一、随机向量

(一) 定义

设 $X_1, X_2, \cdots, X_p$ 为 p 个随机变量，它们组成的向量 $X=[X_1, X_2, \cdots, X_p]^T \in R^p$ 称为随机向量。

(二). 联合概率分布函数 jcdf

记 $F(x_1, x_2, \cdots, x_p) = P(X_1 \le x_1, X_2 \le x_2, \cdots, X_p \le x_p)$ 为联合概率分布函数，或记 $x=[x_1, x_2, \cdots, x_p]^T$，则联合概率分布函数为 $F(x) = P(X \le x)$

(三). 联合概率密度函数 jpdf

若存在非负函数 $f(x_1, x_2, \cdots, x_p)$ 使得对 $\forall x_1, x_2, \cdots, x_p$ 有

$$F(x_1, x_2, \cdots, x_p) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_p} f(t_1, t_2, \cdots, t_p)\, dt_1 dt_2 \cdots dt_p$$

(四) 边际分布

记 $X=[X_1, X_2, \cdots, X_p]^T$ 的 $q\ (q<p)$ 个分量 $X^{(1)} = [X_{i_1}, X_{i_2}, \cdots, X_{i_q}]^T \in \mathbb{R}^q$ 的分布称为边际分布：

$$\begin{aligned} P(X^{(1)} \le u) &= P(X_{i_1} \le u_1, X_{i_2} \le u_2, \cdots X_{i_q} \le u_q) \\ &= F(u_1, u_2, \cdots, u_q, \infty, \infty \cdots \infty) \\ &= F(u) \end{aligned}$$

同样的，边际密度函数为

$$g(u) = \int_{-\infty}^{x_{i_1}} \int_{-\infty}^{x_{i_2}} \cdots \int_{-\infty}^{x_{i_q}} f(t_{i_1}, t_{i_2}, \cdots, t_{i_q}, \cdots)\, dt_{i_1} dt_{i_2} \cdots dt_{i_q}$$

(五) 条件分布

记 X 分块为 $X = [U^T\ V^T]^T$ 的概率密度为 $f(u,v)$ 其中 $U, u \in \mathbb{R}^r$ 而 V $v \in \mathbb{R}^t$ $(r+t=p)$。记 U 的边际密度为 $f_1(u)$ 则 V 在 U 条件下的条件密度：

$$f(v|u) = \frac{f(u,v)}{f_1(u)}$$

(六) 分量间相互独立

随机向量 X 的各分量 $X_1, \cdots, X_p$ 相互独立，当且仅当：

$$F(x_1, x_2, \cdots x_p) = \prod_{i=1}^{p} F_i(x_i)$$

其中 $F_i(\cdot)$ 为 X_i 的边际分布函数。

(七) 随机向量间相互独立

随机向量 X, Y 相互独立，当且仅当：

$$P(X \le x, Y \le y) = P(X \le x) \cdot P(Y \le y)$$

即 $F(x, y) = F_X(x) \cdot F_Y(y)$

二、统计量

(一). 总体均值向量

记 $X = [X_1, \cdots, X_p]^T$ 的均值为 $\mu = [\mu_1, \mu_2, \cdots, \mu_p]^T$ 定义为：

$$\mu = E(X) = [E(X_1), E(X_2), \cdots, E(X_p)]^T \in \mathbb{R}^p$$

$$:= [\mu_1, \mu_2, \cdots, \mu_p]^T \in \mathbb{R}^p$$

其中 $\mu_i = E(X_i) = \int_{\mathbb{R}} x_i f_i(x) dx$ or $\sum x_i p_i(x_i)$

(二). 总体方差-协方差矩阵

记 $X = [X_1, X_2, \cdots, X_p]^T \in \mathbb{R}^p$，其方差-协方差矩阵为

$$\Sigma = E\left[(X-\mu)(X-\mu)^T\right] \in \mathbb{R}^{p \times p}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}$$

(三). 总体相关系数矩阵

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

$$= V^{-\frac{1}{2}} \Sigma V^{-\frac{1}{2}}$$

其中 $\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$ $\quad V = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2)$ $\quad V^{-\frac{1}{2}} = \mathrm{diag}(1/\sigma_1, 1/\sigma_2, \cdots, 1/\sigma_p)$

(四) 分块矩阵

记 $X = [U^T \ V^T]^T \in \mathbb{R}^p$ 其中 $U \in \mathbb{R}^q$ $V \in \mathbb{R}^{p-q}$。记 $\mu = [\mu_1^T \ \mu_2^T]^T$ 其中 $\mu_1 \in \mathbb{R}^q$ $\mu_2 \in \mathbb{R}^{p-q}$。记 $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \in \mathbb{R}^{p\times p}$ 其中 $\Sigma_{11} \in \mathbb{R}^{q\times q}$ $\Sigma_{12}, \Sigma_{21}^T \in \mathbb{R}^{q\times(p-q)}$ $\quad \Sigma_{22} \in \mathbb{R}^{(p-q)\times(p-q)}$。

