SF1901: Sannolikhetslära och statistik
Föreläsning 5.
Flera stokastiska variabler.

Jan Grandell & Timo Koski

31.01.2012

Ofta mäter vi i samma slumpförsök flera storheter, och då beskrivs resultatet av en n-dimensionell stokastisk variabel $(X_1, X_2, \ldots, X_n)$.

Exempel

Slumpförsöket är att vi väljer en person slumpmässigt här i rummet, och sätter
X = *personens vikt;*
Y = *personens längd.*

Vi nöjer oss med att ge detaljer i det två-dimensionella fallet. Låt (X,Y) vara en två-dimensionell s.v.

$F_{X,Y}(x, y) = P(X \leq x,\ Y \leq y)$ kallas (den simultana) fördelningsfunktionen för (X, Y).

$F_X(x) = P(X \leq x) = P(X \leq x,\ Y \leq \infty) = F_{X,Y}(x, \infty)$ kallas den marginella fördelningsfunktionen för X.

$F_Y(y) = F_{X,Y}(\infty, y)$ kallas den marginella fördelningsfunktionen för Y.

KTH Matematik

Definition

X och Y är oberoende stokastiska variabler om

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

obs! Detta bör gälla för ALLA (x, y) .

Flerdimensionella stokastiska variabler

Definition

$(X_1, X_2, \ldots, X_n)$ *är oberoende stokastiska variabler om*

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

$$= F_{X_1}(x_1) \cdots F_{X_n}(x_n).$$

Omvänt gäller att om $X_1, X_2, \ldots, X_n$ är oberoende s.v. så fås den simultana fördelningen enl. definitionen ovan.

Simultan sannolikhetsfunktion

(X,Y) är en diskret två-dimensionell s.v., om

$$F_{X,Y}(x,y) = \sum_{0 \leq j \leq [x]} \sum_{0 \leq k \leq [y]} p_{X,Y}(j,k) dudv$$

där

$$\sum_{j=0}^{\infty} \sum_{k=0}^{\infty} p_{X,Y}(j,k) = 1, \quad p_{X,Y}(j,k) \geq 0.$$

Funktionen $p_{X,Y}(i,k)$ kallas den simultana sannolikhetsfunktionen för (X,Y).

KTH Matematik

Simultan täthet

(X,Y) är en kontinuerlig två-dimensionell s.v., om

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v) dudv$$

där

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dxdy = 1, \quad f_{X,Y}(x,y) \geq 0.$$

Funktionen $f_{X,Y}(x,y)$ kallas den simultana täthetsfunktionen för (X,Y).

KTH Matematik

Marginalfördelningar fört kontinuerliga s.v.er

Låt (X,Y) vara en kontinuerlig två-dimensionell s.v.. Den marginella fördelningsfunktionen för Y är

$$F_Y(y) = F_{X,Y}(\infty, y) = \int_{-\infty}^{y} \int_{-\infty}^{\infty} f_{X,Y}(x, v)\,dx dv$$

och

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,dx$$

är den marginella täthetsfunktionen för Y. Analogt är

$$f_X(x) = \frac{d}{dx} F_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,dy$$

den marginella täthetsfunktionen för X.

