TAMS79: Föreläsning 4
Flerdimensionella stokastiska variabler

Johan Thim (johan.thim@liu.se)

10 november 2018

Vi fokuserar på två-dimensionella variabler. Det är steget från en dimension till två som är det svåraste. Generaliseringar till högre dimensioner följer utan problem i de flesta fall. I $\mathbf{R}^2$ är Borelfamiljen $\mathscr{B}$ den minsta σ-algebra som innehåller alla öppna rektanglar $(a,b)\times(c,d)$.

Stokastisk variabel

Definition. En tvådimensionell stokastisk variabel är en reell-vektorvärd funktion (X,Y) (två komponenter) definierad på ett utfallsrum Ω. Alltså avbildar (X,Y) olika utfall på reella vektorer; $(X,Y)\colon \Omega \to \mathbf{R}^2$. Vi kräver att $(X,Y)^{-1}(B) \in \mathcal{F}$ för alla $B \in \mathscr{B}$. Algebran $\mathcal{F}$ är mängden av alla tillåtna händelser. Om (X,Y) bara antar ändligt eller uppräkneligt många värden så kallar vi (X,Y) för en diskret stokastisk variabel. Om varken X eller Y är diskret kallar vi (X,Y) för kontinuerlig.

Definitionen är analog med envariabelfallet. Observera dock följande: en situation som kan uppstå är att vi får ”halvdiskreta” variabler med ena variabeln diskret och den andra kontinuerlig! Inträffar inte ofta i denna kurs, men det kan vara värt att beakta.

Exempel

(i) Låt (X,Y) vara vikten X och längden Y hos en person. Variabeln är kontinuerlig och $\Omega = (X(\Omega), Y(\Omega)) = [0,\infty)^2 = [0,\infty)\times[0,\infty)$.

(ii) Låt (X,Y) vara resultaten av ett tärningskast (X) respektive ett myntkast (Y), där vi representerar krona med 1 och klave med -1. Variabeln är diskret och:

$$\begin{aligned} \Omega &= \{\,\text{⚀},\text{⚁},\text{⚂},\text{⚃},\text{⚄},\text{⚅}\,\} \times \{\,\text{Krona}, \text{Klave}\,\}, \\ (X(\Omega), Y(\Omega)) &= \{\,1,2,3,4,5,6\,\} \times \{\,-1,1\,\}. \end{aligned}$$

I det 2-dimensionella fallet är vi nu intresserade av delmängder i planet. För att kunna prata om en fördelningsfunktion introducerar vi mängden

$$C(x,y) = \{(u,v)\in\mathbf{R}^2 : u \le x \text{ och } v \le y\}.$$

Vi ser att $P((X,Y)\in C(x,y)) = P(X \le x, Y \le y)$ och gör följande definition.

Fördelningsfunktion

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

Om (X,Y) är diskret låter vi

$$p_{X,Y}(j,k) = P(X = j, Y = k).$$

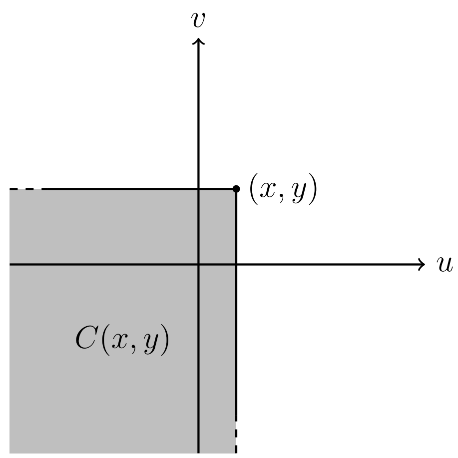
Detta är **sannolikhetsfunktionen** för (X,Y). Fördelningsfunktionen ges då av

