Föreläsning 7

Kovarians och korrelation

Kom ihåg

$$E[g(X)] = \begin{cases} \sum\limits_{\text{alla } k} g(k)\, P(X=k) & \text{om } X \text{ är diskret} \\ \int_{-\infty}^{\infty} g(x) f_X(x)\, dx & \text{om } X \text{ är kontinuerlig} \end{cases} .$$

Sats Låt X, Y vara två stokastiska variabler.
(a) Vidare låt $g : \mathbb{R}^2 \mapsto \mathbb{R}$. Det gäller att

$$E[g(X,Y)] = \begin{cases} \sum\limits_{\text{alla } (k,l)} \sum g(k,l)\, p(k,l) & \text{om } X, Y \text{ är diskreta} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y)\, dx\, dy & \text{om } X, Y \text{är kontinuerliga} \end{cases}$$

där $p(k,l)$ är den simultana sannolikhetsfunktionen (diskret fall) och $f(x,y)$ är den simultana täthetsfunktionen (kontinuerligt fall).

(b) Låt X, Y vara oberoende och $g : \mathbb{R} \mapsto \mathbb{R}$, $h : \mathbb{R} \mapsto \mathbb{R}$. Det gäller att

$$E[g(X)h(Y)] = E[g(X)] \cdot E[h(Y)].$$

I synnerhet,

$$E[XY] = E[X] \cdot E[Y].$$

Definition

$$Cov(X,Y) \equiv C(X,Y) = E[(X - E[X])(Y - E[Y])]$$

kallas *kovarians* av X och Y.

Anmärkningar (1) Ofta fördelaktig är förkortningsformeln

$$Cov(X,Y) = E[XY] - E[X]E[Y]\,.$$

(2) I synnerhet, om X och Y är oberoende så gäller

$$Cov(X,Y) = 0\,.$$

(3) Motriktningen kan vara fel.

Exempel Låt

$$P(X=-1) = P(X=0) = P(X=1) = \frac{1}{3}$$

och

$$Y = \begin{cases} 0 & \text{om} \quad X \neq 0 \\ 1 & \text{om} \quad X = 0 \end{cases}$$

dvs. X och Y är ej oberoende. Men

$$X \cdot Y = 0 \Longrightarrow E[XY] = 0$$

och

$$E[X] = (-1) \cdot P(X = -1) + 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = 0$$

dvs.

$$Cov(X, Y) = E[XY] - E[X] \cdot E[Y] = 0\,.$$

Sats (Räkneregler för Cov)

(a) $Cov(X, X) = Var(X), \quad a \in \mathbb{R}$

(b) $Cov(X, Y) = Cov(Y, X)$

(c) $Cov(aX, Y) = Cov(X, aY) = a\,Cov(X, Y)$

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

Följdsats

(a) $Cov(a, X) = Cov(X, a) = 0, \quad a \in \mathbb{R}$

(b) $Cov\left(\sum_{i=1}^{m} a_i X_i + b, \sum_{j=1}^{n} c_j Y_j + d\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i c_j Cov(X_i, Y_j), \quad a_i, b, c_j, d \in \mathbb{R}$

(c) $Var\left(\sum_{i=1}^{m} X_i\right) = \sum_{i=1}^{m} Var(X_i) + \sum_{j \neq i}^{m} Cov(X_i, X_j)$

Exempel Låt $X \sim U(0, 1)$. Bestäm $Var(X + X^2)$.

Svar:

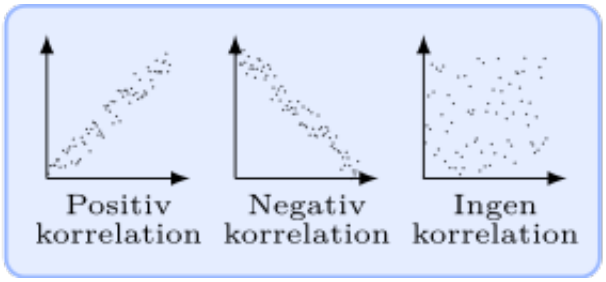
$$\begin{aligned} Var(X + X^2) &= Var(X) + Var(X^2) + 2Cov(X, X^2) \\ &= E[X^2] - (E[X])^2 + E[X^4] - \left(E[X]^2\right)^2 + 2E[X \cdot X^2] - 2E[X] \cdot E[X^2] \\ &= E[X^2] + 2E[X^3] + E[X^4] - (E[X])^2 - \left(E[X]^2\right)^2 - 2E[X] \cdot E[X^2] \\ &= \int_0^1 x^2\,dx + 2\int_0^1 x^3\,dx + \int_0^1 x^4\,dx - \left(\int_0^1 x\,dx\right)^2 - \left(\int_0^1 x^2\,dx\right)^2 \\ &\quad -2\int_0^1 x\,dx \int_0^1 x^2\,dx \\ &= \cdots = \frac{61}{180} = 0.34\,. \end{aligned}$$

