第一章　概率论

1.1　事件的概率

1.1.1　概率的公理化定义

设 Ω 为样本空间，定义事件集类 $\mathscr{J} \subset 2^{\Omega}$。定义 $P : \mathscr{J} \to \mathbb{R}$ 满足三条公理

1. $P(A) \geq 0, \forall A \in \mathscr{J}$
2. $P(\Omega) = 1$
3. $P(\sum_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i), \forall A_i \in \mathscr{J}, A_iA_j = \phi, \forall i \neq j$

则称 P 为概率函数，$(\omega, \mathscr{J}, P)$ 为概率空间

错位排序

A_i 表示第 i 个数恰好在原位上，则每个数**都不在**原位的概率为

$$\begin{aligned}
P &= 1 - P(A_1 + A_2 + \cdots + A_n) \\
&= 1 - \sum_{i_1 \leq \cdots \leq i_r} (-1)^{r+1} P(A_{i_1} \ldots A_{i_r}) \\
&= 1 - \sum_{r=1}^{n} (-1)^{r+1} \binom{n}{r} \frac{(n-r)!}{n!} (= \frac{1}{r!}) \\
&= 1 - [1 - \frac{1}{2!} + \frac{1}{3!} + \cdots + (-1)^{n+1} \frac{1}{(n)!}] (\to \frac{1}{e})
\end{aligned} \tag{1.1}$$

1.1.2　条件概率

设 $A, B \in \mathscr{J}$，定义 条件概率 为

$$P(A|B) \triangleq \frac{P(AB)}{P(B)} \quad (P(B) > 0) \tag{1.2}$$

容易证明，

$$\widetilde{P}(A) = P(A|B) : \mathscr{J} \to \mathbb{R}$$

也是概率函数。

1.1.3　独立事件

两个事件的独立

若 $P(AB)=P(A)P(B)$ 则称 A,B 相互独立。此时有 $P(A)=P(A|B)$，并且能**推出** A **与** B^C **独立**。

多个事件的独立

若对 $A_1,A_2,\dots$ 为可数个事件，若从中**任取有限个事件** $A_{i_1},\dots,A_{i_m}$ 都有

$$P(A_{i_1}\cdots A_{i_m})=P(A_{i_1})\cdots P(A_{i_m})$$

，则称 $A_1,A_2,\dots$ 相互独立。

条件独立

若 $P(AB|E)=P(A|E)P(B|E)$，则称事件 A,B 关于事件 E 条件独立。
注意：条件独立不能推出独立，独立也不能推出条件独立

1.1.4　贝叶斯 (Bayes) 公式

定义 Ω 的一个分割 $\{B\}$ 满足 $\sum_i B_i=\Omega$ 且 $B_iB_j=\phi,\forall i\neq j$。则

$$P(A)=P(A\sum_i B_i)=\sum_i P(B_i)P(A|B_i) \tag{1.3}$$

$$P(B_j|A)=\frac{P(AB_j)}{P(A)}=\frac{P(B_j)P(A|B_j)}{\sum_i P(B_i)P(A|B_i)} \tag{1.4}$$

(1.4) 即为 Bayes 公式。

1.2　随机变量

1.2.1　1 维随机变量

定义实值函数 $X(\omega):\Omega\to\mathbb{R}$，该映射给样本空间中的每个试验结果一个对应的数值，**随机变量** 即定义为试验结果的一个实值函数，也可以通过**随机变量的函数**定义另一个随机变量。(默认用大写字母表示随机变量，小写字母表示实数。)

1.2.2　随机变量的概率

$\forall I\in\mathbb{R}$ 且 I 为可测集，利用数值所对应的事件定义概率函数 $P_X:\mathbb{R}\to\mathbb{R}$

$$P_X(X\in I)\triangleq P(X^{-1}(I))$$

1.2.3　（累积）分布函数 (cdf)

定义为 $F(x) \triangleq P(X \leq x), x \in \mathbb{R}$，则 $P(a < X \leq b) = F(b) - F(a)$。cdf 有以下性质：