$$F_{X,Y}(x,y) = \sum_{j \leq x} \sum_{k \leq y} p_{X,Y}(j,k).$$

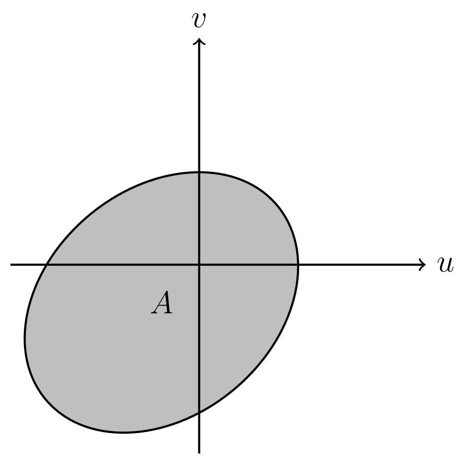
Om det finns en icke-negativ integrerbar funktion $f_{X,Y}$ så att

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

så kallar vi $f_{X,Y}$ för variabelns **simultana täthetsfunktion**. Detta är typfallet för att (X,Y) är en kontinuerlig tvådimensionell stokastisk variabel.
Sannolikheten $F_{X,Y}(x,y)$ och sannolikheten för en mer generell delmängd $A \subset \mathbf{R}^2$ kan grafiskt illustreras genom figuren nedan. Observera att sannolikheten inte är den skuggade arean, utan volymen som uppstår när vi har en funktionsyta definierad ovanför det skuggade området. Arean symboliserar alltså ett integrations- eller summationsområde. Vi ”summerar” (via en integral eller summa) sedan sannolikhetsvärden för de intressanta värderna för variabeln.



Den halvoändliga rektangeln $C(x,y)$.



$P((X,Y) \in A)$ är sannolikheten att få ett resultat (x,y) som ligger i mängden A.

Egenskaper hos sannolikhetsfunktionen

(i) $p_{X,Y}(j,k) \geq 0$ för alla (j,k).

(ii) $\displaystyle\sum_j \sum_k p_{X,Y}(j,k) = 1$.

(iii) Om $A \subset \mathbf{R}^2$ så är $P(X \in A) = \displaystyle\sum\sum_{(j,k)\in A} p_{X,Y}(j,k)$.

Egenskaper hos den simultana täthetsfunktionen

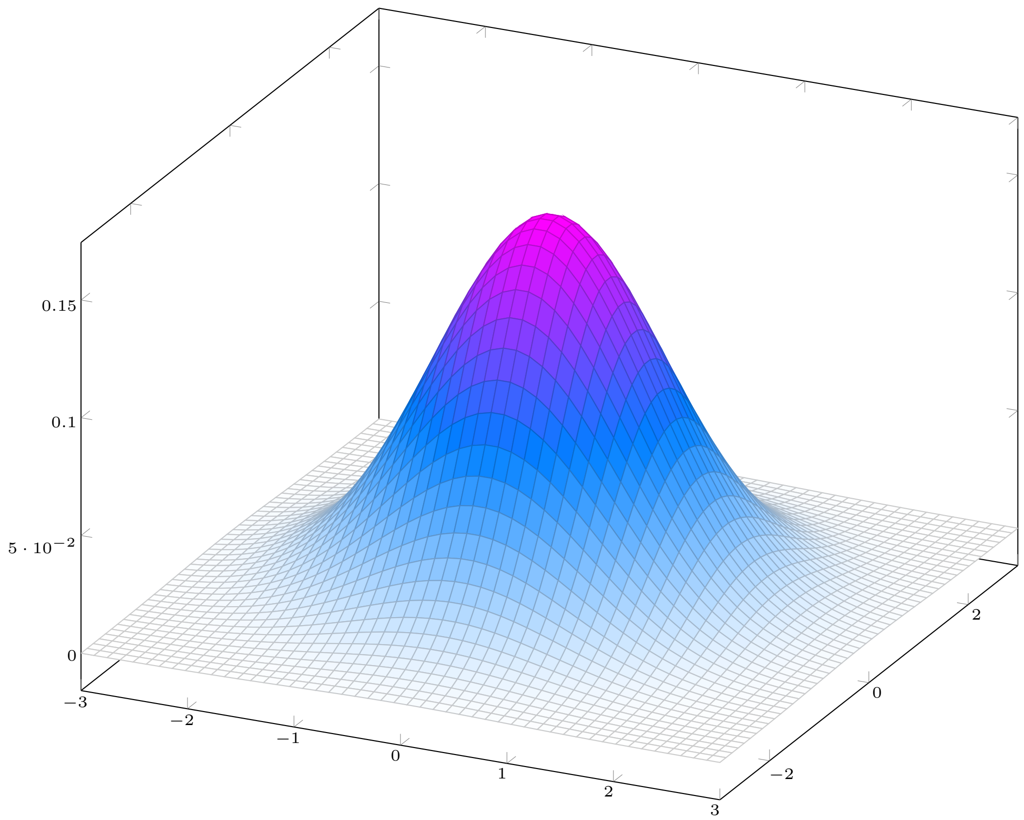
(i) $f_{X,Y}(x,y) \geq 0$ för alla $(x,y) \in \mathbf{R}^2$.

