第二节 · 随机变量

司继春

上海对外经贸大学统计与信息学院

1 一元随机变量

1.1 随机变量的定义

上面介绍了一般的概率空间构建所需要的步骤,即一个概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$,我们需要一个样本空间 Ω,一个性质良好的集合族 $\mathscr{F}$ 以及定义在这个集合族上的概率函数 $\mathscr{P}$。特别地,当我们选取样本空间 $\Omega = \mathbb{R}$ 时,我们有 Borelσ-代数 $\mathscr{B}$ 以及由分布函数 F 定义的概率函数 P 组成的概率空间 $(\mathbb{R}, \mathscr{B}, P)$。

尽管对于一般的样本空间 Ω,我们都可以使用以上的方法构建概率空间,然而很多时候,直接对原始的样本空间 Ω 和集合族 $\mathscr{F}$ 进行分析并不是非常方便。比如,Ω 作为样本空间,我们并没有限制 Ω 具有代数结构,因而一般我们不能对样本点进行加减等运算。再比如如果我们的研究对象为抛 1000 次硬币,那么我们的样本空间有 2^{1000} 个元素,而这些元素不能相加相减(比如不存在正面 + 正面 这样的运算)。为了方便分析,我们一般会把原始的概率空间 Ω 映射到实轴 $\mathbb{R}$ 上进行分析,于是就有了**随机变量**(**random variable**, r.v.)的概念。

定义 1. (**随机变量**)对于概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$,映射 $X : \Omega \to \mathbb{R}^*$ 满足:对于任意的 $B \in \mathscr{B}$,有:

$$X^{-1}(B) \triangleq \{\omega : X(\omega) \in B\} \in \mathscr{F}$$

那么我们称 X 为随机变量。

因而,随机变量实质上是一个从样本空间到 $\mathbb{R}^*$ 上的一个函数。同时,为了使用原始样本空间 Ω 上的概率函数 $\mathscr{P}$ 来定义 $\mathbb{R}^*$ 上的概率函数 P,我们需要额外要求,对于 $\mathbb{R}$ 上的任意一个 Borel 集 B,集合 $\{\omega : X(\omega) \in B\}$ 都在概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 的 σ-代数 $\mathscr{F}$ 中。

例 1. 对于抛硬币的实验,$\Omega = \{H, T\}$,我们可以定义一个随机变量 X 如下:

$$\begin{cases} X(H) = 0 \\ X(T) = 1 \end{cases}$$

对于 $\mathscr{F}=\{\emptyset,\{H\},\{T\},\Omega\}$，我们有 $X^{-1}(\{0\})=\{H\},X^{-1}(\{1\})=\{T\}$，对于其他任何 Borel 集 B，如果 $1\in B$ 则 $T\in X^{-1}(B)$，如果 $0\in B$ 则 $H\in X^{-1}(B)$。如此我们便定义了一个 $\Omega\to\mathbb{R}^*$ 的随机变量 X。

例 2. 对于银行一天之内到达人数的问题，其 $\Omega=\{0,1,2\ldots\}=\mathbb{Z}$，定义 $X(\omega)=\omega,\omega\in\mathbb{Z}$，同上，我们定义了从自然数集合 $\mathbb{Z}\to\mathbb{R}$ 的随机变量 X。

例 3. 电灯泡的寿命的样本空间为 $\Omega=(0,+\infty)$，我们可以定义 $X(\omega)=\omega,\omega\in\Omega$，如此我们定义了从正实数集 $\mathbb{R}^+\to\mathbb{R}$ 的随机变量 X。

实际上，可以证明，集合族 $\left\{X^{-1}(B):B\in\mathscr{B}\right\}$ 是一个 $\sigma-$ 代数，我们通常记为 $\sigma\langle X\rangle$，成为由 X 生成的 $\sigma-$ 代数。$\sigma\langle X\rangle$ 可以看做是随机变量 X 所携带的关于样本空间的信息，即如果我们观察到 X，我们可以对实现的样本点 ω 发生做何种断言。

例 4. 在例 (2) 中，如果我们额外定义两个随机变量：(1) 退化的随机变量 $X_0(\omega)=1$，即无论何种情况发生，X_0 都是常数 1；(2) 一个分类变量：

$$X_1(\omega)=\begin{cases}0 & if\ \ \omega<10\\ 1 & if\ \ \omega\geq 10\end{cases}$$

即当人数到达超过 10 人时 $X_1=1$，否则为 0。那么：

$$\begin{aligned}\sigma\langle X_0\rangle&=\{\emptyset,\mathbb{Z}\}\\ \sigma\langle X_1\rangle&=\{\emptyset,\mathbb{Z},\{\omega:\omega<10\},\{\omega:\omega\geq 10\}\}\\ \sigma\langle X\rangle&=2^{\mathbb{Z}}\end{aligned}$$

其中 $2^{\mathbb{Z}}$ 代表 $\mathbb{Z}$ 的所有子集所组成的 $\sigma-$ 代数。从而：$\sigma\langle X_0\rangle\subset\sigma\langle X_1\rangle\subset\sigma\langle X\rangle$。实际上，$X_0$ 是一个退化的随机变量，并不能给我们带来关于样本空间的任何信息；如果观察到了 X_1，则只能判断实现的样本点究竟是小于 10 还是大于等于 10；而如果我们观察到 X，那么我们可以对实现的样本点做出精确断言。因而，三个随机变量中所携带的信息，X_0 最少，X_1 次之，X 所携带的信息最多。

在此之前我们定义了离散型的样本空间和离散型的分布函数，以上例 (1) 和例 (2) 都属于这种情况。下面我们来定义离散型的随机变量：

定义 2.（**离散型随机变量**）如果存在一个可数集 $B\in\mathscr{B}$，满足 $P(X\in B)=1$，则随机变量 X 成为离散型随机变量。

在得到随机变量的定义之后，我们还需要定义在 $(\mathbb{R},\mathscr{B})$ 上的概率函数才能完成随机变量的概率空间的定义。由于随机变量是定义在一个一般的样本空间 Ω 及其对应的概率空间 $(\Omega,\mathscr{F},\mathscr{P})$ 上，因而自然的想法是使用原概率空间中的概率函数 $\mathscr{P}$ 来定义 $(\mathbb{R},\mathscr{B})$ 上的新的概率函数 P。

定理 1. 对于一个随机变量 $X:\Omega\to\mathbb{R}$，定义

$$P_X(B)=\mathscr{P}\left(X^{-1}(B)\right)=\mathscr{P}(\{\omega:X(\omega)\in B\})$$

则 P_X 为概率函数，$(\mathbb{R},\mathscr{B},P_X)$ 为概率空间，我们称 $(\mathbb{R},\mathscr{B},P_X)$ 为 $(\Omega,\mathscr{F},\mathscr{P})$ 导出的概率空间。

对于任意的定义在概率空间 $(\Omega,\mathscr{F},\mathscr{P})$ 上的随机变量 $X(\omega)\in\mathbb{R}$，根据随机变量的定义，对于任意的 Borel 集合 $B\in\mathscr{B}$，总是可以找到一个集合 $B'\in\mathscr{F}$，使得 $B'=\{\omega:X(\omega)\in B\}$。进而，我们可以使用概率空间 $(\Omega,\mathscr{F},\mathscr{P})$ 上的概率函数 $\mathscr{P}$ 定义 $(\mathbb{R},\mathscr{B})$ 上的概率函数 P：$P(B)=\mathscr{P}(B')$，从而得到 $\mathbb{R}$ 上的概率空间 $(\mathbb{R},\mathscr{B},P_X)$。

例 5. 对于例 (1) 中的随机变量 X，如果原概率空间 $(\Omega,\mathscr{F},\mathscr{P})$ 中定义 $\mathscr{P}(H)=\mathscr{P}(T)=0.5$，则

$$P_X(\{1\})=\mathscr{P}\left(X^{-1}(1)\right)=\mathscr{P}(T)=0.5$$

同理可定义 $P(\{0\})=0.5$，即：

ω	H	T
$\mathscr{P}(\omega)$	0.5	0.5
$X(\omega)$	0	1
$P(X=x)$	0.5	0.5

注意由于随机变量的值域为 $\mathbb{R}$ 而非 $\{0,1\}$，因而我们必须对每一个 Borel 集都定义概率。根据例 (1) 中的随机变量 X 的定义，对于任意其他 Borel 集 B，如果 B 同时包含 0 和 1，那么 $P(B)=1$；如果 B 包含 0 或者 1 中的一个，则 $P(B)=0.5$；否则 $P(B)=0$。从而完成概率空间 $(\mathbb{R},\mathscr{B},P_X)$ 的定义。

1.2 随机变量的分布函数和密度函数

在此之前，我们定义了分布函数，而对于每一个随机变量，由于其值域均为 $\mathbb{R}$，因而其总对应一个分布函数。

定义 3. （累积分布函数）对于一个随机变量 X，函数

$$F_X(x)=P_X(X\le x)=\mathscr{P}\left(X^{-1}((-\infty,x])\right),\forall x\in\mathbb{R}$$

为一个分布函数（满足分布函数定义的要求），我们称其为**累积分布函数**（**cumulative distribution function, c.d.f.**）。

对于随机变量来说，累积分布函数包含了所有概率函数 P_X 的信息，因而使用 P_X 和使用累积分布函数 F_X 是等价的。因而我们通常使用标记 $X\sim F_X(x)$ 表示随机变量 X **服从** F_X **分布**。

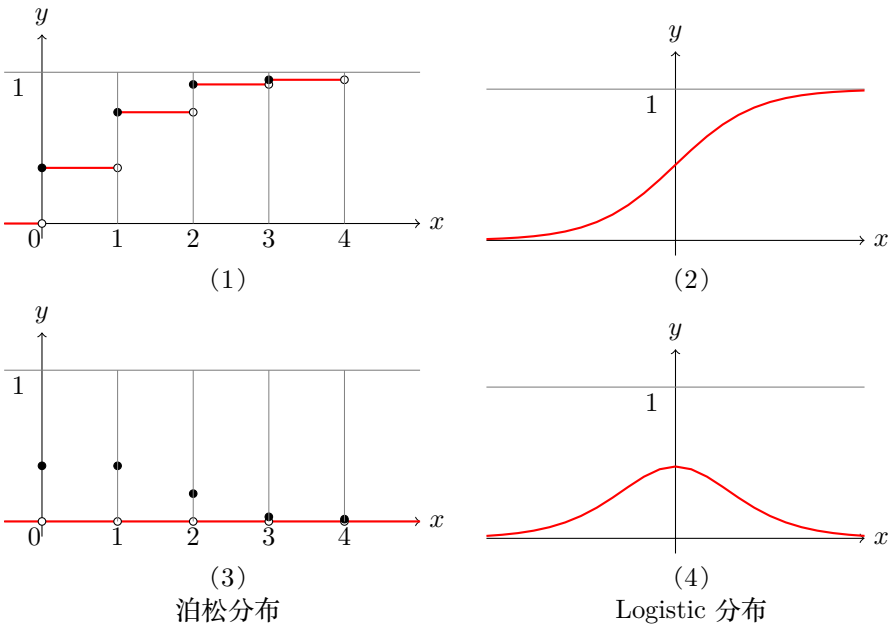


图 1: 累积分布函数与概率密度函数

例 6. （泊松分布累积分布函数）在建模一段时间内的到达次数时，我们经常使用所谓的泊松分布，即到达次数只可能取自然数值，或者 $P(X \in \mathbb{Z}) = 1$。由于自然数集 $\mathbb{Z}$ 为可数集，因而根据定义 (2)，该分布是一个离散分布，且其概率函数为：

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

其中 λ 为这段时间的平均到达次数。其分布函数为：

$$F_X(x) = P(X \leq x) = \sum_{i=0}^{x} \frac{\lambda^i}{i!} e^{-\lambda}$$

如图 (1.1) 所示即为泊松分布的分布函数（$\lambda = 1$）。注意由于分布函数的定义 $F_X(x) = P(X \leq x)$ 中，括号里面为小于等于而非小于号，因而累计分布函数为右连续的。

例 7. （Logistic 分布累积分布函数）如果一个随机变量 X 的分布函数为：

$$F_X(x) = \frac{e^x}{1 + e^x}$$

那么我们称 X 服从 Logistic 分布，其分布函数如图 (1.2) 所示：

定义 4. 如果两个随机变量的累积分布函数 $F_X(x) = F_Y(x)$，则我们称两个随机变量**同分布**（**identically distributed**），记为 $X \sim Y$。

总结一下，图 (2) 回顾了从原概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 定义随机变量 X 并定义

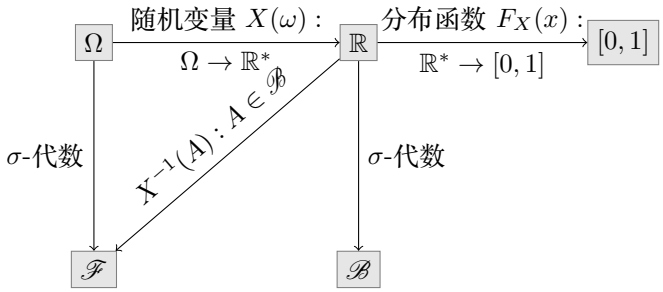


图 2: 随机变量

新的概率空间 $(\mathbb{R}, \mathscr{B}, P_X)$，并在此基础上定义分布函数的整个过程。由于随机变量极大的简化了分析，因此下面的分析中主要以随机变量为主要研究手段研究概率与统计问题。

虽然累计分布函数描述了随机变量的所有特征，然而很多时候，使用概率密度函数可以使分析和计算更为便利。

定义 5. 对于离散随机变量，如果 $P(x \in B) = 1$，B 为可数集，**概率质量函数**（**probability mass function**, p.m.f）定义为：

$$f_X(x) = P(X = x), \text{for all } x \in B$$

例 8.（概率质量函数）例 (6) 中的泊松分布为离散型随机变量，且 $P(X \in \mathbb{Z}) = 1$，那么其概率质量函数为：

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \forall x \in \mathbb{Z}$$

图 (1.3) 描述了泊松分布的概率质量函数（$\lambda = 1$）。

定义 6.（概率密度函数）对于连续型随机变量，**概率密度函数**（**probability density function**, p.d.f），$f_X(x)$ 定义为：

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt, \text{for all } x$$

注意如果分布函数 $F_X(x)$ 可导，那么其密度函数即其导函数。

例 9.（概率密度函数）Logistic 分布的 p.d.f 为其分布函数的导函数：

$$f_X(x) = \frac{e^x}{(1 + e^x)^2}$$

如图 (1.2) 所示。

2 期望及其性质

2.1 期望的定义与性质

数学期望（**mathematical expectation**）的概念在概率论和统计学中有着最广泛的应用。在一个一般的概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 中，数学期望实际上就是在这个概率空间中的积分。下面我们将给出一个在一般的概率空间中期望（积分）的定义。

定义 7.（离散型随机变量的期望）在概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 中，对于正的离散型随机变量 X：

$$X(\omega) = b_j \; if \; \omega \in \Lambda_j$$

其中 Λ_j 为样本空间 Ω 的划分，$b_j \geq 0$。期望定义为：

$$\mathbb{E}(X) = \sum_j b_j \mathscr{P}(\Lambda_j) = \sum_j b_j P(X = b_j)$$

注意在上面的定义中，我们要求随机变量为离散型随机变量，而对于样本空间 Ω 是否离散并没有做假定。

例 10. 抛一枚骰子，其样本空间为 $\Omega = \{1, 2, 3, 4, 5, 6\}$，假设骰子均匀，那么其概率函数可以由：

$$\mathscr{P}(\{w\}) = \frac{1}{6}, \omega \in \Omega$$

来定义。如果定义随机变量 $X(\omega) = \omega$，那么其数学期望为：

$$\mathbb{E}(X) = \sum_{j \in \Omega} X(\omega) \cdot \mathscr{P}(\{\omega\}) = \sum_{j=1}^{6} j \cdot \frac{1}{6} = \frac{7}{2}$$