1. $F(x)$ 单调（非严格）递增
2. $\lim\limits_{x \to -\infty} F(x) = 0, \lim\limits_{x \to +\infty} F(x) = 1$
3. $F(x)$ 右连续

以上性质也是 $F(x)$ 成为 cdf 的充要条件。

1.2.4　离散分布

概率质量函数 (pmf)

定义为 $f(x) = P(X = x), \forall x \in \mathbb{R}$。

期望和方差

期望 $E(X) \triangleq \sum_i x_i f(x_i)$；方差 $Var(X) \triangleq \sum_i (x_i - E(X))^2 f(x_i) = E(X - E(X)^2)$。

1. 若用 X 的函数 $g(X)$ 定义 Y，则 $E(g(X)) = \sum_i g(x_i) f(x_i) \neq g(E(X))$
2. $E(X+Y) = E(X) + E(Y)$；若 X, Y 独立，则 $Var(X+Y) = Var(X) + Var(Y)$

1.2.5　常见离散分布

Bernoulli 分布

- $X = \begin{cases} 1, & p \\ 0, & 1-p \end{cases}$，记为 $X \sim B(p)$
- $E(X) = p$，$Var(X) = p(1-p)$

二项分布

- $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$，记为 $X \sim B(n, p)$
- X 为 n 次伯努利试验中试验成功的次数
- $E(X) = np$，$Var(X) = np(1-p)$

Poisson 分布

- $f(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$，记为 $X \sim P(\lambda)$
- $E(X) = \lambda$，$Var(X) = \lambda$
- 对于 $X \sim B(n, p)$，n 非常大而 np 较小时，X 近似服从 $P(np)$
- $P(\lambda)$ 多用于当 X 表示一定时间内出现的小概率时间的次数时

1.2.6 连续分布

概率密度函数 (pdf)

若存在 $f \geq 0$，使得 $\forall$ 可测集 $I \subset R$ 都有

$$P(X \in I) = \int_I f(x)\mathrm{d}x$$

，则 X 为连续型随机变量，f 为 X 的概率密度函数

1. $P(X = a) \equiv 0, \forall a \in \mathbb{R}$
2. $E(X) = \int_{-\infty}^{\infty} xf(x)\mathrm{d}x$
3. $Var(X) = \int_{-\infty}^{\infty}(x - E(X))^2 f(x)\mathrm{d}x = E(X^2) - E(X)^2$

1.2.7 常见连续分布

均匀分布

- $f(x) = \begin{cases} \frac{1}{b-a}, & x \in (a,b) \\ 0, & otherwise \end{cases}$，记为 $X \sim U(a,b)$
- $E(X) = \frac{a+b}{2}$，$Var(X) = \frac{(b-a)^2}{12}$

正态分布

- $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \forall x \in \mathbb{R}$，记为 $X \sim N(\mu, \sigma^2)$
- $E(X) = \mu, Var(X) = \sigma^2$
- $X \sim N(\mu, \sigma^2) \implies Y = \frac{X-\mu}{\sigma} \sim N(0,1)$

指数分布

- $f = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$
- $E(X) = \frac{1}{\lambda}, Var(X) = \frac{1}{\lambda^2}$
- $F(X) = 1 - e^{-\lambda x}$
- X 一般刻画寿命或者等待时间
- 无记忆性：

$$P(X > x + t | X > x) = \frac{P(X > x + t)}{P(X > x)} = \frac{e^{-\lambda(x+t)}}{e^{-\lambda x}} = e^{-\lambda t} = f(t)$$

1.3　联合分布

1.3.1　随机向量

$(X_1, X_2, \ldots, X_n) : \Omega \to \mathbb{R}^n$，其中 X_i 为随机变量，定义（联合）累积分布函数

$$F(x_1, \ldots, x_n) \triangleq P(X_1 \le x_1, \ldots, X_n \le x_n), (x_1, \ldots, x_n) \in \mathbb{R}^n$$

