Bài số 6

COVARIANCE VÀ CÁC HỆ SỐ TƯƠNG QUAN

I. COVARIANCE(HIỆP PHƯƠNG SAI)

Ta đã biết rằng: Nếu X và Y là các BNN với phân phối xác suất đồng thời là $f(x,y)$ thì kỳ vọng của biến ngẫu nhiên $g(X,Y)$ là

$$\mu_{g(X,Y)} = E[g(X,Y)] = \sum_{x}\sum_{y} g(x,y)f(x,y)$$, nếu X và Y là các BNN rời rạc,

và

$$\mu_{g(X,Y)} = E[g(X,Y)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)f(x,y)dxdy$$, nếu X và Y là các BNN liên tục.

Bây giờ ta xét trường hợp đặc biệt: $g(X,Y) = \left(X-\mu_X\right)\left(Y-\mu_Y\right)$ trong đó $\mu_X = E(X)$ và $\mu_Y = E(Y)$. Khi đó giá trị kỳ vọng của BNN $g(X,Y)$ sẽ được gọi là ***covariance*** của X và Y và được ký hiệu là σ_{XY} hoặc là $\operatorname{cov}\left(X,Y\right)$.

1. Định nghĩa. Cho X và Y là các BNN với phân phối xác suất đồng thời $f(x, y)$. Khi đó ***Covariance*** của X và Y là một đại lượng mà giá trị của nó được xác định bởi:

$$\sigma_{XY} = E[(X - \mu_X)(Y-\mu_Y)] = \sum_{x}\sum_{y}(x-\mu_X)(y-\mu_Y)f(x,y)$$, nếu X và Y là các **BNN rời rạc**,

và

$$\sigma_{XY} = E[(X - \mu_X)(Y-\mu_Y)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-\mu_X)(y-\mu_Y)f(x,y)dxdy$$, nếu X và Y là các **BNN l/ tục**.

Nhận xét.

+ Covariance của hai BNN chính là kỳ vọng của tích các độ lệch của các biến ngẫu nhiên với chính kỳ vọng của chúng.

+ Covariance của hai BNN cho ta biết mối liên hệ khách quan của hai biến ngẫu nhiên.

2. Một số tính chất.

i. Covariance của các biến ngẫu nhiên X và Y với các kỳ vọng ứng là μ_X, μ_Y có thể được xác định bởi công thức:

$$\sigma_{XY} = E(XY) - E(X)E(Y) = E(X,Y) - \mu_X\mu_Y.$$

ii. $\sigma_{XY} = \sigma_{YX}$

iii. $\sigma_{XX} = E\left[\left(X-\mu_X\right)^2\right] = \sigma_X^2$.

iv. Nếu X và Y là hai ***BNN độc lập*** (độc lập thống kê) thì $\sigma_{XY} = 0$. Tuy nhiên, ***điều ngược lại chưa chắc đã đúng.***

Ví dụ 1. Chọn ngẫu nhiên hai ruột bút từ một chiếc hộp. Số ngòi bút xanh X và số ngòi bút đỏ Y là các biến ngẫu nhiên có phân phối xác suất đồng thời được cho ở bảng sau

Y \ X	0	1	2	$g(x)$
0	$3/28$	$3/14$	$1/28$	$5/14$
1	$9/28$	$3/14$		$15/28$
2	$3/28$			$3/28$
$h(y)$	$15/28$	$3/7$	$1/28$	1

Hãy tìm covariance của *X* và *Y*.

Giải: + Ta có:

$$E(XY) = \sum_{x=0}^{2}\sum_{y=0}^{2} xy f(x,y) = (0)(0)f(0,0) + (0)(1)f(0,1) + (0)(2)f(0,2)$$

$$+ (1)(0)f(1,0) + (1)(1)f(1,1) + (2)(0)f(2,0) = f(1,1) = \frac{3}{14}$$

+ Mặt khác: $\mu_X = E(X) = \sum_{x=0}^{2}\sum_{y=0}^{2} xf(x,y) = \sum_{x=0}^{2} xg(x) = (0)\left(\frac{5}{14}\right) + (1)\left(\frac{15}{28}\right) + (2)\left(\frac{3}{28}\right) = \frac{3}{4}$

và $\mu_Y = E(Y) = \sum_{x=0}^{2}\sum_{y=0}^{2} yf(x,y) = \sum_{y=0}^{2} yh(y) = (0)\left(\frac{15}{28}\right) + (1)\left(\frac{3}{7}\right) + (2)\left(\frac{1}{28}\right) = \frac{1}{2}$

+ Do đó: $\sigma_{XY} = E(XY) - \mu_X\mu_Y = \frac{3}{14} - \left(\frac{3}{4}\right)\left(\frac{1}{2}\right) = -\frac{9}{56}$.

