

随机向量及其分布

若随机变量 X_1, X_2, \dots, X_n 定义在同一样本空间 Ω 上, 则称 (X_1, X_2, \dots, X_n) 为 n 维随机变量

二维随机向量及其分布函数

分布函数

类似研究随机变量, 可以为二维随机向量定义其 cdf

设 (X, Y) 是二维随机向量, 对任意实数 x, y , 二元函数

$$F(x, y) = P(X \leq x, Y \leq y)$$

称为二维随机向量的分布函数, 或称为 X, Y 的**联合分布函数**

易得

$$P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1)$$

分布函数 $F(x, y)$ 有如下性质

1. $F(x, y)$ 对于 x, y 都是单调不减的
2. $0 \leq F(x, y) \leq 1$, 且对于固定的 y 有 $F(-\infty, y) = 0$, 对于固定的 x 有 $F(x, -\infty) = 0$ 。 $F(-\infty, -\infty) = 0, F(+\infty, +\infty) = 1$
3. $F(x+0, y) = F(x, y), F(x, y+0) = F(x, y)$, 即 $F(x, y)$ 关于 x, y 都是右连续的
4. 对于任意实数 $x_1 \leq x_2, y_1 \leq y_2$ 有 $F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1) \geq 0$

满足以上性质的二元函数一定是某个二维随机向量的联合分布函数

根据联合分布函数, 也可以得到两个分量各自的分布函数

$$F_X(x) = P(X \leq x) = P(X \leq x, Y \leq +\infty) = F(x, +\infty)$$

同理

$$F_Y(y) = F(+\infty, y)$$

将 X, Y 的分布函数称为**边缘分布函数**

二维离散型随机向量

若二维随机向量 (X, Y) 的每个分量都是离散型随机变量，称其为二维离散型随机向量

设二维离散型随机向量 (X, Y) 取值可能为 $(x_i, y_j), i, j = 1, 2, \dots$ ，则

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

称为 (X, Y) 的联合分布律。

类比离散型随机变量，联合分布律有性质

- $p_{ij} \geq 0, i, j = 1, 2, \dots$
- $\sum_{i=1}^{+\infty} \sum_{j=1}^{+\infty} p_{ij} = 1$

二维离散型随机向量有两种重要的分布：多项分布，多元超几何分布。这两种分布分别是二项分布和超几何分布的推广

多项分布：在 n 次重复独立实验中有三种不同结果，分别为 A_1, A_2, A_3 ，且 $P(A_i) = p_i$ 。以 X_1, X_2 记录 A_1, A_2 发生的概率，则有

$$P(X_1 = k_1, X_2 = k_2) = \frac{n!}{k_1! k_2! (n - k_1 - k_2)!} p_1^{k_1} p_2^{k_2} (1 - p_1 - p_2)^{n - k_1 - k_2}$$

记为 $(X_1, X_2) \sim M(n; p_1, p_2, p_3)$

多元超几何分布：袋子中有三种球，分别有 N_1, N_2, N_3 个，一共有 N 个球，设 X_1, X_2 记录不放回摸取 n 个球时摸到 1, 2 中球的个数，则有

$$P(X_1 = n_1, X_2 = n_2) = \frac{C_{N_1}^{n_1} C_{N_2}^{n_2} C_{N_3}^{n_3}}{C_N^n}$$

二维离散型随机向量的联合分布函数为

$$F(x, y) = P(X \leq x, Y \leq y) = \sum_{x_i \leq x, y_j \leq y} p_{ij}$$

可定义其边缘分布律

$$P(X = x_i) = \sum_{j=1}^{+\infty} p_{ij} = p_{i\cdot}$$
$$P(Y = y_j) = \sum_{i=1}^{+\infty} p_{ij} = p_{\cdot j}$$

二维连续型随机变量

设二维随机向量 (X, Y) 的分布函数为 $F(x, y)$ ，若存在非负可积二元函数 $p(x, y)$ 满足

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y p(u, v) du dv$$

则称 (X, Y) 为二维连续型随机向量，称 $p(x, y)$ 为其联合概率密度函数

易得联合密度有性质

1. $p(x, y) \geq 0$
2. $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x, y) dx dy = 1$

