Ch 6.2 协方差和相关系数

回顾前一次课

若(X,Y)的联合概率密度为$f(x,y)$，则$U=u(X,Y)$和$V=v(X,Y)$的联合密度为$f_{UV}(u,v)=f_{XY}\big(x(u,v),y(u,v)\big)|J|$

n维随机向量的联合分布函数、联合密度函数、边缘分布函数、边缘密度函数、性质、**随机向量X和Y相互独立**

多维正太分布：$X=(X_1,X_2,\cdots,X_n)\sim \mathrm{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$

$X\sim \mathrm{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$分解$\boldsymbol{\Sigma}=\boldsymbol{U}^{\top}\boldsymbol{\Lambda}\boldsymbol{U}$, 则$Y=\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}(X-\boldsymbol{\mu})\sim \mathrm{N}(\boldsymbol{0}_n,I_n)$

设随机向量$X\sim \mathrm{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, 则有$Y=\boldsymbol{A}X+\boldsymbol{b}\sim N(\boldsymbol{A\mu}+\boldsymbol{b},\boldsymbol{A\Sigma A}^{\top})$

多维正太分布重要的性质

定理 5.13 设随机向量 $X=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 和 $Y=(Y_1,Y_2,\cdots,Y_m)^{\mathrm{T}}$, 以及

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{pmatrix}\right),$$

- 随机向量 X 和 Y 的边缘分布分别为 $X\sim\mathcal{N}(\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_{xx})$ 和 $Y\sim\mathcal{N}(\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_{yy})$;
- 随机向量 X 与 Y 相互独立的充要条件是 $\boldsymbol{\Sigma}_{xy}=(\mathbf{0})_{m\times n}$ (元素全为零的 $m\times n$ 矩阵);
- 在 $X=\boldsymbol{x}$ 的条件下随机向量 $Y\sim\mathcal{N}(\boldsymbol{\mu}_y+\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_x),\boldsymbol{\Sigma}_{yy}-\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy})$;
- 在 $Y=\boldsymbol{y}$ 的条件下随机向量 $X\sim\mathcal{N}(\boldsymbol{\mu}_x+\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_y),\boldsymbol{\Sigma}_{xx}-\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx})$.

期望

随机变量$Z = g(X,Y)$的期望为 $E[Z] = \sum_{i,j} g\left(x_i, y_j\right) p_{ij}$

随机变量$Z = g(X,Y)$的期望$E[Z] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y) f(x,y) dxdy$

若随机变量$X \geq Y$, 则有$\boldsymbol{E[X] \geq E[Y]}$

对任意随机变量X, Y有$\boldsymbol{E[X+Y] = E[X] + E[Y]}$

对独立随机变量X和Y, 有$\boldsymbol{E[XY] = E[X]E[Y]}$

对任意随机变量X和Y, 有Cauchy-Schwartz不等式

$$\boldsymbol{|E[XY]| \leq \sqrt{E[X^2]E[Y^2]}}$$

对独立随机变量X和Y, 有

$$\mathrm{Var}(\boldsymbol{X} \pm \boldsymbol{Y}) = \mathrm{Var}(\boldsymbol{X}) + \mathrm{Var}(\boldsymbol{Y})$$

协方差

随机变量的期望或方差仅涉及变量自身的统计信息，没有刻画变量之间的统计信息

协方差：描述随机变量X和Y相互关系的数字特征

定义：设二维随机向量(X,Y)的期望$E[(X-E(X))(Y-E(Y))]$存在，则称其为X和Y的**协方差**，记为

$$\mathrm{Cov}(X,Y)=E\big[\big(X-E(X)\big)\big(Y-E(Y)\big)\big]=E(XY)-E(X)E(Y)$$

协方差是两个随机变量与它们各自期望的偏差之积的期望，由于偏差可正可负，因此协方差可正可负.

协方差的性质

- $\mathrm{Cov}(X, c) = 0$
- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
- $\mathrm{Cov}(X,Y)=\mathrm{Cov}(Y,X)$
- $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\mathrm{Cov}(X, Y)$

对任意常数a和b,随机变量X和Y,有

- $\mathrm{Cov}(aX, bY) = ab\mathrm{Cov}(X, Y)$
- $\mathrm{Cov}(X + a, Y + b) = \mathrm{Cov}(X, Y)$

协方差的性质

对任意常数X_1, X_2和Y, 有

$$\mathrm{Cov}(X_1 + X_2, Y) = \mathrm{Cov}(X_1, Y) + \mathrm{Cov}(X_2, Y)$$

对随机变量$X_1, X_2, \dots, X_n$和$Y_1, Y_2, \dots, Y_m$, 有

$$\mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{Cov}(X_i, Y_j)$$

以及进一步有

$$\mathrm{Var}\left(\sum_{i}^{n} X_i\right) = \sum_{i}^{n} \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j)$$

协方差的性质

若随机变量X与Y独立, 则$\mathrm{Cov}(X,Y)=0$, 反之不成立.