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

(iii) Om $A \in \mathscr{B}$ (så $A \subset \mathbf{R}^2$ är *snäll*) så är $P((X,Y) \in A) = \displaystyle\iint_A f_{X,Y}(x,y)\,dxdy$.

(iv) Talet $f_{X,Y}(x,y)$ anger hur mycket sannolikhetsmassa det finns per areaenhet i punkten (x,y).

Exempel på hur en tvådimensionell täthetsfunktion kan se ut. Det är nu volymen, inte arean, som ska vara ett.



Exempel

Det enklaste exemplet på en 2D-täthetsfunktion är den likformiga fördelningen. Låt A vara rektangeln med hörn i $(0,0)$ och $(2,3)$ och låt (X,Y) vara likformigt fördelad på A. Då ges $f_{X,Y}$ av $f_{X,Y}(x,y) = 1/6$ om $(x,y) \in A$ och $f_{X,Y}(x,y) = 0$ för övrigt. Sannolikheten att $X > Y$ kan vi enkelt beräkna genom att titta på hur stor del av A där detta villkor är uppfyllt, vilket är triangeln med hörn i $(0,0)$, $(2,0)$ och $(2,2)$. Vi erhåller alltså

$$P(X > Y) = \frac{\text{area triangel}}{\text{area rektangel}} = \frac{2}{6} = \frac{1}{3}.$$

Definition. De **marginella** täthetsfunktionerna f_X och f_Y för X och Y i en kontinuerlig stokastisk variabel (X,Y) ges av

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \qquad \text{och} \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx.$$

Motsvarande gäller om (X,Y) är diskret:

$$p_X(j) = \sum_k p_{X,Y}(j,k) \qquad \text{och} \qquad p_Y(k) = \sum_j p_{X,Y}(j,k).$$

Man kan även definiera marginella fördelningsfunktioner genom

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x,y) \qquad \text{och} \qquad F_Y(y) = \lim_{x\to\infty} F_{X,Y}(x,y).$$

Så vad är egentligen dessa *marginella* funktioner? Det vi gör är att vi summerar alla möjligheter för den variabel vi inte är intresserade av och på det sättet skapar något som bara beror på en variabel. Detta leder också till följande sats.

Oberoende variabler

Sats. Om (X,Y) är en stokastisk variabel med simultan täthetsfunktion $f_{X,Y}$ gäller att X och Y är oberoende om och endast om $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. För en diskret variabel är motsvarande villkor $p_{X,Y}(j,k) = p_X(j)p_Y(k)$.

Det är även sant att (i båda fallen) X och Y är oberoende om och endast om

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

Exempel

Låt (X, Y) ha sannolikhetsfunktionen $p_{X,Y}(j,k) = c(jk + k^3)$ för $j = 0, 1, 2, 3$ och $k = 0, 1$. Annars är $p_{X,Y} = 0$.

(i) Vad är c?

(ii) Bestäm $p_X(j)$ och $p_Y(k)$. Är X och Y oberoende?

