)$：没有简单的解析表达式

均匀分布

定义

若 X 的概率密度为

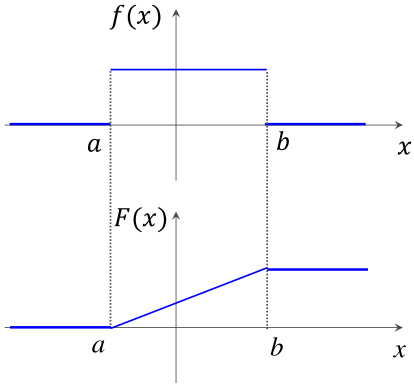
$$f(x)=\begin{cases}\frac{1}{b-a}, & a<x<b\\ 0, & \text{其他}\end{cases}$$

则称 X 服从区间 (a,b) 上的均匀分布，或称 X 服从参数为 a,b 的 均匀分布 。 记作

$$X\sim U(a,b)$$

分布函数为

$$F(x)=\int_{-\infty}^{x}f(t)\mathrm{d}t=\begin{cases}0, & x<a\\ \frac{x-a}{b-a}, & a\le x<b\\ 1, & x\ge b\end{cases}$$



定义

若 X 的概率密度为

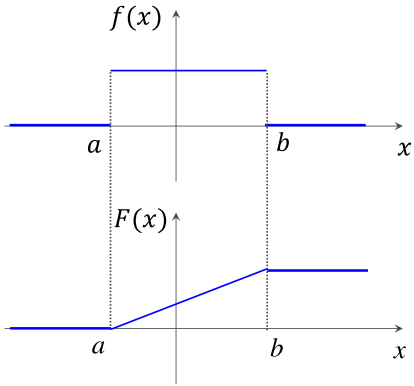
$$f(x)=\begin{cases}\frac{1}{b-a}, & a<x<b\\ 0, & \text{其他}\end{cases}$$

则称 X 服从区间 (a,b) 上的均匀分布，或称 X 服从参数为 a,b 的 均匀分布 。 记作

$$X\sim U(a,b)$$

分布函数为

$$F(x)=\int_{-\infty}^{x}f(t)\mathrm{d}t=\begin{cases}0, & x<a\\ \frac{x-a}{b-a}, & a\le x<b\\ 1, & x\ge b\end{cases}$$



$b \to \infty$ 怎么办?

不失一般地取 $a = 0$，存在

$$X \sim U(0, +\infty)$$

吗?

不存在，$F(x)$ 和 $f(x)$ 都无法定义。

本质问题

无法线性地通向无穷。

- 《庄子》启示：一尺之棰，日取其半，万世不竭

$b \to \infty$ 怎么办?

不失一般地取 $a = 0$，存在

$$X \sim U(0, +\infty)$$

吗?

不存在，$F(x)$ 和 $f(x)$ 都无法定义。

本质问题

无法线性地通向无穷。

- 《庄子》启示：一尺之棰，日取其半，万世不竭

$b \to \infty$ 怎么办?

不失一般地取 $a = 0$，存在

$$X \sim U(0, +\infty)$$

吗?
不存在，$F(x)$ 和 $f(x)$ 都无法定义。

本质问题

无法线性地通向无穷。

- 《庄子》启示：一尺之棰，日取其半，万世不竭

指数分布

定义

若 X 的概率密度为 ($\lambda > 0$ 为常数)

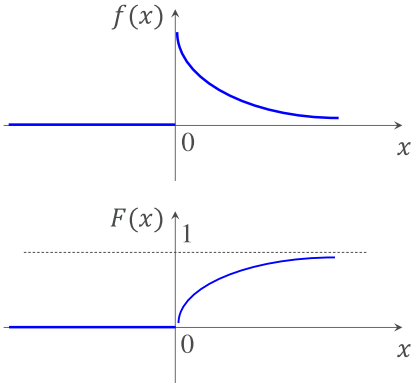
$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \le 0 \end{cases}$$

则称 X 服从参数为 λ 的 指数分布 ，记作

$$X \sim \mathrm{Exp}(\lambda)$$

分布函数为

$$F(x) = \begin{cases} 1-e^{-\lambda x}, & x \ge 0 \\ 0, & x < 0 \end{cases}$$



定义

若 X 的概率密度为 ($\lambda > 0$ 为常数)

