STK1100 våren 2023

Simultane, marginale og betingede fordelinger

Dekker det meste av avsnitt 5.4 og 5.5

Matematisk institutt
Universitetet i Oslo

Simultane og marginale punktsannsynligheter for diskrete stokastiske variabler

For to diskrete stokastiske variabler *X* og *Y* er den simultane punktsannsynligheten gitt ved

$$p(x,y) = P(X = x \text{ og } Y = y)$$

Den marginale punktsannsynligheten for *X* er

$$p_X(x) = P(X = x) = \sum_y p(x,y)$$

Den marginale punktsannsynligheten for *Y* er

$$p_Y(y) = P(Y = y) = \sum_x p(x,y)$$

Betinget punktsannsynlighet for diskrete stokastiske variabler

Den betingete punktsannsynligheten for Y gitt at $X = x$ er gitt ved

$$p_{Y|X}(y \mid x) = P(Y = y \mid X = x)$$

$$= \frac{P(X = x \text{ og } Y = y)}{P(X = x)} = \frac{p(x, y)}{p_X(x)}$$

Den betingete punktsannsynligheten for X gitt at er $Y = y$ er gitt ved

$$p_{X|Y}(x \mid y) = P(X = x \mid Y = y) = \frac{p(x, y)}{p_Y(y)}$$

Eksempel – Fordeling av karakterer

La X være karakteren i norsk og Y karakteren i matematikk for en tilfeldig valgt elev

Simultan og marginal punktsannsynlighet for X og Y er gitt i tabellen (i prosent)

y \ x	1	2	3	4	5	6	$P(Y=y)$
1	1	4	5	3	-	-	**13**
2	1	4	11	6	-	-	**22**
3	1	4	8	8	2	-	**23**
4	-	3	7	6	4	-	**20**
5	-	1	3	6	5	1	**16**
6	-	-	1	2	2	1	**6**
$P(X=x)$	**3**	**16**	**35**	**31**	**13**	**2**	**100**

y \ x	1	2	3	4	5	6	$P(Y=y)$
1	1	4	5	3	-	-	13
2	1	4	11	6	-	-	22
3	1	4	8	8	2	-	23
4	-	3	7	6	4	-	20
5	-	1	3	6	5	1	16
6	-	-	1	2	2	1	6
$P(X=x)$	3	16	35	31	13	2	100

Gitt at $X = 5$, har vi (f. eks.) at

$$p_{Y|X}(4\,|\,5) = \frac{p(5,4)}{p_X(5)} = \frac{4/100}{13/100} = 0.31 \quad (31\%)$$

$$p_{Y|X}(5\,|\,5) = \frac{p(5,5)}{p_X(5)} = \frac{5/100}{13/100} = 0.38 \quad (38\%)$$

Betinget forventning og varians for diskrete stokastiske variabler

Den betingete forventningen for Y gitt at $X = x$ er gitt ved

$$\mu_{Y|X=x} = E(Y \mid X = x) = \sum_y y \cdot p_{Y|X}(y \mid x)$$

Den betingete variansen for Y gitt at $X = x$ er gitt ved

$$\sigma^2_{Y|X=x} = V(Y \mid X = x) = \sum_y (y - \mu_{Y|X=x})^2 \cdot p_{Y|X}(y \mid x)$$

Tilsvarende gjelder for betinget forventning og varians for X gitt at $Y = y$

Eksempel (forts)

y \ x	1	2	3	4	5	6	*P*(*Y*=*y*)
1	1	4	5	3	-	-	13
2	1	4	11	6	-	-	22
3	1	4	8	8	2	-	23
4	-	3	7	6	4	-	20
5	-	1	3	6	5	1	16
6	-	-	1	2	2	1	6
P(*X*=*x*)	3	16	35	31	13	2	100

Gitt at $X = 5$, er forventet verdi av Y

$$\mu_{Y|X=5} = \sum_{y=1}^{6} y \cdot p_{Y|X}(y\,|\,5) = 4.54$$

(Python-kommandoer på neste slide)

Gitt at $X = 3$, er forventet verdi av Y

$$\mu_{Y|X=3} = \sum_{y=1}^{6} y \cdot p_{Y|X}(y\,|\,3) = 2.86$$

Python-kommandoer for å beregne de betingete forventningene på forrige slide

```
tmp = np.arange(1,7)
x = np.tile(tmp,(6,1))
y = x.transpose()
pxy = np.array([[0.01,0.04,0.05,0.03,0.0,0.0],
        [0.01,0.04,0.11,0.06,0.0,0.0],
        [0.01,0.04,0.08,0.08,0.02,0.0],
        [0.0,0.03,0.07,0.06,0.04,0.0],
        [0.0,0.01,0.03,0.06,0.05,0.01],
        [0.0,0.0,0.01,0.02,0.02,0.01]])
px = sum(pxy)
sum(tmp*pxy.transpose()[4]/[px[4]])
sum(tmp*pxy.transpose()[2]/[px[2]])
```