定理： 对任意随机变量X与Y有

$$\left(\mathbf{Cov}(\boldsymbol{X},\boldsymbol{Y})\right)^{2} \leq \mathbf{Var}(\boldsymbol{X})\mathbf{Var}(\boldsymbol{Y})$$

等号成立的充要条件是$\boldsymbol{Y}=\boldsymbol{a}\boldsymbol{X}+\boldsymbol{b}$几乎处处成立, 即$X$与$Y$之间几乎处处存在线性关系

正太分布

若随机向量$(X,Y) \sim \mathrm{N}(\mu_x,\mu_y,\sigma_x^2,\sigma_y^2,\rho)$, 则$\mathrm{Cov}(X,Y) = \rho\sigma_x\sigma_y$

推论： 若随机向量$(X,Y) \sim \mathrm{N}(\mu_x,\mu_y,\sigma_x^2,\sigma_y^2,\rho)$, 则

X和Y相互独立的**充要条件**是协方差$\mathrm{Cov}(X,Y) = 0$

例题

设随机变量$X_1, X_2, \cdots, X_n$相互独立且服从正态分布，方差为σ^2.
记$\bar{X} = \sum_{i=1}^{n} X_i / n$, 讨论$\bar{X}$和$\bar{X} - X_i$的独立性

例题

随机变量(X,Y)联合概率密度为

$$f(x,y)=\begin{cases}(x+y)/8 & 0\le x\le 2, 0\le y\le 2\\ 0 & \text{其它}\end{cases}$$

求$\mathrm{Cov}(X,Y)$和$\mathrm{Var}(X+Y)$.

例题-匹配问题

有n对夫妻参加一次聚会, 将所有参会人员任意分成n组, 每组一男一女, 用X表示夫妻两人被分到一组的对数, 求X的期望和方差

相关系数

两个随机变量之间的关系：独立与非独立

非独立关系：线性关系和非线性关系. 非线性关系较为复杂, 无好办法来处理. 线性相关程度可以通过线性相关系数来刻画

定义： 随机变量X和Y的方差$\mathrm{Var}(X)$,$\mathrm{Var}(Y)$存在且不为0, 称

$$\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

为***X*与*Y*的相关系数**

- 若$\rho_{XY}>0$, 称 ***X*与*Y*正相关**
- 若$\rho_{XY}<0$, 称 ***X*与*Y*负相关**
- 若$\rho_{XY}=0$, 称 ***X*与*Y*不相关**

相关系数性质

- 根据协方差性质可知$|\rho_{XY}| \le 1$
- $|\rho_{XY}| = 1$的充要条件为X与Y几乎处处有线性关系$Y = aX + b$
- 若X与Y相互独立, 则X与Y不相关 ($\rho_{XY} = 0$), 但反之不成立

ρ_{XY}刻画了X与Y的线性相关性, 又称线性相关系数。随机变量X与Y不相关, 仅表示X与Y之间不存在线性关系, 可能存在其他关系

例如，设随机变量$X \sim U(-1/2,1/2)$和$Y = \cos(X)$, 有

$$\mathrm{Cov}\,(X,Y) = E[Xcos(X)] - E(X)E(\cos(X))$$

$$= E[X\cos(X)] = \int_{-1/2}^{1/2} x\cos(x)\,dx = 0$$

相关系数

对方差不为零的随机变量X和Y, 下述条件相互等价:

- $\rho_{XY} = 0$
- $\mathrm{Cov}(X, Y) = 0$
- $E(XY) = E(X)E(Y)$
- $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

若随机向量 $(X, Y) \sim \mathrm{N}(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$, 则$X$与$Y$的相关系数 $\rho_{XY} = \rho$, 因此X与Y独立的充要条件是X与Y不相关

例题

设随机变量$X \sim N(\mu, \sigma^2)$和$Y \sim N(\mu, \sigma^2)$相互独立，求$Z_1 = \alpha X + \beta Y$和$Z_2 = \alpha X - \beta Y$的相关系数$(\alpha, \beta \neq 0)$

例题

设随机向量$(X_1, X_2, \cdots, X_n)$服从多项分布$M(m, p_1, p_2, \cdots, p_n)$，对任意$i \neq j$, 求$X_i$和$X_j$的相关系数