Ch 3. 随机向量及其分布

钟友良

zhongyl@scut.edu.cn

Outline

Outline

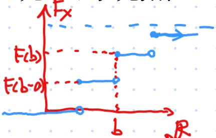
回顾

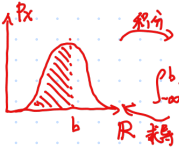
- 概率空间 $(\Omega, \mathcal{F}, P)$ $P: \mathcal{F} \longrightarrow \mathbb{R}_+$
- 随机变量 $X: \Omega \longrightarrow \mathbb{R}$
  可测： $X: \mathcal{F} \longrightarrow \mathcal{B}(\mathbb{R})$
  if $X^{-1}((-\infty, b]) \in \mathcal{F}$
  所有 $(a,b]$ 生成 $(a,b] = (-\infty, b] \setminus (-\infty, a]$
- 分布函数
- 分类
  - 离散型
    - 分布“密度”函数：单点的概率
  - 连续型
    - 分布密度函数：分布函数的导数
- 统一的观点？

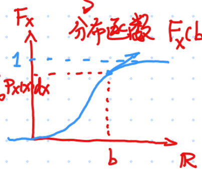
$P_{\mathcal{B}(\mathbb{R})}((-\infty, b])$
$= P(X^{-1}((-\infty, b]))$
$= P(\{\omega \in \Omega \mid X(\omega) \le b\})$
分布函数 $F_X(b)$

$F_X(b)$ 右连续
$\lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon}(F_X(b+\varepsilon) - F_X(b))$

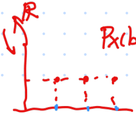
$P(X=b)$
$= P(X \le b) - F(X \le b-0)$
左极限 $\lim_{\varepsilon \to 0} F(X \le b+\varepsilon)$
$= \lim_{n \to +\infty} F(X \le b - \frac{1}{n})$







$p_X(b) = \frac{d}{db}\int_{-\infty}^{b} p_X(x)dx.$



习题

$p_X(x) = \frac{1}{2}$.

$-\sqrt{y-1} \le X \le \sqrt{y-1}$

设 $X \sim U[0,2]$。
求 $Y = X^2 + 1$ 的分布函数和密度函数。

解: $F_Y(y) = P_Y(Y \le y) = P_X(X^2+1 \le y)$

$= \begin{cases} P_X(-\sqrt{y-1} \le X \le \sqrt{y-1}) = P_X(0 \le X \le \sqrt{y-1}), & y \ge 1. \\ P_X(\phi) = 0 & y < 1 \end{cases}$

$= \int_0^{\sqrt{y-1}} \frac{1}{2}\,dx = \frac{\sqrt{y-1}}{2}$.

$p_Y(y) = \frac{d}{dy} F_Y(y) = \begin{cases} \frac{d}{dy}\left(\frac{\sqrt{y-1}}{2}\right) = \frac{1}{4\sqrt{y-1}}, & y \ge 1 \\ 0 & y < 1 \end{cases}$

Outline

动机

有时候仅仅用一个随机变量去描述某个样本点 ω 的信息是不够的，此时就需要引入随机向量。

观察下面的随机试验：

- 研究 $4-6$ 岁年龄段儿童的生长发育情况，考虑每个儿童 (ω) 的身高 ($X_1(\omega)$) 和体重 ($X_2(\omega)$)
- 研究每个家庭的支出情况，考察每个家庭 (ω) 的衣 ($X_1(\omega)$) 食 ($X_2(\omega)$) 住 ($X_3(\omega)$) 行 ($X_4(\omega)$) 的消费情况

随机向量的定义

当一个随机试验的结果是若干并列的随机元素的时候，我们可以用一个数组（向量）来记录随机试验的结果，这就是 随机向量 。

定义 3.1.1

$(\Omega, \mathcal{F}, P)$ 为概率空间，$X_i, i = 1, 2, \cdots, n$ 是定义在同一个样本空间 Ω 上的 n 个随机变量，则称向量 $(X_1, X_2, \cdots, X_n)$ 为 随机向量 或者 n 维随机变量.