1.3.2　离散分布

若 $\forall 1 \le i \le n, X_i$ 为离散型随机变量，则 $(X_1, \ldots, X_n)$ 为离散型随机向量。定义 pmf 为 $f(x_1, \ldots, x_n) \triangleq P(X_1 = x_1, \ldots, X_n = x_n)$。

1.3.3　连续分布

若对一切可测集 $Q \subset \mathbb{R}^n$,$P((X_1, \ldots, X_n) \in Q) = \int_Q f(x_1, \ldots, x_n)\mathrm{d}x_1 \cdots \mathrm{d}x_n$,则 $(X_1, \ldots, X_n)$ 为连续型随机向量，f 为其 pdf。

以 $n = 2$ 为例，$F(a, b) = \int_{-\infty}^{a}\int_{-\infty}^{b} f(x, y)\mathrm{d}x\mathrm{d}y$，反过来有 $f(a, b) = \frac{\partial^2 F}{\partial x \partial y}$

二元正态分布

$X, Y \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho), |\rho| < 1$

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2}\frac{1}{\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-\mu_2}{\sigma_2}\right)^2 - 2\rho\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2}\right]\right\} \tag{1.5}$$

1.3.4　边际分布

对于的连续随机变量，定义 X_i 的边际累积分布函数 (cdf)

$$F_i(x) \triangleq P(X_i \le x) = P(X_i \le x, -\infty < X_j < \infty (j \ne i)) \tag{1.6}$$

以 $n = 2$ 为例，$F_X(x) = \lim_{y\to\infty} P(X \le x, Y \le y) = \lim_{y\to\infty} F(x, y)$。
定义 X_i 的边际密度函数 (pdf)

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\mathrm{d}y$$

1.3.5　条件分布

对于 $n = 2$ 的连续型随机变量，求其条件密度函数

$$P(X \le x | y \le Y \le y + \mathrm{d}y) = \frac{\int_{-\infty}^{x}\left(\int_{y}^{y+\mathrm{d}y} f(t, s)\mathrm{d}q\right)\mathrm{d}p}{\int_{y}^{y+\mathrm{d}y} f_Y(q)\mathrm{d}q}$$

$$P(X=x|y\le Y\le y+\mathrm{d}y)=\frac{\int_y^{y+\mathrm{d}y}f(t,s)\mathrm{d}q}{\int_y^{y+\mathrm{d}y}f_Y(q)\mathrm{d}q}\to\frac{f(x,y)}{f_Y(y)}(\mathrm{d}y\to 0)$$

$$f_X(x|y)\triangleq\frac{f(x,y)}{f_Y(y)} \tag{1.7}$$

定义 $f_X(x|y)$ 的累积分布函数 (cdf) 为 $F(a|y)=\int_{-\infty}^a f_X(x|y)\mathrm{d}x$

(1)（乘法法则）$f(x,y)=f_X(x|y)f_Y(y)=f_Y(y|x)f_X(x)$

(2)（全概率公式）$f_X(x)=\int_{-\infty}^{+\infty}f(x,y)\mathrm{d}y=\int_{-\infty}^{+\infty}f_X(x|y)f_Y(y)\mathrm{d}y$

对于二元正态分布，

$$f_Y(y|x)=\frac{1}{\sqrt{2\pi}\sigma_2}\frac{1}{2\sqrt{1-\rho^2}}\exp\left\{-\frac{[y-(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1))]^2}{2(1-\rho^2)\sigma_2^2}\right\} \tag{1.8}$$

即 $\boldsymbol{Z=(Y|x)\sim N(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1),(1-\rho^2)\sigma_2^2)}$

1.3.6　独立性

若 $F(x_1,\ldots,x_n)=F_1(x_1)\cdots F_n(x_n),\forall x_1,\ldots,x_n\in\mathbb{R}$，则称 $X_1,\ldots,X_n$ 相互独立。该条件完全等价于 $f(x_1,\ldots,x_n)=f_1(x_1)\cdots f_n(x_n)$。