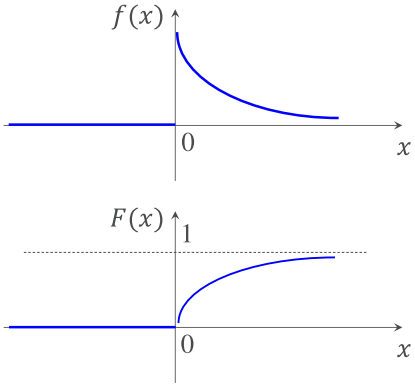
$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \le 0 \end{cases}$$

则称 X 服从参数为 λ 的 指数分布 ，记作

$$X \sim \mathrm{Exp}(\lambda)$$

分布函数为

$$F(x) = \begin{cases} 1-e^{-\lambda x}, & x \ge 0 \\ 0, & x < 0 \end{cases}$$



性质

对于任意的 $0 < a < b$,

$$\begin{aligned} P(a < X < b) &= \int_a^b \lambda e^{-\lambda x} \mathrm{d}x \\ &= F(b) - F(a) \\ &= e^{-\lambda a} - e^{-\lambda b} \end{aligned}$$

应用场合：常作为各种“寿命”分布的近似

- 随机服务系统中的服务时间
- 电话问题中的通话时间
- 不稳定粒子的寿命
- 无线电元件的寿命
- 动物的寿命

性质

对于任意的 $0 < a < b$,

$$\begin{aligned} P(a < X < b) &= \int_a^b \lambda e^{-\lambda x} \mathrm{d}x \\ &= F(b) - F(a) \\ &= e^{-\lambda a} - e^{-\lambda b} \end{aligned}$$

应用场合：常作为各种“寿命”分布的近似

- 随机服务系统中的服务时间
- 电话问题中的通话时间
- 不稳定粒子的寿命
- 无线电元件的寿命
- 动物的寿命

例：指数与泊松

例

假定一大型设备在任何长为 t 的时间内发生故障的次数 $N(t) \sim \pi(\lambda t)$，求第一次发生故障的时间 T 的分布。

$$F_T(t) = P(T \le t)$$

$$= \begin{cases} 0, & t < 0 \\ 1 - P(T > t), & t \ge 0 \end{cases}$$

$$\Longrightarrow P(T > t) = P(N(t) = 0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t}$$

$$\Longrightarrow F_T(t) = \begin{cases} 1 - e^{-\lambda t}, & t \ge 0 \\ 0, & x < 0 \end{cases}$$

所以

$$T \sim \mathrm{Exp}(\lambda)$$

例：指数与泊松

例

假定一大型设备在任何长为 t 的时间内发生故障的次数 $N(t) \sim \pi(\lambda t)$，求第一次发生故障的时间 T 的分布。

$$
\begin{aligned}
F_T(t) =& P(T \le t) \\
=& \begin{cases} 0, & t < 0 \\ 1 - P(T > t), & t \ge 0 \end{cases}
\end{aligned}
$$

$$
\Longrightarrow P(T > t) = P(N(t) = 0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t}
$$

$$
\Longrightarrow F_T(t) = \begin{cases} 1 - e^{-\lambda t}, & t \ge 0 \\ 0, & x < 0 \end{cases}
$$

所以

$$
T \sim \mathrm{Exp}(\lambda)
$$

例：指数与泊松

例

假定一大型设备在任何长为 t 的时间内发生故障的次数 $N(t) \sim \pi(\lambda t)$，求第一次发生故障的时间 T 的分布。

$$
\begin{aligned}
F_T(t) =& P(T \le t) \\
=& \begin{cases} 0, & t < 0 \\ 1-P(T > t), & t \ge 0 \end{cases} \\
\Longrightarrow & P(T > t) = P(N(t) = 0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t} \\
\Longrightarrow & F_T(t) = \begin{cases} 1-e^{-\lambda t}, & t \ge 0 \\ 0, & x < 0 \end{cases}
\end{aligned}
$$

所以

$$T \sim \mathrm{Exp}(\lambda)$$

例：指数与泊松

例

假定一大型设备在任何长为 t 的时间内发生故障的次数 $N(t) \sim \pi(\lambda t)$，求第一次发生故障的时间 T 的分布。

$$
\begin{aligned}
F_T(t) =& P(T \le t) \\
=& \begin{cases} 0, & t < 0 \\ 1 - P(T > t), & t \ge 0 \end{cases} \\
\Longrightarrow & P(T > t) = P(N(t) = 0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t} \\
\Longrightarrow & F_T(t) = \begin{cases} 1 - e^{-\lambda t}, & t \ge 0 \\ 0, & x < 0 \end{cases}
\end{aligned}
$$

