SF1920/SF1921 Sannolikhetsteori och statistik 6,0 hp
Föreläsning 6
Väntevärden
Korrelation och kovarians
Stora talens lag

Jörgen Säve-Söderbergh

Väntevärde för en funktion av en stokastisk variabel

Theorem

Om $Y = g(X)$ *gäller att*

$$E(Y) = \begin{cases} \sum_k g(k)\, p_X(k) & \textit{diskreta s.v.} \\ \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx & \textit{Kontinuerliga s.v.} \end{cases}$$

Vi kan generalisera till flera stokastiska variabler

Theorem

Om $Z = g(X, Y)$ *gäller att*

$$E(Z) = \begin{cases} \sum_j \sum_k g(j,k)\, p_{X,Y}(j,k) & \textit{diskreta s.v.} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)\, f_{X,Y}(x,y)\, dx\, dy & \textit{kontinuerliga s.v.} \end{cases}$$

Intressanta fall av funktionen $g(\cdot)$ är $g(X, Y) = X + Y$, $g(X, Y) = X$ samt $g(X, Y) = XY$

Väntevärdet av en summa av två stokastiska variabler

Theorem

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

Bevis.

Antag X och Y är kontinuerliga stokastiska variabler.

$$\begin{aligned}
E(X+Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y)\, f_{X,Y}(x,y)\, dx\, dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, f_{X,Y}(x,y)\, dx\, dy \ + \\
&+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f_{X,Y}(x,y)\, dx\, dy \\
&= E(X) + E(Y)
\end{aligned}$$

I boken finns ett bevis för diskreta stokastiska variabler. □

Vi behöver inte anta att X och Y är oberoende. Gäller för flera s.v. ...

Väntevärde för en lineärkombination av stokastiska variabler

Theorem

$E(aX + bY + c) = aE(X) + bE(Y) + c$ *(Inget oberoende!)*

Bevis.

(Skiss) Antag att X och Y är kontinuerliga.

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

$\square$

Theorem

Om X och Y är oberoende, gäller att $E(XY) = E(X)E(Y)$

Bevis.

Antag att X och Y är kontinuerliga.

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\ f_{X,Y}(x,y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\ f_X(x)\,f_Y(y)\,dx\,dy \\ &= \left(\int_{-\infty}^{\infty} x\ f_X(x)\ dx\right)\left(\int_{-\infty}^{\infty} y\ f_Y(y)\ dy\right) \\ &= E(X)\,E(Y) \end{aligned}$$

□

Kan generaliseras till flera stokastiska variabler.

Mått på beroendet mellan två stokastiska variabler i en tvådimensionell stokastisk variabel (X, Y)

Vi önskar uttrycka beroendet mellan de två variablerna X och Y i en tvådimensionell stokastisk variabel (X, Y) med ett tal.

Det är motsvarigheten till läges- och spridningsmått som vi söker.

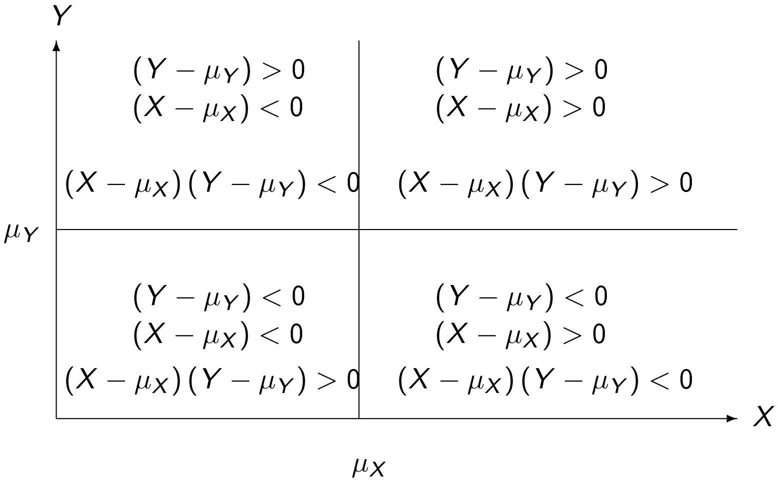
Betrakta en bivariat täthetsfunktion över $\mathbb{R}^2$.