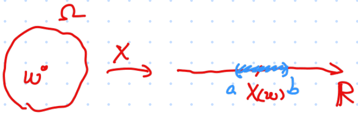
多维随机变量

随机向量 $(X_1, X_2, \cdots, X_n)$ 是 某一个 基本事件空间 Ω 到 n 维实空间的一个映射:

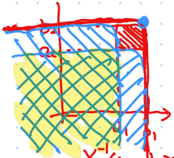
$$\Omega \ni \omega \mapsto (X_1(\omega), X_2(\omega), \cdots, X_n(\omega)) \in \mathbb{R}^n$$

即随机向量的每个分量都是一个随机变量。

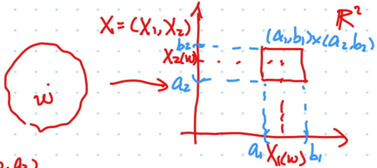
$X^{-1}((-\infty,b)) \in \mathcal{F} \Leftarrow\!\leadsto X^{-1}((a,b)) \in \mathcal{F}.$ $\Leftarrow$ X可测



$\because (a,b) = (-\infty,b) \setminus (-\infty,a]$



$X = (X_1, X_2)$



$X^{-1}((-\infty,b_1)\times(-\infty,b_2)) \leftrightsquigarrow X^{-1}((a_1,b_1)\times(a_2,b_2)) \in \mathcal{F}.$ $\Leftarrow$ X可测

$(a_1,b_1)\times(a_2,b_2) = (-\infty,b_1)\times(-\infty,b_2) - (-\infty,a_1)\times(-\infty,b_2) - (a_1,b_1)\times(-\infty,a_2)$

$= (-\infty,b_1)\times(-\infty,a_2) - (-\infty,a_1)\times(-\infty,a_2)$

联合分布函数

定义 3.1.2

设 $(\Omega, \mathcal{F}, P)$ 为概率空间，$(X_1, X_2, \cdots, X_n)$ 为其上的随机向量，它的 联合分布函数 定义为

$$
\begin{aligned}
& F_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n) \\
& = P(\omega \in \Omega : X_1(\omega) \le x_1, X_2(\omega) \le x_2, \cdots, X_n(\omega) \le x_n) \\
& = P(\omega \in \Omega : \bigcap_{i=1}^{n} \{X_i(\omega) \le x_i\})
\end{aligned}
$$

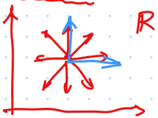
联合分布函数的性质：有界性

对于 $\forall (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n$

$$0 \le F_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n) \le 1$$

联合分布函数的性质：连续性

$F_{X_1,X_2,\cdots,X_n}(x_1,x_2,\cdots,x_n)$ 关于每个变元 x_i 单调递增并且右连续，$i=1,2,\cdots,n$.
固定除了 x_i 以外的所有 x_j,



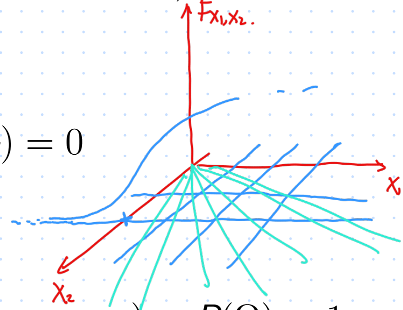
$$
\begin{aligned}
G_{X_i}(x_i) &= F_{X_1,\cdots,X_i\cdots,X_n}(x_1,\cdots,x_i,\cdots,x_n) \\
&= P\left(\omega\in\Omega : \bigcap_{j\neq i}\{X_j(\omega)\leq x_j\}\cap\{X_i(\omega)\leq x_i\}\right)
\end{aligned}
$$

$G_{X_i}(x_i)$ 是一维随机变量 X_i 的分布函数，其单调递增以及右连续性是显然的。

联合分布函数的性质：边缘特性

$$\lim_{x_i \to -\infty} F_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n) = 0, \ \forall i = 1, 2, \cdots, n$$

$$\lim_{x_1, \cdots, x_n \to +\infty} F_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n) = 1.$$



