	Emne	Kompenedium
1	Repetisjon	
2	Kontinuerlige stokastiske variabler	[I] 1.1
3	Forventning og varians	[I] 1.2 - 1.4
4	Median og andre prosentiler	[I] 1.5 - 1.6

	Oppgaver for Forelesning 7	
	Oppgaver fra arbeidsboken	[DA] 7.1 - 7.2
	Oppgaver fra kompendiet	[I] 1.15, 1.19, 1.22 - 1.24

① Repetisjon

$f(\underline{x}) = \underline{x}^T A \underline{x} = \underline{u}^T D \underline{u} = \lambda_1 u_1^2 + \lambda_2 u_2^2 + \dots + \lambda_n u_n^2$

kvadr. form

Variabelskifte:

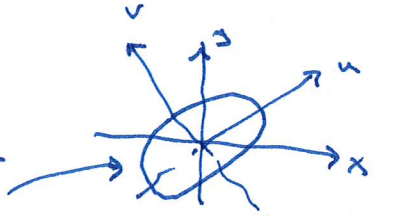
$\underline{x} = P \cdot \underline{u}$

der $P^T A P = D$

er en ortogonal diag.

$f(x,y) = 2x^2 - 2xy + 2y^2 = u^2 + 3v^2 = c$

$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ $\lambda_1 = 1$ $\lambda_2 = 3$



$\frac{u^2}{c} + \frac{v^2}{c/3} = 1$ ellipse når $c > 0$

Optimering:

$f(\underline{x}) = \underline{x}^T A \underline{x} + B\underline{x} + C$

$f'(\underline{x}) = 2A\underline{x} + B^T$

Stasjonære punkt:

$f'(\underline{x}) = 2A\underline{x} + B^T = \underline{0}$

A pos def/pos semidefn: f konvekst

A neg " neg " : f konkav

6.8. Gitt et datasett med N observasjoner av variabelen y som skal forklares og av de n forklaringsvariable $x_1, x_2, \ldots, x_n$, så ønsker vi å finne den lineære likningen

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

som gir beste tilnærming til dataene i datasettet. Vi kan sjelden finne en tilnærming uten feil, og må derfor istedet å gjøre feilen ε gitt ved

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

minst mulig. Vi antar at datasettet er gitt ved følgende tabell (der hver linje svarer til en observasjon):

	x_1	x_2	...	x_n	y
1	x_{11}	x_{21}	...	x_{n1}	y_1
2	x_{12}	x_{22}	...	x_{n2}	y_2
3	x_{13}	x_{23}	...	x_{n3}	y_3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
N	x_{1N}	x_{2N}	...	x_{nN}	y_N

Vi får det lineært likningssystem $\mathbf{y} = X\beta + \varepsilon$ der

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{21} & \ldots & x_{n1} \\ 1 & x_{12} & x_{22} & \ldots & x_{n2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1N} & x_{2N} & \ldots & x_{nN} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

Vi sier at feilen er minst mulig når $||\varepsilon||^2 = \varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_N^2$ er minst mulig, og metoden kalles derfor *minste kvadraters metode*. Vis at problemet har entydig løsning

$$\beta = (X^T X)^{-1} \cdot X^T \mathbf{y}$$

når $\det(X^T X) \neq 0$.

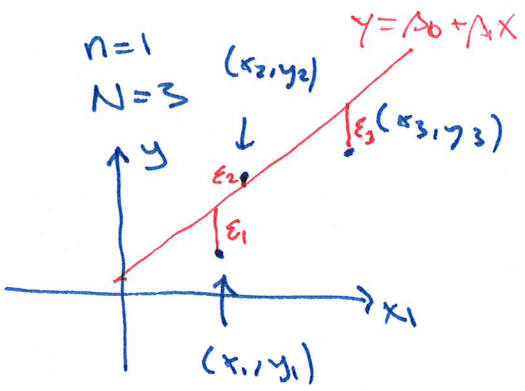
6.9. Estimer β_0, β_1 i den lineære regresjonsmodellen $Y = \beta_0 + \beta_1 x + \varepsilon$ ut fra de tre observasjonene

$$(x_1, y_1) = (0,1)$$
$$(x_2, y_2) = (1,0)$$
$$(x_3, y_3) = (1,1)$$

og beregn feilleddene e_1, e_2 og e_3.