所以

$$T \sim \mathrm{Exp}(\lambda)$$

无记忆性

命题

若 $X \sim \mathrm{Exp}(\lambda)$，则 $P(X > s+t | X > s) = P(X > t)$

$$P(X > s+t|X > s) = \frac{P(X > s+t)}{P(X > s)} = \frac{1-P(X \le s+t)}{1-P(X \le s)}$$
$$= \frac{1-F(s+t)}{1-F(s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}}$$
$$= e^{-\lambda t} = P(X > t)$$

故又把指数分布称为“永远年轻”的分布

无记忆性

命题

若 $X \sim \mathrm{Exp}(\lambda)$，则 $P(X > s+t | X > s) = P(X > t)$

$$P(X > s+t|X > s) = \frac{P(X > s+t)}{P(X > s)} = \frac{1-P(X \le s+t)}{1-P(X \le s)}$$

$$= \frac{1-F(s+t)}{1-F(s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}}$$

$$= e^{-\lambda t} = P(X > t)$$

故又把指数分布称为“永远年轻”的分布

无记忆性

命题

若 $X \sim \mathrm{Exp}(\lambda)$，则 $P(X > s+t | X > s) = P(X > t)$

$$
\begin{aligned}
P(X > s+t | X > s) &= \frac{P(X > s+t)}{P(X > s)} = \frac{1 - P(X \le s+t)}{1 - P(X \le s)} \\
&= \frac{1 - F(s+t)}{1 - F(s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} \\
&= e^{-\lambda t} = P(X > t)
\end{aligned}
$$

故又把指数分布称为“永远年轻”的分布

无记忆性

命题

若 $X \sim \mathrm{Exp}(\lambda)$，则 $P(X > s+t | X > s) = P(X > t)$

$$
\begin{aligned}
P(X > s+t | X > s) &= \frac{P(X > s+t)}{P(X > s)} = \frac{1 - P(X \le s+t)}{1 - P(X \le s)} \\
&= \frac{1 - F(s+t)}{1 - F(s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} \\
&= e^{-\lambda t} = P(X > t)
\end{aligned}
$$

故又把指数分布称为“永远年轻”的分布

例：无记忆性

例

假定一大型设备在任何长为 t 的时间内发生故障的次数 $N(t) \sim \pi(\lambda t)$，求设备已正常运行 8 小时的情况下，再正常运行 10 小时的概率。

由上例，正常运行时间 $T \sim \mathrm{Exp}(\lambda)$。由指数分布的无记忆性，

$$P(T > 18 | T > 8) = P(T > 10) = e^{-10\lambda}.$$

例：无记忆性

例

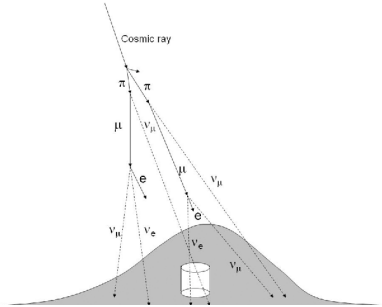
假定一大型设备在任何长为 t 的时间内发生故障的次数 $N(t) \sim \pi(\lambda t)$，求设备已正常运行 8 小时的情况下，再正常运行 10 小时的概率。

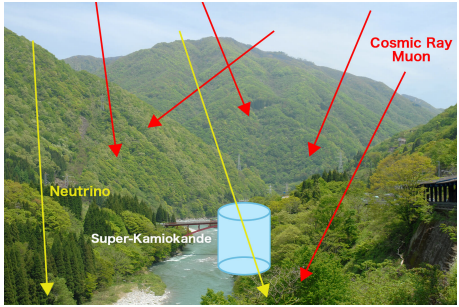
由上例，正常运行时间 $T \sim \mathrm{Exp}(\lambda)$。由指数分布的无记忆性，

$$P(T > 18 | T > 8) = P(T > 10) = e^{-10\lambda}.$$

例：缪子寿命

产生于大气上层的宇宙线 μ 子进入海平面的探测器，其中的一部分在探测器中停止并衰变。进入探测器与衰变的时间差 t 服从指数分布，t 的均值等于 μ 子的平均寿命（近似为 2.2 μs ）。μ 子进入探测器前后存活的时间对于确定平均寿命没有影响。