Definition

$$\rho(X,Y) = \frac{CovX,Y}{\sqrt{Var(X)\cdot Var(Y)}}$$

kallas *korrelation* mellan X och Y.

Anmärkning Korrelationen används som mått på linjärt oberoende.



Google Bild

Punkterna i bilderna motsvarar de olika utfallen av slumpvektorn (X,Y).

Sats (a) Det gäller att

$$|\rho(X,Y)| \leq 1\,.$$

(b) I synnerhet gäller $|\rho(X,Y)| = 1$ om och endast om det finns $a,b,c \in \mathbb{R}$ sådana att

$$aX + bY = c\,.$$

Anmärkning Del (a) är ekvivalent med

$$(Cov(X,Y))^2 \leq Var(X)\cdot Var(Y)\,,$$

den stokastiska versionen av Schwarz' olikhet.

Betingade Fördelningar

Definition Låt X,Y vara två diskreta stokastiska variabler med den simultana sannolikhetsfunktionen p och de marginella sannolikhetsfunktionerna p_X och p_Y. Låt $p_Y(k) > 0$. Då kallas

$$p_{X|Y=k}(j) = \frac{p(j,k)}{p_Y(k)} \quad \text{för alla möjliga värden } j \text{ på } X$$

betingad sannolikhetsfunktion för X *givet* $Y = k$.

Anmärkning Om X och Y är oberoende så gäller

$$p_{X|Y=k}(j) = \frac{p(j,k)}{p_Y(k)} = \frac{p_X(j)\cdot p_Y(k)}{p_Y(k)} = p_X(j)\,.$$

Exempel Låt $X \sim Po(\lambda_1)$ och $Y \sim Po(\lambda_2)$ vara oberoende stokastiska variabler och $n \in \mathbb{N}$. Visa att

$$P(X = k | X + Y = n), \quad n = 0, 1, \ldots$$

är sannolikhetsfunktion av $Bin(n, \frac{\lambda_1}{\lambda_1+\lambda_2})$.

Svar: Vi har

$$\begin{aligned} P(X = k | X + Y = n) \equiv p_{X|X+Y=n}(k) &= \frac{P(X = k, X + Y = n)}{P(X + Y = n)} \\ &= \frac{P(X = k, Y = n - k)}{P(X + Y = n)} = \frac{P(X = k) \cdot P(Y = n - k)}{P(X + Y = n)} \\ &= \frac{e^{-\lambda_1} \frac{\lambda_1^k}{k!} \cdot e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!}}{e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1+\lambda_2)^n}{n!}} = \binom{n}{k} \frac{\lambda_1^k \cdot \lambda_2^{n-k}}{(\lambda_1+\lambda_2)^n} \\ &= \binom{n}{k} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k \cdot \left(1 - \frac{\lambda_1}{\lambda_1+\lambda_2}\right)^{n-k}, \quad k = 0, \ldots, n, \end{aligned}$$

dĕ vi har använt att $X + Y \sim Po(\lambda_1 + \lambda_2)$, se föreläsning 8. Den högra sidan är sannolikhetsfunktionen för $Bin(n, \frac{\lambda_1}{\lambda_1+\lambda_2})$.

Definition Låt X, Y vara två kontinuerliga stokastiska variabler med den simultana täthetsfunktionen f och de marginella täthetsfunktionerna f_X och f_Y. Låt $f_Y(y) > 0$. Då kallas

$$f_{X|Y=y}(x) = \frac{f(x, y)}{f_Y(y)}, \quad x \in \mathbb{R},$$

betingad täthetsfunktion för X givet $Y = y$.