Vi har sett ovan att variansen är väntevärdet av funktionen $g(X) = (X - \mu)^2$.

Vilken funktion $g(X, Y)$ ska vi använda i väntevärdet för att fånga beroendet mellan X och Y?
Vi studerar funktionen

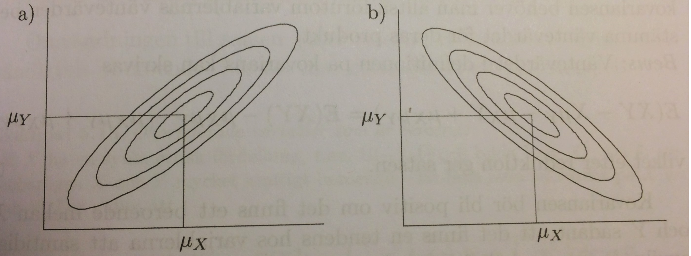
$$g(X, Y) = (X - \mu_X)(Y - \mu_Y).$$

De stokastiska variablerna X och Y har respektive väntevärde μ_X och μ_Y.



Positivt och negativt beroende mellan X och Y

nivåkurvor liknande dem i Figur 5.7, framgår rätt tydligt att det finns ett beroende. I Figur 5.7a har X och Y en tendens att avvika åt samma håll från sina väntevärden μ_X respektive μ_Y, dvs ett högt värde på X har en tendens att följas av ett högt värde även på Y och vice versa. I Figur 5.7b däremot tenderar X och Y att avvika åt olika håll från sina väntevärden.



Figur 5.7 *Nivåkurvor.*

Kovariansen mellan X och Y

Definition

Kovariansen $\mathrm{Cov}\,(X, Y)$ mellan X och Y definieras som

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

Vi ser att

$$\mathrm{Cov}\,(X, X) = E\left[(X - \mu_X)^2\right] = \mathrm{Var}\,(X)$$

Om X och Y har en simultan täthet $f_{X,Y}(x, y)$ gäller att

$$\mathrm{Cov}(X, Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - \mu_X)(y - \mu_Y)f_{X,Y}(x, y)dxdy.$$

Om X och Y är diskreta s.v.

$$\mathrm{Cov}(X, Y) = \sum_i\sum_k(i - \mu_X)(j - \mu_Y)p_{X,Y}(x, y).$$

Beräkningsformel för kovariansen

Om vi använder räknereglerna för $E(\cdot)$ ovan

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

d.v.s.

Theorem

$$Cov(X, Y) = E[XY] - \mu_X\mu_Y.$$

Räkneregler för $\mathrm{Cov}(\cdot)$

Theorem

- $Cov(X, X) = Var(X)$
- $Cov(X, Y) = Cov(Y, X)$
- $Cov(aX + b, cY + d) = acCov(Y, X)$
- $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$

Bevis.

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

Kovariansen är en bilineär funktion

Detta betyder att kovariansen är lineär i båda sina argument.

Theorem

$Cov\,(X+Y,Z+W) =$
$Cov\,(X,Z) + Cov\,(X,W) + Cov\,(Y,Z) + Cov\,(Y,W)$

En lineärkombination av stokastiska variabler $X_1, X_2, \ldots, X_n$ är

$$c_1 X_1 + c_2 X_2 + \cdots + c_n X_n = \sum_{j=1}^{n} c_j X_j,$$

där $c_1, c_2, \ldots, c_n$ är godtyckliga reella tal.

Theorem

$Cov\left(\sum_{j=1}^n a_i X_i, \sum_{j=1}^n b_j X_j\right) = \sum_{i=1}^n \sum_{j=1}^n a_i b_j Cov\,(X_i, X_j)$

Korrelationskoefficienten

Standardavvikelsen för en stokastisk variabel X betecknas $D(X)$.

