TMA4245 Statistikk
Vår 2017

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Anbefalt øving 2
Løsningsskisse

Oppgave 1 En kartong inneholder 10 pakker hvorav 2 er undervektige. Vi skal trekke pakker tilfeldig inntil de to med feil vekt er funnet. Definer:

I_i: pakke trukket som nr i har riktig vekt
I_i': pakke trukket som nr i er undervektig.

a) Ved første trekning er det 10 like sannsynlige utfall, hvorav 2 av utfallene er å trekke en undervektig pakke. Dette gir

$$P(I_1') = \frac{\#\text{gunstige}}{\#\text{mulige}} = \frac{2}{10} = 0.20.$$

b) Dersom den første pakken som trekkes har riktig vekt, gjenstår det to undervektige pakker som kan trekkes. Antallet like sannsynlige utfall er nå 9, da en pakke er trukket.

$$P(I_2'|I_1) = \frac{\#\text{gunstige}}{\#\text{mulige}} = \frac{2}{9} = 0.22.$$

c) Da de to utfallene I_1 og I_1' utgjør en oppdeling av utfallsrommet, kan vi benytte setningen om total sannsynlighet, slik at

$$\begin{aligned}
P(I_2') &= P(I_2'|I_1)\,P(I_1) + P(I_2'|I_1')\,P(I_1') \\
&= \frac{2}{9} \times \frac{8}{10} + \frac{1}{9} \times \frac{2}{10} \\
&= 0.2.
\end{aligned}$$

Alternativt:
Fordi en ikke vet noe mer om den første pakken, vet en like mye om den andre som den første.

$$P(I_2') = P(I_1') = \frac{\#\text{gunstige}}{\#\text{mulige}} = \frac{2}{10} = 0.20.$$

d) Det er finnes to hendelser som gjør at de to undervektige pakkene ikke er funnet når en har trukket fire pakker. En har da enten trukket en undervektig og tre pakker med riktig vekt, eller ingen undervektige og fire pakker med riktig vekt.

$$p = \frac{\#\text{gunstige}}{\#\text{mulige}} = \frac{\binom{2}{1} \times \binom{8}{3} + \binom{8}{4}}{\binom{10}{4}} = \frac{91}{105} = 0.867$$

e) La
X : antall pakker en må kontrollveie for å finne de to med feil.

Vi har at $P(X = x)$ er sannsynligheten for at en må trekke x pakker for å finne de to undervektige. Vi finner, for hver x, $P(X = x)$ ved å se på antall gunstige over antall mulige plasseringer av de to undervektige pakkene i trekkrekkefølgen.

Antall *mulige* plasseringer av de undervektige pakkene er $\binom{10}{2} = 45$. Antall *gunstige* plasseringer av de to undervektige pakkene vil variere med x. Vi tar for oss økende verdier av x for å finne et mønster.

Vi har at $P(X = 1) = \frac{0}{45}$, da det ikke er mulig å kjenne plasseringen til de to undervektige etter bare ett trekk.

Videre har vi at for $X = 2$ trekk vil det finnes 1 gunstig plassering av de to pakkene som vil føre til at en har oppdaget disse etter bare to trekk (nemlig at de er plassert helt først i trekkrekkefølgen).

For $X = 3$ vil det finnes 2 gunstige plasseringer; posisjon 1 og 3, eller posisjon 2 og 3 i rekkeølgen (merk at plassering av undervektige pakker på posisjon 1 og 2, og en normalvektig pakke på posisjon 3 ville ført til $X = 2$ og teller dermed ikke som et gunstig utfall for dette tilfellet).

Vi ser nå et mønster; vi må for en gitt x finne den andre undervektige pakken på posisjon x, mens den første kan ta enhver tidligere plassering. Det gir $(x - 1)$ gunstige plasseringer av de to undervektige pakkene.

Dersom ingen undervektig pakke er funnet etter 8 trekninger vet vi at de to siste pakkene er undervektige. For tilfellet $X = 8$ har vi dermed $(x - 1) + 1 = 8$ gunstige utfall, nemlig at de to undervektige er nummer 8 og et hvilket nummer mindre enn 8 i trekkrekkefølgen, i tillegg til det gunstige utfallet at de to undervektige er de to siste pakkene i trekkrekkefølgen.

For tilfellet $X = 9$ har vi $x - 1 = 8$ gunstige utfall på samme måte som tidligere. I tillegg vet vi at dersom kun en undervektig pakke er funnet i løpet av de 9 første veiingene vil den siste pakken være undervektig, og vi trenger dermed kun å veie 9 pakker. Det gir enda $x - 1 = 8$ gunstige utfall, og totalt sett har vi for $X = 9$, $(x - 1) + (x - 1) = 16$ gunstige utfall.