Anmärkning Om X och Y är oberoende så gäller

$$f_{X|Y=y}(x) = \frac{f(x, y)}{f_Y(y)} = \frac{f_X(x) \cdot f_Y(y)}{f_Y(y)} = f_X(x).$$

Exempel Låt

$$f(x, y) = \begin{cases} \dfrac{e^{-x/y} e^{-y}}{y} & \text{om} \quad 0 < x < \infty, \quad 0 < y < \infty \\ 0 & \text{annars} \end{cases}$$

vara den simultana täthetsfunktionen av X och Y. Bestäm

$$P(X > 1 | Y = y) \quad \text{där } y > 0.$$

Svar: Vi måste först bestämma $f_{X|Y=y}$ eftersom

$$P(X > 1 | Y = y) = \int_1^\infty f_{X|Y=y}(x)\, dx.$$

För att hitta

$$f_{X|Y=y} = \frac{f(x,y)}{f_Y(y)}$$

behöver vi $f_Y(y)$. Vi får

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)\,dy = e^{-y} \int_{x=0}^{\infty} \frac{1}{y}\, e^{-x/y}\,dx \\ &= e^{-y}, \quad y > 0, \end{aligned}$$

där vi har utnyttjat att $\frac{1}{y}\, e^{-x/y}$, $x > 0$, är täthetsfunktionen av $Exp(\frac{1}{y})$. Följaktligen

$$f_{X|Y=y}(x) = \frac{f(x,y)}{f_Y(y)} = \frac{e^{-x/y}\, e^{-y}/y}{e^{-y}} = \frac{1}{y} \cdot e^{-x/y}, \quad x > 0.$$

Alltså

$$P(X > 1|Y = y) = \int_1^{\infty} f_{X|Y=y}(x)\,dx = \int_1^{\infty} \frac{1}{y} \cdot e^{-x/y}\,dx = e^{-\frac{1}{y}}$$

där vi igen har utnyttjat att $\frac{1}{y}\, e^{-x/y}$, $x > 0$, är täthetsfunktionen av $Exp(\frac{1}{y})$.

Betingat väntevärde

Definition Låt X, Y vara två stokastiska variabler.

$$E[X|Y=y] = \begin{cases} \sum\limits_{\text{alla } x} p_{X|Y=y}(x) & \text{om } X, Y \text{ är diskreta} \\ \int_{-\infty}^{\infty} p_{X|Y=y}(x)\,dx & \text{om } X, Y \text{är kontinuerliga} \end{cases}$$

kallas *betingat väntevärde av X givet $Y = y$.*

Sats Låt X, Y vara två stokastiska variabler. Då är väntevärdet av den stokastiska variabeln $E[X|Y = y]$ (en stokastisk variabel som kan anta alla möjliga värden y på Y) lika med $E[X]$, dvs.

$$E[X] = E\left[E[X|Y]\right] = \begin{cases} \sum\limits_{\text{alla } y} E[X|Y=y]\, p_Y(y) & \text{om } Y \text{ är diskret} \\ \int_{-\infty}^{\infty} E[X|Y=y]\, f_Y(y)\,dy & \text{om } Y \text{är kontinuerlig} \end{cases}.$$

Exempel En björn är instängd i en grotta. Det finns två tunnlar. Tunnel 1 startar i grottan och leder björnen till skogen efter 3 timmar. Tunnel 2 startar i grottan och är en loop där björnen tillbringar 4 timmar för att sedan återvända till grottan. Varje gång björnen är i grottan väljer den tunnel 1 med sannolikhet 1/3 och tunnel 2 med sannolikhet 2/3. Beräkna den förväntade tiden för björnen att nå skogen.

Svar: Låt

$$X : \text{restid för björnen}$$

och

$$Y : \text{val av tunnel.}$$

Vi har

$$P(Y = 1) = \frac{1}{3}, \quad P(Y = 2) = \frac{2}{3}\,,$$

och

$$E[X|Y = 1] = 3, \quad E[X|Y = 2] = E[X] + 4\,.$$

Således

$$E[X] = E\left[E[X|Y]\right] = E[X|Y = 1]\,P(Y = 1) + E[X|Y = 2]\,P(Y = 2)$$

dvs.

$$E[X] = 3 \cdot \frac{1}{3} + (E[X] + 4) \cdot \frac{2}{3} \quad \text{som ger} \quad E[X] = 11.$$

Jörg-Uwe Löbus, jorg-uwe.lobus@liu.se