Ví dụ 2. Tỉ lệ X các nam vận động viên và tỉ lệ Y các nữ vận động viên điền kinh hoàn thành bài thi trong cuộ thi marathon được mô tả bằng hàm mật độ đồng thời sau

$$f(x,y) = \begin{cases} 8xy, & (x,y) \in [0;1]\times[0;x] \\ 0, & (x,y) \notin [0;1]\times[0;x] \end{cases}$$

Hãy tìm covariance của *X* và *Y*.

Giải: + Trước tiên, ta phải tìm các hàm mật độ biên duyên:

$$g(x) = \begin{cases} 4x^3, & x \in [0;1] \\ 0, & x \notin [0;1] \end{cases} \quad \text{và} \quad h(y) = \begin{cases} 4y(1-y^2), & y \in [0;1] \\ 0, & y \notin [0;1] \end{cases}$$

+ Khi đó ta có: $\mu_X = E(X) = \int_0^1 4x^4 dx = \frac{4}{5}$, $\mu_Y = E(Y) = \int_0^1 4y^2(1-y^2)dy = \frac{8}{15}$.

+ Mặt khác: $E(XY)=\int_0^1\int_y^1 8x^2y^2dxdy=\dfrac{4}{9}.$

+ Vậy nên covariance cần tìm là: $\sigma_{XY}=E(XY)-\mu_X\mu_Y=\dfrac{4}{9}-\left(\dfrac{4}{5}\right)\left(\dfrac{8}{15}\right)=\dfrac{4}{225}.$

II.HỆ SỐ TƯƠNG QUAN.

Mặc dù covariance của hai biến ngẫu nhiên cung cấp thông tin về ***mối liên hệ khách quan*** giữa hai biến ngẫu nhiên, *nhưng độ lớn của* σ_{XY} ***không cho ta biết về mức độ quan hệ*** của hai biến ngẫu nhiên, bởi vì σ_{XY} còn phụ thuộc vào đơn vị đo. Độ lớn của nó tuỳ thuộc vào đơn vị đo của cả X và Y. Có một phiên bản của covariance mà không phụ thuộc vào đơn vị đo và được sử dụng rộng rãi trong thống kê đó là **hệ số tương quan.**

1. Định nghĩa.

Cho X và Y là các BNN với covariance σ_{XY} và các độ lệch chuẩn tương ứng là σ_X và σ_Y. **Hệ số tương quan** của X và Y là một số thực được xác định bởi:

$$\rho_{XY}=\frac{\sigma_{XY}}{\sigma_X\sigma_Y}.$$

Nhận xét: Dễ thấy rằng ρ_{XY} không phụ thuộc vào đơn vị đo của các biến ngẫu nhiên X và Y.

2. Một số tính chất.

i. Hệ số tương quan thoả mãn bất đẳng thức $-1\le\rho_{XY}\le 1$.

ii. Khi $\sigma_{XY}=0$ thì suy ra ngay rằng $\rho_{XY}=0$.

iii. Nếu X và Y là các **BNN độc lập** thì ta có $\rho_{XY}=0$, tuy nhiên ***điều ngược lại chưa chắc đã đúng***.

iv. Có sự phụ thuộc hàm tuyến tính, tức là $Y=aX+b,\quad a\neq 0$ khi và chỉ khi

$$\rho_{XY}=\begin{cases}1, & \text{nếu } a>0\\ -1, & \text{nếu } a<0\end{cases}$$

v. Ta có: $$\rho_{(aX+c)(bY+d)}=\begin{cases}\rho_{XY}, & \text{nếu } ab>0\\ -1\rho_{XY}, & \text{nếu } ab<0\end{cases}$$

3. Ý nghĩa của hệ số tương quan. Hệ số tương quan đo mức độ phụ thuộc tuyến tính của hai BNN X và Y:

+ Khi $|\rho_{XY}|$ càng gần 1 thì tính chất quan hệ tuyến tính càng chặt.