1. 若 $X_1,\ldots,X_n$ 相互独立，则 $Y_1=g_1(X_1,\ldots,X_m)$ 与 $Y_2=g_2(X_{m+1},\ldots,X_n)$ 相互独立
2. 若 pdf $f(x_1,\ldots,x_n)=g_1(x_1)\cdots g_n(x_n),\forall x_1,\ldots,x_n\in\mathbb{R}$，则 $X_1,X_2,\ldots,X_n$ 相互独立，且 f_i 与 g_i 相差常数倍

1.3.7　多个随机变量的函数

★密度函数变换法

设存在由随机变量 $X1,X2$ 到 Y_1,Y_2 的可逆映射 $\begin{cases}Y_1=g_1(X_1,X_2)\\Y_2=g_2(X_1,X_2)\end{cases}$,逆为 $\begin{cases}X_1=h(Y_1,Y_2)\\X_2=h(Y_1,Y_2)\end{cases}$ 。对于 X_1,X_2 上的区域 A，若其在 Y_1,Y_2 上的映射为 B，则显然有

$$P((X_1,X_2)\in A)=P((Y_1,Y_2)\in B)$$

$$P((Y_1,Y_2)\in B)=\int_B\mathscr{F}(y_1,y_2)\mathrm{d}y_1\mathrm{d}y_2$$

$$P((X_1,X_2)\in A)=\int_A f(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2=\int_B f(h(y_1),h(y_2))\left|\frac{D(x_1,x_2)}{D(y_1,y_2)}\right|\mathrm{d}y_1\mathrm{d}y_2$$

因此

$$\mathscr{F}(Y_1,Y_2)=f(h(Y_1),h(Y_2))\left|\frac{D(x_1,x_2)}{D(y_1,y_2)}\right| \tag{1.9}$$

1.4 随机变量的数学特征

1.4.1 期望

1. 刻画分布的集中趋势
2. $X_1,...,X_n$ 独立时，$E(X_1X_2\cdots X_n)=E(X_1)\cdots E(X_n)$

1.4.2 分位数

$\forall\alpha\in(0,1)$，若 $P(X\leq a)\geq\alpha$ 且 $P(X\geq a)\geq 1-\alpha$，则称 $X=a$ 为 X 的 **下α-分位数**。

1. 若 cdf 连续，则 $F(a)=\alpha$
2. 分位数不一定唯一

1.4.3 方差

1. 刻画分布的集中程度
2. $Var(cX)=c^2Var(X)$
3. $Var(X+Y)=Var(X)+Var(Y)+2E[(X-\mu_1)(Y-\mu_2)]$

1.4.4 协方差与相关系数

定义协方差 $Cov(X,Y)\triangleq E[(X-\mu_1)(Y-\mu_2)]$

1. $Cov(X,X)=Var(X)$
2. $Cov(X,Y)=E(XY)-E(X)E(Y)$
3. $Cov(aX_1+bX_2+c,Y)=aCov(X_1,Y)+bCov(X_2,Y)$

定义相关系数 $Corr(X,Y)\triangleq\frac{Cov(X,Y)}{\sigma_1\sigma_2}=E(\frac{X-\mu_1}{\sigma_1}\frac{Y-\mu_2}{\sigma_2})=\rho$

1. X,Y 独立 $\implies Cov(X,Y)=0$，反之不一定
2. $|Corr(X,Y)|\leq 1$
3. ρ 实际上是**线性**相关系数，不能表达高维的相关关系

1.4.5 矩 (Moment)

定义 将 $E[(X-C)^n]$ 称为 X 关于 C 的 n 阶矩。当 $C=E(X)$ 时，称为中心矩；当 $C=0$ 时，称为原点矩；标准化后的矩 $E[(\frac{x-\mu}{\sigma})^n]$ 称为 n 阶标注矩。

偏度系数

定义 3 阶标准矩又称偏度系数。

1. 偏度系数小于零，左偏；大于零，右偏
2. 相对于 5 阶以上的奇数阶矩，3 阶矩容易计算且噪声的影响小

峰度系数

定义　4 阶标准矩又称峰度系数。

1. 正态分布峰度恒为 3
2. 超值峰度定义为 $E[\left(\frac{X-\mu}{\sigma}\right)^4]$
3. 超值峰度为正，一般相对于正态分布峰更尖，尾部更扁