例 11. 对于例 (6) 中定义的泊松分布随机变量 $X(\omega) = \omega, \omega \in \mathbb{Z}$，其期望为：

$$\mathbb{E}(X) = \sum_{j=0}^{\infty} j \cdot \mathscr{P}(\{j\}) = \sum_{j=0}^{\infty} j \cdot \frac{\lambda^j}{j!} e^{-\lambda} = \lambda \sum_{j=1}^{\infty} \frac{\lambda^{j-1}}{(j-1)!} e^{-\lambda} = \lambda$$

特别的，如果令 $X(\omega) = 1_A(\omega)$，其中 1_A 为指示函数，即：

$$1_A(\omega) = \begin{cases} 1 & if \; \omega \in A \\ 0 & otherwise \end{cases}$$

那么其期望：

$$\mathbb{E}(1_A) = 1 \cdot \mathscr{P}(A) + 0 = \mathscr{P}(A)$$

即指示函数的期望等于该令指示函数等于 1 的集合的概率。

下面我们使用离散型随机变量的数学期望继续定义连续型随机变量的数学期望。令 X 为定义在 $(\Omega, \mathscr{F}, \mathscr{P})$ 上任意的正的随机变量（$X(\omega) \geq 0$），定义集合：

$$\Lambda_{mn} = \left\{ \omega : \frac{n}{2^m} \leq X(\omega) < \frac{n+1}{2^m} \right\} \subset \mathscr{F}$$

如此，对于任意一个 m，我们可以定义一个离散随机变量 X_m：

$$X_m(\omega) = \frac{n}{2^m} \; if \; \omega \in \Lambda_{mn}$$

注意对于任意 m，有：

$$X_m(\omega) \leq X_{m+1}(\omega); 0 \leq X(\omega) - X_m(\omega) < \frac{1}{2^m}, \forall \omega \in \Omega$$

即 X_m 为单调递增的随机变量，从而

$$\lim_{m \to \infty} X_m(\omega) = X(\omega), \forall \omega \in \Omega$$

根据以上离散随机变量的定义，X_m 的期望可以定义为：

$$\mathbb{E}(X_m) = \sum_{n=0}^{\infty} \frac{n}{2^m} \mathscr{P}\left(\left\{ \frac{n}{2^m} \leq X(\omega) < \frac{n+1}{2^m} \right\} \right)$$

如果存在一个 m 使得 $E(X_m) = +\infty$，那么定义 $E(X) = +\infty$，否则，定义：

$$\mathbb{E}(X) = \lim_{m \to \infty} \mathbb{E}(X_m)$$

如果上述极限存在且小于 $+\infty$，我们称随机变量 X 是**可积**（integrable）的。

至此我们定义了任意正随机变量的期望。对于任意的随机变量 X，定义 $X^+(\omega) = \max\{X(\omega), 0\}$，$X^-(\omega) = \max\{-X(\omega), 0\}$，则 $X = X^+ - X^-$，$|X| = X^+ + X^-$。X^+ 和 X^- 都是正的随机变量，如果 $|X|$ 是可积的，即 $\mathbb{E}|X| < \infty$，那么我们称随机变量 X 是**可积**的，此时可以定义随机变量 X 的期望为：

$$\mathbb{E}(X) = \mathbb{E}\left(X^+\right) - \mathbb{E}\left(X^-\right)$$

至此，任意随机变量的期望即定义完毕。由于该期望是由概率函数 $\mathscr{P}$ 定义的，我们一般记期望为：

$$\mathbb{E}(X) = \int_{\Omega} X(\omega) \mathscr{P}(d\omega)$$

更一般的，对于 $A \in \mathscr{F}$，定义随机变量 X 在集合 A 上的积分为：

$$\int_A X(\omega) \mathscr{P}(d\omega) = \mathbb{E}(X \cdot 1_A) = \int_{\Omega} X(\omega) \cdot 1_A(\omega) \mathscr{P}(d\omega)$$

特别的，令 $X(\omega)=1$ 为退化的随机变量，根据以上积分的定义，同样有：

$$\mathscr{P}(A)=\mathbb{E}(1_A)=\int_A \mathscr{P}(d\omega)$$

积分有以下性质：

定理 2. （积分的性质）

1. （线性性）$\int_A (aX(\omega)+bY(\omega))\,\mathscr{P}(d\omega)=a\int_A X(\omega)\,\mathscr{P}(d\omega)+b\int_A Y(\omega)\,\mathscr{P}(d\omega)$
2. （可加性）如果 A_n 不相交，则 $\int_{\bigcup_n A_n} X\mathscr{P}(d\omega)=\sum_n \int_{A_n} X\mathscr{P}(d\omega)$
3. 如果在 A 上，$X\geq 0\ a.s.$，则 $\int_A X\mathscr{P}(d\omega)\geq 0$
4. （单调性）如果在 A 上，$X_1\leq X\leq X_2\ a.s.$，则 $\int_A X_1\mathscr{P}(d\omega)\leq \int_A X\mathscr{P}(d\omega)\leq \int_A X_2\mathscr{P}(d\omega)$
5. （均值定理）如果在 A 上有 $a\leq X\leq b\ a.s.$，那么 $a\mathscr{P}(A)\leq \int_A X\mathscr{P}(d\omega)\leq b\mathscr{P}(A)$
6. $\left|\int_A X\mathscr{P}(d\omega)\right|\leq \int_A |X|\,\mathscr{P}(d\omega)$
7. （有界收敛定理）如果 $\lim_{n\to\infty} X_n=X\ a.s.$，且存在一个常数 M 使得 $\forall n: |X_n|\leq M\ a.s.$，那么

$$\lim_{n\to\infty}\int_A X_n\mathscr{P}(d\omega)=\int_A X\mathscr{P}(d\omega)=\int_A\left(\lim_{n\to\infty}X_n\right)\mathscr{P}(d\omega) \tag{1}$$

8. （单调收敛定理）如果 $X_n\geq 0$ 且 $X_n\uparrow X\ a.s.$，那么 *(1)* 式成立
9. （控制收敛定理）如果 $\lim_{n\to\infty} X_n=X\ a.s.$，且存在一个随机变量 Y 使得 $|X_n|\leq Y\ a.s.$，且 $\int_A Y\mathscr{P}(d\omega)<\infty$，则 *(1)* 式成立。
10. 如果 $\sum_n \int_A |X_n|\mathscr{P}(d\omega)<\infty$，那么 $\sum_n |X_n|<\infty\ a.s.\ on\ A$，从而

$$\int_A\left[\sum_n X_n\right]\mathscr{P}(d\omega)=\sum_n\int_A X_n\mathscr{P}(d\omega)$$

由于期望即定义为 Ω 上的积分，因而以上性质对于期望都成立，比如由积分的线性性可以得到期望的线性性：

$$\mathbb{E}(aX+bY)=a\mathbb{E}(X)+b\mathbb{E}(Y)$$

而定理 (2.7-2.9) 则解决了极限符号和积分符号互换的问题。注意如果不满足以上定理的条件，积分和极限符号不必然可以互换。

例 12. （积分与极限互换）如果令 $\Omega=\mathbb{R}$，分布函数为：

$$F(\omega)=\begin{cases}0 & \omega<-1\\ \frac{1}{2}\omega+\frac{1}{2} & -1\leq\omega\leq 1\\ 1 & \omega>1\end{cases}$$

即在 $[-1,1]$ 上的均匀分布，进而使用此分布函数构建 $\mathbb{R}$ 上的概率测度 P。随机变量

$$X_n(\omega)=\begin{cases}0 & if\ |\omega|>\frac{1}{n}\\ n & if\ |\omega|\leq\frac{1}{n}\end{cases}$$

因而除了在一个点 $\omega=0$ 处之外，$\lim_{n\to\infty}X_n(\omega)=0$，或者 $\lim_{n\to\infty}X_n(\omega)=0\ a.s.$，因而[1]

$$\int_\Omega \lim_{n\to\infty}X_nP(d\omega)=0$$

然而由于 $\int_\Omega X_nP(d\omega)=1$，因而 $\lim_{n\to\infty}\int_\Omega X_nP(d\omega)=1$。因而在这个例子里 $\lim_{n\to\infty}\int_\Omega X_nP(d\omega)\neq\int_\Omega\lim_{n\to\infty}X_nP(d\omega)$。

以上我们介绍了数学期望的定义，然而给定一个随机变量，使用概率空间 $(\Omega,\mathscr{F},\mathscr{P})$ 计算数学期望非常不方便，我们通常希望使用导出的概率空间 $(\mathbb{R},\mathscr{B},P)$ 计算数学期望，因此我们有以下定理：

定理 3. 如果概率空间 $(\Omega,\mathscr{F},\mathscr{P})$ 根据定理 *(1)* 导出了概率空间 $(\mathbb{R},\mathscr{B},P)$，令 g 为一个可测函数，则：

$$\mathbb{E}(g(X))=\int_\Omega g(X(\omega))\,\mathscr{P}(d\omega)=\int_{\mathbb{R}}g(x)\,P(dx)$$

如果等式两边积分都存在。

如果 F 为对应于概率函数 P 的分布函数，则在 $(a,b]$ 上的积分也可以写为：

$$\int_{(a,b]}g(x)\,P(dx)=\int_{(a,b]}g(x)\,dF(x)$$

特别的，根据积分以及分布函数的定义：

$$\int_{(a,b]}dF(x)=\int_{\mathbb{R}}1_{(a,b]}(x)\,dF(x)=P((a,b])=F(b)-F(a)$$

此外，由于随机变量 X 的期望即在样本空间上的积分，因而随机变量 X 的数学期望可以写为：

$$\mathbb{E}(X)=\int_\Omega X(\omega)\,\mathscr{P}(d\omega)=\int_{\mathbb{R}}xP(dx)=\int_{\mathbb{R}}xdF(x)$$

[1]在定义积分时，当遇到 $0\cdot\infty$ 时，定义 $0\cdot\infty=0$。

即我们可以直接使用定义在 $\mathbb{R}$ 上的概率函数或者分布函数计算随机变量 X 的期望。

例 13. 在例 (10) 中，我们定义了样本空间为 $\Omega = \{1,2,3,4,5,6\}$ 及其概率函数：

$$\mathscr{P}(\{w\}) = \frac{1}{6}, \omega \in \Omega$$

即：

ω	1	2	3	4	5	6
$\mathscr{P}(\{\omega\})$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
$(\omega - 3.5)^2$	6.25	2.25	0.25	0.25	2.25	6.25

如果定义一个随机变量 $X(\omega) = (\omega - 3.5)^2$，那么我们可以根据定理 (1) 中定义导出的概率空间 $(\mathbb{R}, \mathscr{B}, P_X)$，其概率质量函数：

$X(\omega)$	6.25	2.25	0.25
$P_X(X = x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

为了计算 $\mathbb{E}(X)$，我们可以使用概率函数 $\mathscr{P}(\cdot)$：

$$\mathbb{E}(X) = \sum_{j \in \Omega} (\omega - 3.5)^2 \mathscr{P}(\{\omega\}) = \sum_{j=1}^{6} (\omega - 3.5)^2 \frac{1}{6} = \frac{33}{4}$$

然而直接使用概率函数 $\mathscr{P}(\cdot)$ 并不方便，因而我们可以直接使用导出的在 $\mathbb{R}$ 上的概率函数 P_X 直接进行计算：

$$\mathbb{E}(X) = \sum_{x} x \cdot P_X(X = x) = \frac{6.25 + 2.25 + 0.25}{3} = \frac{33}{4}$$

根据定理 (3)，以上两种计算方法是等价的。

以上定义的积分我们称之为勒贝格（Lebegue）积分，该积分的定义与常用的黎曼（Riemann）积分并不尽一致，然而在数量上两者是相等的，因而在实际计算中，我们可以直接使用黎曼积分进行计算。特别的，如果随机变量 X 的分布函数 $F(x)$ 有密度函数 $f(x)$，则其期望可以由：

$$\mathbb{E}(X) = \int_{\mathbb{R}} x dF(x) = \int_{\mathbb{R}} x f(x) dx$$

计算。注意在计算期望之前，首先应该保证随机变量是可积的，即 $\mathbb{E}|X| < \infty$。下例给出了一个不可积的随机变量，因而其期望不存在。

例 14. （Cauchy 分布的期望）Cauchy 的密度函数为：

$$f(x) = \frac{1}{\pi(1 + x^2)}$$

如果一个随机变量服从 Cauchy 分布，则：

$$\mathbb{E}(|X|)=\int_{\mathbb{R}}\frac{|x|}{\pi}\frac{1}{1+x^2}dx=\frac{2}{\pi}\int_{[0,\infty)}\frac{x}{1+x^2}dx$$

对于任意正数 M：

$$\int_{[0,M]}\frac{x}{1+x^2}dx=\left.\frac{\log\left(1+x^2\right)}{2}\right|_0^M=\frac{\log\left(1+M^2\right)}{2}$$

因此：

$$\mathbb{E}(|X|)=\frac{1}{\pi}\lim_{M\to\infty}\log\left(1+M^2\right)=\infty$$

因而该随机变量是不可积的，期望不存在。

例 15.（正态分布的期望）给定常数 μ 和 σ^2，如果随机变量 X 的分布函数为：

$$F(x)=\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}dx$$

那么我们称随机变量 X 服从正态分布，记为 $X\sim N\left(\mu,\sigma^2\right)$。注意以上积分内：

$$\phi(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

为其密度函数，且：

$$\begin{aligned}\Phi(\infty)&=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}dx\\&=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}dx\\&=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\sigma d\frac{x-\mu}{\sigma}\\&=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left\{-\frac{z^2}{2}\right\}dz\\&=1\end{aligned}$$

我们可以使用正态分布密度函数求其期望，在计算期望之前，必须要计算 $\mathbb{E}|X|$

保证其可积：

$$\begin{aligned}
\mathbb{E}\,|X| &= \int_{-\infty}^{\infty} |x| \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= 2\int_{0}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= \frac{2}{\sqrt{2\pi}\sigma} \int_{0}^{\infty} x \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} (\mu + \sigma z) \exp\left\{-\frac{z^2}{2}\right\} dz \\
&= \frac{2\mu}{\sqrt{2\pi}} \int_{0}^{\infty} \exp\left\{-\frac{z^2}{2}\right\} dz + \frac{2\sigma}{\sqrt{2\pi}} \int_{0}^{\infty} z \exp\left\{-\frac{z^2}{2}\right\} dz \\
&= \frac{2\mu}{\sqrt{2\pi}} \frac{\sqrt{2\pi}}{2} + \frac{2\sigma}{\sqrt{2\pi}} \int_{0}^{\infty} \exp\left\{-\frac{z^2}{2}\right\} d\frac{z^2}{2} \\
&= \mu + \frac{2\sigma}{\sqrt{2\pi}} < \infty
\end{aligned}$$

因而该随机变量可积，进而可以计算其期望：

$$\begin{aligned}
\mathbb{E}\,(X) &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{\sigma^2}\right\} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + \sigma z) \exp\left\{-\frac{z^2}{2}\right\} dz \\
&= \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{z^2}{2}\right\} dz + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z \exp\left\{-\frac{z^2}{2}\right\} dz \\
&= \mu + \frac{\sigma}{\sqrt{2\pi}} \int_{0}^{\infty} z \exp\left\{-\frac{z^2}{2}\right\} dz + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{0} z \exp\left\{-\frac{z^2}{2}\right\} dz \\
&= \mu
\end{aligned}$$

