STK1100 våren 2023

Forventning, kovarians og korrelasjon

Svarer til avsnitt 5.2 i læreboka

Matematisk institutt
Universitetet i Oslo

Forventningsverdi

La X og Y være diskrete stokastiske variabler med simultan punktsannsynlighet $p(x,y)$ og la $h(x,y)$ være en reell funksjon. Da er

$$E\left[h(X,Y)\right] = \sum_x \sum_y h(x,y)\cdot p(x,y)$$

Hvis X og Y er kontinuerlige stokastiske variabler med simultan synlighetstetthet , $f(x,y)$ så har vi

$$E\left[h(X,Y)\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y)\cdot f(x,y)\,dx\,dy$$

Eksempel – Fordeling av karakterer

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

Punktsannynligheten $p(x,y)$ er gitt ved tabellen (sannsynlighetene er gitt som prosent)

y \ x	1	2	3	4	5	6
1	1	4	5	3	-	-
2	1	4	11	6	-	-
3	1	4	8	8	2	-
4	-	3	7	6	4	-
5	-	1	3	6	5	1
6	-	-	1	2	2	1

Hva er forventet forskjell i karakterene i norsk og matematikk?

Vi er interessert i forskjellen på karakterene i norsk og matematikk, dvs $h(X,Y) = |X - Y|$

Vi finner (se kommandoer på neste slide)

$$E\{|X-Y|\} = \sum_x \sum_y |x-y| \cdot p(x,y)$$

$$= |1-1| \cdot p(1,1) + |2-1| \cdot p(2,1) + ... + |6-6| \cdot p(6,6)$$

$$= 0 \cdot 0.01 + 1 \cdot 0.04 + ... + 0 \cdot 0.01 = 1.07$$

Tilsvarende finner vi

$$E(X) = 3.41 \qquad E(Y) = 3.22$$

Merk at vi også kan finne $E(X)$ og $E(Y)$ av de marginale punktsannsynlighetene

Kommandoer for å beregne forventningsverdiene på forrige slide

Python

```
import numpy as np
import scipy.stats as stats
import math
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]])
sum(sum(abs(x-y)*pxy))
```

Regneregel for forventning

La X og Y være to stokastiske variabler (diskrete eller kontinuerlige)

Da har vi (regner for kontinuerlig)

$$E(aX+bY+c) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(ax+by+c)\cdot f(x,y)\,dx\,dy$$

$$= a\int_{-\infty}^{\infty} x\left[\int_{-\infty}^{\infty} f(x,y)\,dy\right]dx + b\int_{-\infty}^{\infty} y\left[\int_{-\infty}^{\infty} f(x,y)\,dx\right]dy + c\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy$$

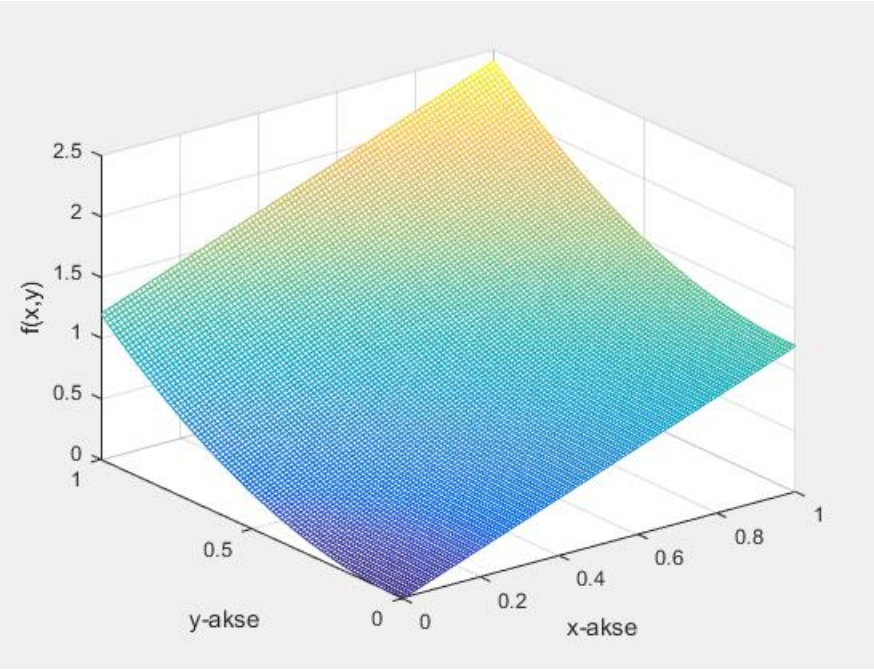
$$= a\int_{-\infty}^{\infty} x\cdot f_X(x)\,dx + b\int_{-\infty}^{\infty} y\cdot f_Y(y)\,dy + c$$

$$= aE(X)+bE(Y)+c$$

Eksempel (forts. av 5.4 i Devore & Berk)

X og Y har simultantetthet

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



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

$$f_X(x) = \frac{6}{5}x + \frac{2}{5} \qquad \text{og} \qquad f_Y(y) = \frac{6}{5}y^2 + \frac{3}{5}$$

Vi har at

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^1 x\left(\frac{6}{5}x + \frac{2}{5}\right) = \frac{3}{5}$$