Sannsynligheten for at $X = 10$ er 0, da det aldri vil være nødvending å veie 10 pakker for å finne de to undervektige.

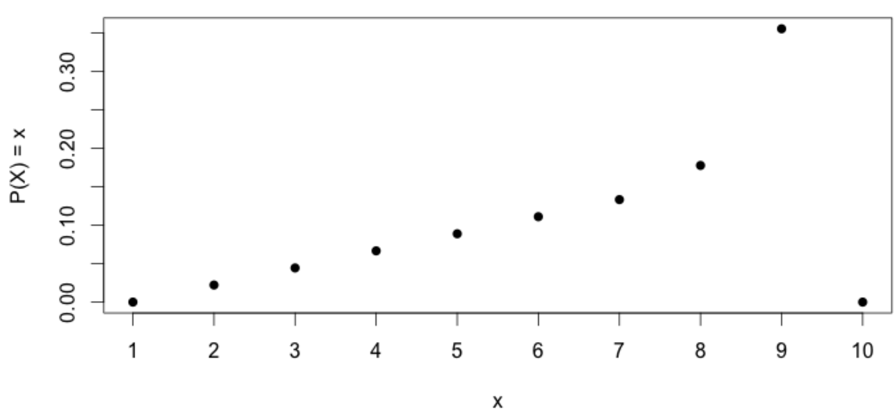
Dette gir

$$P(X=x)=\begin{cases}\frac{x-1}{45} & x=1,...,7\\ \frac{8}{45} & x=8\\ \frac{16}{45} & x=9\\ 0 & x=10\end{cases}.$$

Sannsynlighetsfordelingen til antallet pakker som må kontrollveies kan eventuelt beregnes vha multiplikasjonsmetoden:

$$\begin{aligned}
P(X=1) &= 0\\
P(X=2) &= P(I_1', I_2') = P(I_1')\,P(I_2'|I_1') = \frac{2}{10}\frac{1}{9} = \frac{1}{45}\\
P(X=3) &= P(I_1', I_2, I_3') + P(I_1, I_2', I_3')\\
&= P(P(I_1')\,P(I_2|I_1')\,P(I_3'|I_1', I_2) + P(P(I_1)\,P(I_2'|I_1)\,P(I_3'|I_1, I_2')\\
&= \frac{2}{10}\frac{8}{9}\frac{1}{8} + \frac{8}{10}\frac{2}{9}\frac{1}{8} = \frac{2}{45}\\
&\vdots
\end{aligned}$$

Figur 1 viser punktsannsynlighetene.

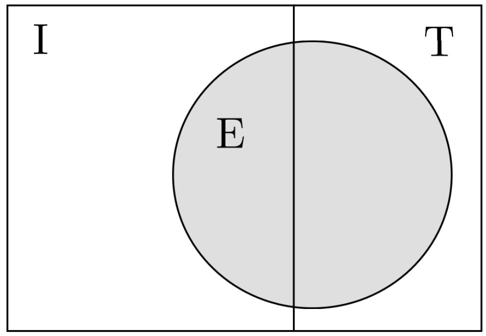


Figur 1: Punktsannsynligheter

Oppgave 2

For en tilfeldig valgt person som turisten støter på, definerer vi de tre hendelsene

- I: personen er innfødt,



Figur 2: Venndiagram for hendelsene I, T og E.

- T: personen er turist,
- E: personen snakker engelsk.

Opplysningene i oppgaveteksten kan da formuleres som følger.

- Hver tiende innfødte snakker engelsk: $P(\mathrm{E}|\mathrm{I}) = 1/10$,
- hver femte person han møter er turist: $P(\mathrm{T}) = 1/5$,
- annenhver turist snakker engelsk: $P(\mathrm{E}|\mathrm{T}) = 1/2$.

a) Vi antar at alle personene turisten møter enten er innfødte, eller er turister selv. Da er I og T komplementære hendelser, slik at

$$P(\mathrm{I}) + P(\mathrm{T}) = 1 \quad \text{og} \quad P(\mathrm{I}) = 1 - P(\mathrm{T}) = 1 - \frac{1}{5} = \frac{4}{5}.$$

I venndiagrammet i figur 2 er dette illustrert ved å dele opp rektangelet som representerer hele utfallsrommet i to deler. Siden vi har engelsktalende både blant de innfødte og blant turistene, plasserer vi regionen som representerer E slik at den overlapper både I og T.