KTH Matematik

Oberoende stokastiska variabler

Definition

X och Y är oberoende stokastiska variabler om

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

KTH Matematik

Normalfördelade stokastiska variabler

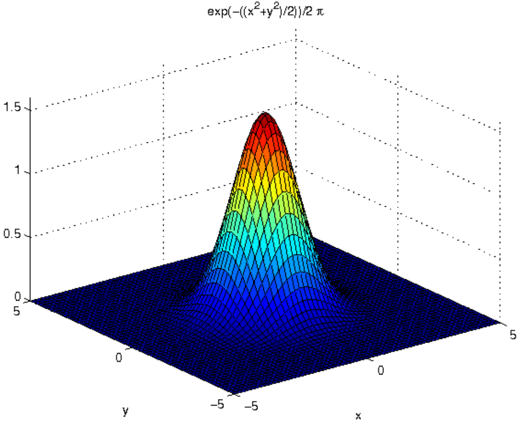
$X \in N(0,1),\ Y \in N(0,1)$

$$\varphi_X(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$$

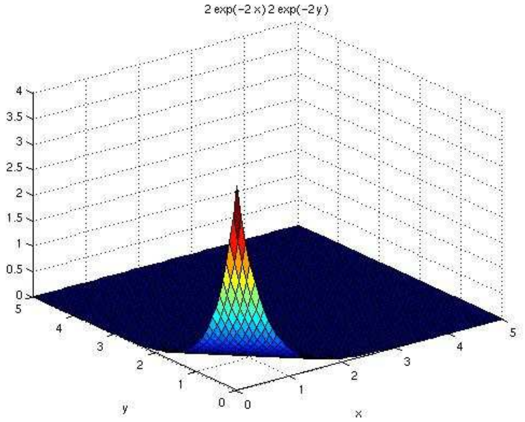
$$\varphi_Y(y) = \frac{1}{\sqrt{2\pi}}\, e^{-y^2/2}$$

$$\varphi_X(x) \cdot \varphi_Y(y) = \frac{1}{2\pi}\, e^{-(x^2+y^2)/2}$$

Produkten $\varphi_X(x) \cdot \varphi_Y(y)$



$X \in \text{Exp}(2), Y \in \text{Exp}(2)$, oberoende



Väntevärdet av $g(X, Y)$

Sats

Låt (X, Y) vara en tvådimensionell s.v. Då gäller

$$E\left[g(X,Y)\right] = \begin{cases} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dxdy & \text{för } (X,Y) \text{ kontinuerlig ,} \\ \sum_{k=0}^{\infty}\sum_{j=0}^{\infty} g(k,j) f_{X,Y}(k,j) & \text{för } (X,Y) \text{ diskret.} \end{cases}$$

$\square$

KTH Matematik

Väntevärdet av $g(X,Y) = X + Y$

Sats

Låt (X,Y) vara en tvådimensionell s.v. Då gäller

$$E[X+Y] = E[X] + E[Y]$$

Bevis. Låt (X,Y) vara en kontinuerlig tvådimensionell s.v.. Den föregående satsen visar med $g(x,y) = x + y$ att

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f_{X,Y}(x,y)dxdy =$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf_{X,Y}(x,y)dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf_{X,Y}(x,y)dxdy$$

$$= \int_{-\infty}^{\infty}x\int_{-\infty}^{\infty}f_{X,Y}(x,y)dydx + \int_{-\infty}^{\infty}y\int_{-\infty}^{\infty}f_{X,Y}(x,y)dxdy$$

$$= \int_{-\infty}^{\infty}xf_X(x)dx + \int_{-\infty}^{\infty}yf_Y(y)dy = E[X] + E[Y]$$

Kovarians och korrelationskoefficient

Låt (X, Y) vara en tvådimensionell s.v. där vi är intresserade av sambandet mellan Xs och Ys variation. Det kan vara natuligt att betrakta variablerna

$$X - \mu_X \quad \text{och} \quad Y - \mu_Y .$$

Vi skiljer på fallen då X och Y "samvarierar" resp. "motverkar varandra", dvs. då

ett stort/litet värde på X gör ett stort/litet värde på Y troligt

resp.

ett stort/litet värde på X gör ett litet/stort värde på Y troligt.

KTH Matematik

Kovariansen mellan X och Y

Betraktar vi nu variabeln

$$(X - \mu_X)(Y - \mu_Y),$$

så innebär detta att den i första fallet, eftersom $+ \cdot + = +$ och $- \cdot - = +$, att den har en tendens att vara positiv. På motsvarande sätt, eftersom $- \cdot + = -$ och $+ \cdot - = -$, har den i andra fallet en tendens att vara negativ. Det som vi, lite slarvigt, har kallat tendens, kan vi ersätta med väntevärde. Vi leds då till följande definition.

Definition

Kovariansen mellan X och Y är

$$C(X, Y) = E[(X - \mu_X)(Y - \mu_Y)],$$

där $\mu_X = E(X)$ och $\mu_Y = E(Y)$.

Kovariansen mellan X och Y

Sats

Kovariansen mellan X och Y är

$$C(X,Y) = E[XY] - \mu_X \cdot \mu_Y,$$

där $\mu_X = E(X)$ och $\mu_Y = E(Y)$.