正态分布

准备 R 语言用于画图

```
library(ggplot2)
options(repr.plot.width=9, repr.plot.height=5, repr.plot.pointsize=24)
theme_set(theme_classic(base_size=20))
```

定义

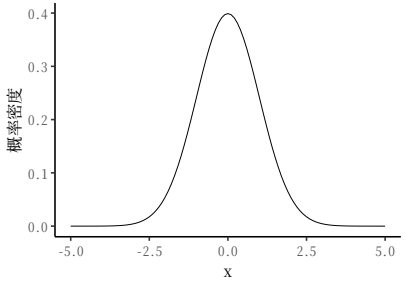
若 X 的概率密度为

$$f(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},-\infty<x<+\infty$$

μ,σ 为常数，$\sigma>0$，则称 X 服从参数为 μ,σ^2 的 正态分布 或 高斯分布 ，记作

$$X\sim N(\mu,\sigma^2)$$

```
x <- seq(-5, 5, by=0.1)
f <- dnorm(x, mean=0, sd=1)
qplot(x, f, ylab="概率密度", geom="line")
```



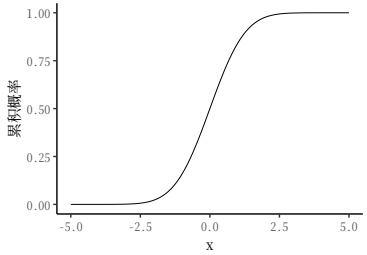
累积分布函数

- 没有初等函数的表示

$$F(x)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}e^{-\frac{(t-\mu)^2}{2\sigma^2}}\mathrm{d}t$$

- $\sigma=1,\mu=0$ 的 $F(x)$，叫作 误差函数 ，记为 $\mathrm{erf}(x)$

```
x <- seq(-5, 5, by=0.1)
f <- pnorm(x, mean=0, sd=1)
qplot(x, f, ylab="累积概率", geom="line")
```



性质

- 图形关于直线 $x=\mu$ 对称, 即

$$f(\mu+x)=f(\mu-x)$$

- 在 $x=\mu$ 时, $f(x)$ 取得最大值 $\frac{1}{\sqrt{2\pi}\sigma}$

应用场景极为广泛

- 各种测量的误差； 人体的生理特征；
- 工厂产品的尺寸； 农作物的收获量；
- 海洋波浪的高度； 金属线抗拉强度；
- 热噪声电流强度； 学生的考试成绩。

根源参考第 b 节，中心极限定理。

性质

- 图形关于直线 $x=\mu$ 对称, 即

$$f(\mu+x)=f(\mu-x)$$

- 在 $x=\mu$ 时, $f(x)$ 取得最大值 $\frac{1}{\sqrt{2\pi}\sigma}$

应用场景极为广泛

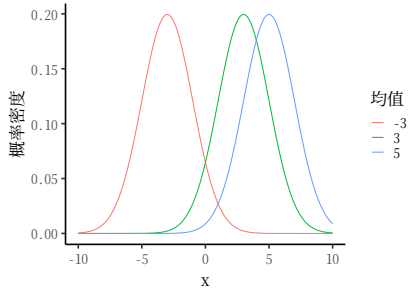
- 各种测量的误差； 人体的生理特征；
- 工厂产品的尺寸； 农作物的收获量；
- 海洋波浪的高度； 金属线抗拉强度；
- 热噪声电流强度； 学生的考试成绩。

根源参考第 b 节，中心极限定理。

位置参数 μ

固定 σ, 改变 μ，曲线沿 x-轴平移。设 $\sigma^2 = 2$

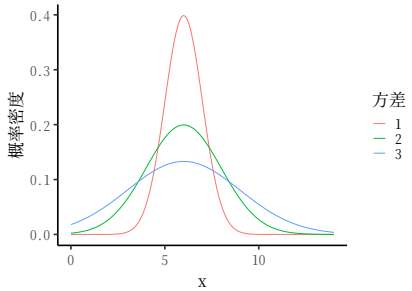
```
x <- seq(-10, 10, by=0.1)
normal_mu <- data.frame(x=c(), pd=c(), mu=c())
for (mu in c(-3, 3, 5)) {
    normal_mu <- rbind(normal_mu,
                       data.frame(x=x, pd=dnorm(x, mean=mu, sd=2), mu=mu))
}
normal_mu$mu <- as.factor(normal_mu$mu)
plot <- ggplot(normal_mu, aes(x=x, y=pd, color=mu)) + geom_line()
print(plot + labs(y="概率密度", color="均值"))
```