因而正态分布 $N\left(\mu, \sigma^2\right)$ 中，参数 μ 即其期望。

2.2 方差、偏度、峰度

对于任意的正数 p，如果 $\mathbb{E}\,(|X|^p) < \infty$，则记 $X \in L^p = L^p\,(\Omega, \mathscr{F}, \mathscr{P})$。对于整数 $r < p$，随机变量 X 的 r **阶矩**被定义为 $\mathbb{E}\,(X^r)$。一阶矩即为随机变量 X 的期望。此外，随机变量 X 的 r **阶中心矩**被定义为 $\mathbb{E}\,([X - E\,(X)]^r)$。特别的，当 $r = 2$ 时，2 阶中心矩即为随机变量的**方差**（**variance**），记为 $\mathrm{Var}\,(X)$ 或者 $\sigma^2\,(X)$，**标准差**（**standard deviation**）定义为 $\sigma\,(X) = \sqrt{\mathrm{Var}\,(X)}$。$X$

的方差可以使用一阶和二阶矩计算得到：

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}\left(\left[X-\mathbb{E}(X)\right]^2\right)=\mathbb{E}\left(X^2-2\mathbb{E}(X)\cdot X+\mathbb{E}(X)^2\right)\\ &= \mathbb{E}\left(X^2\right)-2\mathbb{E}(X)^2+\mathbb{E}(X)^2=\mathbb{E}X^2-(\mathbb{E}X)^2\end{aligned}$$

注意 $\mathbb{E}X^2\geq(\mathbb{E}X)^2$。此外，根据方差定义，常数（退化的随机变量）的方差为零，且：

$$\begin{aligned}\mathrm{Var}(aX+b)&=\mathbb{E}\left[aX+b-\mathbb{E}(aX+b)\right]^2\\ &=\mathbb{E}\left[aX+b-a\mathbb{E}(X)-b\right]^2\\ &=a^2\mathbb{E}(X-\mathbb{E}X)^2\\ &=a^2\mathrm{Var}(X)\end{aligned}$$

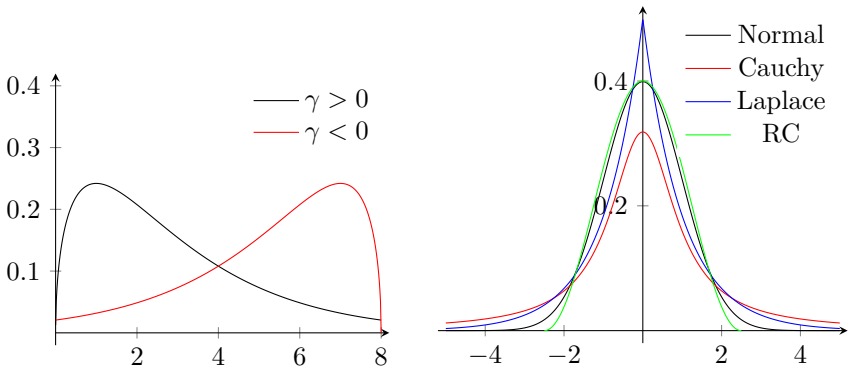
即一个随机变量的线性函数的方差等于该随机变量方差乘以斜率的平方。

例 16. （正态分布方差）为了计算 $X\sim N\left(\mu,\sigma^2\right)$ 的方差，我们首先必须保证 X^2 可积，即 $\mathbb{E}\left|X^2\right|<\infty$。由于 $X^2>0$ 且 $\mathrm{Var}(X)=\mathbb{E}\left(X^2\right)-(\mathbb{E}X)^2$，因而我们只要计算 $\mathbb{E}X^2$ 即可。由于：

$$\begin{aligned}\mathbb{E}\left(X^2\right)&=\int_{\mathbb{R}}x^2\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{\sigma^2}\right\}dx\\ &=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}x^2\exp\left\{-\frac{(x-\mu)^2}{\sigma^2}\right\}dx\\ &=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(\mu+\sigma z)^2\exp\left\{-\frac{z^2}{2}\right\}dz\\ &=\frac{1}{\sqrt{2\pi}}\mu^2\int_{-\infty}^{\infty}\exp\left\{-\frac{z^2}{2}\right\}dz+\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}z^2\exp\left\{-\frac{z^2}{2}\right\}dz\\ &+\underbrace{\frac{2\mu\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty}z\exp\left\{-\frac{z^2}{2}\right\}dz}_{=0}\\ &=\mu^2-\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}z\,d\exp\left\{-\frac{z^2}{2}\right\}\\ &=\mu^2-\frac{\sigma^2}{\sqrt{2\pi}}\underbrace{\left[z\exp\left\{-\frac{z^2}{2}\right\}\right]_{-\infty}^{\infty}}_{=0}+\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left\{-\frac{z^2}{2}\right\}dz\\ &=\mu^2+\sigma^2<\infty\end{aligned}$$

因而正态分布的方差 $\mathrm{Var}(X)=\mathbb{E}\left(X^2\right)-(\mathbb{E}X)^2=\sigma^2$，即正态分布参数中 σ^2 为该正态分布的方差。

除了前两阶矩之外，经常我们还会关心更高阶的矩。其中，**偏度**（**skewness**）



(a) 左偏与右偏 (b) 尖峰厚尾

图 3: 偏度与峰度

和**峰度**（**kurtosis**）是最经常被关心的高阶矩。其中偏度为随机变量的三阶中心矩，即定义：

$$\gamma = \mathbb{E}\left(\left[\frac{X - \mathbb{E}(X)}{\sigma(X)}\right]^3\right) = \frac{\mathbb{E}[X - \mathbb{E}(X)]^3}{\sigma^3(X)}$$

如果 X 为对称分布，那么必然有 $\gamma = 0$。顾名思义，偏度与分布的不对称性有关，如图 (3.a) 所示，当 $\gamma > 0$ 时，分布函数右边的尾巴比较厚，我们称其分布为右偏（negative skew）分布，反之为左偏（positive skew）分布。

而峰度则是随机变量的四阶中心矩，即：

$$\mathrm{Kurt}(X) = \mathbb{E}\left(\left[\frac{X - \mathbb{E}(X)}{\sigma(X)}\right]^4\right) = \frac{\mathbb{E}\left([X - \mathbb{E}(X)]^4\right)}{[\mathrm{Var}(X)]^2}$$

尽管我们一般将其称为「峰度」，然而这一称呼并不准确，更准确的称呼应该为「尾厚度」。如图 (3.a) 所示，其中 RC 代表 Raised cosine 分布。如果比较正态分布，会发现正态分布的尾巴比 Raised cosine 分布要厚，而相应的正态分布的峰要「尖」一点，所以我们经常说「尖峰厚尾」。实际上，正态分布的峰度为 3，而图中 Raised cosine 分布的峰度约为 $2.40623 < 3$。

然而注意到，峰度大的并不一定代表峰更「尖」，而仅仅是尾巴更「厚」。如图中 Laplace 分布和 Cauthy 分布相比，Laplace 分布的峰更尖，但是其峰度小于 Cauthy 分布的峰度。实际上，Cauthy 分布的峰度为 $+\infty$。因而判断峰度时，不能顾名思义只看其峰的尖锐程度，而应该看尾巴的厚度。

我们一般会把峰度与正态分布的峰度相比，定义超额峰度（excess kurtosis）为峰度减 3，因而如果超额峰度 > 0，那么其尾巴比正态分布的尾巴要厚，而如果超额峰度 < 0，那么其尾巴要比正态分布的尾巴要薄。

2.3 积分与导数互换

在实际应用中，我们经常需要对积分进行求导，比如：

$$\frac{d}{d\theta}\int_{\mathbb{R}} g(x,\theta)\,dF(x)$$

然而积分常常难以计算，经常我们希望使用：

$$\int_{\mathbb{R}} \frac{dg(x,\theta)}{d\theta}dF(x)$$

来计算。然而这一计算方法是有条件的。

定理 4. （积分与微分互换）如果函数 $g(x,\theta)$ 在 $\theta=\theta_0$ 处可微，即对于任意 x，极限

$$\lim_{\delta\to 0}\frac{g(x,\theta_0+\delta)-g(x,\theta_0)}{\delta}=\frac{\partial}{\partial\theta}g(x,\theta)\Big|_{\theta=\theta_0}$$

存在，并且存在一个函数 $h(x,\theta_0)$ 以及一个常数 δ_0，有

1. 对于任意的 x 和 $|\delta|\le\delta_0$，有：$\left|\frac{g(x,\theta_0+\delta)-g(x,\theta_0)}{\delta}\right|\le h(x,\theta_0)$
2. $\int_{\mathbb{R}} h(x,\theta_0)\,dF(x)<\infty$

那么：

$$\frac{d}{d\theta}\int_{\mathbb{R}} g(x,\theta)\,dF(x)\Big|_{\theta=\theta_0}=\int_{\mathbb{R}}\left[\frac{dg(x,\theta)}{d\theta}\Big|_{\theta=\theta_0}\right]dF(x)$$

该定理即控制收敛定理的应用。该定理表明，在一定的条件下，积分与微分操作可以交换顺序。在该条件成立下，进一步我们有莱布尼茨法则：

定理 5. （***Leibnitz*** 法则）如果 $g(x,\theta),a(\theta),b(\theta)$ 对 θ 可微，那么：

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} g(x,\theta)\,dx=g(b(\theta),\theta)\frac{d}{d\theta}b(\theta)-g(a(\theta),\theta)\frac{d}{d\theta}a(\theta)+\int_{a(\theta)}^{b(\theta)}\frac{\partial}{\partial\theta}g(x,\theta)\,dx$$

例 17. 如果令函数 $g(x,\theta)=\theta x^2+\theta^2 x$，我们希望求：

$$\int_{\frac{\theta}{2}}^{\theta} g(x,\theta)\,dx=\int_{\frac{\theta}{2}}^{\theta}\theta x^2+\theta^2 x dx$$

对 θ 的导函数，由于：

$$\int_{\frac{\theta}{2}}^{\theta}\theta x^2+\theta^2 x dx=\left[\frac{\theta}{3}x^3+\frac{\theta^2}{2}x^2\right]_{\frac{\theta}{2}}^{\theta}=\frac{\theta^4}{3}+\frac{\theta^4}{2}-\frac{\theta}{3}\frac{\theta^3}{8}-\frac{\theta^2}{2}\frac{\theta^2}{4}=\frac{2}{3}\theta^4$$

因而：

$$\frac{d}{d\theta}\int_{\frac{\theta}{2}}^{\theta} g(x,\theta)\,dx=\frac{d\left(\frac{2}{3}\theta^4\right)}{d\theta}=\frac{8}{3}\theta^3$$

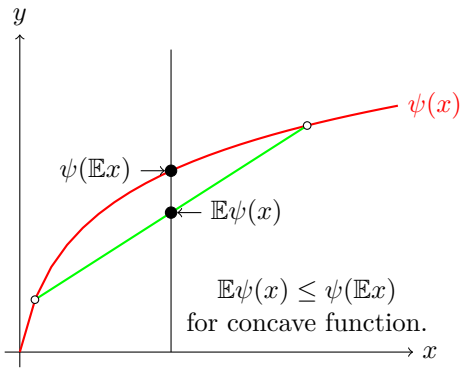


图 4: Jensen 不等式

或者，可以根据 Leibnitz 法则：

$$\begin{aligned}\frac{d}{d\theta}\int_{\frac{\theta}{2}}^{\theta}\theta x^2+\theta^2 x dx &= 2\theta^3\cdot 1-\left(\frac{\theta^3}{4}+\frac{\theta^3}{2}\right)\cdot\frac{1}{2}+\int_{\frac{\theta}{2}}^{\theta}x^2+2\theta x dx\\ &= 2\theta^3-\frac{3}{8}\theta^3+\left[\frac{x^3}{3}+\theta x^2\right]_{\frac{\theta}{2}}^{\theta}\\ &= 2\theta^3-\frac{3}{8}\theta^3+\frac{\theta^3}{3}+\theta^3-\frac{\theta^3}{24}-\frac{\theta^3}{4}\\ &= \frac{8}{3}\theta^3\end{aligned}$$

2.4 常用不等式

下面我们介绍几个常用的不等式。

定理 6. （***Chebyshev*** 不等式）如果函数 ψ 满足：$\psi(u)=\psi(-u)\geq 0$，且在 $(0,\infty)$ 上单调递增，X 为随机变量，且 $\psi(X)<\infty$，则对于 $u>0$，有：

$$P(|X|\geq u)\leq\frac{\mathbb{E}[(\psi(X))]}{\psi(u)}$$

证明. 根据均值定理：

$$\mathbb{E}[\psi(X)]=\int_{\mathbb{R}}\psi(X)P(dX)\geq\int_{\{|X|\geq u\}}\psi(X)P(d\omega)\geq\psi(u)P(|X|\geq u)$$

□

特别的，令 $Y=\frac{[X-\mathbb{E}(X)]}{\sigma(X)}$，则 $\mathbb{E}(Y)=0$，$\mathbb{E}(Y^2)=1$。令 $\psi(x)=x^2$，有：

$$P(|Y|\geq u)\leq\frac{1}{u^2}$$

定理 7.（***Jensen*** **不等式**）如果 ψ 为凹函数，且随机变量 X 和 $\psi(X)$ 可积，则：

$$\psi(\mathbb{E}(X)) \geq \mathbb{E}[\psi(X)]$$

Jensen 不等式表明，对于一般的非线性函数，期望的函数与函数的期望并不相等，如图 (4) 所示。

作为 Jensen 不等式的一个应用，考虑 $0 < r < s$，并令 $p = \frac{s}{r} > 1$，注意 $\psi(x) = |x|^p$ 为凸函数，因而根据 Jensen 不等式，有：

$$\mathbb{E}(|X|^{rp}) = \mathbb{E}(|X|^s) \geq [\mathbb{E}(|X|^r)]^p = [\mathbb{E}(|X|^r)]^{\frac{s}{r}}$$

整理之后可以得到：

$$\mathbb{E}(|X|^r) \leq [\mathbb{E}(|X|^s)]^{\frac{r}{s}}$$

以上不等式被称为 **Liapounov 不等式**。

定理 8.（***Cauchy-Schwarz*** **不等式**）对于两个随机变量 X, Y，若满足可积性，有：

$$|\mathbb{E}(XY)| \leq \mathbb{E}(|XY|) \leq \left[\mathbb{E}\left(|X|^2\right)\right]^{\frac{1}{2}} \left[\mathbb{E}\left(|Y|^2\right)\right]^{\frac{1}{2}}$$

3 常用分布

3.1 离散分布

3.1.1 离散均匀分布

随机变量 X 服从离散均匀分布如果 $P(X = x) = \frac{1}{N}, x \in (a_1, a_2, ..., a_N)$。

- 期望：$\mathbb{E}(X) = \frac{1}{N}\sum_{i=1}^{N} a_i$
- 方差：$\mathrm{Var}(X) = \frac{1}{N}\sum_{i=1}^{N}(a_i - \mathbb{E}X)^2$

