

TMS136

Föreläsning 6

Kovarians och korrelation

- För två s.v. X och Y är man ofta intresserad av hur ”de samverkar”
- Ett sätt att mäta deras **linjära** beroende är kallas *kovarians*
- För att definiera kovarians behöver vi kunna beräkna väntevärdet av en funktion av de två s.v.

Väntevärdet av en funktion

- För X och Y med gemensam mass-/täthetsfunktion $f_{XY}(x, y)$ och $h(x, y)$ någon reellvärd funktion har vi

$$E(h(X, Y)) = \begin{cases} \sum_{\text{alla } x} \sum_{\text{alla } y} h(x, y) f_{XY}(x, y) \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) f_{XY}(x, y) dx dy \end{cases}$$

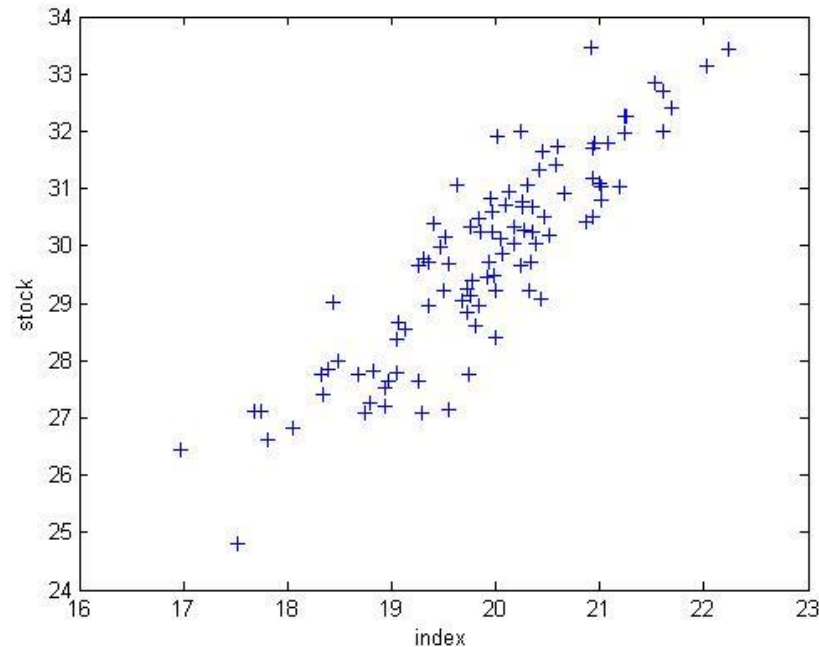
Kovarians

- *Kovariansen* mellan X och Y är ett mått på deras linjära beroende och definieras som (verifiera sista likheten)

$$\begin{aligned}\sigma_{XY} &= cov(X, Y) = E((X - \mu_X)(Y - \mu_Y)) \\ &= E(XY) - \mu_X\mu_Y\end{aligned}$$