(iii) Beräkna $P(X \leq 2, Y \geq 0.5)$.

Lösning:

(i) Vi måste välja $c > 0$ för att $p_{X,Y}$ ska kunna vara en sannolikhetsfunktion. För att finna c summerar vi över alla j och k:

$$\sum_{j=0}^{3}\sum_{k=0}^{1} c(jk + k^3) = \sum_{j=0}^{3} c(j+1) = c(1+2+3+4) = 10c,$$

så $c = 1/10$ är nödvändigt och tillräckligt.

(ii) De marginella sannolikhetsfunktionerna fås ur definitionen enligt

$$p_X(j) = \sum_{k=0}^{1} c(jk + k^3) = c(j+1), \quad j = 0, 1, 2, 3.$$
$$p_Y(k) = \sum_{j=0}^{3} c(jk + k^3) = c(k^3 + k + k^3 + 2k + k^3 + 3k + k^3) = 2c(3k + 2k^3), \quad k = 0, 1.$$

Vi testar även att

$$\sum_{j=0}^{3} p_X(j) = c(1+2+3+4) = 1,$$
$$\sum_{k=0}^{1} p_Y(k) = 2c(3+2) = 1,$$

så p_X och p_Y är sannolikhetsfunktioner. Vi undersöker oberoendet:

$$p_X(j)p_Y(k) = 2c^2(j+1)(3k + 2k^3).$$

Här kan man kanske direkt tro att X och Y är beroende, men det skulle vara en gissning. Vi undersöker explicit:

$p_{X,Y}$	$j=0$	$j=1$	$j=2$	$j=3$
$k=0$	0	0	0	0
$k=1$	1/10	2/10	3/10	4/10

$p_X p_Y$	$j=0$	$j=1$	$j=2$	$j=3$
$k=0$	0	0	0	0
$k=1$	$\frac{1\cdot2\cdot5}{100}$	$\frac{2\cdot2\cdot5}{100}$	$\frac{3\cdot2\cdot5}{100}$	$\frac{4\cdot2\cdot5}{100}$

Vi ser här att alla siffror matchar varandra och att därmed $p_{X,Y}(j,k) = p_X(j)p_X(k)$ för *alla* j och k. Detta trots att uttrycken såg väldigt olika ut från början. Var försiktiga med att dra slutsatser utan att undersöka ordentligt!

(iii) Vi kan numrera de tillåtna värderna (där $X = j \leq 2$ och $Y = k \geq 0.5$) på (j,k): $(0,1)$, $(1,1)$, $(2,1)$. Sannolikheten blir då

$$P(X \leq 2, Y \geq 0.5) = p_{X,Y}(0,1) + p_{X,Y}(1,1) + p_{X,Y}(2,1) = c(1+2+3) = 6/10.$$

Exempel

Låt (X,Y) ha den simultana täthetsfunktionen

$$f_{X,Y}(x,y) = \begin{cases} c(x^2y + xy^2), & 0 < x < 1,\ 0 < y < 1, \\ 0, & \text{annars.} \end{cases}$$

Lös följande problem.

(i) Bestäm c;

(ii) Vad är sannolikheten för händelsen att $X > 1/2$ och att $Y < 1/2$? Det vill säga, beräkna sannolikheten $P(X > 1/2,\ Y < 1/2)$;

(iii) Beräkna $P(X > 1/2)$;

(iv) Beräkna $P\big(Y < 1/2 \mid X > 1/2\big)$;

(v) Beräkna $f_X(x)$ och $f_Y(y)$. Är X och Y oberoende?

(vi) Beräkna $F_{X,Y}(x,y)$.

(vii) Vad blir $\dfrac{\partial^2}{\partial_x \partial_y} F_{X,Y}(x,y)$?

Lösning:

(i) Vi räknar ut följande

$$1 = \iint_{\mathbf{R}^2} f_{X,Y}(x,y)\,dxdy = \int_0^1 \int_0^1 c(x^2y + xy^2)\,dxdy = c\int_0^1 \left(\frac{y}{3} + \frac{y^2}{2}\right) dy = \frac{c}{3},$$

så $c = 3$ är nödvändigt för att få en täthetsfunktion.