若一个二元函数满足上述性质，则可以保证 $F(x, y)$ 满足分布函数的性质，故 $p(x, y)$ 是某个二维连续型随机向量的联合密度

同样可以得到对于任意实数 $x_1 \leq x_2, y_1 \leq y_2$ 有

$$P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = \int_{x_1}^{x_2} \int_{y_1}^{y_2} p(u, v) du dv$$

更一般的，设 G 为平面任一区域，则

$$P((X, Y) \in G) = \iint_G p(x, y) dx dy$$

重积分的计算可以使用累次积分或是换元法

联合密度也可以看作是联合分布的导函数。若 $p(x, y)$ 在点 (x_0, y_0) 处连续，则

$$\left. \frac{\partial^2 F(x, y)}{\partial x \partial y} \right|_{(x_0, y_0)} = p(x_0, y_0)$$

对于任意有限实数 x ，令 $G = (-\infty, x] \times (-\infty, +\infty)$ ，则可以得到 X 的边缘分布

$$F_X(x) = P((X, Y) \in G) = \int_{-\infty}^x \int_{-\infty}^{+\infty} p(x, y) dx dy$$

求导，可以得到 X 的概率密度为

$$p_X(x) = \int_{-\infty}^{+\infty} p(x, y) dy$$

同理， Y 的概率密度为

$$p_Y(y) = \int_{-\infty}^{+\infty} p(x, y) dx$$

这两个概率密度称为 X, Y 的**边缘概率密度**

连续型随机向量的分量都是连续型随机变量，但连续型随机变量组成的向量不一定是连续型随机向量

平面上有两种常见的二维连续型分布：均匀分布，二元正态分布

均匀分布：设 G 是 \mathbb{R}^2 上的有限区域，其面积 $S > 0$ ，由密度函数

$$p(x, y) = \begin{cases} \frac{1}{S} & (x, y) \in G \\ 0 & (x, y) \notin G \end{cases}$$

给出的分布称为 G 上的均匀分布。若 G 为矩形 $[a, b] \times [c, d]$ ，则联合密度为

$$p(x, y) = \begin{cases} \frac{1}{(b-a)(d-c)} & (x, y) \in G \\ 0 & (x, y) \notin G \end{cases}$$

则可以得到 X 的边缘分布

$$F_X(x) = P(X \leq x) = \frac{x-a}{b-a}, x \in [a, b]$$

其密度函数为

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

即 X 服从区间 $[a, b]$ 上的均匀分布。同理 Y 服从区间 $[c, d]$ 上的均匀分布

二元正态分布：若二维随机向量 (X, Y) 的联合密度函数为

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

其中 $\mu_1, \mu_2 \in \mathbb{R}, \sigma_1, \sigma_2 > 0, |\rho| < 1$ 均为常数，则称 (X, Y) 服从二元正态分布。记为

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

易得 X 的概率密度为

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}$$

即 $X \sim N(\mu, \sigma_1^2)$ ，同理可得 $Y \sim N(\mu, \sigma_2^2)$

二维正态分布的边缘分布是正态分布，且其不依赖于参数 ρ ，从另一角度也说明了根据 X, Y 的边缘分布**一般不能确定联合分布**

相互独立的随机变量

设 $F(x, y)$ 是二维随机向量 (X, Y) 的联合分布, $F_X(x), F_Y(y)$ 为其边缘分布, 若对所有 x, y 都有

$$\begin{aligned}P(X \leq x, Y \leq y) &= P(X \leq x)P(Y \leq y) \\F(x, y) &= F_X(x)F_Y(y)\end{aligned}$$

