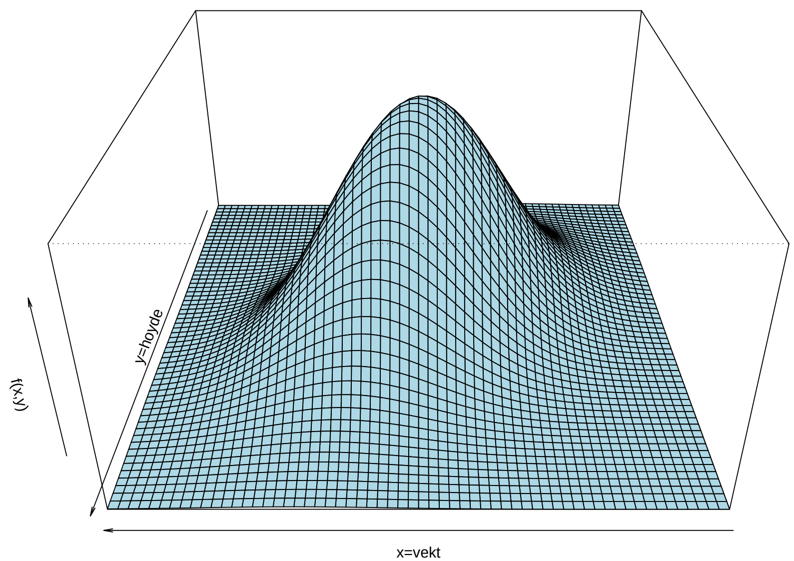
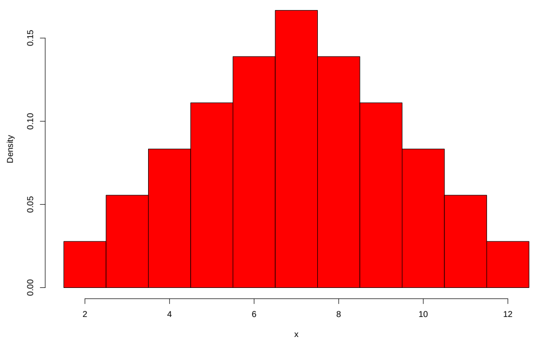
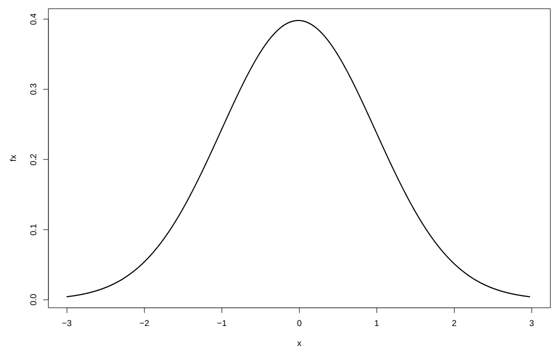
Kapittel 3: Stokastiske variable og sannsynlighetsfordelinger

TMA4240 Statistikk (F2 og E7)

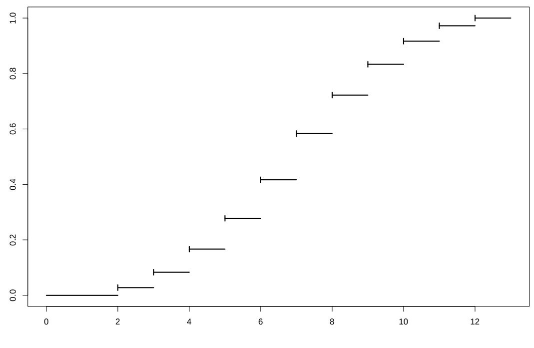
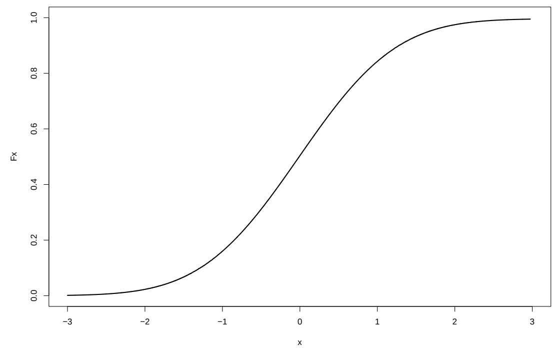
3.4: Foreleses mandag 30.august



Oppsummering 3.1-3.3

Diskret stokastisk variabel	Kontinuerlig stokastisk variabel
X Mulige verdier x: Endelig eller tellbart mange	X Mulige verdier x: Intervall eller hele R
Eksempel: $\{0, 1, \ldots, n\}$	Eksempel: $[0, 1]$ eller $[0, \infty)$
Sannsynlighetsfordeling: $f(x) = P(X = x)$ for alle mulige x	Sannsynlighetsfordeling: $f(x)$ definert for alle reelle x ved $P(a < X < b) = \int_a^b f(x)dx$
	

Oppsummering 3.1-3.3, forts.