$$y = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \quad X = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$$

[DAI] 6.8 $\quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon$



Minste kvadraters metode

Velg $\beta_0, \beta_1, \dots, \beta_n$ slik at

$\varepsilon_1^2 + \varepsilon_2^2 + \dots + \varepsilon_N^2$ blir minst mulig.

Dvs:

$$\min_{(\beta_0,\dots,\beta_n)} \varepsilon_1^2 + \varepsilon_2^2 + \dots + \varepsilon_N^2$$

	x_1	x_2	...	x_n	y
1	x_{11}	x_{21}		x_{n1}	y_1
2	x_{12}	x_{22}		x_{n2}	y_2
⋮	⋮	⋮		⋮	⋮
N					

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \dots + \beta_n x_{n1} + \varepsilon_1$$
$$y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \dots + \beta_n x_{n2} + \varepsilon_2$$
$$\vdots$$

$$\underline{y} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{21} & \dots \\ 1 & x_{12} & x_{22} & \dots \\ & & & \end{pmatrix}}_{X} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

Likn. på matriseform:

$$\underline{y} = X \cdot \underline{\beta} + \underline{\varepsilon} \quad \leftarrow \quad \underline{\varepsilon} = \underline{y} - X\underline{\beta}$$

$$\|\underline{\varepsilon}\|^2 = \underline{\varepsilon} \cdot \underline{\varepsilon} = \underline{\varepsilon}^T \underline{\varepsilon} = (\underline{y} - X\underline{\beta})^T \cdot (\underline{y} - X\underline{\beta}) = (\underline{y}^T - \underline{\beta}^T X^T)(\underline{y} - X\underline{\beta})$$

prikk-produkt; matrise-produkt

$$= \underline{y}^T \cdot \underline{y} - \underline{\beta}^T X^T \underline{y} + \underline{y}^T (-X\underline{\beta}) + \underline{\beta}^T X^T X \underline{\beta}$$

$$= \underline{\beta}^T (X^T X) \underline{\beta} - \underline{y}^T X \underline{\beta} - (\underline{\beta}^T X^T \underline{y})^T + \underline{y}^T \underline{y}$$

1×1-matrise

$1 \times (n+1) \quad (n+1) \times N \quad N \times 1$

A; B; $y^T X \beta$; C

$$\|\underline{\varepsilon}\|^2 = \underline{\beta}^T (X^T X)\underline{\beta} - 2(y^T X)\underline{\beta} + y^T y = \underline{\beta}^T A \underline{\beta} + B\underline{\beta} + C$$

$A = X^T X$, $B = -2y^T X$, $C = y^T y$

$$\min \varepsilon_1^2 + \varepsilon_2^2 + \dots + \varepsilon_N^2 \rightsquigarrow \min f(\beta) = \underline{\beta}^T (X^T X)\underline{\beta} - 2y^T X \underline{\beta} + y^T y$$
$(\beta_0, \dots, \beta_n)$

Stasjonære punkt: $f'(\underline{\beta}) = 2(X^T X)\underline{\beta} + (-2y^T X)^T$

$$= 2X^T X \underline{\beta} - 2X^T y = \underline{0} \quad | : 2$$

$$X^T X \underline{\beta} - X^T y = \underline{0}$$

$$(X^T X) \cdot \underline{\beta} = X^T \cdot y$$

$(n+1)\times(n+1)$ $\quad$ $(n+1)\times 1$

Bruker:

i) $X^T \cdot X$ er pos. semidefin.

ii) $|X^T X| \neq 0 \Rightarrow X^T X$ invertibel

$$\underline{\beta} = (X^T X)^{-1} \cdot X^T y$$

f konvekst $\Rightarrow \beta = (X^T X)^{-1} \cdot (X^T y)$ er global min for $f(\underline{\varepsilon}) = \varepsilon_1^2 + \dots + \varepsilon_N^2$