3.1.2 伯努利分布

伯努利分布**即伯努利试验**（**Bernoulli trial**）的结果。伯努利试验即试验结果只有两种可能性（成功，失败），独立重复 N 次的试验结果。一个随机变量 X 定义为 $X = 1\ if$ 成功，$= 0\ if$ 失败。如果成功的概率为 p，那么**伯努利分布**（**Bernoulli Distribution**）$X \sim Ber(p)$ 即：

$$X = \begin{cases} 1 & with\ probability\ p \\ 0 & with\ probability\ 1-p \end{cases} \quad 0 \leq p \leq 1$$

- 期望：$\mathbb{E}(X) = p$
- 方差：$\mathrm{Var}(X) = p(1-p)$

3.1.3 二项分布

二项分布即独立重复 N 次伯努利实验，成功次数 $Y = \sum_{i=1}^{N} X_i, X_i \sim Ber(p)$ 的分布。简单计算可以得到，随机变量 Y 的概率质量函数：

$$P(Y = y|N,p) = \begin{pmatrix} N \\ y \end{pmatrix} p^y (1-p)^{N-y}, y = 1, 2, ..., N$$

我们成 Y 服从**二项分布**（**Binomial Distribution**），记为 $Y \sim Bi(N,p)$。由于对于任意的实数 a, b 以及任意整数 $N \geq 0$，有：$(a+b)^N = \sum_{i=0}^{N} \begin{pmatrix} N \\ i \end{pmatrix} a^i b^{n-i}$，因而：

$$\sum_{y=1}^{N} P(Y = y|N,p) = \sum_{y=0}^{N} \begin{pmatrix} N \\ y \end{pmatrix} p^y (1-p)^{N-y} = (p+1-p)^N = 1$$

- 期望：$\mathbb{E}(Y) = Np$
- 方差：$\mathrm{Var}(Y) = Np(1-p)$

3.1.4 负二项分布

二项分布关心在 N 次伯努利实验中成功的次数，而有时我们会关心为了达到 r 次成功所必须要做的试验的次数。记 Z 为为了达到 r 次成功所需的试验次数，那么事件 $Z = z$ 即 z 次试验达到了 r 次成功，这一事件可以进一步分解为在前 $z-1$ 次试验中得到了 $r-1$ 次成功，而第 r 次也成功了。所以其概率质量函数为：

$$\begin{aligned} P(Z = z|r,p) &= \begin{pmatrix} z-1 \\ r-1 \end{pmatrix} p^{r-1} (1-p)^{z-r} \cdot p \\ &= \begin{pmatrix} z-1 \\ r-1 \end{pmatrix} p^r (1-p)^{z-r}, z = r, r+1, ... \end{aligned}$$

我们称随机变量 Z 服从**负二项分布**（**Negative Binomial Distribution**），记为 $Z \sim NB(r,p)$。

下面计算 Z 的期望：

$$\begin{aligned}
E(Z) &= \sum_{z=r}^{\infty} z \cdot \binom{z-1}{r-1} p^r (1-p)^{z-r} \\
&= \sum_{z=r}^{\infty} z \cdot \frac{(z-1)!}{(r-1)!(z-r)!} p^r (1-p)^{z-r} \\
&= \frac{r}{p} \sum_{z=r}^{\infty} \frac{z!}{r!(z-r)!} p^{r+1} (1-p)^{z-r} \\
&\overset{z'=z+1, r'=r+1}{=} \frac{r}{p} \sum_{z'=r+1}^{\infty} \frac{(z'-1)!}{(r'-1)!(z'-r')!} p^{r'} (1-p)^{z'-r'} \\
&\overset{NB(r+1,p)\text{的质量函数和为 } 1}{=} \frac{r}{p}
\end{aligned}$$

同理可计算其方差。

- 期望：$\mathbb{E}(Z) = \frac{r}{p}$
- 方差：$\mathrm{Var}(Z) = \frac{r(1-p)}{p^2}$

3.1.5 几何分布

几何分布（**Geometric Distribution**）是最简单形式的负二项分布，如果一个随机变量 $V \sim NB(1,p)$，则随机变量 V 服从几何分布：

$$P(V = v|p) = p(1-p)^{v-1}$$

我们记 $V \sim G(p)$。几何分布具有无记忆性，即：

$$P(V > s|V > t) = P(V > s-t), s > t$$

即在给定我们为了等待成功已经等了 t 次的情况下，还需要等待的次数跟已经等待的次数是没有关系的。比如北京的车牌摇号，10 次没有中签的人跟 100 次没有中签的人，其需要继续等待时间的分布是一样的。因而此分布不适于建模带有生命长度的问题。

- 期望：$\mathbb{E}(V) = \frac{1}{p}$
- 方差：$\mathrm{Var}(V) = \frac{1-p}{p^2}$

3.1.6 超几何分布

即在一个 N 个试验组成的总体中，有已知 K 次成功，从这 N 个总体中抽取出 n 个样本，抽到的成功次数 $M = k$ 的概率。我们称随机变量 M 服从**超几**

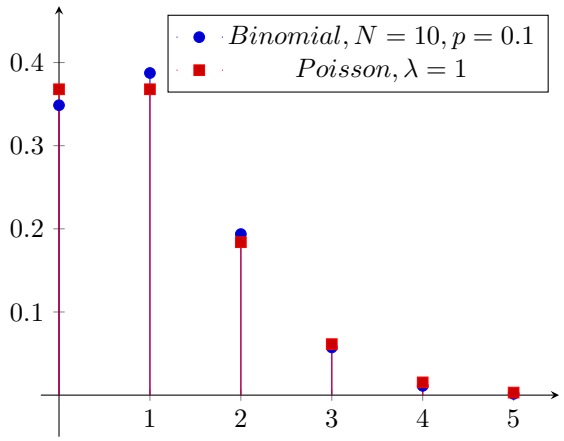


图 5: 泊松分布与二项分布

何分布（**Hypergeometric Distribution**），其概率质量函数为：

$$P(M=k|N,K,n)=\frac{\begin{pmatrix}K\\k\end{pmatrix}\begin{pmatrix}N-K\\n-k\end{pmatrix}}{\begin{pmatrix}N\\n\end{pmatrix}}$$

- 期望：$\mathbb{E}(M)=\frac{n}{N}\cdot K$
- 方差：$\mathrm{Var}(M)=n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}$

3.1.7 泊松分布

在第一节中我们研究了到达次数问题，相应的，如果随机变量 X 的概率质量函数为：

$$P(X=x)=\frac{\lambda^x}{x!}e^{-\lambda},x=0,1,2,...$$

那么我们称随机变量 X 服从**泊松分布**（**Poisson Distribution**），记为 $X\sim P(\lambda)$。泊松分布经常被用来建模次数问题。

- 期望：$\mathbb{E}(X)=\lambda$
- 方差：$\mathrm{Var}(X)=\lambda$

根据第一节的例题中，我们已经知道如果一个随机变量 $Y\sim Bi(N,p)$，令 $\lambda=Np$，并令 $N\to\infty$，那么随机变量 Y 近似服从泊松分布，因而对于相对较大的 N，可以使用泊松分布去近似二项分布。

例 18. 如果一个人打字出错的概率为 $p=0.001$，那么他写 1000 字出错的概率应该服从二项分布。记出错的字数为 Y，则 $Y\sim Bi(1000,0.001)$。比如写 1000

字出错两个以内的概率为::

$$P(Y \leq 2) = \sum_{y=0}^{2} \begin{pmatrix} 1000 \\ y \end{pmatrix} 0.001^y 0.999^{1000-y} \approx 0.91979$$

以上计算比较繁琐，如果使用泊松逼近，令 $X \sim P(1)$:

$$P(Y \leq 2) \approx P(X \leq 2) = \sum_{y=0}^{2} \frac{1^y}{y!} e^{-1} = 0.91970$$

两者非常接近。

3.2 连续分布

3.2.1 均匀分布

如果随机变量 X 在区间 $[a,b]$ 内的概率密度函数为常数，即:

$$f(x|a,b) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & otherwise \end{cases}$$

则称随机变量 X 服从区间 $[a,b]$ 上的**均匀分布**（**Uniform Distribution**），记为 $X \sim U(a,b)$。

- 分布函数:
$$F(x|a,b) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a,b] \\ 1 & x > b \end{cases}$$
- 期望: $\mathbb{E}(X) = \frac{a+b}{2}$
- 方差: $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$

3.2.2 指数分布

若随机变量 X 的概率密度函数为:

$$f(x|a,\beta) = \frac{1}{\beta} e^{-\frac{x-a}{\beta}}, x > a$$

那么我们称 X 服从**指数分布**（**Exponential Distribution**），其密度函数如图(6) 所示，记为 $X \sim E(a,\beta)$。

- 分布函数: $F(x) = 1 - e^{-\frac{x-a}{\beta}}, x > a$
- 期望: $\mathbb{E}(X) = \beta + a$

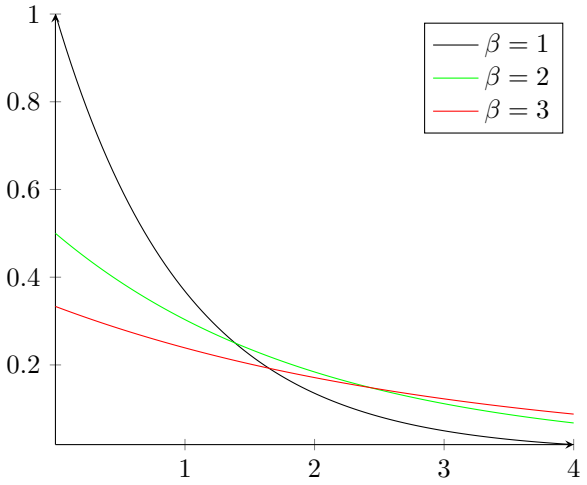


图 6: 指数分布密度函数

- 方差：$\mathrm{Var}(X)=\beta^2$

例 19. （指数分布）如果一个随机变量 $Y\sim P(\lambda)$，其中参数 λ 代表一段时间 T 内到达的平均次数，所以对于时间 t 内，平均到达的人数即 $\lambda\frac{t}{T}$，因而在 t 时间内只有 0 次到达的概率为 $e^{-\frac{\lambda t}{T}}$。如果在时间 t 内有 0 次到达，意味着等待时间大于 t，因而等待时间 X 大于 t 的概率 $P(X>t)=e^{-\frac{\lambda t}{T}}$，所以 $P(X\leq t)=1-e^{-\frac{\lambda t}{T}}$，因而等待时间 X 服从指数分布，即 $X\sim E\left(0,\frac{T}{\lambda}\right)$。

同样，指数分布也具有无记忆性，即如果令 $\beta=\frac{T}{\lambda}$，那么 $X\sim E(0,\beta)$，那么

$$P(X>s|X>t)=P(X>s-t),s>t$$

意味着为了等待第一个人到达，如果已经等了 t 时间，那么继续等待的时间仍然服从指数分布：给定已经等了五分钟的情况下，继续等待五分钟有到达的概率与第一个五分钟有到达的概率是一样的。

现在考虑，如果一件产品的使用寿命为 T，其分布函数为 $F(t)$，那么寿命大于 t 的概率为 $S(t)=1-F(t)$，我们称之为**生存函数**（**Survival function**）。进一步，我们可以研究产品「死亡」的风险。如果一件产品已经使用了时间 t，在 $(t,t+dt)$ 一小段时间内死亡的概率即为在这一小段时间内死亡的「风险」，而如果令 $dt\to 0$，可以认为是瞬时的死亡风险。可以计算：

$$\lambda(t)=\lim_{dt\to 0}\frac{P(t\leq T<t+dt)}{dt\cdot S(t)}=\frac{f(t)}{S(t)}=-\frac{S'(t)}{S(t)}=-\frac{d}{dt}\ln S(t)$$

我们称 $\lambda(t)$ 为**风险函数**（**hazard function**）。指数分布的风险函数为：

$$\lambda(t)=-\frac{d}{dt}\ln S(t)=-\frac{d}{dt}\ln\left(e^{-\frac{x}{\beta}}\right)=\frac{1}{\beta}$$

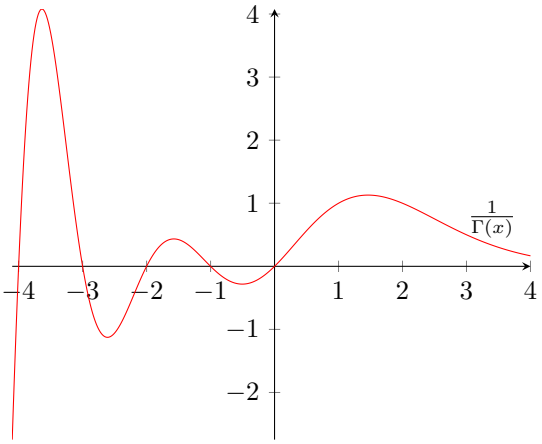


图 7: $\Gamma^{-1}(\alpha)$

因而指数函数的风险函数为常数，即其死亡（到达）的可能性不随着时间的增加而增加（或者减少）。

3.2.3 伽马分布

在介绍伽马分布之前，我们需要先介绍伽马函数（Γ function）。我们定义 Γ 函数为：

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$$

易知当 $\alpha > 0$ 时，$\Gamma(\alpha) < \infty$，而当 $\alpha \leq 0$ 时，该积分不一定有限。图 (7) 展示了 Γ 函数的逆函数。

根据 Γ 函数的定义，可知 $\Gamma(1) = 1$，而当 $\alpha > 0$ 时，有：

$$\begin{aligned}
\Gamma(\alpha+1) &= \int_0^\infty t^\alpha e^{-t}dt \\
&= -\int_0^\infty t^\alpha de^{-t} \\
&= -t^\alpha e^{-t}\big|_0^\infty + \int_0^\infty e^{-t}dt^\alpha \\
&= \alpha\int_0^\infty t^{\alpha-1}e^{-t}dt \\
&= \alpha\Gamma(\alpha)
\end{aligned}$$

结合 $\Gamma(1) = 1$，可以得到对于任意的正整数 n，$\Gamma(n) = (n-1)!$。此外，$\Gamma\left(\frac{1}{2}\right) =$

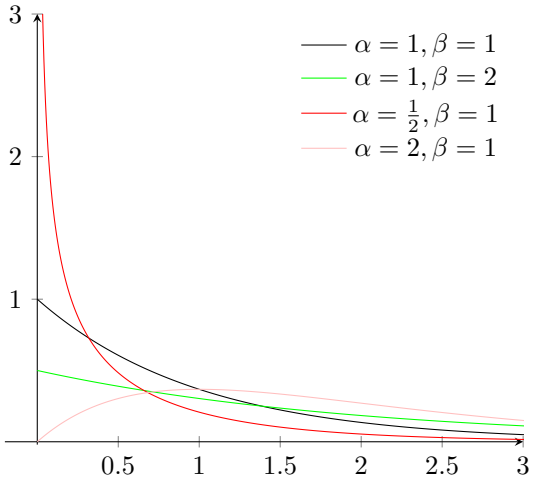


图 8: $\Gamma(\alpha, \beta)$ 分布的密度函数

$\int_0^\infty \frac{1}{\sqrt{t}} e^{-t} dt = \sqrt{\pi}$，因而：

$$\begin{aligned} \int_0^\infty \frac{1}{\sqrt{t}} e^{-t} dt & \overset{t=\frac{x^2}{2}}{=} \int_0^\infty \frac{\sqrt{2}}{x} e^{-\frac{x^2}{2}} d\frac{x^2}{2} \\ & = \sqrt{2} \int_0^\infty e^{-\frac{x^2}{2}} dx \\ & = \sqrt{\pi} \end{aligned}$$

