Chap 2 随机变量

1 1 维随机变量

- **定义(随机变量)** 样本空间上的实值函数.

$$X : \Omega \to \mathbb{R},\ \omega \to X(\omega).$$

- **例**

试验	样本空间	随机变量	像集 (新样本空间)
随机调查 50 人对某议题支持与否	$\{(1, 0, \cdots), \cdots\}$	$X_1 = 1$ 的个数	$\{0, 1, 2, \cdots, 50\}$
随机抽取一个北京市成年公民	所有北京市成年公民之集	$X_2 =$ 其 2022 年的收入	$(-\infty, +\infty)$

- **定义(事件)** $X_1 = 30$, $X_2 > 100,000$.
- **注**
  - 概括作用: 提供了试验结果的数值摘要;
  - 事件 **v.s.** 变量, 静态 **v.s.** 动态.
- **分类**
  - 离散型: 至多可数个取值;
  - 连续型: 区间型取值 (定义不严格);
  - 其他.
- **定义** $\forall I \subset \mathbb{R}$, 令 $X^{-1}(I)$ 表示 I 在 X 下的原像集, $X^{-1}(I) \subset \Omega$, 例如

$$X^{-1}((a, b)) = \{\omega \in \Omega \mid a < X(\omega) < b\}.$$

- **定义**

$$P_X(X \in I) = P(X^{-1}(I)),\ \forall I \subset \mathbb{R}\text{ 可测}.$$

需要 $X^{-1}(I) \in \mathscr{F}$, 一般记 P_X 为 P.

- **定义(累积分布函数)(CDF)**

$$F(x) := P(X \leq x),\ \forall x \in \mathbb{R}.$$

我们有

$$P(a < X \leq b) = F(b) - F(a).$$

- **性质**

- $0 \le F(x) \le 1$, 单调增(未必严格);
- $\lim\limits_{x\to+\infty} F(x) = 1$, $\lim\limits_{x\to-\infty} F(x) = 0$;
- 右连续 (PS. 若定义 $F(x) := P(X < x), \forall x \in \mathbb{R}$, 则有 $F(x)$ 左连续).

- **注**
  - 随机要素体现在样本点 ω 的不确定性;
  - 随机变量的直观意义往往出现在样本空间的直观意义之前;

- **辨析**

$$X_i = \begin{cases} 1, \text{第}i\text{次抛硬币正面向上;} \\ 0, \text{第}i\text{次抛硬币正面向下.} \end{cases}$$

其中 $i = 1, 2$. 那么 $X_1 + X_2$ 的样本空间为

$$\{\text{正正}, \text{正反}, \text{反正}, \text{反反}\}.$$

因为随机变量可视作函数, 需要满足定义域相同, 因此 X_1, X_2 的定义域同上.

- **注**
  - $aX + bY$, XY, $\dfrac{X}{Y}(Y \ne 0)$, $g(X, Y)$ 为随机变量, 其中 X, Y 样本空间相同;
  - 需要有 $X^{-1}(I) \in \mathscr{F}$, 从而 $P(X^{-1}(I))$ 有意义.
- **定义(同分布)** X_1, X_2 的 **CDF** 分别为 $F_1(x), F_2(x)$, 那么

$$\begin{aligned} X_1, X_2\text{同分布} &\Leftrightarrow P(X_1^{-1}(I)) = P(X_2^{-1}(I)), \forall I \subset \mathbb{R}\text{ 可测} \\ &\Leftrightarrow F_1(x) = F_2(x), \forall x \in \mathbb{R}. \end{aligned}$$

- **注**
  - X_1, X_2 同分布 $\not\Rightarrow X_1 = X_2$.
  - 考虑掷一次硬币, $X_1 =$ 正面向上的次数, $X_2 =$ 反面向上的次数, 这两个随机变量是同分布的.
  - 随机变量是函数!

2 离散型随机变量

- **定义(概率质量函数)(PMF)**

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

- **注**
  - $f(x_i) = p_i, \sum\limits_i P_i = 1$;
  - **CDF** 为阶梯函数.
- **定义(期望与方差)**

$$E(X)=\sum_i x_i p_i=\sum_i x_i f(x_i)=\mu$$
$$Var(X)=\sum_i (x_i-\mu)^2 p_i=\sum_i (x_i-\mu)^2 f(x_i)=\sigma^2$$

我们有

$$Var(X)=E((X-E(X))^2)=E(X^2)-E^2(X).$$

- **注**
  - 算数均值即期望:
  $$\frac{1}{n}\sum_{i=1}^{n} a_i=\frac{1}{n}\sum_{i=1}^{n} k_i x_i=\sum_{i=1}^{n} p_i x_i=\mu;$$
  - 期望存在 $\Leftrightarrow \sum\limits_i |x_i| p_i<+\infty$;
  - $E(g(X))=\sum\limits_i g(x_i)p_i$;
  - $E(X), Var(X)$ 为随机变量 X 的分布的特征, 分别刻画了随机变量的集中趋势和分散程度.

