Statistikk

Jo Eidsvik

Matematiske fag, NTNU

Simultane fordelinger

Stokastiske variable X og Y med endelige eller tellbare utfallsrom.
Punktsannsynlighet $f(x, y) = P(X = x, Y = y)$.
KRAV:

$$\sum_x \sum_y f(x, y) = 1, \qquad f(x, y) \geq 0.$$

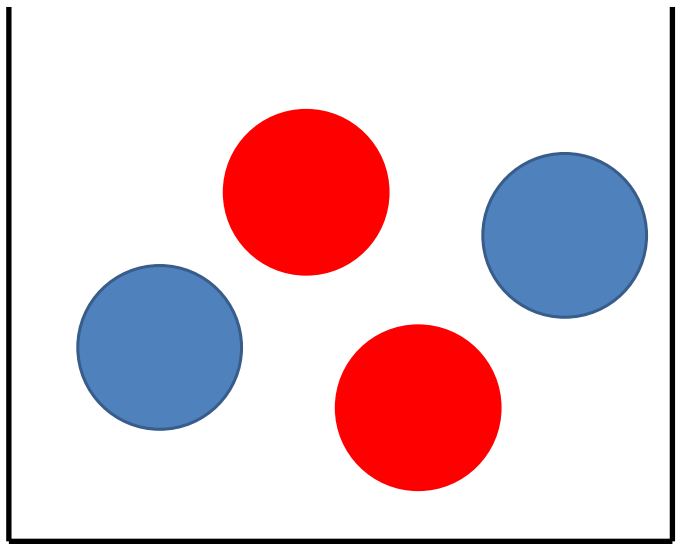
Kontinuerlige fordelinger

Stokastiske variable X og Y med kontinuerlige utfallsrom.
Sannsynlighetstetthet $f(x, y)$.
KRAV:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1, \qquad f(x,y) \geq 0.$$

Her er $P(a < X < b, c < Y < d) = \int_a^b \int_c^d f(x,y)dxdy$.

Eksempel: Trekning **uten** tilbakelegging



En urne har 2 røde og 2 blå baller.

$X =$ antall røde kuler i første trekning, $x \in \{0, 1\}$.
$Y =$ antall røde kuler i andre trekning, $y \in \{0, 1\}$.

Eksempel: Trekning **uten** tilbakelegging

Hva er sannsynlighet for at $X = 1$ og $Y = 1$?

$$P(X = 1, Y = 1) = \frac{2}{4}\frac{1}{3} = 1/6$$

Annen regnemetode
(gunstige: ingen av blå og to kuler av rød, mulige: 2 av 4):

$$P(X = 1, Y = 1) = \frac{\binom{2}{0}\binom{2}{2}}{\binom{4}{2}} = 1/6$$

Eksempel: Trekning **uten** tilbakelegging

Tilsvarende

$$P(X = 0, Y = 0) = \frac{\binom{2}{2}\binom{2}{0}}{\binom{4}{2}} = 1/6$$

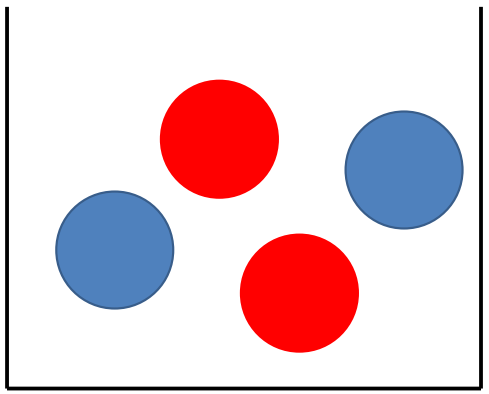
$$P(X = 1, Y = 0) = \frac{2}{4}\frac{2}{3} = 1/3$$

$$P(X = 0, Y = 1) = \frac{2}{4}\frac{2}{3} = 1/3$$

Marginal sannsynlighet

Diskret tilfelle:
$g(x) = \sum_y f(x, y)$, $h(y) = \sum_x f(x, y)$
Kontinuerlig tilfelle:
$g(x) = \int f(x, y)dy$, $h(y) = \int f(x, y)dx$

Eksempel: Marginal sannsynlighet



En urne har 2 røde og 2 blå baller.

$X =$ antall røde kuler i første trekning, $x \in \{0, 1\}$.
$Y =$ antall røde kuler i andre trekning, $y \in \{0, 1\}$.
$P(X = 1) = g(1) = \sum_y f(1, y) = 1/3 + 1/6 = 1/2,$
$P(Y = 1) = h(1) = \sum_x f(x, 1) = 1/6 + 1/3 = 1/2$

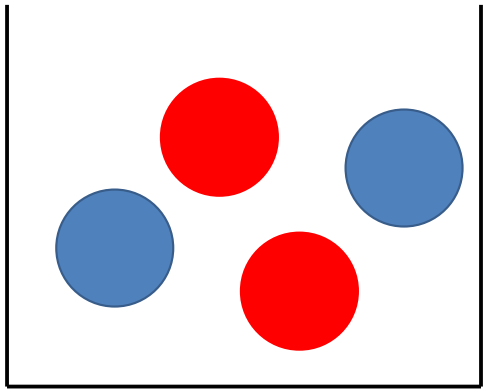
Betinget sannsynlighet

Diskret og kontinuerlig tilfelle:

$$f(x|y) = P(X = x|Y = y) = \frac{f(x,y)}{h(y)}$$

$$f(y|x) = P(Y = y|X = x) = \frac{f(x,y)}{g(x)}$$

Eksempel: Betinget sannsynlighet



En urne har 2 røde og 2 blå baller.

$X =$ antall røde kuler i første trekning, $x \in \{0, 1\}$.
$Y =$ antall røde kuler i andre trekning, $y \in \{0, 1\}$.
$f(1|X = 1) = \frac{f(1,1)}{g(1)} = \frac{1/6}{1/2} = 1/3$, $f(1|X = 0) = \frac{f(0,1)}{g(0)} = \frac{1/3}{1/2} = 2/3$,

Eksempel: Machine learning

X = evaluering / score på hotell A i Oslo, $x \in \{1, 2, 3\}$.
Y = evaluering / score på hotell B i Trondheim, $x \in \{1, 2, 3\}$.
Simultan punktsannsynlighet fra data (x i rekker, y i kolonner):

	1	2	3
1	0.1	0.05	0.05
2	0.1	0.05	0.1
3	0.05	0.15	0.35

Skal vi anbefale hotell B for en som har gitt score '3' på hotell A?
$f(3|X = 3) = \frac{f(3,3)}{g(3)} = \frac{0.35}{0.55} = 0.64$. $h(3) = 0.5$.
Kanskje ikke tydelig nok at han/hun vil like anbefaling.

Uavhengighet

Stokastiske variable X og Y er **uavhengige** dersom:

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

Generell formel er

$$f(x, y) = g(x)f(y|x) = h(y)f(x|y)$$

Dersom de er uavhengige så er $f(x|y) = g(x)$. Utfall y betyr ingenting.
Dersom de er uavhengige så er $f(y|x) = h(y)$. Utfall x betyr ingenting.

Generalisering til mange variable

Stokastiske variable $X_1, \ldots, X_n$ er **uavhengige** dersom:

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\ldots f(x_n)$$

De marginale punktsannsynlighetene eller sannsynlighetstetthetene ganges sammen!

Generell formel er

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2|x_1)\ldots f(x_n|x_{n-1}, \ldots, x_2, x_1)$$

Dersom de er uavhengige, så betyr 'historien' ingenting.