Exempel

- I bilden nedan ser vi parvisa observationer av att index och en aktie

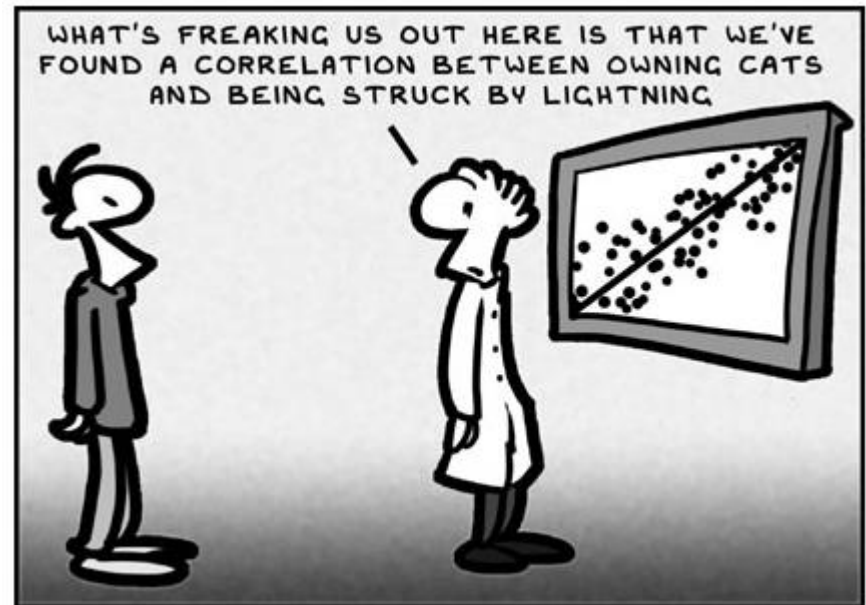


- Det verkar finnas ett linjärt beroende...

Korrelation

- En kovarians kan bli 3000 utan att det linjära sambandet är särskilt starkt
- För att få ett mått som bättre förklarar hur starkt det linjära sambandet är använder man oftast *korrelation*

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$



Egenskaper

- Man kan visa (övning) att $-1 \leq \rho_{XY} \leq 1$
- Om $\rho_{XY} = -1$ gäller att $Y = aX + b$, där $a < 0$
- Om $\rho_{XY} = 1$ gäller att $Y = aX + b$, där $a > 0$
- Så korrelationer nära -1 eller 1 betyder starkt linjärt samband mellan X och Y

VIKTIGT!

- Om X och Y är oberoende har vi att $\rho_{XY} = 0$ eftersom $E(XY) = E(X)E(Y)$ (varför?)
- Att $\rho_{XY} = 0$ betyder i allmänhet inte att X och Y är oberoende
- Detta kan finnas beroenden mellan X och Y som inte är linjära...

Exempel

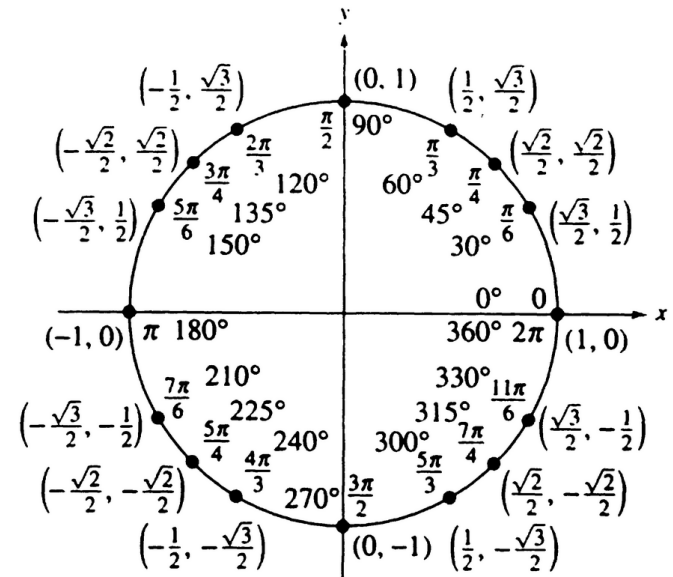
- Låt W vara likformigt fördelad på $(0, 2\pi)$ och låt $X = \cos(W)$ och $Y = \sin(W)$

- Paret (X, Y) är alltså en punkt på enhetscirkeln

- Observera att $E(X) = E(Y) = 0$ av symmetriskäl och att

$$E(XY) = \frac{1}{2} E(\sin(2W)) = 0 \text{ så } \rho_{XY} = 0$$

- Men $X = \pm\sqrt{1 - Y^2}$ så om $Y = 1$ så är $X = 0$ vilket betyder att X och Y inte är oberoende



Räkneregler och observationer

- Per definition är $cov(X, X) = V(X)$
- Vi har att

$$V(X \pm Y) = cov(X \pm Y, X \pm Y)$$

$$= cov(X, X) \pm cov(X, Y) \pm cov(Y, X) + cov(Y, Y)$$

$$= V(X) + V(Y) \pm 2cov(X, Y)$$

Räkneregler och observationer

- Om X och Y är oberoende har vi

$$V(X \pm Y) = V(X) + V(Y)$$

- Observera att det **ALDRIG NÅGONSIN** blir ett minus mellan varianserna i högerledet
- Varianser är per definition icke-negativa...