Simultane og marginale sannsynlighetstettheter for kontinuerlige stokastiske variabler

For to kontinuerlige stokastiske variabler X og Y har vi en simultan synlighetstetthet $f(x,y)$

De marginale sannsynlighetstetthetene for X og Y er

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

Betingete sannsynlighetstettheter for kontinuerlige stokastiske variabler

Den betingede sannsynlighetstettheten for Y gitt at $X = x$ er gitt ved

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

Den betingede sannsynlighetstettheten for X gitt at $Y = y$ er gitt ved

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

Eksempel (fra læreboka)

X og Y har simultantetthet

$$f(x,y) = \begin{cases} \frac{6}{5}(x+y^2) & 0 \leq x \leq 1, \quad 0 \leq y \leq 1 \\ 0 & \text{ellers} \end{cases}$$

De marginale tetthetene er (for $0 \leq x \leq 1, \quad 0 \leq y \leq 1$)

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_0^1 \frac{6}{5}(x+y^2)\,dy = \frac{6}{5}x + \frac{2}{5}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \int_0^1 \frac{6}{5}(x+y^2)\,dx = \frac{6}{5}y^2 + \frac{3}{5}$$

Den betingete tettheten for *Y* gitt *X* = *x* er (for $0 \leq y \leq 1$)

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

$$= \frac{\frac{6}{5}(x + y^2)}{\frac{6}{5}x + \frac{2}{5}}$$

$$= \frac{3\,(x + y^2)}{3x + 1}$$

Et tilsvarende uttrykk gjelder for den betingete tettheten for *X* gitt *Y* = *y*

Betinget forventning og varians for kontinuerlige stokastiske variabler

Den betingete forventningen for Y gitt at $X = x$ er gitt ved

$$\mu_{Y|X=x} = E(Y \mid X = x) = \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y \mid x)\,dy$$

Den betingete variansen for Y gitt at $X = x$ er gitt ved

$$\sigma^2_{Y|X=x} = V(Y \mid X = x) = \int_{-\infty}^{\infty} (y - \mu_{Y|X=x})^2 \cdot f_{Y|X}(y \mid x)\,dy$$

Tilsvarende gjelder for betinget forventning og varians for X gitt at $Y = y$

Eksempel (forts)

Den betingete tettheten for Y gitt $X = x$ er (for $0 \leq y \leq 1$)

$$f_{Y|X}(y|x) = \frac{3\,(x + y^2)}{3x + 1}$$

Den betingete forventningen til Y gitt $X = x$ er

$$\mu_{Y|X=x} = \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y|x)dy$$

$$= \int_0^1 y \cdot \frac{3\,(x + y^2)}{3x + 1} dy = \frac{2x + 1}{4\left(x + \frac{1}{3}\right)}$$

Uavhengige stokastiske variabler

To stokastiske variabler X og Y er uavhengige hvis vi for alle (x,y) har at

$$p(x,y) = p_X(x) \cdot p_Y(y) \qquad \text{(diskret)}$$

$$f(x,y) = f_X(x) \cdot f_Y(y) \qquad \text{(kontinuerlig)}$$

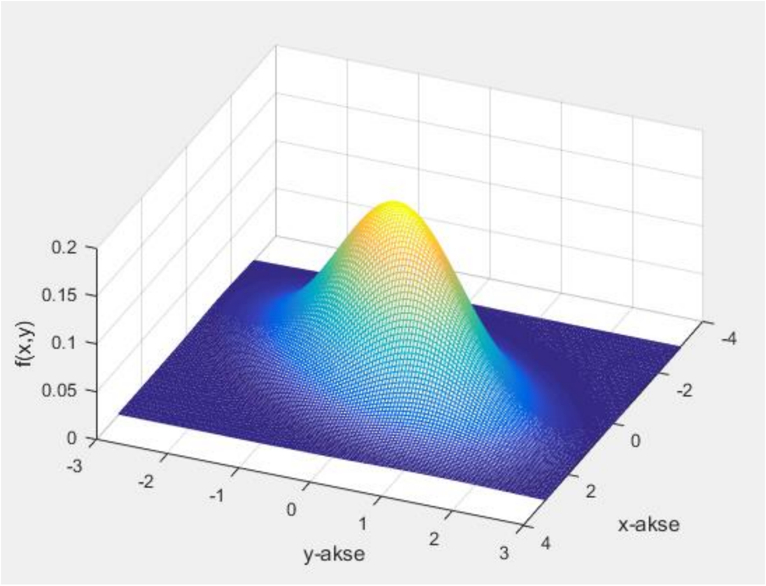
X og Y er uavhengige hvis og bare hvis de betingete punktsannsynlighetene / tetthetene er lik de marginale punktsannsynlighetene / tetthetene