根据第二条，

$$\lim_{x_i \to -\infty} G_{X_i}(x_i) = 0$$

另外

$$\lim_{x_1, x_2, \cdots, x_n \to +\infty} F_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n) = P(\Omega) = 1$$

联合分布函数的性质：事件概率

对二维随机向量，任意 $(x_1, x_2) \in^n$ 和 $h_i > 0$, $i = 1, 2$ 有

$$
\begin{aligned}
&P(X_1 \in (x_1, x_1 + h_1], X_2 \in (x_2, x_2 + h_2]) \\
&= F_{X_1,X_2}(x_1 + h_1, x_2 + h_2) - F_{X_1,X_2}(x_1, x_2 + h_2) \\
&- F_{X_1,X_2}(x_1 + h_1, x_2) + F_{X_1,X_2}(x_1, x_2) \geq 0
\end{aligned}
$$

对 $n = 2$ 的情形

$$
\begin{aligned}
0 &\leq P(x_1 < X_1 \leq x_1 + h_1, x_2 < X_2 \leq x_2 + h_2) \\
&= (P(X_1 \leq x_1 + h_1, X_2 \leq x_2 + h_2) - P(X_1 \leq x_1, X_2 \leq x \\
&- (P(X_1 \leq x_1 + h_1, X_2 \leq x_2) - P(X_1 \leq x_1, X_2 \leq x_2)) \\
&= ({\color{red}F_{X_1,X_2}(x_1 + h_1, x_2 + h_2) - F_{X_1,X_2}(x_1, x_2 + h_2)}) \\
&- ({\color{blue}F_{X_1,X_2}(x_1 + h_1, x_2) - F_{X_1,X_2}(x_1, x_2)})
\end{aligned}
$$

柯尔莫哥洛夫存在性定理

Kolmogorov 存在性定理

若有定义在 $\mathbb{R}^n$ 上的实函数 F 满足定理 3.1.1 的所有性质，那么就可以定义一个概率空间 $(\Omega, \mathcal{F}, P)$ 和其上的随机向量 $(X_1, X_2, \cdots, X_n)$，使得

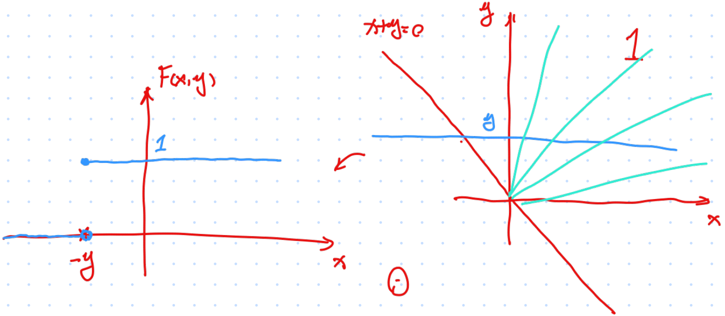
$$F_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n) = F(x_1, x_2, \cdots, x_n),$$

$\forall (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n$

讨论

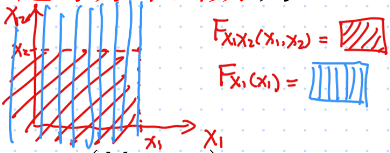
请说明下面的函数能不能成为某个二维随机向量的分布函数，为什么?

$$F(x,y)=\begin{cases}0, & x+y<0\\ 1, & x+y\geq 0\end{cases}$$



边缘分布函数

定义 X_i， $i = 1, \cdots, n$， 的 边缘分布函数 为 $F_{X_i}(x_i)$， 那么



$F_{X_1,\cdots,X_n}$

$$
\begin{aligned}
F_{X_i}(x_i) &= P(\underline{X_i \le x_i}) \\
&= P\big((X_1 \le +\infty) \cap \cdots \cap \underline{(X_i \le x_i)} \cap \cdots \\
&\qquad \cap (X_n \le +\infty)\big) \\
&= \lim_{x_j \to +\infty, j \ne i} F_{X_1,\cdots,X_i,\cdots,X_n}(x_1, \cdots, x_i, \cdots, x_n)
\end{aligned}
$$