Tilsvarende er $E(Y) = \frac{3}{5}$

Vi finner:

$$E(1 + X + 2Y) = 1 + E(X) + 2E(Y)$$

$$= 1 + \frac{3}{5} + 2\frac{3}{5} = \frac{14}{5} = 2.8$$

Uavhengige stokastiske variabler

La X og Y være uavhengige stokastiske variabler

Da har vi for alle (x,y) at

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

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

Vi kan da vise at (jf. forelesningen)

$$E\{g(X) \cdot h(Y)\} = E\{g(X)\} \cdot E\{h(Y)\}$$

Spesielt har vi at

$$E(X \cdot Y) = E(X) \cdot E(Y)$$

Kovarians

La X og Y være stokastiske variabler med simultan punktsannsynlighet $p(x,y)$ eller simultantetthet $f(x,y)$

Forventningene er $\mu_X = E(X)$ og $\mu_Y = E(Y)$

Da er kovariansen mellom X og Y gitt ved

$$\mathrm{Cov}(X,Y) = E\{(X-\mu_X)\cdot(Y-\mu_Y)\}$$

$$= \begin{cases} \sum_x \sum_y (x-\mu_X)\cdot(y-\mu_Y)\cdot p(x,y) & \text{(diskret)} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)\cdot(y-\mu_Y)\cdot f(x,y)\,dx\,dy & \text{(kontinuerlig)} \end{cases}$$

Merk at $\mathrm{Cov}(X,X) = E\{(X-\mu_X)^2\} = V(X)$

Hvis X og Y er uavhengige, så er $\mathrm{Cov}(X,Y) = 0$ (men ikke omvendt)

Eksempel – kovarians mellom karakterer

Vi vil bestemme kovariansen mellom karakterene i norsk (X) og matematikk (Y)

Vi har funnet at

$$\mu = E(X) = 3.41 \qquad \mu_Y = E(Y) = 3.22$$

Vi får at

$$\mathrm{Cov}(X,Y) = \sum_x \sum_y (x-3.41)\cdot(y-3.22)\cdot p(x,y)$$

$$= (1-3.41)\cdot(1-3.22)\cdot p(1,1) + \ldots.$$

$$\ldots\ldots + (6-3.41)\cdot(6-3.22)\cdot p(6,6)$$

$$= 0.710$$

Regneregel for kovarians

La X og Y være to stokastiske variabler (diskrete eller kontinuerlige)

Forventningene er

$$\mu_X = E(X) \qquad \mu_Y = E(Y)$$

Da har vi

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= E\{(X-\mu_X)\cdot(Y-\mu_Y)\} \\
&= E(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) \\
&= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y \\
&= E(XY) - \mu_X \cdot \mu_Y
\end{aligned}$$

Eksempel (forts. av 5.4 i Devore & Berk)

X og Y har simultan tetthet

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

Vi har at $\mu_X = E(X) = \frac{3}{5}$ og $\mu_y = E(Y) = \frac{3}{5}$

Nå har vi at

$$E(XY) = \int_{-\infty}^{\infty}\int xyf(x,y)dxdy = \int_{0}^{1}\int xy\frac{6}{5}(x+y^2)dxdy = \frac{7}{20}$$

Det gir at

$$Cov(X,Y) = E(XY) - \mu_X\mu_Y = \frac{7}{20} - \frac{3}{5}\cdot\frac{3}{5} = -\frac{1}{100}$$

Regneregel for kovarians

La X og Y være to stokastiske variabler med

$\mu_X = \mathrm{E}(\mathrm{X})$ og $\mu_Y = \mathrm{E}(\mathrm{Y})$

Da har vi

$$\begin{aligned}
&\mathrm{Cov}(aX+b, cY+d) \\
&= E\{[aX+b-(a\mu_X+b)]\cdot[cY+d-(c\mu_Y+d)]\} \\
&= E\{ac\cdot(X-\mu_X)\cdot(Y-\mu_Y)\} \\
&= acE\{(X-\mu_X)\cdot(Y-\mu_Y)\} \\
&= ac\,\mathrm{Cov}(X,Y)
\end{aligned}$$

Korrelasjon

Kovariansen avhenger av hvilken måleenhet vi bruker

Et mål for samvariasjon som ikke avhenger av måleenhet er korrelasjonskoeffisienten

La $\sigma_X = SD(X)$ og $\sigma_Y = SD(Y)$

Da er korrelasjonskoeffisienten gitt ved

$$\rho_{X,Y} = \text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

Merk at

$$\text{Corr}(X,Y) = \text{Cov}\left(\frac{X-\mu_X}{\sigma_X}, \frac{Y-\mu_Y}{\sigma_Y} \right)$$

Eksempel – Korrelasjon mellom karakterer

Vi har funnet at

$$\mathrm{Cov}(X,Y) = 0.710$$

Vi har videre at

$$\sigma_X = \sqrt{\mathrm{Var}(X)} = 1.06$$

$$\sigma_Y = \sqrt{\mathrm{Var}(Y)} = 1.44$$

Dermed er

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{0.710}{1.06 \cdot 1.44} = 0.47$$

Egenskaper til korrelasjonskoeffisienten

For korrelasjonskoeffisienten gjelder følgende

1) Hvis a og c begge er positive, eller begge er negative, er

$$\text{Corr}(aX+b, cY+d) = \text{Corr}(X,Y)$$

2) $$-1 \le \text{Corr}(X,Y) \le 1$$

3) $\text{Corr}(X,Y) = \pm 1$ hvis og bare hvis $Y = aX + b$ der $a \neq 0$

4) Hvis X og Y er uavhengige, er $\text{Corr}(X,Y) = 0$ (det omvendte resultatet gjelder ikke)