Diskret stokastisk variabel	**Kontinuerlig stokastisk variabel**
Kumulativ fordeling: $F(x) = P(X \leq x) = \sum_{t \leq x} f(t)$ definert for alle reelle x	Kumulativ fordeling: $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$ definert for alle reelle x
	
$P(a < X \leq b) = \sum_{x \in (a,b]} f(t)$ $= F(b) - F(a)$	$P(a < X \leq b) = \int_a^b f(x)dx$ $= F(b) - F(a)$
Hvis mulige verdier er heltall: $f(x) = F(x) - F(x-1)$	Hvis f er kontinuerlig i x: $f(x) = F'(x)$

Flere stokastiske variabler?

- Vi har til nå kun sett på en stokastisk variabel om gangen (univariate fordelinger). Hva skjer om vi ser på to eller flere stokastiske variabler samtidig (bivariate-/multivariate fordelinger)?
- Bør vi se på dem hver for seg, eller mister vi informasjon hvis vi ikke ser på dem sammen?

Simultanfordeling, $f(x, y)$

DEF 3.8: Funksjonen $f(x, y)$, er simultanfordelingen til to diskrete stokastiske variable X og Y, dersom

1. $f(x, y) \geq 0$ for alle (x, y)
2. $\sum_x \sum_y f(x, y) = 1$
3. $P(X = x \cap Y = y) = P(X = x, Y = y) = f(x, y)$ (punktsannsynlighet)

For enhver region A i xy-rommet så er $P[(X, Y) \in A] = \sum_A \sum f(x, y)$.

DEF 3.9: Funksjonen $f(x, y)$, er simultan sannsynlighetstetthet til to kontinuerlige stokastiske variable X og Y, dersom

1. $f(x, y) \geq 0$ for alle (x, y)
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$
3. $P[(X, Y) \in A] = \int_A \int f(x, y) dx dy$ for enhver region A i (x, y)-planet.

Kast med to terninger

- Definerer:
  - X = maksimum antall øyne
  - Y = absoluttverdi av differanse i antall øyne
- Utfallsrom og verdier av X og (Y) (i parentes)

	1	**2**	**3**	**4**	**5**	**6**
1	1 (0)	2 (1)	3 (2)	4 (3)	5 (4)	6 (5)
2	2 (1)	2 (0)	3 (1)	4 (2)	5 (3)	6 (4)
3	3 (2)	3 (1)	3 (0)	4 (1)	5 (2)	6 (3)
4	4 (3)	4 (2)	4 (1)	4 (0)	5 (1)	6 (2)
5	5 (4)	5 (3)	5 (2)	5 (1)	5 (0)	6 (1)
6	6 (5)	6 (4)	6 (3)	6 (2)	6 (1)	6 (0)

Kast med to terninger (forts.)