边缘分布反映的是随机向量每个分量 单独 的分布。

二维随机变量

二维随机向量的边缘分布函数定义

$$F_{X_1}(x_1) = \lim_{x_2 \to +\infty} F_{X_1,X_2}(x_1, x_2) = F_{X_1,X_2}(x_1, +\infty)$$

$$F_{X_2}(x_2) = \lim_{x_1 \to +\infty} F_{X_1,X_2}(x_1, x_2) = F_{X_1,X_2}(+\infty, x_2)$$

Outline

二维离散型随机向量 $X: \Omega \to \mathbb{R}^2$

取值为有限个或者可列个的随机向量为离散型随机向量

设二维离散型随机向量 (X, Y) 的取值为 (x_i, y_j) ，$i, j = 1, 2, \cdots$ ， 其联合分布列 $\{P(X = x_i, Y = y_j)\}$ 为

$$P(X = x_i, Y = y_j) = p_{ij}, \ \ i, j = 1, 2, \cdots .$$

显然有

- $p_{ij} \geq 0$
- $\sum_{i=1}^{+\infty} \sum_{j=1}^{+\infty} p_{ij} = 1$

二维离散型随机向量

此时，(X, Y) 的联合分布函数为

$$F_{X,Y}(x, y) = \sum_{x_i \le x, y_j \le y} p_{ij}$$

且对任意的 $a < b$，$c < d$，有

$$P(a < X \le b, c < Y \le d) = \sum_{a < x_i \le b, c < y_j \le d} p_{ij}$$

即 (X, Y) 的统计特性完全由概率分布 $\{p_{ij}, i, j = 1, 2, \cdots\}$ 确定。

二维离散型随机向量的分布列

$X \sim \begin{pmatrix} x_1 & \cdots & x_n & \cdots \\ p_1 & \cdots & p_n & \cdots \end{pmatrix}$

$P_{i\cdot} = \sum_{j=1}^{+\infty} P_{ij}$

表: **二维离散随机向量的分布列**

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$	$p_{i\cdot}$
x_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$	$\sum_j p_{1j}$
x_2	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$	$\sum_j p_{2j}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
x_i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$	$\sum_j p_{ij}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
$p_{\cdot j}$	$\sum_i p_{i1}$	$\sum_i p_{i2}$	$\cdots$	$\sum_i p_{ij}$	$\cdots$	1

二维离散型随机向量的分布列

注意上面的表格中的最右列和最下行,

$$p_{i\cdot} = \sum_j p_{ij} = \sum_j P(X = x_i, Y = y_j)$$

$$= P\left(X = x_i, \bigcup_{j=1}^{+\infty}(Y = y_j)\right) = P(X = x_i), i = 1, 2, \cdots .$$

$$p_{\cdot j} = \sum_i p_{ij} = \sum_i P(X = x_i, Y = y_j)$$

$$= P\left(\bigcup_{i=1}^{+\infty}(X = x_i), Y = y_j\right) = P(Y = y_j), j = 1, 2, \cdots .$$

即 $\{p_{i\cdot}, i = 1, 2, \cdots\}$ 和 $\{p_{\cdot j}, j = 1, 2, \cdots\}$ 分别为 X 和 Y 的边缘分布。

例 3.2.1

已知 10 件产品中有 3 件一等品，5 件二等品，2 件三等品。现从中任取 4 件，求其中一等品件数 X 与二等品件数 Y 的联合分布。

解：根据题目条件以及问题，$X = 0, 1, 2, 3$, $Y = 0, 1, 2, 3, 4$，且 $2 \le i + j \le 4$，按照古典概型计算联合分布

$$p_{ij} = P(X = i, Y = j) = \frac{\binom{3}{i}\binom{5}{j}\binom{2}{4-i-j}}{\binom{10}{4}}$$

其中 $i = 0, 1, 2, 3$, $j = 0, 1, 2, 3, 4$, $2 \le i + j \le 4$.