三、线性变换

(一). 随机向量的线性组合

① 线性组合:

记 $c = [c_1, c_2, \cdots c_p]^T \in \mathbb{R}^p$，则 X 的一个线性组合为:

$$c^T X = \sum_{i=1}^{p} c_i X_i \in \mathbb{R}$$

② 均值

$$E(c^T X) = \sum_{i=1}^{p} c_i E(X_i) = \sum_{i=1}^{p} c_i \mu_i \in \mathbb{R}$$

$$= c^T \cdot E(X) = c^T \cdot \mu \in \mathbb{R}$$

③ 方差

$$\mathrm{var}(c^T X) = c^T \cdot \mathrm{cov}(X) \cdot c = c^T \Sigma c \in \mathbb{R}$$

(二). 随机向量的多个线性组合

记 $X=[X_1,X_2,\cdots,X_p]^T\in\mathbb{R}^p$，有 q 个向量 $C_i\in\mathbb{R}^p$ 其中 $C_i=[C_{i1},C_{i2},\cdots,C_{ip}]^T\in\mathbb{R}^p$，于是 X 的 q 个线性组合 Z_j 为

$$Z_j = C_j^T X \qquad j=1,2,\cdots,q$$

组合为

$$Z = CX$$

其中 $Z=[Z_1,Z_2,\cdots,Z_q]\in\mathbb{R}^q$ $\quad X=[X_1,\cdots,X_p]\in\mathbb{R}^p$

$$C=\begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1p} \\ C_{21} & C_{22} & \cdots & C_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ C_{q1} & C_{q2} & \cdots & C_{qp} \end{bmatrix} = \begin{bmatrix} C_1^T \\ C_2^T \\ \vdots \\ C_q^T \end{bmatrix} \in\mathbb{R}^{q\times p}$$

Z 成为一个新的随机向量

① 均值向量

$$E(Z) = C\cdot E(X)$$

② 方差-协方差矩阵

$$\mathrm{cov}(Z) = C\Sigma C^T$$

其中 $\Sigma=\mathrm{cov}(X)$

四. 多元随机抽样

(一). 多元随机抽样

对于 p 个特征，事先预定 n 个样本，在未抽样前这 $n\times p$ 个值仍为随机量：

$$X_{n\times p} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \in \mathbb{R}^{n\times p}$$

此时 X_{ij} 为随机量（未抽样前）

(二). 定理

设 $X_1, X_2, \cdots, X_n \in \mathbb{R}^p$ 是来自均值为 $\mu \in \mathbb{R}^p$，协方差为 $\Sigma \in \mathbb{R}^{p\times p}$ 的总体的一个随机样本 (iid) [注：此处 X_i 为随机向量，与之前记法不同]

则有 $\bar{X}$ 为 μ 的无偏估计，样本方差 S 与 S_n 有：

$$E(\bar{X}) = \mu$$

$$\mathrm{cov}(\bar{X}) = \frac{1}{n}\Sigma$$

$$E\left(\frac{n}{n-1}S_n\right) = \Sigma$$

其中 $S_n \triangleq \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^T \in \mathbb{R}^{p\times p}$ $\quad S = \frac{n}{n-1}S_n$

注意：这里符号与之前有异，$X_1, X_2, \cdots, X_n \in \mathbb{R}^p$ 代表的是样本

Proof: $E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \frac{1}{n}\left\{\sum_{i=1}^{n}E(X_i)\right\}$

$\because X_i$ iid $\sim (\mu, \Sigma)$ 总体 $\Rightarrow \forall\ E(X_i) = \mu$

故 $E(\bar{X}) = \frac{1}{n}\cdot n\mu = \mu \quad \square$

$$\mathrm{cov}(\bar{X}) = E\left\{(\bar{X}-E\bar{X})(\bar{X}-E\bar{X})^T\right\}$$

$$= E\left\{(\bar{X}-\mu)(\bar{X}-\mu)^T\right\}$$

$$= E\left\{\left(\frac{1}{n}\sum X_i - \mu\right)\left(\frac{1}{n}\sum X_i - \mu\right)^T\right\}$$

$$= E\left\{ \frac{1}{n}\left[\sum (X_i-\mu)\right] \cdot \frac{1}{n}\left[\sum (X_i-\mu)^T\right] \right\}$$

$$= \frac{1}{n^2} E\left\{ \sum (X_i-\mu) \sum (X_i-\mu)^T \right\}$$

$\because$ X_i, X_j 相互独立：

$$E\{(X_i-\mu)(X_j-\mu)\} = E(X_i-\mu)\cdot E(X_j-\mu) = 0$$

故

$$\begin{aligned} cov(\bar{X}) &= \frac{1}{n^2} E\left\{ \sum (X_i-\mu)(X_i-\mu)^T \right\} \\ &= \frac{1}{n^2} \sum \underline{E(X_i-\mu)(X_i-\mu)^T} \quad = cov(X_i) = \Sigma \\ &= \frac{1}{n^2} \cdot n\Sigma \quad = \frac{1}{n}\Sigma \quad \square \end{aligned}$$

$$S_n = \frac{1}{n}\sum_{i=1}^{n} (X_i-\bar{X})(X_i-\bar{X})^T \qquad S = \frac{n}{n-1} S_n$$