因而

$$\int_0^\infty e^{-\frac{x^2}{2}} dx = \sqrt{\frac{\pi}{2}} \tag{2}$$

根据 Γ 函数的定义，函数：

$$f(t|\alpha) = \frac{t^{\alpha-1} e^{-t}}{\Gamma(\alpha)}, t > 0$$

为一个密度函数。如果令 $T \sim f(t|\alpha)$，那么对于任意的 $\beta > 0$，我们称 $X = \beta T$ 服从**伽马分布**（**Gamma Distribution**），记为 $X \sim \Gamma(\alpha, \beta)$，其密度函数为：

$$f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha) \beta^\alpha} x^{\alpha-1} e^{-x/\beta}, x > 0, \alpha > 0, \beta > 0$$

图 (8) 展示了不同参数下 $\Gamma(\alpha, \beta)$ 分布的密度函数。可知，指数分布为 $\alpha = 1$ 时

Γ 分布的特例。其期望：

$$\begin{aligned} E(X) &= \int_0^\infty x f(x|\alpha,\beta)\,dx \\ &= \int_0^\infty x \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^\alpha e^{-x/\beta} dx \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \Gamma(\alpha+1)\beta^{\alpha+1} \\ &= \alpha\beta \end{aligned}$$

类似地，可以计算其方差：

- 期望：$\mathbb{E}(X) = \alpha\beta$
- 方差：$\mathrm{Var}(X) = \alpha\beta^2$

3.2.4 Beta 分布

如果一个随机变量 $X \in (0,1)$，其分布函数为：

$$f(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0<x<1, \alpha>0, \beta>0$$

那么我们称 X 服从 **Beta 分布**（**Beta Distribution**）。其中

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

图 (9) 给出了 Beta 分布的密度函数。当 $\alpha>1, \beta=1$ 时，Beta 分布的密度函数为单调递增的；当 $\alpha=1, \beta>1$ 时，Beta 分布的密度函数为单调递减的；当 $\alpha<1, \beta<1$ 时，Beta 分布的密度函数为 U 型的；当 $\alpha>1, \beta>1$ 时，Beta 分布的密度函数为钟型的；当 $\alpha=1, \beta=1$ 时，Beta 分布的密度函数退化为均匀分布。由于 Beta 分布的取值范围在 $(0,1)$ 范围内，因而 Beta 分布经常被用于研究比率等问题，或者在贝叶斯分析中作为概率的先验。

Beta 分布的 r 阶矩可以计算为：

$$\begin{aligned} E(X^r) &= \frac{1}{B(\alpha,\beta)} \int_0^1 x^r x^{\alpha-1}(1-x)^{\beta-1}\,dx \\ &= \frac{1}{B(\alpha,\beta)} \int_0^1 x^{r+\alpha-1}(1-x)^{\beta-1}\,dx \\ &= \frac{B(r+\alpha,\beta)}{B(\alpha,\beta)} \frac{1}{B(r+\alpha,\beta)} \int_0^1 x^{r+\alpha-1}(1-x)^{\beta-1}\,dx \\ &= \frac{B(r+\alpha,\beta)}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+r)\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+r)\Gamma(\alpha)} \end{aligned}$$

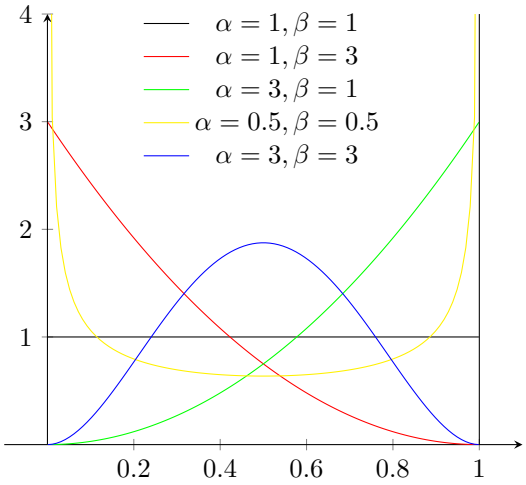


图 9: 指数分布密度函数

从而:

- 期望: $\mathbb{E}(X) = \frac{\alpha}{\beta+a}$
- 方差: $\mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

3.2.5 正态分布

如果一个随机变量 X 潜在受到非常多的独立因素的影响，即 $X = f_1 + f_2 + \cdots$，而每个 f_i 又不能单独对 X 有非常大的影响，那么一般来说 X 将会服从**正态分布**（**Normal Distribution**）或者**高斯分布**（**Gaussian Distribution**）。正态分布的密度函数为:

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

记为 $X \sim N(\mu, \sigma^2)$。

- 期望: $\mathbb{E}(X) = \mu$
- 方差: $\mathrm{Var}(X) = \sigma^2$

如果 $X \sim N(\mu, \sigma^2)$，如果令 $Z = \frac{X-\mu}{\sigma}$，则 $E(Z) = \frac{E(X)-\mu}{\sigma} = 0$，$\mathrm{Var}(Z) = \frac{1}{\sigma^2} \cdot \mathrm{Var}(X) = 1$，随机变量 Z 服从 $\mu = 0, \sigma = 1$ 的正态分布，即 $Z \sim N(0, 1)$，我们称之为**标准正态分布**（**Standard Normal Distribution**）。标准正态分布的密度函数为:

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

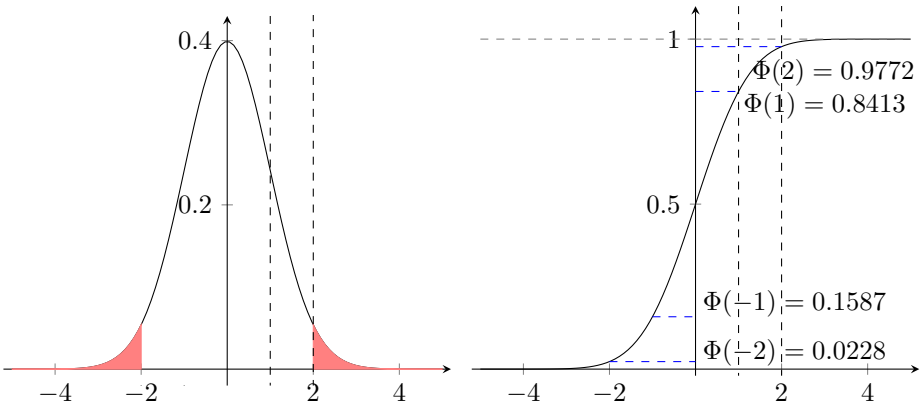


图 10: 标准正态分布密度函数与分布函数

根据公式 (2) 可知，其密度函数在 $\mathbb{R}$ 上的积分为 1。而其分布函数为:

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)\,dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

由于 $\Phi(x)$ 没有初等函数表示，因而一般我们用 $\Phi(z)$ 来代表标准正态的分布函数。

图 (10) 列出了标准正态分布的密度函数以及分布函数。由于正态分布为对称分布，即 $\phi(x) = \phi(-x)$，因而分布函数 $\Phi(x) = 1 - \Phi(-x)$。如果 $X \sim N(\mu, \sigma^2)$，那么 $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$，根据标准正态的分布函数 $\Phi(x)$ 可以计算 X 在区间内取值的概率，比如:

$$P(|X-\mu| \le \sigma) = P(|Z| < 1) = \Phi(1) - \Phi(-1) = 2\Phi(1) - 1 = 0.6827$$

同理，有:

$$\begin{aligned}
P(|X-\mu| \le 1.65\sigma) &= P(|Z| < 1.65) \approx 0.90 \\
P(|X-\mu| \le 1.96\sigma) &= P(|Z| < 1.96) \approx 0.95 \\
P(|X-\mu| \le 2\sigma) &= P(|Z| < 2) = 0.9545 \\
P(|X-\mu| \le 2.58\sigma) &= P(|Z| < 2.58) \approx 0.99 \\
P(|X-\mu| \le 3\sigma) &= P(|Z| < 3) = 0.9973 \\
P(|X-\mu| \le 5\sigma) &= P(|Z| < 5) \ge 1 - 10^{-6} \\
P(|X-\mu| \le 6\sigma) &= P(|Z| < 6) \ge 1 - 10^{-8}
\end{aligned}$$

正态分布有很多优良的性质，因而在建模中是最经常被使用的一种分布。比如，如果两个随机变量 X, Y 独立，且 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$，那么 $V = X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$，即两个独立的正态分布的和仍然是正态分

布。在下一节多元随机变量中，我们还将介绍更多的正态分布独有的性质。

3.2.6 对数正态分布

若随机变量 $Y = e^X, X \sim N\left(\mu, \sigma^2\right)$，则我们称随机变量 Y 服从**对数正态分布**（**Lognormal Distribution**），记为 $Y \sim LN\left(\mu, \sigma^2\right)$。

- 期望：$\mathbb{E}(Y) = e^{\mu + \frac{\sigma^2}{2}}$
- 方差：$\mathrm{Var}(Y) = e^{2\left(\mu+\sigma^2\right)} - e^{2\mu+\sigma^2}$

3.2.7 逻辑斯蒂分布

若随机变量 X 的分布函数为：

$$F(x|\mu,\sigma) = \frac{e^{\frac{x-\mu}{\sigma}}}{1+e^{\frac{x-\mu}{\sigma}}}$$

那么我们称 X 服从**逻辑斯蒂分布**（**Logistic Distribution**），其密度函数为：

$$f(x|\mu,\sigma) = \frac{1}{\sigma}\frac{e^{-\frac{x-\mu}{\sigma}}}{\left[1+e^{-\frac{x-\mu}{\sigma}}\right]^2}$$

记为：$X \sim LG(\mu,\sigma)$。Logistic 分布被广泛运用在**离散选择**（**Discrete Choice**）模型或者**分类器**（**Classifier**）中。

- 期望：$\mathbb{E}(X) = \mu$
- 方差：$\mathrm{Var}(X) = \sigma^2\frac{\pi^2}{3}$

3.2.8 柯西分布

若随机变量 X 的概率密度函数为：

$$f(x|\mu,\sigma) = \frac{1}{\pi\sigma}\left[1+\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-1}$$

那么我们称 X 服从**柯西分布**（**Cauchy Distribution**），记为 $X \sim C(\mu,\sigma)$。

- 期望：不存在
- 方差：不存在

由于柯西分布的一阶矩不存在，因而通常作为不可积的分布的例子。此外，可以得到，两个独立的标准正态分布 X, Y，其比例 $U = \frac{Y}{X}$ 刚好服从柯西分布。

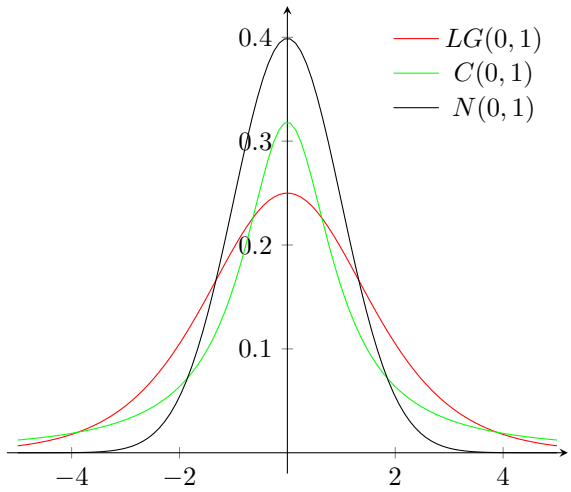


图 11: Logistic 分布、Cauchy 分布与正态分布密度函数

3.2.9 χ^2 分布

K 个独立的正态分布的平方和的分布被称为**卡方分布**（**Chi-square Distribution**）或者 χ^2 **分布**。即，如果 $X_1, X_2, ..., X_K$ 为 K 个**独立**的**标准**正态分布，那么

$$X = \sum_{i=1}^{K} X_i^2 \sim \chi^2(K)$$

其中参数 K 为卡方分布的**自由度**（**degrees of freedom**）。χ^2 分布的密度函数为：

$$f(x|K) = \frac{1}{\Gamma(K/2)\, 2^{K/2}} x^{K/2-1} e^{-x/2}, x > 0$$

因而，χ^2 分布实际是 Γ 分布的特殊形式（$\alpha = K/2, \beta = 2$）。

- 期望：$\mathbb{E}(X) = K$
- 方差：$\mathrm{Var}(X) = 2K$

χ^2 分布在假设检验中非常常用。

3.2.10 学生氏 t 分布

如果存在一个标准正态分布 $Z \sim N(0,1)$，以及一个 χ^2 分布 $X \sim \chi^2(K)$，且 Z 和 X **独立**，那么随机变量 $T = \frac{Z}{\sqrt{X/K}}$ 即服从**学生氏** t **分布**（**Students'**t**-distribution**）或者简称 t **分布**（t**-distribution**），记为 $T \sim t(K)$，参数 K 为**自由度**。t 分布的密度函数为：

$$f(x|K) = \frac{\Gamma[(K+1)/2]}{\sqrt{K\pi}\Gamma(K/2)} \left(1 + \frac{x^2}{K}\right)^{-\frac{K+1}{2}} = \frac{1}{\sqrt{K} B(1/2, K/2)} \left(1 + \frac{x^2}{K}\right)^{-\frac{K+1}{2}}$$

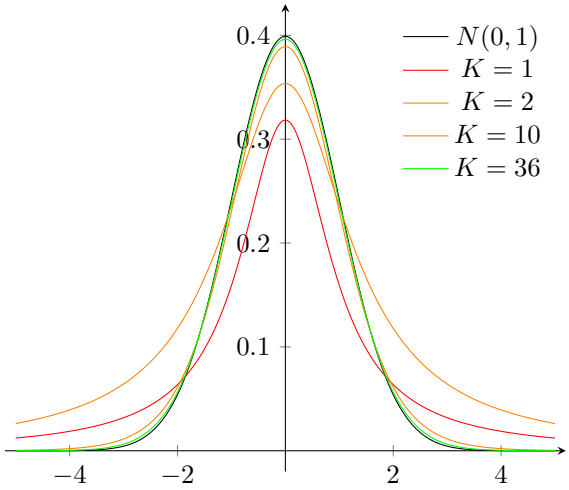


图 12: t 分布与正态分布密度函数

当 $K=1$ 时，t 分布即退化为柯西分布，而当 $K\to\infty$ 时，t 分布趋向于正态分布。图 (12) 显示了不同自由度下的 t 分布与正态分布的比较。

- 期望：$\mathbb{E}(X)=0, K>1$
- 方差：$\mathrm{Var}(X)=\frac{K}{K-2}, K>2$

t 分布在假设检验中也扮演着至关重要的地位。

3.2.11 F 分布

如果存在两个**独立**的 χ^2 分布 $X_1\sim\chi^2(n), X_2\sim\chi^2(m)$，那么随机变量 $F=\frac{X_1/n}{X_2/m}$ 即服从 F **分布**（F-**distribution**），记为 $T\sim F(n,m)$，参数 n,m 为**自由度**。F 分布的密度函数为：

$$f(x|K)=\frac{n^{n/2}m^{m/2}\Gamma\left[(n+m)/2\right]}{\Gamma(n/2)\Gamma(m/2)(m+nx)^{(n+m)/2}}x^{\frac{n}{2}-1}, x>0$$

图 (13) 显示了不同自由度下的 χ^2 分布与 F 分布的密度函数。

- 期望：$\mathbb{E}(X)=\frac{m}{m-2}, m>2$
- 方差：$\mathrm{Var}(X)=\frac{2m^2(m+n-2)}{n(m-2)^2(m-4)}, m>4$