例 3.2.1

写出 X, Y 的联合分布表

表: **例题 3.2.1**

X \ Y	0	1	2	3	4	$p_{i\cdot}$
0	0	0	$10/210$	$20/210$	$5/210$	$35/210$
1	0	$15/210$	$60/210$	$30/210$	0	$105/210$
2	$3/210$	$30/210$	$30/210$	0	0	$63/210$
3	$2/210$	$5/210$	0	0	0	$7/210$
$p_{\cdot j}$	$5/210$	$50/210$	$100/210$	$50/210$	$5/210$	1

例 3.2.2

设随机试验只有 A, B, C 三个结果，各结果出现的概率分别为 $p, q, 1-p-q$。现将该随机试验独立的做 n 次，记 X 和 Y 分别为 n 次试验中 A 和 B 发生的次数，试求 (X, Y) 的联合分布和边缘分布。

解：X 和 Y 的取值只可能是 $0, 1, 2, \cdots, n$，并且 $X+Y \leq n$，由于试验是独立的，按独立试验概型计算，联合分布

$$P(X=i, Y=j) = \binom{n}{i}\binom{n-i}{j} p^i q^j (1-p-q)^{n-i-j},$$

$0 \leq i+j \leq n$。

例 3.2.2

记 $p_{ij} = P(X = i, Y = j)$，那么，

$$\sum_{i=0}^{n-j} p_{ij} = b(j; n, q), \quad \sum_{j=0}^{n-i} p_{ij} = b(i; n, p)$$

称上述 (X, Y) 服从三项分布，其边缘分布可以证明

$$X \sim B(n, p), \quad Y \sim B(n, q)$$

练习

设二维随机向量 (X_1, X_2) 的分布为
$P(X_i = -1) = P(X_i = 1) = 0.25$， $P(X_i = 0) = 0.5$，
且满足 $P(X_1X_2 = 0) = 1$， 求 (X_1, X_2) 的联合分布
以及 $P(X_1 = X_2)$ 。

$P(X_1X_2 \neq 0) = 0$

$X_2 \backslash X_1$	-1	0	1
-1	0	$\frac{1}{4}$	0
0	$\frac{1}{4}$	0	$\frac{1}{4}$
1	0	$\frac{1}{4}$	0

练习

假设随机变量 $X_1, \ldots, X_4$ 相互独立且同分布，
$P(X_i = 1) = 0.6$, $P(X_i = 0) = 0.4$。
求行列式 $\begin{vmatrix} X_1 & X_2 \\ X_3 & X_4 \end{vmatrix}$ 的概率分布

二维离散型随机向量条件分布列

已知二维随机向量 (X, Y) 的分布列为

$$P(X = a_i, Y = b_j) = p_{ij}, i, j = 1, 2, \cdots .$$

在事件 $\{Y = b_j\}$ 发生的前提下 X 的分布列称为条件分布列，记为 $P(X = a_i | Y = b_j)$，$i = 1, 2, \cdots$。

条件分布列

由条件概率定义有，在事件 $\{Y=b_j\}$ 发生条件下 X 的条件分布列为

$$P(X=a_iY=b_j)=\frac{P(X=a_i,Y=b_j)}{P(Y=b_j)}=\frac{p_{ij}}{p_{\cdot j}},i=1,2,\cdots.$$

类似的，在事件 $\{X=a_i\}$ 发生条件下 Y 的条件分布列为

$$P(Y=b_jX=a_i)=\frac{P(X=a_i,Y=b_j)}{P(X=a_i)}=\frac{p_{ij}}{p_{i\cdot}},j=1,2,\cdots.$$

课堂练习

在例 3.2.1 中，计算 $\{Y=1\}$ 条件下 X 的条件分布列以及 $\{X=2\}$ 条件下 Y 的条件分布列

表: **例题 3.2.1**

X \ Y	0	1	2	3	4	$p_{i\cdot}$
0	0	0	10/210	20/210	5/210	35/210
1	0	15/210	60/210	30/210	0	105/210
2	3/210	30/210	30/210	0	0	63/210
3	2/210	5/210	0	0	0	7/210
$p_{\cdot j}$	5/210	50/210	100/210	50/210	5/210	1