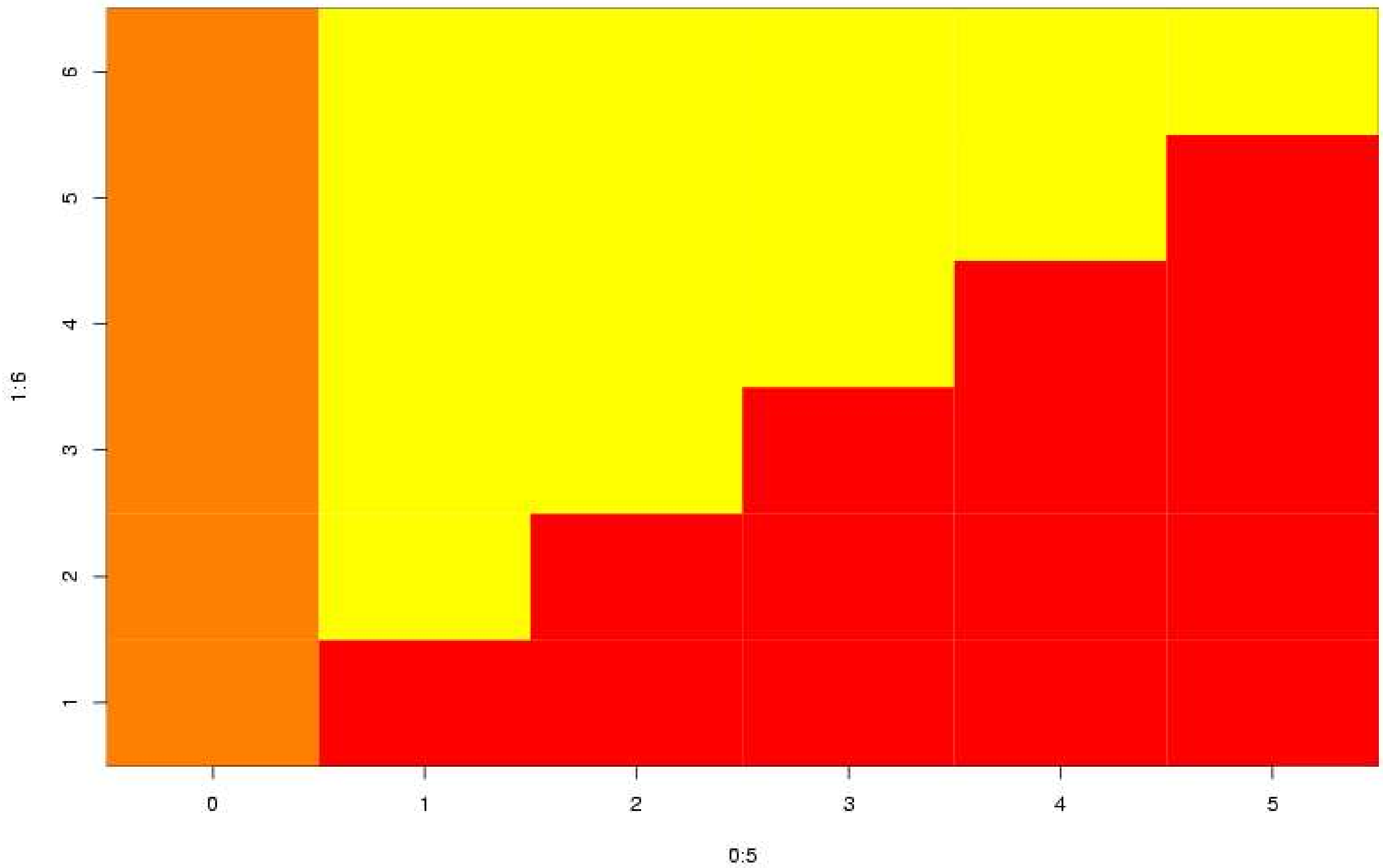
- Utfallsrom og verdier av X og (Y) (i parentes)

	1	2	3	4	5	6
1	1 (0)	2 (1)	3 (2)	4 (3)	5 (4)	6 (5)
2	2 (1)	2 (0)	3 (1)	4 (2)	5 (3)	6 (4)
3	3 (2)	3 (1)	3 (0)	4 (1)	5 (2)	6 (3)
4	4 (3)	4 (2)	4 (1)	4 (0)	5 (1)	6 (2)
5	5 (4)	5 (3)	5 (2)	5 (1)	5 (0)	6 (1)
6	6 (5)	6 (4)	6 (3)	6 (2)	6 (1)	6 (0)