F 分布与其他多种分布有关系：

- 如果 $X_k\sim\Gamma(\alpha_k,\beta_k)$ 且独立，那么 $\frac{\alpha_2\beta_1X_1}{\alpha_1\beta_2X_2}\sim F(2\alpha_1,2\alpha_2)$
- 如果 $X\sim B(n/2,m/2)$，那么 $\frac{mX}{n(1-X)}\sim F(n,m)$，反之亦成立
- 如果 $X\sim F(n,m)$，那么 $\lim_{m\to\infty}nX\sim\chi^2(n)$

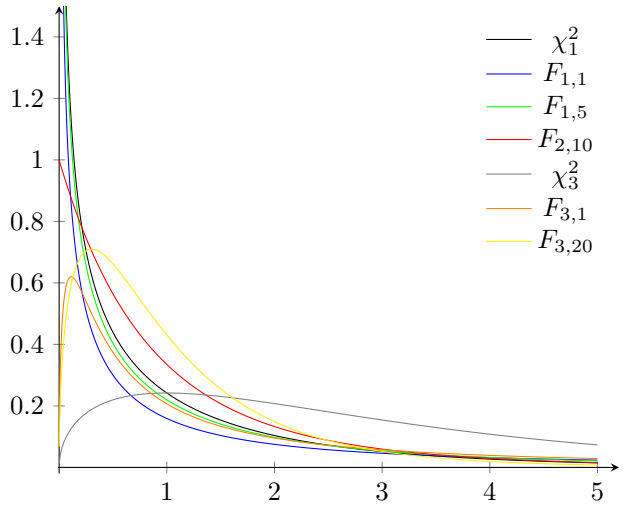


图 13: F 分布与 χ^2 分布密度函数

- 如果 $T \sim t(K)$，那么 $T^2 \sim F(1, K)$

3.3 分布族

给定任意的一个 $(\Omega, \mathcal{F})$，在这个空间里面我们可以定义各种不同的概率函数，统计学需要解决的问题就是使用样本数据去推断总体的概率函数 $\mathscr{P}$。然而一般情况下，概率函数 $\mathscr{P}$ 的可能性太多，因而我们经常把研究的重点放在一个概率函数的可能集合内，并在此集合内部对总体的概率函数做出推断。为此我们可以定义参数族的概念。

定义 8. 如果对于每一个已知的 $\theta \in \Theta, \Theta \subset \mathbb{R}^d$，$P_\theta$ 为在 $(\Omega, \mathcal{F})$ 上的一个已知的概率函数，那么 $\{P_\theta, \theta \in \Theta\}$ 即被称为**参数族**（**Parametric family**），其中 Θ 为**参数空间**（**Parametric space**），正整数 d 为参数空间的维数。

其中**位置尺度族**（**Location-scale families**）和**指数分布族**（**Exponent families**）是两类最为特殊且重要的参数族，我们这节将分别介绍这两类分布族。

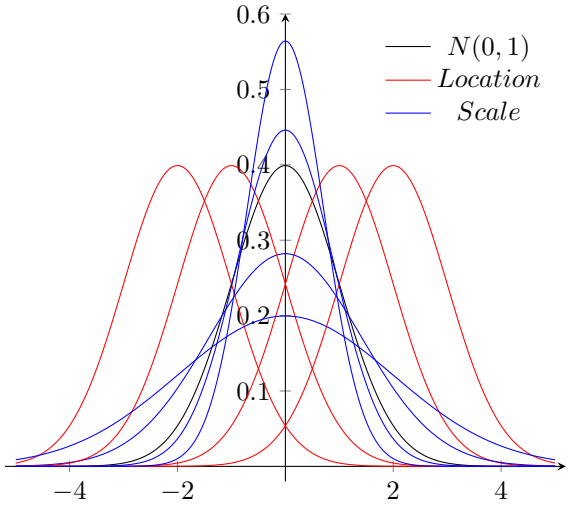


图 14: 正态分布的位置尺度族的密度函数

3.3.1 位置尺度族

对于一个随机变量 X，我们令 $Y = \sigma X + \mu, \sigma > 0$，那么其分布函数：

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(\sigma X + \mu \leq y) \\ &= P\left(X \leq \frac{y-\mu}{\sigma}\right) \\ &= F_X\left(\frac{y-\mu}{\sigma}\right) \end{aligned}$$

所以其密度函数满足：

$$f_Y(y) = \frac{1}{\sigma} f_X\left(\frac{x-\mu}{\sigma}\right)$$

同时，其期望和方差满足：$\mathbb{E}(Y) = \sigma\mathbb{E}(X) + \mu$，$\mathrm{Var}(Y) = \sigma^2 \mathrm{Var}(X)$。

令 $f(x)$ 为任意的密度函数，那么形如 $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$ 的密度函数就形成了以 $\theta = (\mu, \sigma)$ 为参数的参数族，我们称之为位置尺度族。位置尺度族即对于任意的随机变量做位移和数乘所得到的随机变量的分布组成的分布族。其中参数 μ 一般称为**位置参数**（**Location parameter**），而 σ 一般称为**尺度参数**（**Scale parameter**）。

例 20. 标准正态分布的密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

对于任意的 $\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$，其位置尺度族的密度函数可以写为：

$$f(x|\mu,\sigma) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma}}$$

即正态分布的密度函数。因而正态分布族 $\{P_{\mu,\sigma}, \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$ 为一个位置尺度族。图 (14) 展示了正态分布的位置族和尺度族。

上面介绍的包括指数分布、Γ 分布等都是各种类型的位置尺度族。正如正态分布一样，很多时候我们会将一个随机变量变换为期望为 0、方差为 1 的随机变量，即令随机变量 $Z = \frac{X-\mu}{\sigma}$，易得 $\mathbb{E}(Z) = 0, \mathrm{Var}(Z) = 1$，我们称之为**标准化**（**standardized**）。由于 $P(X \le x) = P(Z \le \sigma x + \mu)$，因而应用中对于位置尺度族，经常首先研究标准化之后的随机变量，进而推广到其位置尺度族。

3.3.2 单参数指数分布族

指数分布族是非常常用的分布族，其包含了我们上面介绍的多数分布。在此我们首先讨论单参数的指数分布族。其定义如下：

定义 9. （指数分布族）对于一个参数族 $\{P_\theta, \theta \in \Theta\}$，如果其概率密度（质量）函数可以写成如下形式：

$$f(x|\theta) = h(x) \cdot \exp\{\eta(\theta) \cdot T(x) - B(\theta)\} \tag{3}$$

那么我们称 $\{P_\theta, \theta \in \Theta\}$ 为**单参数指数分布族**（**One-parameter exponential family**）。

由于很多随机变量的取值范围不是 $\mathbb{R}$，然而随机变量的值域为 $\mathbb{R}$，因而我们一般使用指示函数（indicator function）来表示随机变量的取值范围，即定义

$$1_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

例如 $1_{(-\infty,0]}(x)$ 即当 $x <= 0$ 时，$1_{(-\infty,0]}(x) = 1$；当 $x > 0$ 时，$1_{(-\infty,0]}(x) = 0$。

我们之前遇到的很多分布都属于指数分布族。例如：

例 21. 令 N 为已知的正整数，如果随机变量 $X \sim Bi(N,p)$，参数 $\theta = p \in \Theta =$

$(0,1)$，其质量函数：

$$\begin{aligned} P(x|p) &= \left(\begin{array}{c} N \\ x \end{array} \right) p^x (1-p)^{N-x} \\ &= \left(\begin{array}{c} N \\ x \end{array} \right) \exp\{x \ln(p) + (N-x)\ln(1-p)\} \\ &= \left(\begin{array}{c} N \\ x \end{array} \right) \exp\{x[\ln(p) - \ln(1-p)] + N\ln(1-p)\} \\ &= \left(\begin{array}{c} N \\ x \end{array} \right) \exp\left\{x\left[\ln\left(\frac{p}{1-p}\right)\right] + N\ln(1-p)\right\} \end{aligned}$$

故令 $h(x) = \left(\begin{array}{c} N \\ x \end{array} \right)$，$\eta(\theta) = \ln\left(\frac{p}{1-p}\right)$，$T(x) = x$，$B(\theta) = -N\ln(1-p)$，所以二项分布属于指数分布族。

例 22. 泊松分布参数为 $\theta = \lambda \in \Theta = (0,\infty)$，其质量函数为：

$$\begin{aligned} P(x|\lambda) &= \frac{\lambda^x}{x!} e^{-\lambda} \\ &= \frac{1}{x!} \exp\{x\ln(\lambda) - \lambda\} \end{aligned}$$

故令 $h(x) = \frac{1}{x!}$，$\eta(\theta) = \ln(\lambda)$，$T(x) = x$，$B(\theta) = \lambda$，所以泊松分布属于指数分布族。

例 23. 指数分布的密度函数为：

$$\begin{aligned} f(x|\beta) &= \frac{1}{\beta} e^{-\frac{x}{\beta}} \cdot 1_{(0,\infty)}(x) \\ &= 1_{(0,\infty)}(x) \exp\left\{-x \cdot \frac{1}{\beta} - \ln\beta\right\} \end{aligned}$$

因而令 $h(x) = 1_{(0,\infty)}(x)$，$\eta(\theta) = -\frac{1}{\beta}$，$T(x) = x$，$B(\theta) = \ln\beta$，故指数分布属于指数分布族。

然而并非所有带指数的密度函数都属于指数分布族，比如：

例 24. 若某一密度函数为：

$$f(x|\beta) = \frac{1}{\beta} \exp\left\{1 - \frac{x}{\beta}\right\}, x > \beta > 0$$

可知 $f(x|\beta)$ 为密度函数。然而：

$$\begin{aligned} f(x|\beta) &= \frac{1}{\beta}\exp\left\{1-\frac{x}{\beta}\right\}\cdot 1_{(\beta,\infty)}(x) \\ &= 1_{(\beta,\infty)}(x)\cdot\exp\left\{-\frac{x}{\beta}-\ln\beta+1\right\} \end{aligned}$$

由于 $1_{(\beta,\infty)}(x)$ 不仅仅依赖于 x，而且依赖于 β，因而这一分布不属于指数分布族。

很多时候为了方便起见，我们会把密度函数重新参数化，即对于指数分布族

$$f(x|\theta)=h(x)\cdot\exp\{\eta(\theta)\cdot T(x)-B(\theta)\}$$

我们令 $\lambda=\eta(\theta)$，那么指数分布族可以写为：

$$f(x|\lambda)=h(x)\cdot\exp\{\lambda\cdot T(x)-C(\lambda)\} \tag{4}$$

我们将指数分布族重新参数化为式 (4) 的形式，并将这种形式成为**规范形式**（**Canonical form**）。

例 25. 在例 (21) 中，如果令 $\lambda=\ln\left(\frac{p}{1-p}\right)$，那么

$$P(x|\lambda)=\begin{pmatrix} N \\ x \end{pmatrix}\exp\left\{\lambda\cdot x-N\ln\left(1+e^{\lambda}\right)\right\}$$

即为二项分布的质量函数的规范形式。

对于指数分布族的规范形式，我们有如下定理：

定理 9. 如果随机变量 X 的概率密度（质量）函数为式 *(4)* 的形式，那么 $\mathbb{E}[T(X)]=C'(\lambda)$，$Var[T(X)]=C''(\lambda)$。

证明. 对于规范形式 $f(x|\lambda)=h(x)\cdot\exp\{\lambda\cdot T(x)-C(\lambda)\}$，由于其为密度函数，因而：

$$1=\int_{\mathbb{R}} f(x|\lambda)\,dx=\int_{\mathbb{R}} h(x)\cdot\exp\{\lambda\cdot T(x)-C(\lambda)\}\,dx$$

从而：

$$e^{C(\lambda)}=\int_{\mathbb{R}} h(x)\cdot\exp\{\lambda\cdot T(x)\}\,dx$$

两边对 λ 求导可得:

$$\begin{aligned} e^{C(\lambda)}C'(\lambda) &= \frac{\partial}{\partial\lambda}\int_{\mathbb{R}} h(x)\cdot\exp\{\lambda\cdot T(x)\}\,dx \\ &= \int_{\mathbb{R}} h(x)\cdot\frac{\partial}{\partial\lambda}\exp\{\lambda\cdot T(x)\}\,dx \\ &= \int_{\mathbb{R}} h(x)\cdot T(x)\exp\{\lambda\cdot T(x)\}\,dx \end{aligned}$$

整理可得:

$$C'(\lambda) = \int_{\mathbb{R}} T(x)\cdot h(x)\cdot\exp\{\lambda\cdot T(x) - C(\lambda)\}\,dx = \mathbb{E}[T(X)]$$

方差可以类似证明。 □

例 26. 在例 (25) 中, $T(x) = x$, 因而 $\mathbb{E}[T(X)] = \mathbb{E}(X) = C'(\lambda) = \frac{Ne^{\lambda}}{1+e^{\lambda}} = Np$, $\mathrm{Var}[T(X)] = \mathrm{Var}(X) = C''(\lambda) = \frac{Ne^{\lambda}}{(1+e^{\lambda})^2} = Np(1-p)$。

4 随机变量的变换

对于一个可测函数 $g(\cdot)\,\mathbb{R}\to\mathbb{R}$, $Y = g(X)$ 也是一个随机变量。如果我们已知 X 的分布, 如何获得新的随机变量 Y 的分布呢?

首先仿照随机变量的定义, 我们定义函数 $g(\cdot)$ 对于一个集合的逆为:

$$g^{-1}(A) = \{x\in\mathbb{R} : g(x)\in A\}$$

因而对于单点集 $g^{-1}(\{y\}) = \{x\in\mathbb{R} : g(x) = y\}$。

对于离散型的随机变量 X, $g(X)$ 也是离散型的, 因而随机变量 Y 的 p.m.f. 为:

$$f_Y(y) = P(Y=y) = \sum_{x\in g^{-1}(\{y\})} P(X=x) = \sum_{x\in g^{-1}(\{y\})} f_X(x)$$

由此可以确定随机变量 $Y = g(X)$ 的概率质量函数。

例 27. 若假设 X 为一次抛骰子的随机试验的结果, 其概率质量函数为:

X	1	2	3	4	5	6
$P(X=x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

如果我们希望得到 $Y = g(X) = (X-3.5)^2$ 的概率质量函数, 可以通过以上步骤, 比如:

$$g^{-1}(\{6.25\}) = \left\{x\in\mathbb{R} : (x-3.5)^2 = 6.25\right\} = \{1,6\}$$

因而 $P(Y=6.25) = P(X\in\{1,6\}) = \frac{1}{3}$。以此类推, 我们得到 Y 的概率质量函数:

Y	6.25	2.25	0.25
$P(X=x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

对于一个一般的随机变量 X，随机变量 $Y=g(X)$ 的累积分布函数可以使用如下定义计算：

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(g(X) \le y) \\ &= P(\{x : g(x) \le y\}) \\ &= \int_{\{x:g(x)\le y\}} dF_X \end{aligned}$$

特别的，如果 $g(\cdot)$ 严格单调递增，则：

$$F_Y(y) = F_X\left(g^{-1}(y)\right)$$

如果 $g(\cdot)$ 严格单调递减，则：

$$F_Y(y) = 1 - F_X\left(g^{-1}(y)\right)$$

例 28. 如果 $U \sim U(-1,1)$，欲求 $Y=U^2$ 的密度函数，可以首先求其累计分布函数，对于 $y \ge 0$，有：

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P\left(U^2 \le y\right) \\ &= P(\{-\sqrt{y} \le U \le \sqrt{y}\}) \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{2} du \\ &= \sqrt{y} \end{aligned}$$