+ Khi $|\rho_{XY}|$ càng gần 0 thì sự phụ thuộc tuyến tính càng ít, càng lỏng lẻo.

+ Khi $\rho_{XY} = 0$ ta nói X và Y là không tương quan.

III. MỘT SỐ TÍNH CHẤT CỦA PHƯƠNG SAI.

Cho X là BNN, và các hằng số $a, b, a_1, a_2, \ldots$ Khi đó ta có

i. $\sigma^2_{\mathrm{aX+b}} = a^2\sigma^2_X = a^2\sigma^2.$

ii. Khi $a = 1$, ta được: $\sigma^2_{X+b} = \sigma^2_X = \sigma^2.$

iii. Khi $b = 0$, ta được

$$\sigma^2_{aX} = a^2\sigma^2_X = a^2\sigma^2$$

iv. Nếu X và Y là các biến ngẫu nhiên với phân phối xác suất là $f(x,y)$ thì:

$$\sigma^2_{aX+bY} = a^2\sigma^2_X + b^2\sigma^2_Y + 2ab\sigma_{XY}$$

v. Nếu X và Y là ***hai biến ngẫu nhiên độc lập***, thì ta có:

$$\sigma^2_{aX+bY} = a^2\sigma^2_X + b^2\sigma^2_Y.$$

$$\sigma^2_{aX-bY} = a^2\sigma^2_X + b^2\sigma^2_Y.$$

vi. Nếu $X_1, X_2, \ldots, X_n$ là **các biến ngẫu nhiên độc lập**, thì

$$\sigma^2_{a_1X_1+a_2X_2+\cdots a_nX_n} = a_1^2\sigma^2_{X_1} + a_2^2\sigma^2_{X_2} + \cdots + a_n^2\sigma^2_{X_n}.$$

Ví dụ 3. Cho X và Y là hai biến ngẫu nhiên độc lập, có các phương sai tương ứng là $\sigma^2_X = 2,\ \sigma^2_Y = 4$ và covariance $\sigma_{XY} = -2$. Hãy tìm phương sai của biến ngẫu nhiên $Z = 3X - 4Y + 8$.

Giải:

+ Theo Định lý trên, ta được

$$\sigma^2_Z = \sigma^2_{3X-4Y+8} = \sigma^2_{3X-4Y} = 9\sigma^2_X + 16\sigma^2_Y - 24\sigma_{XY}$$
$$= (9)(2) + (16)(4) - (24)(-2) = 130.$$

Ví dụ 4. Gọi X và Y là lượng hai loại tạp chất trong một lô của một loại sản phẩm hoá học nào đó. Giả sử rằng X và Y là các biến ngẫu nhiên độc lập với các phương sai lần lượt là

$$\sigma^2_X = 2,\ \sigma^2_Y = 3$$

Hãy tìm phương sai của biến ngẫu nhiên:

$$Z = 3X - 2Y + 5$$

Giải: Ta được: $\sigma^2_Z = \sigma^2_{3X-2Y+5} = \sigma^2_{3X-2Y} = 9\sigma^2_X + 4\sigma^2_Y = (9)(2) + (4)(3) = 30.$

NỘI DUNG ÔN TẬP CHUẨN BỊ KIỂM TRA GIỮA KỲ MÔN TOÁN 5

(Theo lịch trình sẽ Kiểm tra vào ngày 9/9/2011)

I. Xác suất của một biến cố và phân phối xác suất của biến ngẫu nhiên.

+ Tính xác suất của một biến cố

+ Tính xác suất theo quy tắc cộng, quy tắc nhân, quy tắc Bayes, xác suất có điều kiện, định lý xác suất đầy đủ

+ Tìm phân phối xác suất của biến ngẫu nhiên rời rạc một chiều hoặc hai chiều.

II. Biến ngẫu nhiên

+ Biến ngẫu nhiên một chiều: hàm phân phối tích lũy, tính xác suất

+ Biến ngẫu nhiên hai chiều: phân phối đồng thời, phân phối biên duyên, tính xác suất

III. Các số đặc trưng biến ngẫu nhiên

+ Kỳ vọng và tính chất.

+ Phương sai, độ lệch chuẩn và tính chất.

Về nhà:

Tự đọc: Mục

Bài tập: Tr. **128**

Đọc trước các Mục trong **Chương 5 và 6** chuẩn bị cho **Bài số 7 :**

Một số phân phối xác suất thường gặp