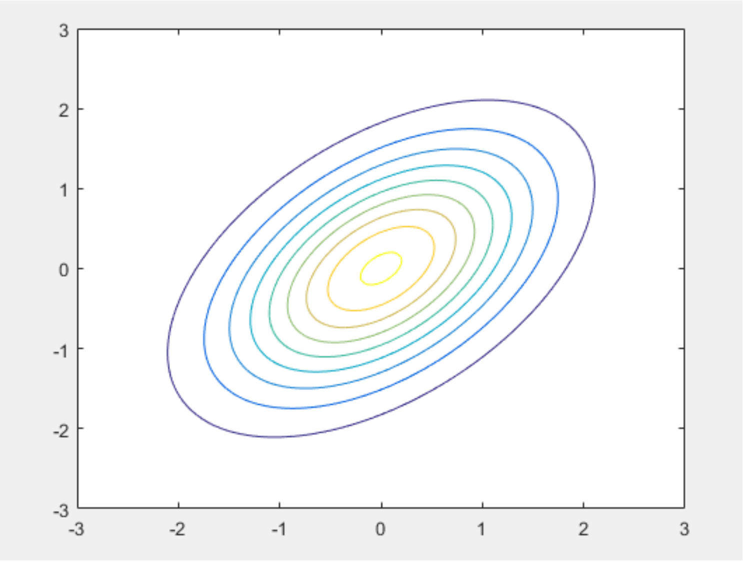
Binormalfordeling

X og *Y* er binormalt fordelt med simultantetthet

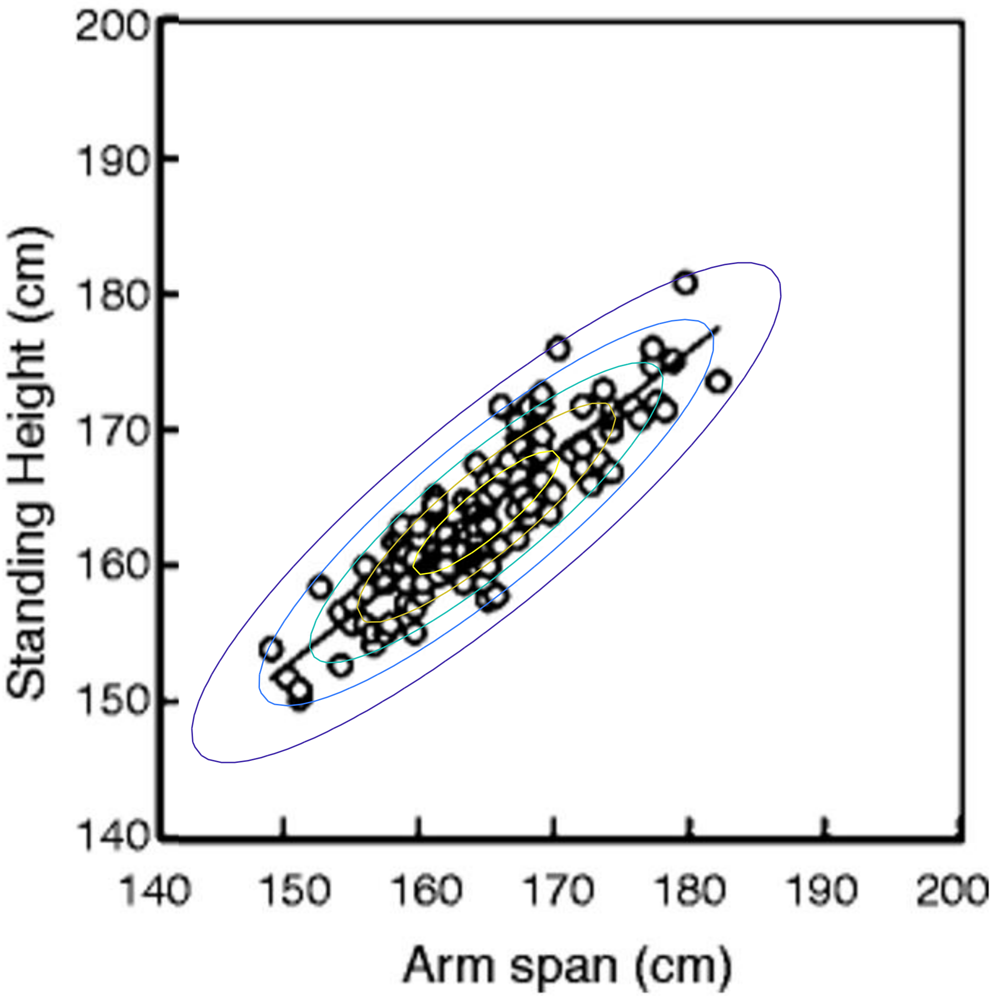
$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2-2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right)+\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\Big/\left[2(1-\rho^2)\right]\right\}$$