因而其密度函数为：

$$f_Y(y) = \frac{1}{2\sqrt{y}}$$

5 随机数生成

随着计算机的不断发展，统计学发展了很多基于模拟（simulation）方法，比如蒙特卡洛法（Monte Carlo）、马尔可夫链蒙特卡洛方法（Markov Chain Monte Carlo, MCMC）等等。而这些方法的基础是使用计算机产生随机数。

5.1 均匀分布随机数生成

实际上，目前单独的计算机无法生成真正的随机数，但是计算机可以使用一些算法生成一些**伪随机数**（**pseudorandom number**），即本身这些数字不随机，而是根据一定的算法（比如 KISS 算法）生成的确定性的数字，只不过这些所生成的数字的统计特性接近于随机。一般这些算法都可以产生一些统计特性比较像均匀分布的随机数，而服从其他分布的随机数都由均匀分布来生成。

在绝大多数计算机语言中，都有生成均匀分布的指令，比如在 (ANSI) C 语言中，可以使用标准库 <stdlib.h> 中的 rand() 函数生成均匀分布。在伪随机数算法中，一般有一个所谓的「种子（seed）」，即如果给定这个数字，则接下来生成的一系列随机数在每次程序运行时都是一样的。在 C 语言中，使用 srand() 函数设置 seed，使用 rand() 函数生成一个 0 到常数 RAND_MAX 的整数，因而除以 RAND_MAX 即得到了 (0,1) 上的均匀分布（伪）随机数：

代码 1: C 语言中生成均匀分布

```c
// file: uniform.c
#include<stdio.h>
#include<stdlib.h> // 使用标准库<stdlib.h>, 包含rand()
int main(int argc, char const *argv[]) {
  // 打印出rand()所生成的最大整数
  printf("%d\n",RAND_MAX);
  // 设置seed
  srand(505);
  // 生成(0,1)上的10个随机数
  int i;
  for (i=0; i<5; ++i){
    double x=(double)rand()/RAND_MAX;
    printf("%f␣␣",x);
  }
  printf("\n");
  // 如果每次都设置seed生成(0,1)上的5个随机数
  //那么每次生成的随机数都相同
  for (i=0; i<5; ++i){
    srand(505);
    double x=(double)rand()/RAND_MAX;
    printf("%f␣␣",x);
  }
  printf("\n");
  return 0;
}
```

显示结果：

2147483647
0.592173 0.654869 0.385925 0.240955 0.727209
0.592173 0.592173 0.592173 0.592173 0.592173

在 Python 中，可以使用 Python 标准库的 random 包，或者 NumPy 的 random 包，比如：

代码 2: Python 中生成均匀分布

```python
#!/usr/bin/python3
## file: uniform.py

# 导入random包
import random as rd
# 设置seed
rd.seed(505)
# 获取5个均匀分布随机数
x=[rd.random() for i in range(5)]
print(x)
# 导入numpy中的random包
import numpy.random as nprd
# 设置seed
nprd.seed(505)
# 获取5个均匀分布随机数
x=nprd.random(5)
print(x)
```

显示结果：

[0.14960684073609387, 0.8443100623742851, 0.2805013532789453, 0.7016520386211358, 0.2938865835498684]
[0.8506181 0.29254118 0.22308853 0.19771466 0.37789012]

而在 Stata 中，可以使用 runiform() 函数来生成 $(0,1)$ 上的均匀分布：

代码 3: Stata 中生成均匀分布

```
// file: uniform.do
// 关闭—more—
set more off
// 清除工作区内所有数据
clear
// 设置样本量
```

```
set obs 100
// 设置seed
set seed 505
// 产生随机数
gen x=runiform()
```

运行以上命令后，就可以在数据浏览器中看到数据集中新产生了一个变量 x，其值服从 $(0,1)$ 上的均匀分布。

前面提到，使用计算机生成的所谓随机数并非真正的随机，而是伪随机，我们也许会担心其随机性是否满足。比如，如果一个随机数序列 $\{X_i, i = 1, ..., N\}$ 的确是随机的，那么 X_i 与 X_{i-1} 之间应该是独立的，两者之间不存在任何关系。换句话说，(X_{i-1}, X_i) 的实现应该均匀的铺满 $[0,1] \times [0,1]$ 的平面，在 Python 中，我们可以使用以下的图直观上检查前后两个随机数之间是否的确是无关的：

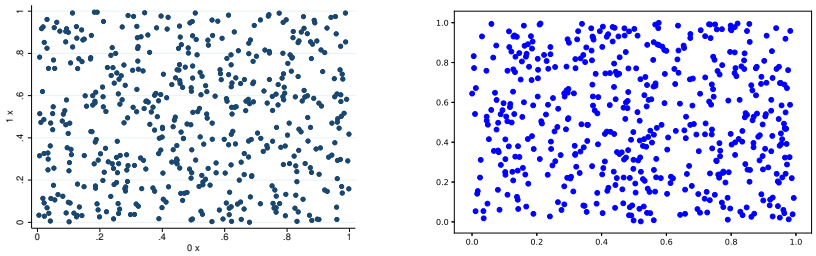
代码 4: Python 中生成均匀分布

```
#!/usr/bin/python3
## file: check_random.py
import numpy as np
import numpy.random as nprd
# 获取序贯的1000个均匀分布随机数
x=nprd.random(1000)
# 将1000个均匀分布随机数按照奇偶数分成两个序列x0 和x1
x0=np.array([x[i] for i in range(1000) if i%2==0])
x1=np.array([x[i] for i in range(1000) if i%2==1])
# 画图
import matplotlib.pyplot as plt
# 设定图像大小
plt.rcParams['figure.figsize'] = (8.0, 5.0)
plt.scatter(x0,x1,color='blue') ## 画出散点图
plt.savefig("check_random_py.eps")
```

在 Stata 中可以先产生一系列随机数，进而使用 reshape 命令将奇数和偶数分成两个变量：

代码 5: Stata 中生成均匀分布

```
// check_random.do
set more off
clear
set obs 1000
// 行号是奇数还是偶数
gen variable_id=mod(_n,2)
```



(1)Stata (2)Python

图 15: 检查随机数生成的随机性

```
gen group=ceil(_n/2)
// 产生序贯的随机数
gen x=runiform()
// 将奇数观测作为x1，偶数观测作为x0
reshape wide x, i(group) j(variable_id)
// 画散点图
scatter x1 x0 , graphr(fcolor(white) color(white))
// 保存图片
graph export check_random_stata.eps, replace
```

得到的结果图 (15) 所示。可以看到，生成的相邻的随机数对均匀的分布在 $[0,1]\times[0,1]$ 这个平面内，因而直观上，通过算法得到的伪随机数性质比较接近于随机。当然，这仅仅是一个直观的检查，我们还可以通过更多更严格的方法检验随机性，比如使用回归工具、检验分布的工具等等。

5.2 逆变换方法

现在，假设我们有一个随机变量 $U\sim Uniform(0,1)$，即 $(0,1)$ 区间上的均匀分布，而 $F(\cdot)$ 为一个分布函数，我们可以定义一个新的随机变量 $X=F^{-1}(U)$。

注意如果分布函数 $F(\cdot)$ 存在「平台」，即 $F(x)=c,\ for\ a\le x<b$，那么定义 $F^{-1}(u)=\inf\{x:F(x)\ge u\}$。对于一个存在「平台」的分布函数 F，$F(x)=c,\ for\ a\le x<b$，$P(X\le x)=P(X\le a),\ for\ a\le x<b$，也就是说如果某个随机变量的分布函数为 F，那么其取值在 $[a,b)$ 范围内的概率为 0。根

据以上推理，随机变量 $X = F^{-1}(U)$ 的分布函数为：

$$\begin{aligned} F_X(x) &= \int_{F^{-1}(u)\leq x} dG(u) \\ &= \int_{F^{-1}(u)\leq x} du \\ &= \int_{u\leq F(x)} du \\ &= F(x) \end{aligned}$$

其中第二个等号由于均匀分布的分布函数 $G(u) = u, 0 \leq u \leq 1$。这也就意味着，如果我们有分布函数 $F(\cdot)$，可以生成一个 $(0,1)$ 区间内的随机变量 U，进而生成 $X = F^{-1}(U)$，那么我们就生成了一个新的随机变量，其分布函数为 F。

例 29. （指数分布）若 $U \sim Uniform(0,1)$，令 $F(x) = 1 - e^{-x}, x > 0$，即指数分布的分布函数。令 $X = F^{-1}(U)$，则随机变量 X 的分布函数为：

$$\begin{aligned} F_X(x) &= \int_{F^{-1}(u)\leq x} dG(u) \\ &= \int_{F^{-1}(u)\leq x} du \\ &= \int_{u\leq(1-e^{-x})} du \\ &= 1 - e^{-x} \end{aligned}$$

因而为了生成指数分布的随机变量，只要生成均匀分布 U，并令 $U = F(X)$，由于：

$$1 - e^{-X} = U \Leftrightarrow X = -\ln(1-U)$$

从而，为了生成 $X \sim F(x)$，只要令 $X = -\ln(1-U)$ 即可。例如在 Python 中，我们可以使用如下代码产生指数分布的随机数：

代码 6: Python 中生成指数分布

```
#!/usr/bin/python3
## file: exponential.py

# 导入math包，使用math中的log()函数
import math
import random as rd
# 获取5个均匀分布随机数
x=[-1*math.log(rd.random()) for i in range(5)]
print(x)
# 导入numpy包，使用其中的log()函数
```

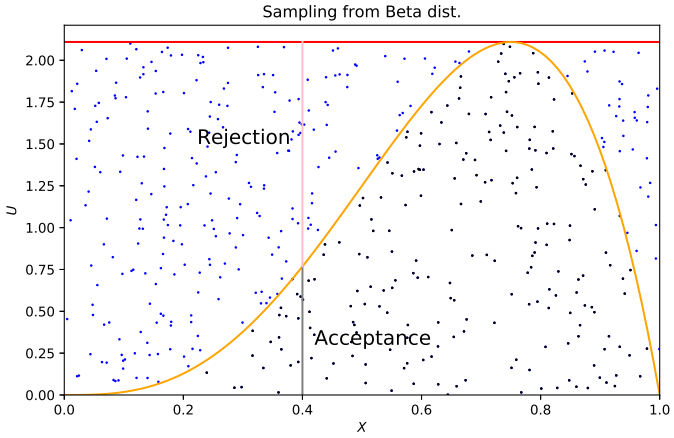


图 16: Beta 分布的拒绝采样法

```python
import numpy as np
import numpy.random as nprd
# 获取5个均匀分布随机数
x=-1*np.log(nprd.random(5))
print(x)
```

或者在 Stata 中：

代码 7: Stata 中生成指数分布

```
// file: exponential.do
set more off
clear
set obs 100
// 产生随机数
gen x=-1*log(runiform())
```

5.3 拒绝采样法

以上介绍了使用分布函数的逆函数对任意分布 $F(\cdot)$ 进行抽样的方法。然而很多时候，分布函数非常难以计算，分布函数的逆函数就更加难以计算，因而使用以上分布函数逆函数的方法进行抽样经常是非常困难的。然而随机变量的密度函数通常可以非常方便的计算出，那么能否使用密度函数完成特定分布的抽样呢？下例展示了如何使用密度函数对 Beta 分布进行抽样。

例 30. Beta 分布的密度函数为：

$$f(x|\alpha,\beta)=\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1},0<x<1$$

我们不妨假设 $\alpha>1,\beta>1$，因而其密度函数为先增后减的钟型。求以上密度函数的最大值，可以得到当 $x=\frac{1-\alpha}{2-\alpha-\beta}$ 时，密度函数取得最大值 M。图 (16) 给出了当 $\alpha=4,\beta=2$ 时的密度函数及其最大值。

现在考虑生成两个独立的均匀分布：$X\sim U(0,1),U\sim U(0,M)$，因而 (X,U) 就是在 $(0,1)\times(0,M)$ 二维平面的均匀分布，图 (16) 中所有的蓝色、黑色散点画出了 (X,U) 的抽样。

仿照直方图的想法，如果我们只保留密度函数以下的点（即图中黑色的点），那么这些黑色的点的直方图应该是任意逼近于其密度函数的，因而如果只保留图中黑色的点，那么 x 的分布应该就是服从 Beta 分布的。

为了证明以上直觉，我们不妨计算 $X|U\leq f(X)$ 的分布函数。由于 X 和 U 独立，因而密度函数 $f(x,u)=\frac{1}{M}$，从而：

$$\begin{aligned}F_{X|U\leq f(x|\alpha,\beta)}(x)&=P(X\leq x|U\leq f(X|\alpha,\beta))\\&=\frac{P(X\leq x,U\leq f(X|\alpha,\beta))}{P(U\leq f(X|\alpha,\beta))}\\&=\frac{\int_0^x\int_0^{f(t|\alpha,\beta)}\frac{1}{M}dudt}{\int_0^1\int_0^{f(t|\alpha,\beta)}\frac{1}{M}dudt}\\&=\frac{\frac{1}{M}\int_0^x f(t|\alpha,\beta)\,dt}{\frac{1}{M}\int_0^1 f(t|\alpha,\beta)\,dt}\\&=\int_0^x f(t|\alpha,\beta)\,dt\end{aligned}$$

注意以上得到的积分即 Beta 分布的分布函数，从而根据以上算法得到的 $X|U\leq f(X)$ 即服从 Beta 分布。

因而，我们可以用以下算法产生 Beta 分布的随机数：

1. 产生两个独立的均匀分布：$X\sim U(0,1),U\sim U(0,M)$，其实现为 x 和 u
2. 计算 $f(x|\alpha,\beta)$，如果 $u\leq f(x|\alpha,\beta)$ 则接受 x 为产生的随机变量；否则回到第 1 步。

以上算法的 Python 代码如下：

代码 8: Beta 分布的拒绝采样法

```
#!/usr/bin/python3
## file: rejection_beta.py
import numpy as np
import numpy.random as nprd
```

```
import scipy.special as scisp
# 设定参数
alpha=4
beta=2
# beta分布密度函数
f=lambda x: 1/scisp.beta(alpha,beta)* \
    x**(alpha-1) * (1-x)**(beta-1)
# 计算密度函数最大值
M=f((1-alpha)/(2-alpha-beta))
print(M)
# 随机抽两个均匀分布，一个为(0,1)，一个为(0,M)，抽500个
N=500
x=nprd.random(N)
u=nprd.random(N)*M
# 挑出使得u<beta密度函数的x
accepted=[i for i in range(N) if u[i]<=f(x[i])]
rand_beta=x[accepted] #生成的Beta分布随机数
rand_U_selected=u[accepted]
# 画图
import matplotlib.pyplot as plt
# 设定图像大小和坐标范围
plt.rcParams['figure.figsize'] = (8.0, 5.0)
plt.xlim(0,1)
plt.ylim(0,M+0.1)
# 横线和竖线
x_grid=np.linspace(0,1,100)#(0,1)均匀的100个点
y_M=np.ones(100)*M
beta_dens=f(x_grid)
y1=np.linspace(0,f(0.4),20)
y2=np.linspace(f(0.4),M,20)
x_hline=np.ones(20)*0.4
plt.xlabel(r'$X$')
plt.ylabel(r'$U$')
plt.title('Sampling from Beta dist.') # 标题
## 画出散点图
plt.scatter(x,u,color='blue',s=0.8)
plt.scatter(rand_beta,rand_U_selected,color='black',s=0.8)
plt.plot(x_grid,y_M,color='red') ## 最大值
plt.plot(x_grid,beta_dens,color='orange') ## 密度函数
plt.plot(x_hline,y1,color='grey') ## 接受区域
```

```
plt.text(0.42,0.3,"Acceptance",fontsize=15,
    horizontalalignment="left")
plt.plot(x_hline,y2,color='pink') ## 拒绝区域
plt.text(0.38,1.5,"Rejection",fontsize=15,
    horizontalalignment="right")
plt.savefig("rejection_beta.eps")
```