b) Sannsynligheten for at en tilfeldig person turisten møter er engelsktalende, er gitt ved loven om total sannsynlighet,

$$\begin{aligned} P(\mathrm{E}) &= P(\mathrm{E} \cap \mathrm{I}) + P(\mathrm{E} \cap \mathrm{T}) \\ &= P(\mathrm{E}|\mathrm{I})P(\mathrm{I}) + P(\mathrm{E}|\mathrm{T})P(\mathrm{T}) \\ &= \frac{1}{10} \cdot \frac{4}{5} + \frac{1}{2} \cdot \frac{1}{5} = \underline{\underline{\frac{9}{50}}}. \end{aligned}$$

c) For å finne den betingede sannsynligheten for at en person er innfødt, gitt at vedkommende snakker engelsk, bruker vi definisjonen av betinget sannsynlighet, samt sannsynligheten $P(\mathrm{E})$ fra **b)**,

$$P(\mathrm{I}|\mathrm{E}) = \frac{P(\mathrm{I} \cap \mathrm{E})}{P(\mathrm{E})} = \frac{P(\mathrm{E}|\mathrm{I})P(\mathrm{I})}{P(\mathrm{E})} = \frac{\frac{1}{10} \cdot \frac{4}{5}}{\frac{9}{50}} = \underline{\underline{\frac{4}{9}}}.$$

Oppgave 3 Undervannssøk (Bayesiansk oppdatering)

a) For å beregne verdien av apriorifordelingen i hver gridcelle, kan vi bruke en dobbel for-løkke som itererer over rader og kolonner.

```
clear; close all; clc

nx = 12;
ny = 12;

pr = NaN(nx,ny);
z = NaN(nx,ny);
for i = 1:nx
    for j = 1:ny
        pr(i,j) = exp(-0.5*((i-(nx+1)/2)^2 + (j-(ny+1)/2)^2)/8);
    end
end
pr = pr./sum(sum(pr));
pr0 = pr;
```

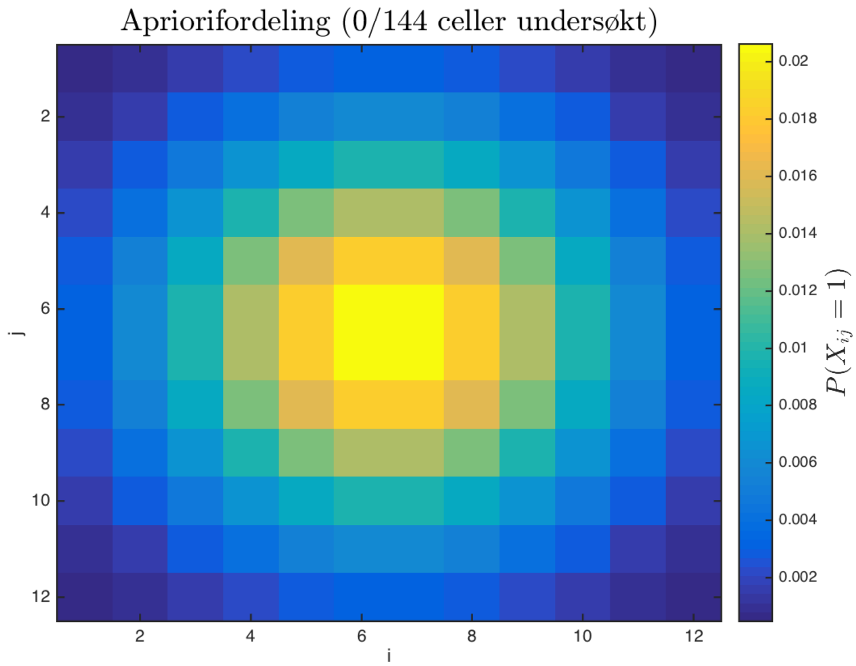
En måte å visualisere fordelingen på er ved å bruke `imagesc`. Dette gir et bilde hvor funksjonsverdi representeres av farge. Legger vi til en colorbar ved siden av selve bildet, blir figuren lett å tolke.

```
figure(1)
imagesc(pr0)
title(sprintf('Apriorifordeling (0/%i celler unders{\\o}kt)',nx*ny), ...
'FontSize',18,'Interpreter','Latex')
cbh1 = colorbar;
xlabel('i')
ylabel('j')
ylabel(cbh1,'$P(X_{ij} = 1)$','Interpreter','Latex','FontSize',16)
```

b) Vi vil finne sannsynligheten for at $X_{1,1} = 1$ gitt at $Y_{1,1} = 0$. Ved Bayes' setning er denne lik