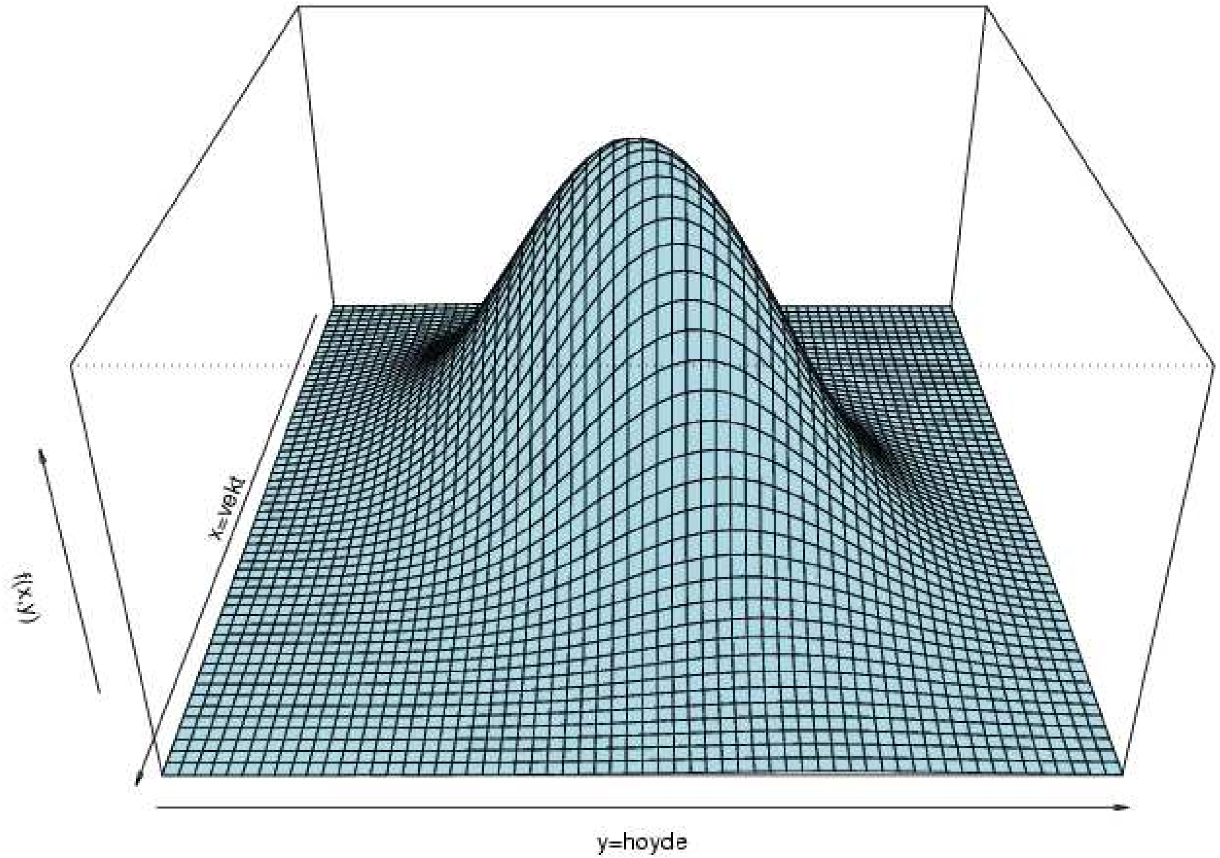
- Simultan sannsynlighetsfordeling

	1	2	3	4	5	6
0	1/36	1/36	1/36	1/36	1/36	1/36
1	0	2/36	2/36	2/36	2/36	2/36
2	0	0	2/36	2/36	2/36	2/36
3	0	0	0	2/36	2/36	2/36
4	0	0	0	0	2/36	2/36
5	0	0	0	0	0	2/36