1.4.6　矩母函数

定义　(moment generating function, mgf) $M_X(t) \triangleq E(e^{tX})$ 若在 $t=0$ 的某个邻域内 $M_X(t)$ 存在，则称其为 X 的矩母函数，否则称 X 的矩母函数不存在。**注意标明 t 的取值范围。**

对于 $X \sim Exp(\lambda)$，

$$M_X(t) = \int_0^\infty e^{tx}\lambda e^{-\lambda x}\mathrm{d}x = \frac{\lambda}{\lambda - t}, t < \lambda \tag{1.10}$$

对于 $X \sim N(\mu, \sigma^2)$，

$$M_X(t) = e^{\frac{\sigma^2 t^2}{2} + \mu t} \tag{1.11}$$

对于 $Y = aX + b$，

$$M_Y(t) = E(e^{aX+b}) = e^{tb}E(e^{taX}) = e^{tb}M_X(ta) \tag{1.12}$$

定理　矩母函数确定矩：

$$E(X^n) = M_X^{(n)}(0) \tag{1.13}$$

证明.

$$M_X(t) = \sum_{n\geq 0} M_X^{(n)}(0)\frac{t^n}{n!}$$

$$M_X(t) = E(e^{tx}) = E\left(\sum_{n\geq 0}\frac{(tx)^n}{n!}\right) = \sum_{n\geq 0} E(t^n)\frac{x^n}{n!}$$

由于 $M_X(t)$ 泰勒展开唯一，故 $M_X^{(n)} = E(t^n)$ □

例　对于 $X \sim N(0,1)$，

$$M_X(t) = e^{\frac{t^2}{2}} = \sum_{n\geq 0}\frac{t^{2n}}{2^n n!} = \sum_{n\geq 0}\frac{(2n)!}{2^n n!}\frac{t^{2n}}{(2n)!}$$

故 $E(X^{2n}) = \frac{(2n!)}{2^n n!}, E(X^{2n-1}) \equiv 0$

定理　矩母函数确定分布。若存在 $a > 0$，使得 $M_X(t) = M_Y(t), t \in (-a, a)$，则 X, Y 同分

布。

REMARK　矩存在的时候，矩母函数不一定存在（即泰勒展式不收敛），因此各阶矩完全相同也无法说明两个随机变量同分布。

例　取服从对数正态分布的变量 Y 于另一变量 Z

$$
\begin{aligned}
y = f_1(x) &= \frac{1}{\sqrt{2\pi x}}e^{-\frac{\log^2 x}{2}}, x > 0 \\
z = f_2(x) &= f_1(x)[1+\sin(2\pi\log x)], x > 0
\end{aligned}
$$

则它们的 n 阶矩存在以下关系：

$$E(Z^n) - E(Y^n) = \int_0^\infty x^n f_1(x)\sin(2\pi\log x)\mathrm{d}x = T$$

令 $t = \log x - n$，则 $x = e^{t+n}$

$$
\begin{aligned}
T &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{n(t+n)}e^{-\frac{1}{2}(t+n)}e^{-\frac{1}{2}(t+n)^2}\sin(2\pi(t+n))e^{t+n}\mathrm{d}t \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left[\frac{1}{2}(n+t)(n+t+1)\right]\mathrm{d}t \\
&= \frac{e^{-\frac{1}{8}}}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left[\frac{1}{2}(n+t+\frac{1}{2})^2\right]\mathrm{d}t = 0
\end{aligned}
$$

因此对一切 n，Y, Z 的 n 阶矩相等。但是由于

$$e^{tx}f_1(x) = \frac{1}{\sqrt{2\pi}}e^{tx-\frac{\log^2 x}{2}}$$

对于一切的 $t > 0$ 都有 $\lim\limits_{x\to\infty} e^{tx}f_1(x) = +\infty$，因此无法积分，$M_Y(t)$ 不存在。故即使各阶矩完全相同，实际上也不是同分布。