(2) Kontinuerlige Stokastiske variabler

Stokastiske forsøk:

- mengden av mulige utfall (utfallsrommet) er kjent
- det enkelte utfall er ukjent

Stokastisk variabel: X variabel med verdi som avhenger av utfallet i et stokastisk forsøk

Eks: Vi kaster to terninger

X = Summen (diskret)

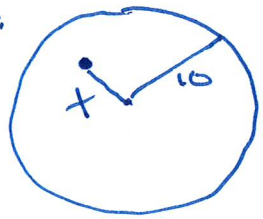
Y = max

$U = \{(1,1),(1,2),\ldots,(6,6)\}$

Mulige X-verdier: $2,\ldots,12$

Ser på tilfellet med en kontinuerlig stokastisk variabel

Eks:

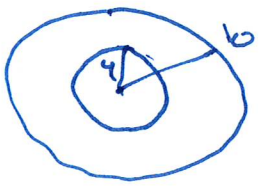


Vi kaster pil på en blink

X = avstand til sentrum

Mulige X-verdier: $[0,10]$

$$P(X \le 4) = \frac{\pi \cdot 4^2}{\pi \cdot 10^2} = \frac{16}{100}$$



$$P(X \le b) = \frac{\pi \cdot b^2}{\pi \cdot 10^2} = \frac{b^2}{100}$$

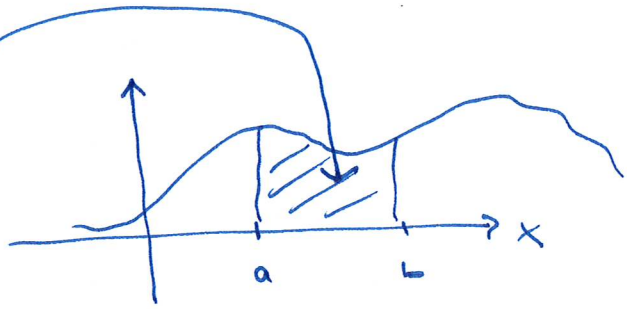
$(0 \le b \le 10)$

$F(b) = b^2/100$ $\quad F(x) = x^2/100$

$\Rightarrow f(x) = F'(x) = 2x/100 = x/50$

For en kontinuerlig stokastisk variabel finnes det en funksjon $f(x)$ som kalles sannsynlighetstetthetsen til X som har egenskapen

$$P(a \le X \le b) = \int_a^b f(x)\,dx$$

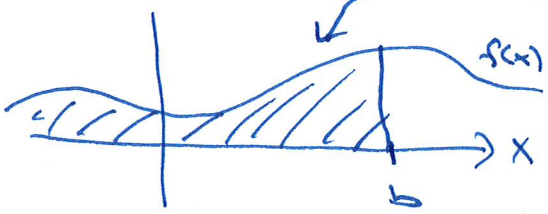


Krav til tetthetsfunksjon:

i) $f(x) \ge 0$ for alle x

ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$

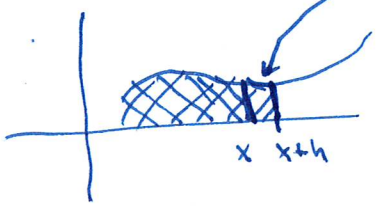
Kumulativ fordelingsfunksjon for en kontinuerlig stokastisk variabel X: $F(x) = P(X \le x)$ $F(b) = P(X \le b)$



$$F(b) = \int_{-\infty}^{b} f(x)\,dx \qquad F(x) = \int_{-\infty}^{x} f(x)\,dx$$

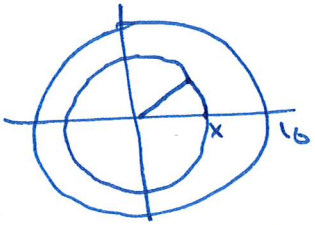
Integralregningens fundamentalsetning: $F'(x) = f(x)$