Illustrasjon for $\mu_1=\mu_2=0$, $\sigma_1=\sigma_2=1$ og $\rho=0.50$





Eksempel: Rekkevidde og høyde for unge kvinner (australske data)



Vi har to kontinuerlige stokastiske variabler
X = rekkevidde
Y = høyde

X og Y er binormalt fordelt med

$$\mu_1 = \mu_2 = 164$$
$$\sigma_1 = \sigma_2 = 10$$
$$\rho = 0.90$$

Den marginale tettheten for X er (krever en del regning)

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy \;= \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right\}$$

Altså er $X \sim N(\mu_1, \sigma_1)$

Tilsvarende er $\mathrm{Y} \sim N(\mu_2, \sigma_2)$

Videre har vi at $\rho = \mathrm{corr}(X,Y)$

Hvis $\rho = 0$, er X og Y uavhengige

Den betingete tettheten til *Y* gitt *X* = *x* er (krever en del regning)

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

$$= \frac{1}{\sigma_2\sqrt{1-\rho^2}\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma_2^2(1-\rho^2)}\left(y-\mu_2-\rho\sigma_2\frac{x-\mu_1}{\sigma_1}\right)^2\right\}$$

Altså er

$$[Y|X=x] \sim N\left(\mu_2+\rho\sigma_2\frac{x-\mu_1}{\sigma_1}, \sigma_2^2(1-\rho^2)\right)$$

Regresjon mot gjennomsnittet

For binormalfordeling har vi at

$$\mu_{Y|X=x} = \mu_2 + \rho\sigma_2 \frac{x-\mu_1}{\sigma_1}$$

som vi kan skrive

$$\frac{\mu_{Y|X=x} - \mu_2}{\sigma_2} = \rho \frac{x-\mu_1}{\sigma_1}$$

Eksempel

Høyde av mødre og døtre (jf læreboka 333-334)

Dobbelforventning

Den betingete forventningen for Y gitt at $X = x$ er gitt ved (for diskrete tilfellet)

$$\mu_{Y|X=x} = E(Y \mid X = x) = \sum_y y \cdot p_{Y|X}(y \mid x)$$

Vi har at $\mu_{Y|X=x} = h(x)$ er en funksjon av x

Merk at

$$E\left[E(Y \mid X)\right] = Eh(X) = \sum_x \left(\sum_y y \cdot p_{Y|X}(y \mid x) \right) p_X(x)$$

$$= \sum_y \sum_x y \cdot p(x,y) = \sum_y y \cdot p_Y(y) = E(Y)$$

Resultatet gjelder tilsvarende for det kontinuerlige tilfellet

Eksempel

Simultan og marginal punktsannsynlighet for X = karakter i norsk og Y = karakter i matematikk er gitt i tabellen

y \ x	1	2	3	4	5	6	$P(Y=y)$
1	1	4	5	3	-	-	13
2	1	4	11	6	-	-	22
3	1	4	8	8	2	-	23
4	-	3	7	6	4	-	20
5	-	1	3	6	5	1	16
6	-	-	1	2	2	1	6
$P(X=x)$	3	16	35	31	13	2	100

Vi finner

x	1	2	3	4	5	6
$\mu_{Y\|X=x}$	2.00	2.56	2.86	3.39	4.54	5.50

Det gir

$$E(E(Y \mid X)) = 2.00 \cdot 0.03 + ... + 5.50 \cdot 0.02 \quad = 3.22 = E(Y)$$