Linjärkombinationer av s.v.

- Låt $X_1, \dots, X_p$ vara s.v. och $c_1, \dots, c_p$ vara konstanter. Vi kallar

$$Y = c_1X_1 + \dots + c_pX_p$$

en *linjärkombination* av $X_1, \dots, X_p$

Väntevärde/Varians

- Ur definitionen av väntevärde följer (verifiera) att

$$E(Y) = c_1 E(X_1) + \cdots + c_p E(X_p)$$

- Ur definitionen av varians/kovarians följer (verifiera) att

$$\begin{aligned} V(Y) = & c_1^2 V(X_1) + \cdots + c_p^2 V(X_p) \\ & + 2 \sum_{i=1}^{p-1} \sum_{j=i+1}^p c_i c_j \operatorname{cov}(X_i, X_j) \end{aligned}$$

Varians forts

- Om $X_1, \dots, X_p$ är oberoende gäller att

$$V(Y) = c_1^2 V(X_1) + \dots + c_p^2 V(X_p)$$

- Minns att oberoende implicerar okorrelerad (men inte tvärtom) så kovarianstermerna försvinner om de stokastiska variablerna är oberoende

Av speciellt intresse är

- Om $X_1, \dots, X_p$ är oberoende och likafördelade (stickprov) är man ofta intresserad av stickprovsmedelvärdet

$$\bar{X} = \frac{1}{p} \sum_{i=1}^p X_i$$

Väntevärde/varians

- Vi har att

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

och

$$V(\bar{X}) = V\left(\frac{1}{p} \sum_{i=1}^p X_i\right) = \frac{1}{p^2} \sum_{i=1}^p V(X_i) = \frac{V(X_i)}{p}$$

Normalfördelning

- Om $X_1, \dots, X_p$ är oberoende och normalfördelade med väntevärden $\mu_1, \dots, \mu_p$ och standardavvikelser $\sigma_1, \dots, \sigma_p$ gäller att

$$Y = c_1 X_1 + \dots + X_p c_p$$

är normalfördelad med väntevärde

$c_1 \mu_1 + \dots + c_p \mu_p$ och varians $c_1^2 \sigma_1^2 + \dots + c_p^2 \sigma_p^2$

Exempel

- Låt X och Y vara oberoende där X är normalfördelad med väntevärde 2 och varians 1 och Y är normalfördelad med väntevärde 3 och varians 4
- Då har vi för $W = 3X + 5Y$ att W är normalfördelad med väntevärde

$$E(W) = 3E(X) + 5E(Y) = 21$$

och

$$Var(W) = 3^2 Var(X) + 5^2 Var(Y) = 109$$

Exempel

- Burkfyllningsvolymen i en läskfabrik är normalfördelad med väntevärde 33 cl och standardavvikelse 0.2 cl



- Vad är (under antagande om att volymerna är oberoende) sannolikheten att medelvolymer av tio burkar överstiger 33.1 cl?

Exempel forts

- Enligt ovan är medelvolymen normalfördelad med väntevärde 33cl och standardavvikelse $0.2/\sqrt{10}$.
- Vi har då att

$$P(\bar{X} > 33.1) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{33.1 - \mu}{\sigma/\sqrt{n}}\right)$$

$$= P\left(Z > \frac{33.1 - 33}{0.2/\sqrt{10}}\right) \approx P(Z > 1.58)$$

$$= 1 - \Phi(1.58) \approx 0.06$$