3　常见离散分布

- **定义(Bernoulli 分布)**

$$X=\begin{cases}1, \text{事件成功}, p\\ 0, \text{事件不成功}, 1-p\end{cases}$$

记为 $X\sim B(p)$. 我们有

$$E(X)=p, Var(X)=p(1-p).$$

- **定义(二项分布)**

记 X 为 n 次独立 **Bernoulli** 试验的成功次数. 满足

$$P(X=k)=C_n^k p^k(1-p)^{n-k},\ k=0,1,\cdots,n.$$

记为 $X\sim B(n,p)$. 我们有

$$E(X)=np, Var(X)=np(1-p).$$

- **定义(Poisson 分布)**

满足

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

记为 $X \sim P(\lambda)$. 我们有

$$E(X)=\lambda, Var(X)=\lambda.$$

- **例** 观察时间 $[0,1)$ 某路口发生的交通事故数 X.
    - $l_i=[\frac{i-1}{n},\frac{i}{n}), i=1,2,\cdots,n$.
    - n 充分大.
    - 假设:
        - l_i 上至多发生一起事故;
        - l_i 上恰发生一次事故的概率 $p=\frac{\lambda}{n}$, 与时长成正比;
        - l_i 各段相互独立.
    - 此时

$$\begin{aligned} P(X=k) &= C_n^k p^k(1-p)^{n-k} \\ &= \frac{n!}{k!(n-k)!}(\frac{\lambda}{n})^k(1-\frac{\lambda}{n})^{n-k} \\ &\to \frac{\lambda^k}{k!}e^{-\lambda}, \text{当}\, n\to\infty. \end{aligned}$$

- **注**
    - 若 $X\sim B(n,p)$, p 很小, n 很大, np 不太大, 则 $X\sim P(\lambda)$, $\lambda=np$.
    - 误差最多为 $\min(p,np^2)$.
    - **Poisson** 分布多用于一定时间或空间内小概率事件发生次数的场景.
- **例** 某医院平均每小时出生婴儿 λ 名, 接下来 t 小时出生婴儿数的分布.
- **解答** 我们有

$$P(N(t)=k)=\frac{(\lambda t)^k}{k!}e^{-\lambda t},\, k=0,1,2,\cdots.$$

其中 λ 为均值.

- **注** **Bernoulli** 试验不独立, 但弱相依条件下仍为较好近似.
- **例(配对问题)**
- **解答** 弱相依条件下:

$$P(A_i)=\frac{1}{n}\simeq P(A_i\mid A_j)=\frac{1}{n-1}$$

恰有 k 个人拿到自己的帽子的概率:

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}=\frac{e^{-1}}{k!}$$

- **常规解答**

  设 $E=$ 指定的 k 个人拿到了自己的帽子.

  设 $F=$ 其余的 $n-k$ 个人未拿到自己的帽子.

  我们有:

  $$P(EF)=P(F\mid E)P(E)=P_{n-k}\frac{(n-k)!}{n!}$$

  进而有:

  $$P(X=k)=C_n^kP(EF)=\frac{1}{k!}P_{n-k}\to\frac{e^{-1}}{k!}.$$

4 连续随机变量

- **定义(概率密度函数)(PDF)** 若存在 $f\geq 0$, 使得 $\forall I\subset\mathbb{R}$ 可测, 都有

  $$P(X\in I)=\int_I f(x)dx$$

  则称 X 为连续型随机变量, f 为 X 的概率密度函数 (**PDF**).

- **性质**
  - $\int_{-\infty}^{\infty}f(x)dx\equiv 1$;
  - $P(a<X\leq b)=\int_a^b f(x)dx=P(a\leq X\leq b)=P(a\leq X<b)=P(a<X<b)$;
  - $P(X=a)=0,\forall a\in\mathbb{R}$;
  - $P(x_0-\delta<X\leq x_0+\delta)=\int_{x_0-\delta}^{x_0+\delta}f(x)dx=2\delta f(x_0)$, 要求 f 在 x_0 处连续;
  - $F(x)=P(X\leq x)=\int_{-\infty}^{x}f(t)dt$ 连续, $F'(x)=f(x)$ (f 在 x 处连续);
  - **PDF** 与 **PMF** 实质上可以统一; **PDF** 若存在, 则不唯一.
- **定义(期望与方差)**

  $$E(X)=\int_{-\infty}^{\infty}xf(x)dx=\mu$$
  $$Var(X)=\int_{-\infty}^{\infty}(x-\mu)^2f(x)dx=\sigma^2$$

  我们有

  $$Var(X)=E((X-E(X))^2)=E(X^2)-E^2(X).$$

- **约定** $E(X)$ 存在 $\Leftrightarrow E(X)<\infty$.