Simultanfordeling

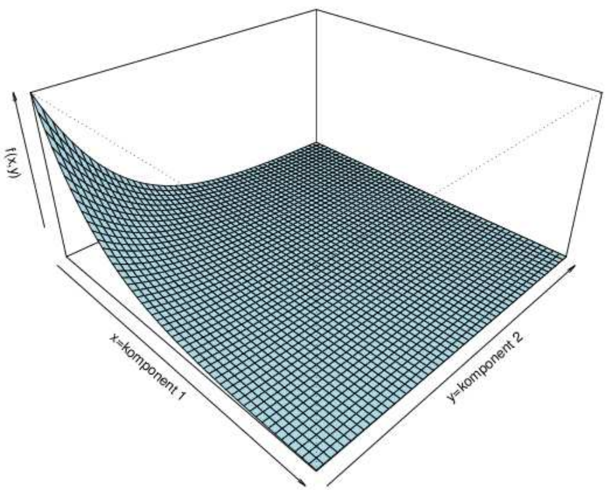


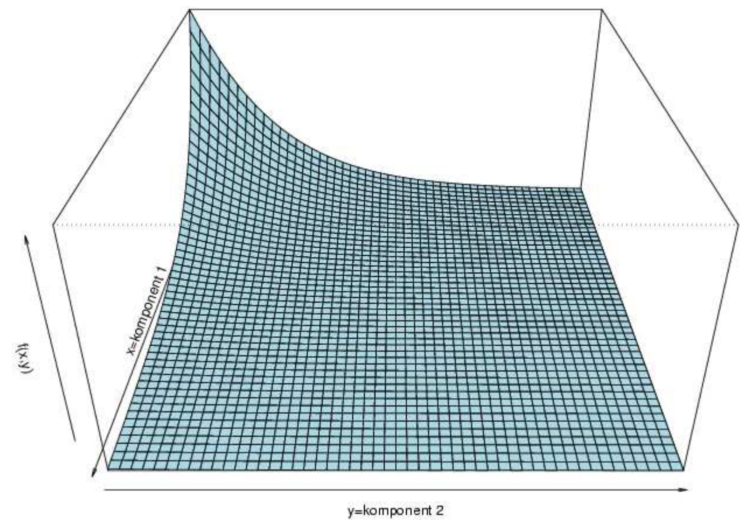
Høyde og vekt av studenter



Elektroniske komponenter

- Levetiden til to elektriske komponenter, X og Y:
- $f(x,y) = e^{-(x+y)}$ for $x > 0$ og $y > 0$ og 0 ellers.





Marginalfordelinger

DEF 3.10: Marginalfordelingen for X alene og for Y alene er

- for det diskrete tilfellet:

$$g(x) = \sum_y f(x,y) \quad \text{og} \quad h(y) = \sum_x f(x,y)$$

- og for det kontinuerlinge tilfellet:

$$g(x) = \int_{-\infty}^{\infty} f(x,y)dy \quad \text{og} \quad h(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

Kast med to terninger (igjen)

- X = maksimum antall øyne
- Y = absoluttverdi av differanse i antall øyne
- Simultan sannsynlighetsfordeling

	1	**2**	**3**	**4**	**5**	**6**	$h(y)$
0	1/36	1/36	1/36	1/36	1/36	1/36	6/36
1	0	2/36	2/36	2/36	2/36	2/36	10/36
2	0	0	2/36	2/36	2/36	2/36	8/36
3	0	0	0	2/36	2/36	2/36	6/36
4	0	0	0	0	2/36	2/36	4/36
5	0	0	0	0	0	2/36	2/36
$g(x)$	1/36	3/36	5/36	7/36	9/36	11/36	36/36

Betingede fordelinger

DEF 3.11: La X og Y være to stokastiske variable, diskrete eller kontinuerlige.

- Den betingede fordelingen for den stokastiske variablen Y gitt at $X = x$ er

$$f(y|x) = \frac{f(x,y)}{g(x)}, \qquad g(x) > 0$$

- Den betingede fordelingen for den stokastiske variablen X gitt at $Y = y$ er

$$f(x|y) = \frac{f(x,y)}{h(y)}, \qquad h(y) > 0$$

Uavhengighet

DEF 3.12: La X og Y være to stokastiske variable, diskrete eller kontinuerlige, med simultan sannsynlighetsfordeling $f(x, y)$, og marginale fordelinger $g(x)$ og $h(y)$. Da er X og Y *uavhengige* hvis og bare hvis

$$f(x, y) = g(x) \cdot h(y)$$

for alle (x, y).

DEF 3.13: La $X_1, X_2, ..., X_n$ være n stokastiske variable, diskrete eller kontinuerlige, med simultan sannsynlighetsfordeling $f(x_1, x_2, ..., x_n)$, og marginale fordelinger $f_1(x_1), f_2(x_2), ..., f_n(x_n)$. Da er $X_1, X_2, ..., X_n$ innbyrdes *uavhengige* hvis og bare hvis

$$f(x_1, x_2, ..., x_n) = f_1(x_1) \cdot f_2(x_2) \cdots f_n(x_n)$$

for alle $(x_1, x_2, ..., x_n)$.

Togforsinkelsen, forts.

- La X (minutter) betegne togets forsinkelse på en tilfeldig valgt hverdag. Vi antar at X er en stokastisk variabel med sannsynlighetstetthet

$$g(x) = \begin{cases} 4xe^{-2x} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

der $4 > 0$ er en konstant.