形状参数 σ

固定 μ, 缩小 σ，曲线变得越尖，因而 X 落在 μ 附近的概率越大。设 $\mu = 6$

```
x <- seq(0, 14, by=0.1)
normal_variance <- data.frame(x=c(), pd=c(), var=c())
for (var in c(1,2,3)) {
    normal_variance <- rbind(normal_variance,
                             data.frame(x=x, pd=dnorm(x, mean=6, sd=var), var=var))
}
normal_variance$var <- as.factor(normal_variance$var)
plot <- ggplot(normal_variance, aes(x=x, y=pd, color=var)) + geom_line()
print(plot + labs(y="概率密度", color="方差"))
```



标准正态分布 $N(0,1)$

定义

$N(0,1)$ 的密度函数

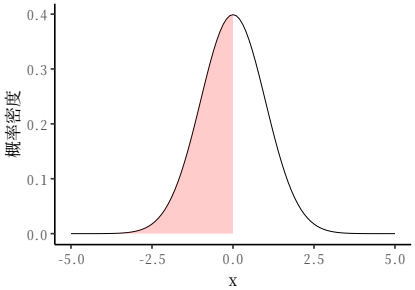
$$\phi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}},-\infty<x<+\infty$$

是偶函数，分布函数是 $\mathrm{erf}(x)$ 常记为

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}\mathrm{d}t,-\infty<x<+\infty$$

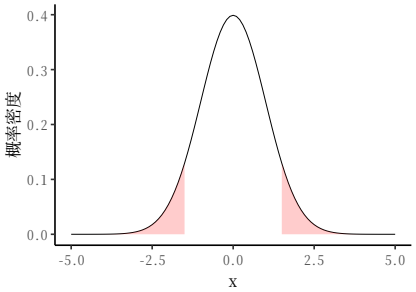
$\Phi(0) = 0.5$

```
mm <- 5
x=seq(-mm, mm, by=0.1)
df_line <- data.frame(x=x, 概率密度=dnorm(x, mean=0, sd=1))
p <- ggplot(df_line, aes(x=x, y=概率密度)) + geom_line()
df_poly_under <- rbind(subset(df_line, x <= 0),
                       data.frame(x=c(0, -mm), 概率密度=c(0, 0)))
q <- p + geom_polygon(data=df_poly_under, fill="red", alpha=1/5)
print(q)
```



$\Phi(-x) = 1-\Phi(x)$

```
poly_l <- rbind(subset(df_line, x <= -1.5),
                data.frame(x=c(-1.5, -mm), 概率密度=c(0, 0)))
poly_r <- rbind(subset(df_line, x >= 1.5),
                data.frame(x=c(mm, 1.5), 概率密度=c(0, 0)))
q <- p + geom_polygon(data=poly_l, fill="red", alpha=1/5)
q <- q + geom_polygon(data=poly_r, fill="red", alpha=1/5)
print(q)
```



中心面积

设 $a > 0$

$$
\begin{aligned}
&P(|X| < a) \\
=&P(-a < X < a) \\
=&2P(0 < X < a) \\
=&2[\Phi(a) - \Phi(0)] \\
=&2\Phi(a)-1
\end{aligned}
$$