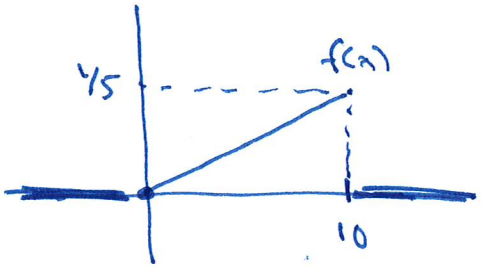
$$F'(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h}$$



$$\frac{h \cdot f(x)}{h} = f(x)$$

Eks (pilkast) $F(x) = x^2/100$ $f(x) = \frac{2x}{100} = \frac{x}{50}$





Krav til kumulativ fordelingsfunksjon:

i) F voksende

ii) $\lim_{x\to\infty} F(x) = 1$ og $\lim_{x\to-\infty} F(x) = 0$

Skrivemåte: $f(x) = f_X(x)$, $F(x) = F_X(x)$ sannsynlighetstetthet til X, kum. fordelingsfunksjon -"-

(3) Forventning og varians

Def: $E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$ forventningsverdien til X

Eks: Pilkast

$f(x) = x/50$, $0 \le x \le 10$

$$E(X) = \int_0^{10} x \cdot f(x)\,dx = \int_0^{10} x \cdot \frac{x}{50}\,dx = \left[\frac{1}{50}\cdot\frac{1}{3}x^3\right]_0^{10}$$

$$= \frac{1}{150}\left(10^3 - 0^3\right) = \frac{1000}{150} = \frac{100}{15} = \frac{20}{3} \simeq \underline{6.67}$$

Hvis $h(X)$ er en avledet stokastisk variabel:

$h(X) = X^2$ eller $h(X) = X^3 - X$

Forventningsverdien til $h(X)$:

$$E(h(X)) = \int_{-\infty}^{\infty} h(x) \cdot f(x)\,dx$$

Eks: Pilkast

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx = \int_0^{10} x^2 \cdot \tfrac{1}{50}\,dx = \left[\tfrac{1}{50} \cdot \tfrac{1}{4}x^4\right]_0^{10}$$

$$= \frac{1}{200}\left(10^4 - 0^4\right) = \frac{10000}{200} = 50$$

$$E(X)^2 = (20/3)^2 = \frac{400}{9} \neq E(X^2) = 50$$

Def: Varians til X, X stokastisk variabel med $E(X) = \mu$

$$Var(X) = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\,dx$$

$$= \int_{-\infty}^{\infty} (x^2 - 2\mu x + \mu^2) f(x)\,dx = \int_{-\infty}^{\infty} x^2 f(x)\,dx - 2\mu \int_{-\infty}^{\infty} x f(x)\,dx$$

$$+ \mu^2 \int_{-\infty}^{\infty} f(x)\,dx = E(X^2) - 2\mu E(X) + \mu^2 \cdot 1$$

$$= E(X^2) - 2\mu \cdot \mu + \mu^2 = E(X^2) - \mu^2$$

$$\boxed{Var(X) = E(X^2) - E(X)^2}$$

Eks: Pilkast

$E(X) = \frac{20}{3}$ $E(X^2) = 50$ $\Rightarrow$ $Var(X) = 50 - \left(\frac{20}{3}\right)^2 = \frac{50 \cdot 9 - 400}{9}$

$= 50/9$

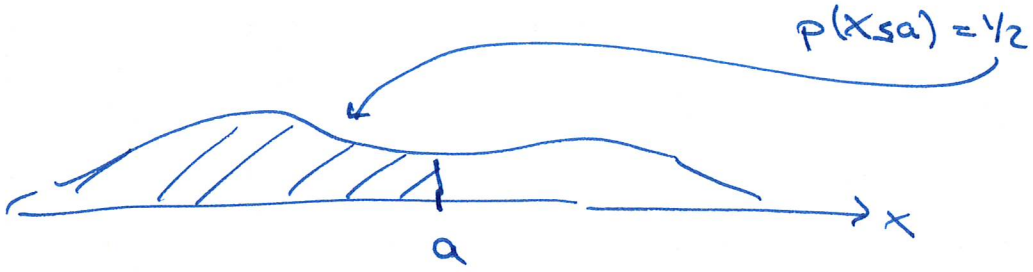
Generelt: $\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx \geq 0$