- La Y (minutter) være den tiden toget står på stasjonen. Oppholdstiden Y vil være influert av forsinkelsen, og vi antar at den betingede sannsynlighetstetthet $f(y|x)$ for Y, gitt at forsinkelsen X er lik x (> 0), er gitt ved

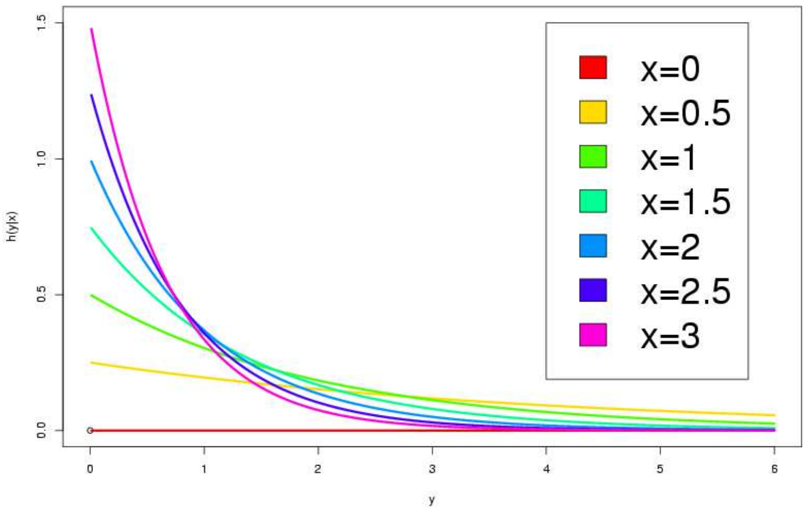
$$f(y|x) = \begin{cases} (x/2)\, e^{-xy/2} & \text{for } y > 0 \\ 0 & \text{for } y \leq 0 \end{cases}$$

iii) Sett opp simultantettheten $f(x, y)$ for X og Y.

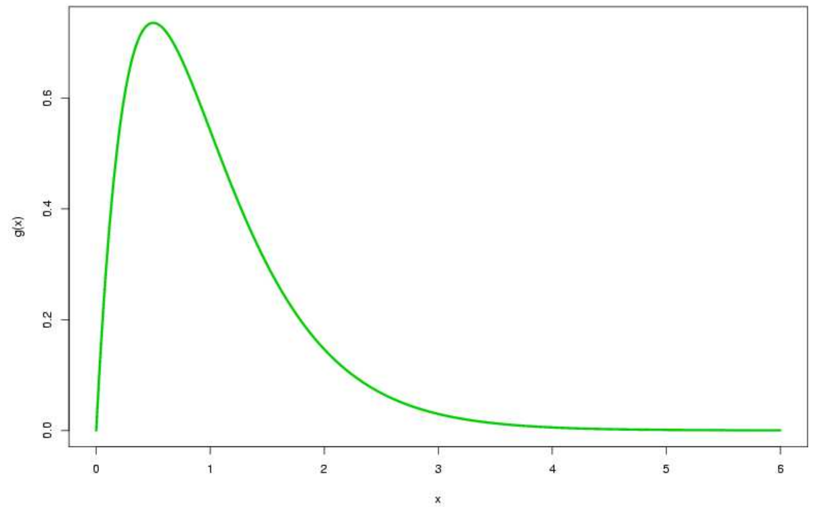
iv) Finn sannsynlighetstettheten $h(y)$ for oppholdstiden Y.