以上方法可以进一步推广，von Neumann (1961) 提出了一个非常一般化的使用密度函数进行抽样的方法，即拒绝采样法 (Rejection sampling)。其思想是，我们可以将以上过程中的 $X \sim U(0,1)$ 替换为其他分布，如果该分布的密度函数与欲抽样的密度函数差距变小，就可以大大增加接受的概率。

如果我们希望抽象随机数使其密度函数为 $\pi(x)$，令 $l(x) = c \cdot \pi(x)$，其中 c 可以是未知常数。如果我们可以从另一个密度函数 $g(x)$ 进行抽样，同时存在一个常数 M 使得：

$$Mg(x) \geq l(x)$$

那么我们可以使用如下算法对 $\pi(x)$ 进行抽样：

1. 从 $g(\cdot)$ 中得到抽样 x，并计算比例：

$$r(x) = \frac{l(x)}{Mg(x)} \leq 1$$

2. 以 $r(x)$ 的概率接受 x，否则拒绝 x 重新返回第一步继续抽样。即，产生一个 $U \sim (0,1)$ 上的均匀分布随机数 u，如果 $u \leq r(x)$，那么接受 x 作为抽样；否则重新返回第 1 步重新抽样。

由此得到的 x，其密度函数即为 $\pi(x)$。

在以上步骤中，如果记 $I = 1$ 为接受第 1 步的随机变量，那么给定 x，接受的概率为

$$P(I=1|X=x) = P(U \leq r(x)|X=x) = P\left(U \leq \frac{l(x)}{Mg(x)}\right) = \frac{l(x)}{Mg(x)} = \frac{c\pi(x)}{Mg(x)}$$

从而接受概率为（全概率公式）：

$$P(I=1) = \int_{\mathbb{R}} P(I=1|X=x)\, g(x)\, dx = \int_{\mathbb{R}} \frac{c\pi(x)}{Mg(x)} g(x)\, dx = \frac{c}{M}\int_{\mathbb{R}} \pi(x)\, dx = \frac{c}{M}$$

仿照之前对于 Beta 分布拒绝抽样的证明方法，我们可以计算给定接受 ($I = 1$)，生成的 x 的分布函数：

$$F_{X|I=1}(x) = \frac{P(X \leq x, I=1)}{P(I=1)} = \frac{\int_{-\infty}^{x} g(t)\left[\int_0^{\frac{c\pi(t)}{Mg(t)}} du\right] dt}{\frac{c}{M}} = \int_{-\infty}^{x} \pi(t)\, dt$$

从而接受的 x 其分布的密度函数即 $\pi(x)$。

注意由于接受的概率为 $\frac{c}{M}$，通常 c 是未知的，因而我们可以通过找到一个好的分布 $g(x)$，使得 M 尽量变小，从而增大接受的概率，从而在产生 $\pi(x)$ 分布样本数量一定的情况下，减少产生随机变量 x 的次数。

例 31. 如果我们希望抽取一个截断的正态分布（truncated normal distribution），即如果 $X^* \sim N(0,1)$，我们希望得到 X^* 的取值范围只在某个区间的随机变量。比如给定一个常数 $c>0$，我们希望得到 $X = X^*|X^* \geq c$。可以计算其分布函数为：

$$\begin{aligned} F_X(x) &= P(X^* < x | X^* > c) \\ &= \frac{\Phi(x) - \Phi(c)}{P(X^* > c)} \\ &= \begin{cases} \frac{\Phi(x)-\Phi(c)}{1-\Phi(c)} & x \geq c \\ 0 & x < c \end{cases} \end{aligned}$$

因而其密度函数为：

$$f_X(x) = \begin{cases} \frac{\phi(x)}{1-\Phi(c)} & x \geq c \\ 0 & x < c \end{cases}$$

首先，我们可以通过不断生成正态分布 x，如果 $x \geq c$，则接受该样本，否则拒绝该样本。由于 $P(X>c) = 1-\Phi(c)$，因而以上方法结构的概率为 $1-\Phi(c)$。

这个方法实际上就是一个拒绝采样，其中 $\pi(x) = f_X(x)$，$l(x) = \pi(x)$，因而 $c=1$；$g(x) = \phi(x)$，由于我们需要保证：

$$M \geq \frac{l(x)}{g(x)} = \frac{f_X(x)}{\phi(x)}$$

而 $\frac{f_X(x)}{\phi(x)}$ 的最大值为 $\frac{1}{1-\Phi(c)}$，因而 $M = \frac{1}{1-\Phi(c)}$。进而，平均的拒绝率为 $\frac{c}{M} = 1-\Phi(c)$。

当然，我们还可以使用其他的 $g(x)$。当 c 比较大时，正态分布的特性导致随机产生大于 c 的可能性非常低，因而我们可以通过一些比较厚尾的分布当做 $g(x)$，保证 $g(x)$ 能够以比较大的概率产生大于 c 的随机数，从而提高接受率。比如我们可以使用指数分布，即令

$$g(x) = \begin{cases} \lambda e^{-\lambda(x-c)} & x \geq c \\ 0 & x < c \end{cases}$$

该分布可以非常简单的使用分布函数的逆函数得到，由于其分布函数为 $G(x) = 1-e^{-\lambda(x-c)}$，因而可以通过产生均匀分布 $U \sim U(0,1)$，并计算：

$$X = c - \frac{\ln U}{\lambda}$$

得到该指数分布。

由于拒绝采样需要满足：

$$M \geq \frac{l(x)}{g(x)} = \frac{f_X(x)}{g(x)} = \frac{\frac{\phi(x)}{1-\Phi(c)}}{\lambda e^{-\lambda(x-c)}} = \frac{e^{-\frac{x^2}{2}+\lambda(x-c)}}{\lambda(1-\Phi(c))\sqrt{2\pi}}, x \geq c$$

可以解得，当 $x = \lambda$ 时，不等式右边达到最大，因而给定 λ，M 最小可以取到：

$$M(\lambda) = \frac{e^{-\frac{\lambda^2}{2}+\lambda(\lambda-c)}}{\lambda(1-\Phi(c))\sqrt{2\pi}} = \frac{e^{\frac{\lambda^2}{2}-\lambda c-\ln\lambda}}{(1-\Phi(c))\sqrt{2\pi}}$$

我们不妨取 $\lambda = 2.5$，以下 Stata 程序对以上拒绝采样法进行了模拟：

代码 9: Stata 中截断正态的模拟

```
// rejection_trunc.do
clear
set more off
set obs 1000
// 参数
scalar c=3
// 计算lambda和M
scalar pi=3.1415926
scalar temp_cons=sqrt(2*pi)*(1-normal(c))
// 最优lambda
scalar lam=(c+sqrt(c^2+4))/2
// 实验lambda
scalar lam=2.5
scalar M=(exp(lam^2/2-lam*c-log(lam)))/temp_cons
// 产生均匀分布和指数分布
gen x=c-log(runiform())/lam
gen u=runiform()
// 截断正态分布密度函数
gen density=normalden(x)/(1-normal(c))
label variable density "f(x)"
// 指数分布密度函数 (用M标准化，即 M*g(x))
gen density_exp=M*lam*exp(-1*lam*(x-c))
label variable density_exp "M*g(x)"
// 挑出使得u<l/(M*g)的x,
gen selected=u<=(density/density_exp)
// 画图
sort x
```

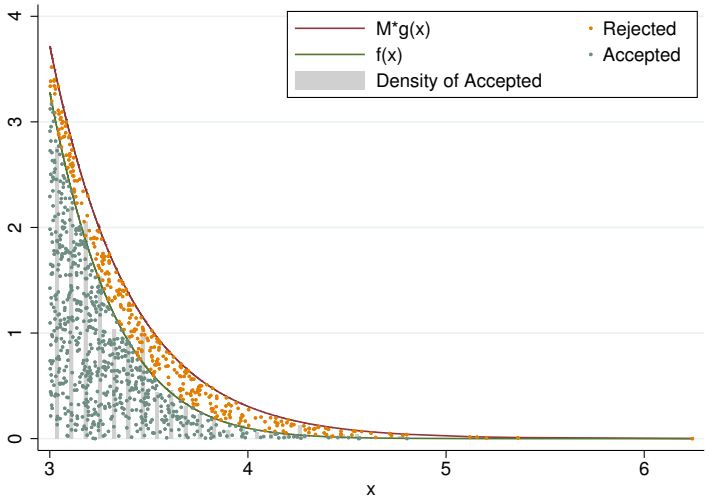


图 17: 截断正态的模拟

```
// 显示方便，散点图y轴用标准化的密度函数值乘以u
gen u_y=u*density_exp
twoway (hist x if selected==1, ///
      color(gs13) fcolor(gs13) barwidth(0.02)) ///
      (line density_exp x) ///
      (line density x) ///
      (scatter u_y x if selected==0, msize(tiny)) ///
      (scatter u_y x if selected==1, msize(tiny)) ///
      , legend(on ring(0) pos(1) order(2 4 3 5 1) ///
      label(4 "Rejected") label(5 "Accepted") ///
      label(1 "Density of Accepted")) ///
      graphr(fcolor(white) color(white))
graph export rejection_trunc.eps, replace
```

图 (17) 展示了拒绝采样的示意图。图中红色的线为 $M \cdot g(x)$，而绿色的线为欲生成的截断正态分布的密度函数 $f_X(x)$，墨绿色的点为接受的点，橙色的点为拒绝的点。图中所有的点（橙色 + 墨绿色）服从指数分布，其密度函数为 $g(x)$，而我们使用了拒绝采样，只有墨绿色的点被保留，通过保留的样本的直方图可以看到，其分布与截断正态分布一致。

根据以上 $\lambda = 2.5$ 的设定，我们通过模拟发现，其接受率大概为 67.5%。更进一步，我们可以通过选取不同的 λ 使得 M 进一步减小。最小化 $M(\lambda)$，得到：

$$\lambda = \frac{c + \sqrt{c^2 + 4}}{2}$$

从而我们得到了最优的 M 和 λ。在以下程序中，我们比较了当 $c=3$ 时以上两种算法的接受率：

代码 10: 拒绝采样法示例

```
#!/usr/bin/python3
## file: rejection.py
import numpy as np
import numpy.random as nprd
import scipy.special as scisp
## 设定参数
c=3 #抽取大于5的正态分布样本
N=1000 #抽取200个
## 直接使用正态分布进行拒绝抽样
times=0 #产生了多少次正态分布
accept=0 #接受了多少个
sample1=[] #结果
while accept<N:
    x=nprd.normal()
    if x>=c:
        sample1.append(x)
        accept=accept+1
    times=times+1
print("接受率=",N/times)

## 使用指数分布进行拒绝抽样
times=0 #产生了多少次指数分布
accept=0 #接受了多少个
sample2=[] #结果
# 计算参数
lam=(c+np.sqrt(c**2+4))/2
M=np.exp((lam**2-2*lam*c)/2)/(np.sqrt(2*np.pi)* \
    lam*(1-scisp.ndtr(c)))
normal_m=1/np.sqrt(2*np.pi) #正态分布密度函数前面的常数
# 截断正态分布密度函数
f_trunc_normal = lambda x: \
    normal_m*np.exp(-0.5*x**2)/(1-scisp.ndtr(c))
# 指数分布密度函数
f_exponential = lambda x: lam*np.exp(-1*lam*(x-c))
while accept<N:
    ## 产生指数分布
```

```
37      x=c-np.log(nprd.uniform())/lam
38      ## 接受概率
39      r=f_trunc_normal(x)/(M*f_exponential(x))
40      if nprd.uniform()<=r:
41          sample2.append(x)
42          accept=accept+1
43      times=times+1
44  print("接受率=",N/times)
```

以上程序结果，直接使用正态分布的算法的接受率大约为 0.13%，而第使用了最优的 λ 的指数分布的算法的接受概率大约为 96%，这无疑大大提高了抽样的速度。

习题

练习 1. 对于定义在概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的一个随机变量 X，证明 $\sigma\langle X\rangle = \left\{X^{-1}(B) : B \in \mathscr{B}\right\}$ 是一个 $\sigma-$ 代数。

练习 2. 在研究中，对收入等变量取对数是非常常见的处理手段。如果 $X > 0$ 代表总体收入，那么 $\mathbb{E}(X)$ 和 $\exp[\mathbb{E}(\ln X)]$ 哪一个更大?

练习 3. 如果 $X \sim N\left(\mu, \sigma^2\right)$，求 $\mathbb{E}\left(e^X\right)$。

练习 4. 如果 r.v.$X \sim Binomial(n, p)$，求随机变量 $Y = n - X$ 的概率质量函数。

练习 5. 求对数正态分布（$Y = e^X, X \sim N(0, 1)$）的概率密度函数。

练习 6. 求负二项分布的方差。

练习 7. 证明几何分布的无记忆性。

练习 8. 计算泊松分布的方差。

练习 9. 计算 $\Gamma(\alpha, \beta)$ 分布的方差。

练习 10. 求标准正态分布的偏度、峰度。

练习 11. Pareto 分布的密度函数为:

$$f(x|a, \beta) = \beta \cdot a^{\beta} x^{-(\beta+1)}, x > a$$

那么 Pareto 分布是否属于指数分布族?

练习 12. 请用定理 (9) 计算指数分布、泊松分布的期望和方差。

练习 13. 求 Logistic 分布的期望。

练习 14. 若 $X \sim N(0,1)$，求 $Y = |X|$ 的密度函数。

练习 15. 证明:对于一个随机变量 $X \sim F_X$,随机变量 $Y = F_X(X) \sim Uniform(0,1)$。

练习 16. 使用任何编程语言，通过均匀分布生成 100 个 Logistic 分布（分布函数 $\frac{e^x}{1+e^x}$）的随机数，并将理论的分布函数及其经验分布函数画在一张图中。其中经验分布函数的定义为:

$$\hat{F}(x) = \frac{1}{N}\sum_{i=1}^{N} 1(X_i \leq x)$$

即样本中小于 x 的样本的 s 比例。观察两者是否贴近? 生成 1000 个数据，重复以上步骤，并比较两条曲线的差异。

练习 17. 使用任何编程语言，通过均匀分布生成 100 个服从泊松分布（$\lambda = 1$）的随机数，并计算均值。

练习 18. 使用拒绝采样，选取适当的 $g(x)$ 和 M，对自由度 $K > 2$ 的 χ^2 分布进行抽样。

参考文献

[1] Ash, R.B., Doleans-Dade, C., 2000. Probability and measure theory. Academic Press.

[2] Athreya, K.B., Lahiri, S.N., 2006. Measure Theory and Probability Theory. Springer, New York.

[3] Bickel, P.J., Doksum, K.A., 2001. Mathematical Statistics: Basic Ideas and Selected Topics. Prentice-Hall, Inc, New Jersey.

[4] Casella, G., Berger, R.L., 2002. Statistical inference. Duxbury Pacific Grove, CA.

[5] Chung, K.L., 2001. A Course in Probability Theory, 3rd editio. ed. Elsevier Ltd., Singapore.

[6] Shao, J., 2007. Mathematical Statistics, 2nd ed. Springer, New York.

[7] Schervish, M.J., 1995. Theory of Statistics. Springer-Verlag, New York.

[8] Von Neumann, J., 1961. Various techniques used in connection with random digits, Paper No. 13 in in "monte carlo method."NBS Appl. Math. Ser.