变换：化为标准正态分布

引理

若 $X \sim N(\mu, \sigma^2)$，$Z = \frac{X-\mu}{\sigma}$，则 $Z \sim N(0,1)$

令 $\frac{x-\mu}{\sigma} = y \implies x = \mu + \sigma y, \mathrm{d}x = \sigma \mathrm{d}y$，$Z = \frac{X-\mu}{\sigma}$ 的分布函数为

$$F(z) = P(Z \le z) = P\left(\frac{X-\mu}{\sigma} \le z\right) = P(X \le \mu + \sigma z)$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\mu+\sigma z} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x$$

$$\implies F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{y^2}{2}} \mathrm{d}y = \Phi(z)$$

$$\implies Z \sim N(0,1)$$

变换：化为标准正态分布

引理

若 $X \sim N(\mu, \sigma^2)$，$Z = \frac{X-\mu}{\sigma}$，则 $Z \sim N(0,1)$

令 $\frac{x-\mu}{\sigma} = y \implies x = \mu + \sigma y, \mathrm{d}x = \sigma \mathrm{d}y$，$Z = \frac{X-\mu}{\sigma}$ 的分布函数为

$$F(z) = P(Z \le z) = P\left(\frac{X-\mu}{\sigma} \le z\right) = P(X \le \mu + \sigma z)$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\mu+\sigma z} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x$$

$$\implies F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{y^2}{2}} \mathrm{d}y = \Phi(z)$$

$$\implies Z \sim N(0,1)$$

变换：化为标准正态分布

引理

若 $X \sim N(\mu, \sigma^2)$，$Z = \frac{X-\mu}{\sigma}$，则 $Z \sim N(0,1)$

令 $\frac{x-\mu}{\sigma} = y \implies x = \mu + \sigma y, \mathrm{d}x = \sigma \mathrm{d}y$，$Z = \frac{X-\mu}{\sigma}$ 的分布函数为

$$
\begin{aligned}
F(z) =& P(Z \le z) = P\left(\frac{X-\mu}{\sigma} \le z\right) = P(X \le \mu + \sigma z) \\
=& \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\mu+\sigma z} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x \\
\implies & F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{y^2}{2}} \mathrm{d}y = \Phi(z) \\
\implies & Z \sim N(0,1)
\end{aligned}
$$

变换：化为标准正态分布

引理

若 $X \sim N(\mu, \sigma^2)$，$Z = \frac{X-\mu}{\sigma}$，则 $Z \sim N(0,1)$

令 $\frac{x-\mu}{\sigma} = y \implies x = \mu + \sigma y, \mathrm{d}x = \sigma \mathrm{d}y$，$Z = \frac{X-\mu}{\sigma}$ 的分布函数为

$$
\begin{aligned}
F(z) =& P(Z \le z) = P\left(\frac{X-\mu}{\sigma} \le z\right) = P(X \le \mu + \sigma z) \\
=& \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\mu+\sigma z} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x \\
\implies & F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{y^2}{2}} \mathrm{d}y = \Phi(z) \\
\implies & Z \sim N(0,1)
\end{aligned}
$$

变换：计算分布函数

若 $X \sim N(\mu, \sigma^2)$

$$F(x) = P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

$$\Longrightarrow \begin{cases} P(a < X < b) & = F(b) - F(a) & = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \\ P(X > a) & = 1 - F(a) & = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{cases}$$

变换：计算分布函数

若 $X \sim N(\mu, \sigma^2)$

$$F(x) = P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

$$\Longrightarrow \begin{cases} P(a < X < b) & = F(b) - F(a) & = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \\ P(X > a) & = 1 - F(a) & = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{cases}$$

$n\sigma$ 原理

```
for (n in seq(1,5)) {
    print(sprintf("-%d sigma 至 %d sigma 的概率为 %.8f", n, n, pnorm(n)-pnorm(-n)))
}
```

```
[1] "-1 sigma 至 1 sigma 的概率为 0.68268949"
[1] "-2 sigma 至 2 sigma 的概率为 0.95449974"
[1] "-3 sigma 至 3 sigma 的概率为 0.99730020"
[1] "-4 sigma 至 4 sigma 的概率为 0.99993666"
[1] "-5 sigma 至 5 sigma 的概率为 0.99999943"
```

服从正态分布的随机变量虽然取值在 $(-\infty,+\infty)$，但其值

```
for (n in c(3,5,6)) {
    print(sprintf("落在(-%d sigma, +%d sigma)之外的概率只有 %.2g", n, n, 2-2*pnorm(n)))
}
```

```
[1] "落在(-3 sigma, +3 sigma)之外的概率只有 0.0027"
[1] "落在(-5 sigma, +5 sigma)之外的概率只有 5.7e-07"
[1] "落在(-6 sigma, +6 sigma)之外的概率只有 2e-09"
```

$n\sigma$ 原理

```
for (n in seq(1,5)) {
    print(sprintf("-%d sigma 至 %d sigma 的概率为 %.8f", n, n, pnorm(n)-pnorm(-n)))
}
```