Togforsinkelsen: fordelinger

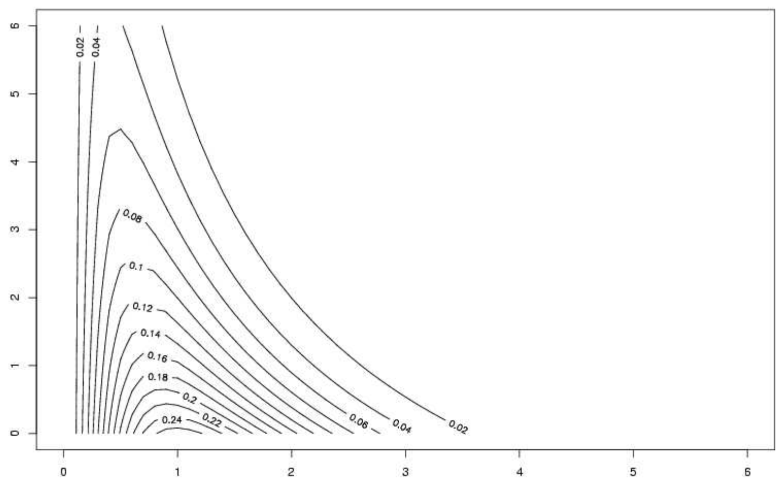
$f(y|x)$



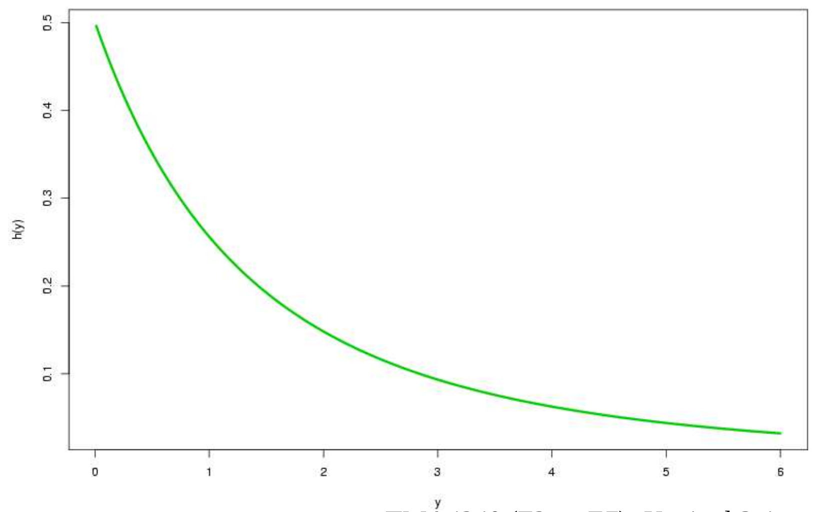
$g(x)$



$f(x,y)$

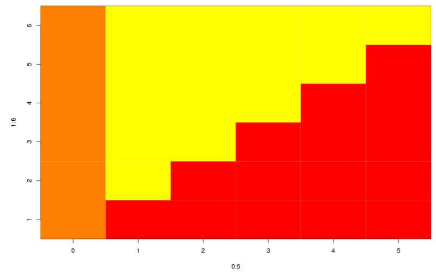
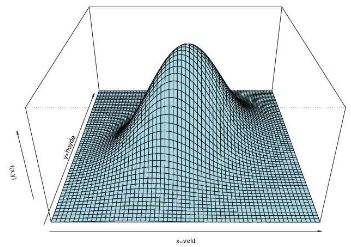


$h(y)$



Oppsummering 3.4

- Funksjonen $f(x, y)$, er simultan sannsynlighetsfordeling for X og Y.

- Marginalfordelinger:

$$
\begin{aligned}
g(x) &= \textstyle\sum_y f(x,y) \quad \text{og} \quad h(y) = \sum_x f(x,y) && \text{diskret} \\
g(x) &= \textstyle\int_{-\infty}^{\infty} f(x,y)dy \quad \text{og} \quad h(y) = \int_{-\infty}^{\infty} f(x,y)dx && \text{kontinuerlig}
\end{aligned}
$$

- Betingede fordelinger:

$$
\begin{aligned}
f(y|x) &= f(x,y)/g(x), && g(x) > 0 \\
f(x|y) &= f(x,y)/h(y), && h(y) > 0
\end{aligned}
$$

- Uavhengighet: X og Y *uavhengige* hvis og bare hvis

$$f(x,y) = g(x) \cdot h(y) \quad \text{for alle } (x,y).$$

Eksempel: Kabel

- Vi ser på en kabel med lengde L. Det er typer feil på kabelen, A og B. Vi vet at det er en A feil og en B feil på kabelen. Vi definerer to stokastiske variabler
  X - posisjon for feil A
  Y - posisjon for feil B
- Anta at X og Y er uniformt fordelt på kabelen, og at feilene oppstår uavhengig av hverandre, dvs.:

$$X \sim g(x) = \begin{cases} \frac{1}{L} \; ; \; 0 < x < L \\ 0 \; ; \; ellers \end{cases}$$

$$Y \sim h(y) = \begin{cases} \frac{1}{L} \; ; \; 0 < y < L \\ 0 \; ; \; ellers \end{cases}$$

$$(X, Y) \sim f(x, y) = g(x)h(y) = \begin{cases} \frac{1}{L^2} \; ; \; 0 < x < L; 0 < y < L \\ 0 \; ; \; ellers \end{cases}$$

- Hva er:
  i) Sannsynligheten for at begge feil er siste halvdel av kabelen?
  ii)Sannsynligheten for at feil B skjer foran feil A?