则称随机变量 X, Y 是相互独立的。

对于连续型随机向量 (X, Y) 而言, 设其联合密度为 $p(x, y)$, 且 X, Y 的边缘密度为 $p_X(x), p_Y(y)$, 则 X, Y 独立等价于

$$p(x, y) = p_X(x)p_Y(y)$$

在平面上几乎处处成立

几乎处处成立意为除去平面上面积为 0 的集合, 处处成立

对于离散型随机向量 (X, Y) 而言, X, Y 相互独立等价于对于所有 x_i, y_j 有

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$$

实际中根据分布律或密度函数判断独立性比使用分布函数要方便

事实上, 对于连续型随机向量 (X, Y) , X, Y 独立 \iff 存在函数 $g_1(x), g_2(y)$, 满足

$$p(x, y) = g_1(x)g_2(y)$$

对于二维正态分布来说, 其分量独立 \iff 参数 $\rho = 0$

随机向量函数的分布

问题: 设 (X, Y) 是一个二维随机向量, $Z = g(X, Y)$ 是一个一维随机变量, 由 (X, Y) 的分布计算 Z 的分布

二维离散型随机向量

设 (X, Y) 是二维离散型随机变量, 则 $Z = g(X, Y)$ 最多可以取可数个值, Z 也是离散型随即变量。其分布律为

$$P(Z = z_k) = P(g(X, Y) = z_k) = \sum_{g(x_i, y_j) = z_k} p_{ij}$$

考虑几种特殊的 $Z = g(X, Y)$

随机变量和的分布

设 $Z = X + Y$ ，则其分布律为

$$P(Z = z_k) = \sum_{x_i} P(X = x_i, Y = z_k - x_i)$$

特别的，若 X, Y 独立

$$P(Z = z_k) = \sum_{x_i} P(X = x_i)P(Y = z_k - x_i)$$

上述公式被称为**卷积公式**

离散型分布的可加性

- 泊松分布：两个独立的泊松分布 $X \sim P(\lambda_1), Y \sim P(\lambda_2)$ ，其和 $Z = X + Y \sim P(\lambda_1 + \lambda_2)$
- 二项分布：两个独立的二项分布 $X \sim B(n_1, p), Y \sim B(n_2, p)$ ，其和 $Z = X + Y \sim B(n_1 + n_2, p)$

随机变量商的分布

设 $P(Y = 0) = 0, Z = X/Y$ ，则其分布律为

$$P(Z = z_k) = \sum_{y_j} P(X = z_k \cdot y_j, Y = y_j)$$

对于随机变量的乘积，也可也看作是商的分布

二维连续型随机向量

对于二维连续型随机向量 (X, Y) ，其联合密度函数为 $p(x, y)$ ，其函数 $Z = g(X, Y)$ 也是一个连续型随机变量，计算其密度函数可以使用**分布函数法**

$$F_Z(z) = P(Z \leq z) = \iint_{g(x,y) \leq z} p(x, y) dx dy$$

求出其分布函数 $F_Z(z)$ 后求密度函数只要对其求导即可

瑞利 (Rayleigh) 分布： X, Y 独立同分布于 $N(0, 1)$ ，则 $Z = \sqrt{X^2 + Y^2}$ 的密度函数为

$$p_Z(z) = \begin{cases} ze^{-\frac{z^2}{2}} & z \geq 0 \\ 0 & z < 0 \end{cases}$$

分布函数法是最普适的求连续型随机向量分布的方法，对于一些特殊的函数，有

随机变量和的分布

设 (X, Y) 有联合密度函数 $p(x, y)$, $Z = X + Y$ 是连续型随机变量, 其密度函数为

$$p_Z(z) = \int_{-\infty}^{+\infty} p(x, z-x) dx$$

特别的, 当 X, Y 独立, 其边缘密度为 $p_X(x), p_Y(y)$ 时, 有

$$p_Z(z) = \int_{-\infty}^{+\infty} p_X(x)p_Y(z-x) dx$$

这个公式也被称为 $p_X(x)$ 和 $p_Y(y)$ 的**卷积公式**

注意在积分时上下限的选择要使得密度函数不为 0

随机变量商的分布

设 (X, Y) 有联合密度函数 $p(x, y)$, 且 $Y \neq 0$, 则 $Z = X/Y$ 是连续型随机变量, 其密度函数为

$$p_Z(x) = \int_{-\infty}^{+\infty} |y|p(x/y, y) dy$$

Proof. Z 的分布函数为

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = \iint_{x/y \leq z} p(x, y) dx dy \\ &= \int_0^{+\infty} \int_{-\infty}^{zy} p(x, y) dx dy + \int_{-\infty}^0 \int_{zy}^{+\infty} p(x, y) dx dy \end{aligned}$$