Bevis. :

$$C(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E[XY - X\mu_Y - \mu_X Y + \mu_X\mu_Y]$$

$$= E[XY] - E[X\mu_Y] - E[\mu_X Y] - \mu_X\mu_Y]$$

$$= E[XY] - \mu_Y E[X] - \mu_X E[Y] - \mu_X\mu_Y = E[XY] - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y$$

$$= E[XY] - \mu_Y\mu_X$$

Kovariansen mellan oberoende X och Y

Sats

Om X och Y är oberoende, så

$$C(X,Y) = 0.$$

Bevis. : Vi tittar på $C(X,Y) = E[XY] - \mu_X \cdot \mu_Y$. Låt oss betrakta det kontinuerliga fallet. Med funktionen $g(x,y) = x \cdot y$ får vi

$$E[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \cdot y f_{X,Y}(x,y)dxdy$$

och antagandet om oberoende, $f_{X,Y}(x,y) = f_X(x)f_Y(y)$, ger

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \cdot y f_X(x) f_Y(y) dxdy = \int_{-\infty}^{\infty} x f_X(x)dx \int_{-\infty}^{\infty} y f_Y(y)dy = \mu_X \cdot \mu_Y.$$

Kovariansen kan sägas ha fel sort. Det verkar rimligt att ett mått på ett så abstrakt begrepp som samvariation skall vara "sortfritt". Det vanligaste måttet är korrelationskoefficienten.

Definition

Korrelationskoefficienten mellan X och Y är

$$\rho = \rho(X, Y) = \frac{C(X, Y)}{D(X)D(Y)}.$$

Man kan visa att $|\rho| \leq 1$, där $\rho = \pm 1$ betyder att det finns ett perfekt linjärt samband, dvs. $Y = aX + b$.

Sats

Om X och Y är oberoende så är de okorrelerade, dvs. $\rho(X, Y) = 0$.

Omvändningen gäller *ej*, dvs. okorrelerade variabler kan vara beroende.

Exempel

Den simultana sannolikhetsfunktionen för stokastiska variablerna X och Y med $p_{X,Y}(j,k)$

X/Y	0	1	2	3
0	0.2	0	0	0
1	0	0.1	0.1	0
2	0	0.1	0.1	0
3	0	0	0	0.4

- Marginalfördelning för X:

$$p_X(0) = 0.2 + 0 + 0 + 0 = 0.2,$$

$$p_X(1) = 0 + 0.1 + 0.1 + 0 = 0.2$$

$$p_X(2) = 0 + 0.1 + 0.1 + 0 = 0.2,$$

$$p_X(3) = 0 + 0 + 0 + 0.4 = 0.4$$

Exempel (forts.)

- På samma sätt fås marginalfördelning för Y:

$$p_Y(0) = 0.2, p_Y(1) = 0.2$$

$$p_Y(2) = 0.2, p_Y(3) = 0.4.$$

- X och Y är INTE oberoende, ty, t.ex.,

$$p_X(0) \cdot p_Y(0) = 0.2 \cdot 0.2 = 0.04 \neq 0.2 = p_{X,Y}(0,0)$$

KTH Matematik

Exempel (forts.):

$$E(X) = 0 \cdot 0.2 + 1 \cdot 0.2 + 2 \cdot 0.2 + 3 \cdot 0.4 = 1.8.$$

samt

$$E(Y) = 1.8.$$

$$E\left(X^2\right) = 0^2 \cdot 0.2 + 1^2 \cdot 0.2 + 2^2 \cdot 0.2 + 3^2 \cdot 0.4 = 4.6.$$

$$V(X) = E\left(X^2\right) - E\left(X\right)^2 = 4.6 - 1.8^2 = 1.36 = V(Y).$$

Exempel (forts.): Kovarians och korrelation

- $$E(X \cdot Y) = 0 \cdot 0 \cdot 0.2 + 1 \cdot 1 \cdot 0.1$$
  $$+1 \cdot 2 \cdot 0.1 + 2 \cdot 1 \cdot 0.1 + 2 \cdot 2 \cdot 0.1 + 3 \cdot 3 \cdot 0.4 = 4.5$$
- Kovariansen
  $$\mathrm{C}(X, Y) = E(X \cdot Y) - E(X)\,E(Y) = 4.5 - 1.8 \cdot 1.8 = 1.26$$
- Korrelationskoefficienten
  $$\rho(X, Y) = \frac{\mathrm{C}(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{1.26}{\sqrt{1.36}\sqrt{1.36}}$$
  $$= \frac{1.26}{1.36} = 0.926.$$

Vad säger detta ?