Definition

Korrelationskoefficienten för X och Y definieras som

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

- ρ är ett mått på det **lineära** beroendet mellan X och Y.
- ρ saknar dimension.
- Varför delar vi $\mathrm{Cov}(X, Y)$ med produkten av $D(X)$ och $D(Y)$?

  Det gör att

$$-1 \leq \rho \leq 1.$$

Korrelationskoefficienten

Definition

Om $\mathrm{Cov}\,(X,Y) = 0$ säges X och Y vara *okorrelerade*.

Då $\mathrm{Cov}\,(X,Y) = E\,(XY) - E\,(X)\,E\,(Y)$ är två s.v. okorrelerade om och endast om $E\,(XY) = E\,(X)\,E\,(Y)$.

Theorem

Om X och Y är oberoende så är de också okorrelerade.

Bevis.

Om X och Y är oberoende $E\,(XY) = E\,(X)\,E\,(Y)$. Alltså är X och Y okorrelerade. $\square$

Motsatsen är inte sann; Beroende variabler kan vara okorrelerade.

Exempel 5.13

En lineärkombination av stokastiska variabler $X_1, X_2, \ldots, X_n$ är

$$c_1 X_1 + c_2 X_2 + \cdots + c_n X_n = \sum_{j=1}^{n} c_j X_j,$$

där $c_1, c_2, \ldots, c_n$ är godtyckliga reella tal.

Theorem

$E\left(\sum_{j=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i E\left(X_i\right)$

$Var\left(\sum_{j=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i^2 \, Var\left(X_i\right) + 2 \sum_{i<j} c_i c_j Cov\left(X_i, X_j\right)$

Om $X_1, X_2, \ldots, X_n$ är oberoende

$Var\left(\sum_{j=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i^2 \, Var\left(X_i\right)$

Välj $c_i = \frac{1}{n}$, vilket ger

$$c_1 X_1 + \cdots + c_n X_n = \frac{1}{n} X_1 + \cdots + \frac{1}{n} X_n = \frac{X_1 + \cdots + X_n}{n} = \overline{X}.$$

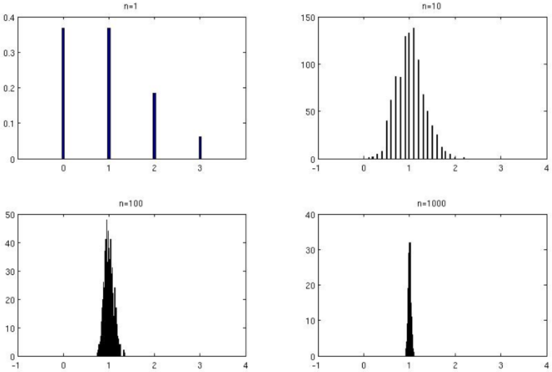
Theorem

Om $X_1, X_2, \ldots, X_n$ är oberoende s.v. med samma väntevärde μ och standardavvikelsen σ

$$E\left(\overline{X}\right) = \frac{1}{n} \sum_{i=1}^{n} \mu = n \frac{\mu}{n} = \mu$$

$Var\left(\overline{X}\right) = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \sigma^2 = n \frac{\sigma^2}{n^2} = \frac{\sigma^2}{n}$
$D\left(\overline{X}\right) = \frac{\sigma}{\sqrt{n}}$

Stora talens lag



FIGUR: Sannolikhetsfunktionen för medelvärdet $\bar{X}_n = (X_1 + \ldots X_n)/n$, där de enskilda observationerna är $Po(1)$, för $n = 1, 10, 100$ och 1000. Den diskreta fördelningen koncentreras alltmer runt värdet 1, dvs väntevärdet.

Stora talens lag

Låt $X_1, X_2, \ldots$ vara oberoende och likafördelade stokastiska variabler, var och en med väntevärdet μ, och sätt

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Då gäller, för alla $\epsilon > 0$, att

$$P\left(\mu - \epsilon < \overline{X}_n < \mu + \epsilon\right) \to 1 \text{ då } n \to \infty.$$