- **注**
  - $E(X)$ 存在 $\Leftrightarrow \int\limits_{-\infty}^{\infty}|x|f(x)dx<\infty$;

- 一般地, $E(g(X)) = \int\limits_{-\infty}^{\infty} g(x)f(x)dx$.

5 常见连续分布

- **定义(连续分布)**

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

  记为 $X \sim U(a,b)$. 我们有

$$E(X) = \frac{a+b}{2}, Var(X) = ?.$$

- **注** $X \sim U(0,1)$ 称为随机数.
- **定义(正态分布)**

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}.$$

  记为 $X \sim N(\mu, \sigma^2)$. 我们有

$$E(X) = \mu, Var(X) = \sigma^2.$$

- **注**
  - $X \sim N(\mu, \sigma^2) \Leftrightarrow Y = \dfrac{X-\mu}{\sigma} \sim N(0,1)$;
  - $N(0,1)$ 标准正态;
  - 经验法则.
- **定义(指数分布)**

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

  记为 $X \sim Exp(\lambda)$. 我们有

$$E(X) = \frac{1}{\lambda}, Var(X) = \frac{1}{\lambda^2}.$$

- **注**
  - 有的软件取参数为 $\beta = \dfrac{1}{\lambda}$;
  - 通常刻画寿命或等待时间.
- **例** 观察到有婴儿出生, 接下来 t 小时有婴儿出生的概率为?
- **解答**

$$
\begin{aligned}
P(X \le t) &= 1 - P(X > t) \\
&= 1 - P(N(t) = 0) \\
&= 1 - \frac{(\lambda t)^0}{0!} e^{-\lambda t} \\
&= 1 - e^{-\lambda t}.
\end{aligned}
$$

这是一个 **Poisson** 过程, 数量是 **Poisson** 分布, 间隔是指数分布.

- **定义** 假设 $X > 0$ 连续, 其 **CDF** 为 $F(x)$, 满足 $F(0) = 0$. 考虑

$$
\begin{aligned}
&P(x < X < x + dx \mid X > x) \\
=&\frac{P(x < X < x + dx)}{P(X > x)} \\
=&\frac{F(x + dx) - F(x)}{1 - F(x)} \\
\approx&\frac{F'(x)}{1 - F(x)} dx.
\end{aligned}
$$

视为年龄为 x 的元件失效的条件概率密度 (瞬时失效率/危险率).

- **注**
  - 令 $\dfrac{F'(x)}{1 - F(x)} = \lambda(x) \Rightarrow F(x) = 1 - e^{-\int_0^x \lambda(t) dt}, x > 0;$
  - 若 $\lambda(x) \equiv \lambda$ (无老化假设), 则 $F(x) = 1 - e^{-\lambda x}$,

    $\Rightarrow P(X > t + s \mid X > s) = \dfrac{P(X > t + s)}{P(X > s)} = \dfrac{1 - F(s + t)}{1 - F(s)} = e^{-\lambda t}$ (无记忆性);
  - 改进 $\lambda(x) = \alpha \dfrac{x^{\alpha - 1}}{\beta^\alpha}, \alpha, \beta > 0$, 则 $F(x) = 1 - e^{-(\frac{x}{\beta})^\alpha} \Rightarrow$ **Weibull** 分布.

6 随机变量的函数

- **定义** $Y = g(X)$
- **例** $X \sim Exp(\lambda)$.
- **解答**

$$
Y = \begin{cases} 1,\ x > t_0 \\ 0,\ x \le t_0 \end{cases} \quad t_0 > 0 \text{给定}.
$$

那么有 $P(Y = 0) = 1 - e^{-\lambda t_0}$, $P(Y = 1) = e^{-\lambda t_0}$.

- **例** X 连续, 其 **PDF** 为 $f(x)$, 且 $Y = X^2$.
- **解答** $\forall y > 0$, 我们有

$$
\begin{aligned}
P(Y \le y) &= P(X^2 \le y) \\
&= P(-\sqrt{y} \le X \le \sqrt{y}) \\
&= \int_{-\sqrt{y}}^{\sqrt{y}} f(x)dx \\
&= \int_0^y l(t)dt.
\end{aligned}
$$

其中 Y 的 **PDF** 为 $l(y) = \dfrac{1}{2\sqrt{y}}(f(\sqrt{y}) + f(-\sqrt{y}))$.

- **例** $X \sim N(0,1)$, 且 $Y = X^2$.
- **解答** Y 的 **PDF** 为

$$
l(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-\dfrac{y}{2}}
$$

这是自由度为 1 的 **Chi-Square** 分布 $\chi^2(1)$.