(ii) Vi har

$$P(X > 1/2 \text{ och } Y < 1/2) = \int_{1/2}^1 \int_0^{1/2} 3(x^2y + xy^2)\,dydx = \frac{5}{32}.$$

(iii) Vi har endast ett krav på x, så vi integrerar över alla y:

$$P(X > 1/2) = \int_{1/2}^1 \int_0^1 3(x^2y + xy^2)\,dydx = \frac{13}{16}.$$

(iv) Enligt definitionen av betingad sannolikhet,

$$P\big(Y < 1/2 \mid X > 1/2\big) = \frac{P(X > 1/2,\ Y < 1/2)}{P(X > 1/2)} = \frac{5/32}{13/16} = \frac{5}{26}.$$

(v) De marginella tätheterna beräknas direkt från definitionen:

$$f_X(x) = \int_0^1 3(xy^2 + x^2y)\,dy = x + \frac{3}{2}x^2, \quad 0 \le x \le 1,$$
$$f_Y(y) = \int_0^1 3(xy^2 + x^2y)\,dx = y + \frac{3}{2}y^2, \quad 0 \le y \le 1.$$

Vi ser tydligt att $f_X(x)f_Y(y) \neq f_{X,Y}(x,y)$ för många val av punkter (x,y); ta till exempel $(x,y) = (1,1)$. Variablerna är beroende.

(vi) Låt $(x,y) \in [0,1] \times [0,1]$. Då blir

$$F_{X,Y}(x,y) = \int_0^x \int_0^y 3(uv^2 + u^2v)\,dvdu = \frac{x^3y^2 + x^2y^3}{2}.$$

Om $x < 0$ *eller* $y < 0$, måste $F_{X,Y}(x,y) = 0$, och om $x > 1$ *och* $y > 1$, så måste $F_{X,Y}(x,y) = 1$. Övriga fall täcks av

$$F_{X,Y}(x,y) = \frac{y^2 + y^3}{2}, \quad 0 \le y \le 1 \text{ och } x > 1,$$
$$F_{X,Y}(x,y) = \frac{x^3 + x^2}{2}, \quad 0 \le x \le 1 \text{ och } y > 1.$$

(vii) Om $0 \le x \le 1$ och $0 \le y \le 1$,

$$\frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) = \frac{\partial}{\partial x}\left(\frac{2x^3y + 3x^2y^2}{2}\right) = \frac{6x^2y + 6xy^2}{2} = f_{X,Y}(x,y).$$

Övriga kombinationer av x och y kommer att ge $\dfrac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) = 0$.

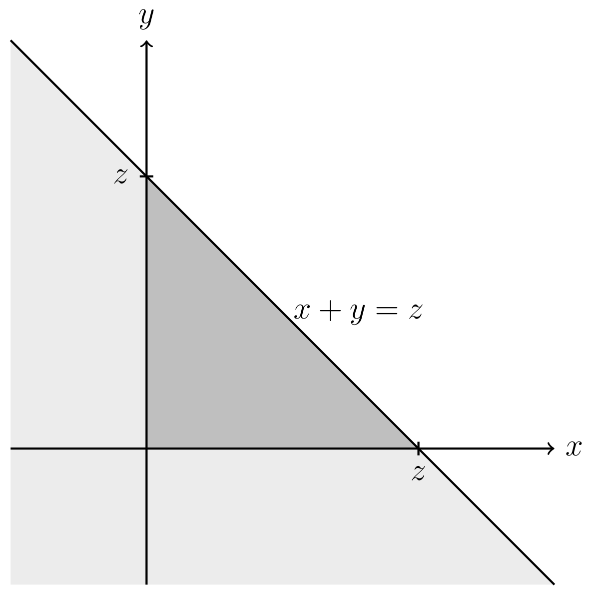
4.1 Funktioner av flera stokastiska variabler

Om vi har den simultana täthets- eller sannolikhetsfunktionen så kan vi hitta fördelningar för funktioner av flera stokastiska variabler. Låt oss betrakta ett par vanliga exempel.