$= E(X^2) - E(X)^2 \geq 0$

$\Rightarrow$ Standardavvikelt til X: $\sigma_X = \sqrt{\mathrm{Var}(X)}$

(4) Median.

Hvis X er en kontinuerlig stokastisk variabel med kumulativ fordelingsfunksjon $F(x)$, så er medianen til X det tallet a slik at $F(a) = 1/2$.



$p(X \leq a) = 1/2$

$F(a) = 1/2 \iff \int_{-\infty}^{a} f(x)\,dx = 1/2$

Eks: Pilkast

$F(b) = b^2/100 = 1/2$

$b^2 = 100 \cdot 1/2 = 50$

$b = \sqrt{50}$

median: $\sqrt{50} \simeq 7$

$E(X)$: $20/3 \simeq 6.67$

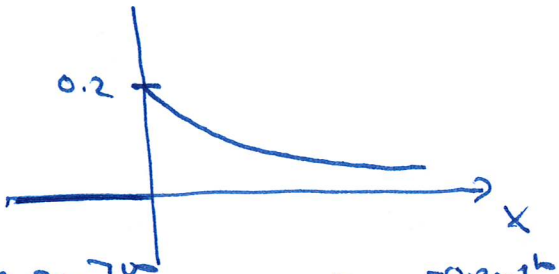
90% prosentil: $F(b) = b^2/100 = 0.9$

$b^2 = 90$ $\quad b = \sqrt{90}$

Eks: X eksponensialfordelt med $f(x) = 0.2 \cdot e^{-0.2x}$, $x \geq 0$

Sjekk: i) $f(x) \geq 0$ ok.

ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$ ok.



$$\int_0^{\infty} 0.2e^{-0.2x}\,dx = \left[-e^{-0.2x}\right]_0^{\infty} = \lim_{b\to\infty}\left[-e^{-0.2x}\right]_0^{b}$$

$$= \lim_{b\to\infty}\left(-e^{-0.2b} + 1\right) = 1 \quad \text{ok}$$

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx = \int_0^{\infty} x \cdot 0.2e^{-0.2x}\,dx$$

$$\int x e^{-0.2x}\,dx = \left[\begin{array}{l} u = \frac{e^{-0.2x}}{-0.2} \quad v = x \\ u' = e^{-0.2x} \quad v' = 1 \end{array}\right] = \frac{1}{-0.2}e^{-0.2x} \cdot x - \int \frac{1}{-0.2}e^{-0.2x} \cdot 1\,dx$$

$$= -5xe^{-0.2x} + 5\int e^{-0.2x}\,dx$$

$$= -5xe^{-0.2x} + 5\cdot\left(\frac{1}{-0.2}\right)e^{-0.2x} + C$$

$$= -5xe^{-0.2x} - 25e^{-0.2x} + C$$

$$\overset{0.2\cdot}{=} \lim_{b\to\infty}\left[-5xe^{-0.2x} - 25e^{-0.2x}\right]_0^{b} \overset{0.2\cdot}{=} \lim_{b\to\infty}\left(-5be^{-0.2b} - 25e^{-0.2b} + 25\right)$$

$$\overset{0.2\cdot}{=} \left(\lim_{b\to\infty}\frac{-5b}{e^{0.2b}} + \lim_{b\to\infty}\frac{-25}{e^{0.2b}} + 25\right) = 25 \cdot 0.2 = 5$$

($\lim_{b\to\infty}\frac{-5b}{e^{0.2b}} = 0$, $\lim_{b\to\infty}\frac{-25}{e^{0.2b}} = 0$)