独立随机变量和的分布

对于相互独立的 $X_1, X_2, \ldots, X_n$，有

$$M_{X_1+\cdots+X_n}(t) = M_{X_1}(t)\cdots M_{X_n}(t) \tag{1.14}$$

对于联合分布 $(X_1, X_2, \ldots, X_n)$，定义矩母函数为多元函数

$$M_{X_1,\ldots,X_n}(t_1, t_2, \ldots, t_n) = E(e^{t_1X_1+\cdots+t_nX_n}) \tag{1.15}$$

若该函数在原点的邻域 $B(0,\sigma), \sigma > 0$ 内有定义，则称联合分布的矩母函数存在；若两组联合分布的矩母函数在 $B(0,\sigma)$ 内恒等，则这两组联合分布同分布。

1.4.7 条件期望

定义

$$E(Y|X=x)=\begin{cases}\sum_i y_i P(Y=y_i|X=x)\\ \int_{-\infty}^{\infty} f_Y(y|X=x)\mathrm{d}y\end{cases}$$

定理 $E(Y)=E(E(Y|X))$

定理 $E((Y-g(X))^2)\geq E[(Y-E(Y|X))^2]$，这意味着 $E(Y|X)$ 是均方误差意义下的最优预测。均方最优：$E[(Y-c)^2]\geq E[(Y-E(Y))^2]$，这是因为 $\frac{\partial E[(Y-c)^2]}{\partial c}=2E[Y-c]\big|_{c=E(Y)}=0$。

1.5 不等式与极限定理

1.5.1 概率不等式

Markov 不等式

若 $Y\geq 0$，则 $\forall a>0$，有

$$P(Y\geq a)\leq\frac{E(Y)}{a} \tag{1.16}$$

证明. 令 $I=\begin{cases}1,Y\geq a\\ 0,Y<a\end{cases}$ ，则不论 Y 的取值，均有 $I\leq\frac{Y}{a}$

故 $P(Y\geq a)=E(I)\leq E(\frac{Y}{a})=\frac{E(Y)}{a}$ □

Chebyshef 不等式

若 $Var(Y)$ 存在，则 $\forall a>0$

$$E(|Y-E(Y)|\geq a)\leq\frac{Var(Y)}{a^2} \tag{1.17}$$

证明. 在(1.16)中代入 $P[|Y-E(Y)|\geq a]=P[(Y-E(Y))^2\geq a^2]=E[Var(Y)\geq a^2]$ 即可。□

Chernoff 不等式

$\forall a>0,t>0$ 有

$$P(Y\geq a)\leq\frac{E(e^{tY})}{e^{ta}} \tag{1.18}$$

1.5.2 大数定律

设 $X_1,X_2,\ldots,X_n$ 独立同分布，$\bar{X}=\frac{1}{n}\sum_{k=1}^{n}X_k$ ，则

Khinchin 弱大数定律

$\forall\varepsilon > 0$ 有

$$\lim_{n\to\infty} P(|\bar{X} - \mu| < \varepsilon) = 1 \tag{1.19}$$

证明.

$$P(|\bar{X} - \mu| \geq \varepsilon) \leq \frac{Var(\bar{X})}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0(n \to \infty)$$

□

若 $P(|\bar{X} - \mu| \geq \varepsilon) \leq \alpha$，则称 α 为置信水平，ε 为精度。

Kolmogorov 强大数定律

$\bar{X}$ 几乎必然收敛到 μ

$$\forall\varepsilon > 0,\ P(\lim_{n\to\infty} |\bar{X} - \mu| < \varepsilon) = 1 \tag{1.20}$$

1.5.3　中心极限定理 (CLT)

若 $E(X_i), Var(X_i)$ 均存在，则

$$\lim_{n\to\infty} P(\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} \leq x) = \Phi(X) \sim N(0, 1), \forall x \in \mathbb{R} \tag{1.21}$$