Exempel

Låt $X \sim \text{Exp}(1)$ och $Y \sim \text{Exp}(2)$ vara oberoende. Beräkna täthetsfunktionen för $Z = X + Y$.

Klart att $f_{X,Y}(x,y) = f_X(x)f_Y(y) = 2 \cdot 1 \cdot e^{-x-2y}$ för $x \ge 0$ och $y \ge 0$ (annars $f_{X,Y} = 0$). Vi söker $f_Z(z)$. Det är klart att om $z < 0$ så är $f_{X,Y}(x,y) = 0$. Antag att $z \ge 0$. Vi ställer upp fördelningsfunktionen $F_Z(z)$ för Z, och för det behöver vi ha klart för oss vilket område vi integrerar över:



Fördelningsfunktionen blir nu, för $z \geq 0$,

$$
\begin{aligned}
F_Z(z) = P(X + Y \leq z) &= \int_{-\infty}^{z} \int_{-\infty}^{z-x} f_{X,Y}(x, y)\, dydx \\
&= 2 \int_0^z e^{-x} \int_0^{z-x} e^{-2y}\, dydx = \int_0^z e^{-x}(1 - e^{-2(z-x)})\, dx \\
&= \int_0^z \left(e^{-x} - e^{-2z+x}\right) dx = 1 + e^{-2z} - 2e^{-z},
\end{aligned}
$$

och vi kan derivera fram $f_Z(z) = F_Z'(z) = 2e^{-z} - 2e^{-2z}$. Kontrollera att $f_Z(z) \geq 0$ samt att $\int_0^\infty f_Z(z)\, dz = 1$.

Faltningssatsen

Sats. Om X och Y är oberoende kontinuerliga stokastiska variabler så ges täthetsfunktionen f_Z för $Z = X + Y$ av

$$
f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\, dx, \quad z \in \mathbf{R}.
$$

Motsvarande gäller för diskreta variabler:

$$
p_Z(k) = \sum_j p_X(j) p_Y(k - j).
$$

Min och max

Vad blir fördelningsfunktionerna för maximum respektive minimum av två oberoende stokastiska variabler?

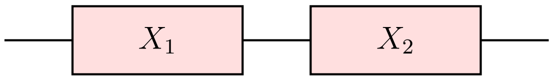
Låt $Y = \max(X_1, X_2)$. Då blir

$$F_Y(y) = P(\max(X_1, X_2) \leq y) = P(X_1 \leq y \text{ och } X_2 \leq y) = P(X_1 \leq y)P(X_2 \leq y)$$
$$= F_{X_1}(y)F_{X_2}(y).$$

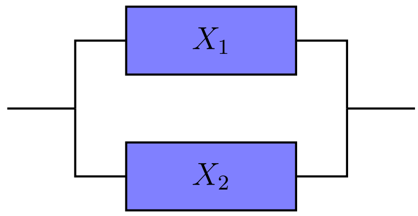
För $Y = \min(X_1, X_2)$ så erhåller vi

$$F_Y(y) = P(\min(X_1, X_2) \leq y) = 1 - P(\min(X_1, X_2) > y) = 1 - P(X_1 > y \text{ och } X_2 > y)$$
$$= 1 - P(X_1 > y)P(X_2 > y) = 1 - (1 - P(X_1 \leq y))(1 - P(X_2 \leq y))$$
$$= 1 - (1 - F_{X_1}(y))(1 - F_{X_2}(y)).$$

En vanlig situation där dessa uttryck dyker upp är i parallell- och seriekoppling av system. Schematiskt kan man beskriva dessa situationer med blockscheman.



Seriekoppling. Båda kanalerna måste fungera. Ger minimum av livslängderna.



Parallellkoppling. Räcker att en kanal fungerar. Ger maximum av livslängderna.