则其密度函数为

$$\begin{aligned} p_Z(z) &= F'_Z(z) = \left(\int_0^{+\infty} \int_{-\infty}^{zy} p(x, y) dx dy \right)' + \left(\int_{-\infty}^0 \int_{zy}^{+\infty} p(x, y) dx dy \right)' \\ &= \int_0^{+\infty} yp(zy, y) dy + \int_{-\infty}^0 -yp(zy, y) dy \\ &= \int_{-\infty}^{+\infty} |y|p(zy, y) dy \end{aligned}$$

随机变量极值的分布

设 X, Y 相互独立, 则考虑 $M = \max(X, Y), N = \min(X, Y)$

对于 $M = \max(X, Y)$, 其分布函数为

$$F_M(z) = P(M \leq z) = P(\max(X, Y) \leq z) = P(X \leq z, Y \leq z)$$

由于 X, Y 独立, 有

$$F_M(z) = P(X \leq z)P(Y \leq z) = F_X(z)F_Y(z)$$

对于 $N = \min(X, Y)$, 其分布函数为

$$F_N = P(\min(X, Y) \leq z) = 1 - P(\min(X, Y) > z) = 1 - P(X > z, Y > z)$$

由于 X, Y 独立, 有

$$F_N(z) = 1 - P(X > z)P(Y > z) = 1 - (1 - F_X(z))(1 - F_Y(z))$$

上述结论可以推广至有限个相互独立的随机变量。

进而, 当 X, Y 是连续型随机变量时, M, N 也是连续型随机变量, 对其分布函数求导, 有

$$\begin{aligned} p_M(z) &= p_X(z)F_Y(z) + F_X(z)p_Y(z) \\ p_N(z) &= p_X(z)(1 - F_Y(z)) + p_Y(z)(1 - F_X(z)) \end{aligned}$$

对于一个串并联系统来说, 设两个元件寿命的随机变量为 X, Y , 则整个系统的寿命的随机变量为

- 串联: $\min(X, Y)$
- 并联: $\max(X, Y)$
- 备用: $X + Y$

一般情况下随机变量函数的分布

首先考虑 (X, Y) 的两个函数的联合分布。

设 (X, Y) 的联合密度为 $p(x, y)$, 随机变量 U, V 都是 (X, Y) 的函数

$$U = g(X, Y), V = h(X, Y)$$

这里的函数 g, h 满足:

存在唯一的逆变换

$$\begin{cases} x = x(u, v) \\ y = y(u, v) \end{cases}$$

存在连续的一阶偏导数, 且雅可比行列式

$$J(u, v) = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} \neq 0$$

则其联合分布为

$$F_{(U,V)}(u, v) = P(U \leq u, V \leq v) = \iint_{g(x,y) \leq u, h(x,y) \leq v} p(x, y) dx dy$$

进行变量替换, 可得

$$F_{(U,V)}(u, v) = \int_{-\infty}^v \int_{-\infty}^u p(x(z_1, z_2), y(z_1, z_2)) |J(z_1, z_2)| dz_1 dz_2$$

则 (U, V) 的联合密度为

$$l(u, v) = p(x(u, v), y(u, v)) |J(u, v)|$$

若求 $Z = g(X, Y)$ 的分布时, 只需求上述联合分布的边缘分布即可, 即

$$p_Z(z) = \int_{-\infty}^{+\infty} l(z, v) dv$$

此时一般另一个函数选择 $V = x$ 等易于计算的函数。