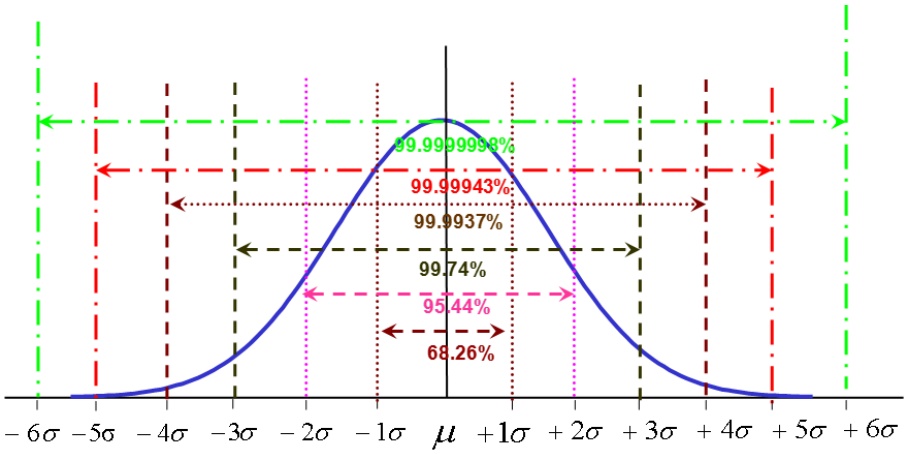
```
[1] "-1 sigma 至 1 sigma 的概率为 0.68268949"
[1] "-2 sigma 至 2 sigma 的概率为 0.95449974"
[1] "-3 sigma 至 3 sigma 的概率为 0.99730020"
[1] "-4 sigma 至 4 sigma 的概率为 0.99993666"
[1] "-5 sigma 至 5 sigma 的概率为 0.99999943"
```

服从正态分布的随机变量虽然取值在 $(-\infty, +\infty)$，但其值

```
for (n in c(3,5,6)) {
    print(sprintf("落在(-%d sigma, +%d sigma)之外的概率只有 %.2g", n, n, 2-2*pnorm(n)))
}
```

```
[1] "落在(-3 sigma, +3 sigma)之外的概率只有 0.0027"
[1] "落在(-5 sigma, +5 sigma)之外的概率只有 5.7e-07"
[1] "落在(-6 sigma, +6 sigma)之外的概率只有 2e-09"
```

$n\sigma$ 原理


99.9999998%
99.99943%
99.9937%
99.74%
95.44%
68.26%
-6σ -5σ -4σ -3σ -2σ -1σ μ $+1\sigma$ $+2\sigma$ $+3\sigma$ $+4\sigma$ $+5\sigma$ $+6\sigma$

例：“迹象”与“发现”

例

粒子物理与核物理实验中，假如观测到了某个现象 A。在给出结论之前需要评估一个概率 p：如果现象 A 不是真的，而是有统计涨落出现的小概率事件，这个概率是多大。

- 如果 $p > 2.7 \times 10^{-3}$，即落在 $\pm 3\sigma$ 之内，则认为这个现象很可能是 统计涨落 (fluctuation)；
- 如果 $5.7 \times 10^{-7} < p < 2.7 \times 10^{-3}$，即落在 $\pm(3\text{–}5)\sigma$ 之间，则称观测到现象 A 的 迹象 (evidence)；
- 如果 $p < 5.7 \times 10^{-7}$，即落在 $\pm 5\sigma$ 之外，则称 发现 了现象 A （ discovery 或 observation ）。

例

练习

已知 $X \sim N(2, \sigma^2)$，且 $P(2 < X < 4) = 0.3$，求 $P(X < 0)$.

正态分布是概率论中最重要的分布

- 在自然现象和社会现象中，大量随机变量都服从或近似服从正态分布。如人的身体特征指标 (身高、体重)，学习成绩，产品的数量指标等等都服从正态分布。
- 许多较复杂的指标，只要在受到的大量因素作用下每个因素的影响都不显著，且因素相互独立，也可认为近似服从正态分布。又如二项分布、泊松分布在 n 很大时，也以正态分布为极限分布。因此，可以说正态分布是最重要的分布。
- 根源，参考第 b 节，中心极限定理。