Mer om väntevärden

Sats

Låt (X, Y) *vara en tvådimensionell s.v. Då gäller*

(1) $E(aX + bY) = aE(X) + bE(Y)$;

(2) $V(aX + bY) = a^2V(X) + b^2V(Y) + 2abC(X, Y)$.

Bevis. (1) har visats ovan.

(2) fås av följande

$$V(aX + bY) = E[(aX + bY - a\mu_X - b\mu_Y)^2] = E[(aX - a\mu_X + bY - b\mu_Y)^2]$$

$$= E[a^2(X - \mu_X)^2 + b^2(Y - \mu_Y)^2 + 2ab(X - \mu_X)(Y - \mu_Y)]$$

$$= a^2V(X) + b^2V(Y) + 2abC(X, Y).$$

□

KTH Matematik

Mer om väntevärden

Följdsats

Låt X och Y vara två oberoende (okorrelerade räcker) s.v. Då gäller

$$E(X+Y) = E(X) + E(Y) \quad V(X+Y) = V(X) + V(Y)$$
$$E(X-Y) = E(X) - E(Y) \quad V(X-Y) = V(X) + V(Y).$$

KTH Matematik

Detta går att utvidga till godtyckligt många variabler:

Sats

Låt $X_1, \ldots, X_n$ vara oberoende (okorrelerade räcker) s.v. och sätt

$$Y = c_1 X_1 + \ldots + c_n X_n.$$

Då gäller

$$E(Y) = c_1 E(X_1) + \ldots + c_n E(X_n)$$

och

$$V(Y) = c_1^2 V(X_1) + \ldots + c_n^2 V(X_n)$$

Arimetiskt medelvärde

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Sats

Låt $X_1, X_2, \ldots, X_n$ vara oberoende och likafördelade s.v. med väntevärde μ och standardavvikelse σ. Då gäller att

$$E(\overline{X}) = \mu, \quad V(\overline{X}) = \frac{\sigma^2}{n} \quad \textit{och} \quad D(\overline{X}) = \frac{\sigma}{\sqrt{n}}.$$

Uttrycket "$X_1, X_2, \ldots, X_n$ är likafördelade" betyder att de stokastiska variablernas fördelningar, dvs. att de stokastiska variablernas *statistiska egenskaper*, är identiska. *Utfallen* av variablerna varierar dock.

KTH Matematik

Stora talens lag

$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

Sats

Stora talens lag *För varje* $\varepsilon > 0$ *gäller*

$$P(|\overline{X} - \mu| > \varepsilon) \to 0 \quad \text{då } n \to \infty.$$

Bevis. Enl. Tjebysjovs olikhet gäller

$$P(|\overline{X} - \mu| > \varepsilon) \leq \frac{V(\overline{X})}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0$$

då $n \to \infty$. $\square$

KTH Matematik

Summans fördelning

Låt X och Y vara två oberoende kontinuerliga stokastiska variabler med tätheter $f_X(x)$ och $f_Y(y)$.
Sätt $Z = X + Y$. Då gäller

$$F_Z(z) = P(X + Y \leq z) = P((X, Y) \in \{(x, y); x + y \leq z\})$$

$$= \int_{x+y\leq z} f_X(x) f_Y(y)\, dx\, dy$$

(fixera x och integrera över y)

$$= \int_{-\infty}^{\infty} f_X(x) \left(\int_{-\infty}^{z-x} f_Y(y)\, dy \right) dx$$

$$= \int_{-\infty}^{\infty} f_X(x) F_Y(z - x)\, dx.$$

Summans fördelning

Z är också en kontinuerlig stokastisk variabel. Derivation map. z ger

$$f_Z(z) = F_Z'(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,dx.$$

Denna operation kallas *faltning*.

KTH Matematik