即证：

$$E(S) = \Sigma$$

$$E(S) = \frac{1}{n-1} E\left\{ \sum (X_i-\bar{X})(X_i-\bar{X})^T \right\}$$

化简

$$\begin{aligned} \sum (X_i-\bar{X})(X_i-\bar{X})^T &= \sum (X_i-\bar{X})X_i^T - \sum (X_i-\bar{X})\bar{X}^T \\ &= \sum X_iX_i^T - \bar{X}\sum X_i^T - \left(\sum X_i - n\bar{X}\right)\bar{X}^T \\ &= \sum X_iX_i^T - n\bar{X}\bar{X}^T - 0 \end{aligned}$$

然后求期望：

$$E\left\{ \sum X_iX_i^T - n\bar{X}\bar{X}^T \right\} = \sum \left\{ E(X_iX_i^T) - E(\bar{X}\bar{X}^T) \right\}$$

【引理】 对于 $\forall$ 随机向量 V，记 $E(V) = \mu_v$ $\quad cov(V) = \Sigma_v$ 则有

$$E(VV^T) = \Sigma_v + \mu_v\mu_v^T$$

Proof: 写出 $cov(V) = \Sigma_v$

$$= E\{(V-\mu_v)(V-\mu_v)^T\}$$

$$= E\{VV^T - \mu_v V^T - V\mu_v^T + \mu_v\mu_v^T\}$$

$$= E(VV^T) - \mu_v\mu_v^T - \mu_v\mu_v^T + \mu_v\mu_v^T$$

$$= E(VV^T) - \mu_v\mu_v^T \quad \square$$

回到定理的证明:

$$E(X_iX_i^T) = \Sigma_{x_i} + \mu_{x_i}\mu_{x_i}^T = \Sigma + \mu\mu^T$$

$$E(\bar{X}\bar{X}^T) = \Sigma_{\bar{x}} + \mu_{\bar{x}}\mu_{\bar{x}}^T = \frac{1}{n}\Sigma + \mu\mu^T$$

于是:

$$E(S) = \frac{1}{n-1}\Sigma\{E(X_iX_i^T) - E(\bar{X}\bar{X}^T)\}$$

$$= \frac{1}{n-1}\cdot\sum_{i=1}^{n}\{\Sigma + \mu\mu^T - \frac{1}{n}\Sigma - \mu\mu^T\}$$

$$= \frac{1}{n-1}\cdot\sum_{i=1}^{n}\{\frac{n-1}{n}\Sigma\}$$

$$= \Sigma \quad \square$$

(三). 随机样本的线性组合

定理 线性组合

$$b^TX = b_1X_1 + b_2X_2 + \cdots + b_pX_p$$

$$c^TX = c_1X_1 + c_2X_2 + \cdots + c_pX_p$$

的样本均值与方差为:

$$b^T\bar{X} \qquad c^T\bar{X}$$

$$b^TSb \qquad c^TSc$$

而 b^TX 与 c^TX 的样本协方差为：

$$\cancel{Cov}(b^TX, c^TX) = b^TSc$$

其中 $\bar{X}$ 为 X 的均值， $S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^T \in \mathbb{R}^{p\times p}$

Proof：设 $X = [X_1, X_2, \cdots, X_p]^T$ 的样本结果为

$$X_{n\times p} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$$

其中 $x_i \in \mathbb{R}^p$

提有第 i 个观测值为 $b^Tx_i = b_1x_{i1} + b_2x_{i2} + \cdots + b_px_{ip}$

故：
$$\begin{aligned}\overline{b^TX} &= \frac{1}{n}(b^Tx_1 + b^Tx_2 + \cdots + b^Tx_n) \\ &= \frac{1}{n}\cdot b^T\cdot \sum_{i=1}^{n} x_i \\ &= b^T\bar{X}\end{aligned}$$

□

其中 $$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}x_i = \begin{bmatrix} \frac{1}{n}(x_{11}+x_{21}+\cdots+x_{n1}) \\ \frac{1}{n}(x_{12}+x_{22}+\cdots+x_{n2}) \\ \vdots \\ \frac{1}{n}(x_{1p}+x_{2p}+\cdots+x_{np}) \end{bmatrix} \in \mathbb{R}^p$$

下证：$Cov(b^TX, c^TX)$

$$\begin{aligned} &= \frac{1}{n-1}\cdot\sum_{i=1}^{n}(b^Tx_i - b^T\bar{x})(c^Tx_i - c^T\bar{x}) \\ &= \frac{1}{n-1}\cdot\sum_{i=1}^{n} b^T(x_i-\bar{x})(x_i-\bar{x})^Tc \\ &= b^T\cdot\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T\right]\cdot c \\ &= b^TSc \end{aligned}$$

□