$$
\begin{aligned}
P(X_{1,1} = 1|Y_{1,1} = 0) &= \frac{P(X_{1,1} = 1)P(Y_{1,1} = 0|X_{1,1} = 1)}{P(Y_{1,1} = 0)} \\
&= \frac{P(X_{1,1} = 1)P(Y_{1,1} = 0|X_{1,1} = 1)}{P(Y_{1,1} = 0|X_{1,1} = 0)P(X_{1,1} = 0) + P(Y_{1,1} = 0|X_{1,1} = 1)P(X_{1,1} = 1)} \\
&= \frac{(1-\ell)p_0(1,1)}{1 \cdot (1 - p(1,1)) + (1-\ell)p_0(1,1)} \\
&= \frac{p_0(1,1)(1-\ell)}{1 - p_0(1,1) + p_0(1,1) - \ell p_0(1,1)} \\
&= \frac{p_0(1,1)(1-\ell)}{1 - p_0(1,1)\ell}.
\end{aligned}
$$

Videre vil vi finne sannsynligheten for at $X_{ij} = 1$ gitt at $Y_{1,1} = 0$, der (i, j) ikke er lik



Figur 3: Apriorifordelingen $p_0(i, j)$.

$(1, 1)$. Igjen bruker vi Bayes' setning og får

$$\begin{aligned} P(X_{ij} = 1 | Y_{1,1} = 0) &= \frac{P(X_{ij} = 1)P(Y_{1,1} = 0 | X_{ij} = 1)}{P(Y_{1,1} = 0)} \\ &= \frac{P(X_{ij} = 1) \cdot 1}{P(Y_{1,1} = 0 | X_{1,1} = 0)P(X_{1,1} = 0) + P(Y_{1,1} = 0 | X_{1,1} = 1)P(X_{1,1} = 1)} \\ &= \frac{p_0(i, j)}{1 - p_0(1, 1)\ell}. \end{aligned}$$

Her bruker vi antakelsen om at det kun finnes ett vrak, og at det kun befinner seg i en enkelt gridcelle, til å si at $P(Y_{1,1} = 0 | X_{ij} = 1) = 0$. Merk at uttrykket for den marginale sannsynligheten $P(Y_{1,1} = 0)$ i nevneren er det samme i begge tilfellene.

c) Her kan vi gå fram på samme måte som i **b)**, men med (i_k, j_k) i stedet for $(1, 1)$, og med $p_k(i,j)$ i stedet for $p_0(i,j)$. Se først på tilfellet $(i,j) = (i_k, j_k)$. Der har vi

$$
\begin{aligned}
p_k(i_k,j_k) &= P(X_{i_k,j_k} = 1 | Y_{1,1} = \ldots = Y_{i_{k-1},j_{k-1}} = Y_{i_k,j_k} = 0) \\
&= \frac{P(X_{i_k,j_k} = 1, Y_{1,1} = \ldots = Y_{i_{k-1},j_{k-1}} = Y_{i_k,j_k} = 0)}{P(Y_{1,1} = \ldots = Y_{i_{k-1},j_{k-1}} = Y_{i_k,j_k} = 0)} \\
&= \frac{P(X_{i_k,j_k} = 1, Y_{1,1} = \ldots = Y_{i_{k-1},j_{k-1}} = 0)}{P(Y_{1,1} = \ldots = Y_{i_{k-1},j_{k-1}} = 0)} \\
&\quad \cdot \frac{P(Y_{i_k,j_k} = 0 | X_{i_k,j_k} = 1)}{P(Y_{i_k,j_k} = 0 | Y_{1,1} = \ldots = Y_{i_{k-1},j_{k-1}} = Y_{i_k,j_k} = 0)} \\
&= p_{k-1}(i_k,j_k) \cdot P(Y_{i_k,j_k} = 0 | X_{i_k,j_k} = 1) / \\
&\quad (P(Y_{i_k,j_k} = 0 | X_{i_k,j_k} = 0) P(X_{i_k,j_k} = 0 | Y_{1,1} = \ldots = Y_{i_{k-1},j_{k-1}} = 0) \\
&\quad + P(Y_{i_k,j_k} = 0 | X_{i_k,j_k} = 1) P(X_{i_k,j_k} = 1 | Y_{1,1} = \ldots = Y_{i_{k-1},j_{k-1}} = 0)) \\
&= p_{k-1}(i_k,j_k) \cdot \frac{1-\ell}{1 \cdot (1 - p_{k-1}(i_k,j_k)) + (1-\ell) p_{k-1}(i_k,j_k)} \\
&= p_{k-1}(i_k,j_k) \cdot \frac{1-\ell}{1 - p_{k-1}(i_k,j_k) + p_{k-1}(i_k,j_k) - \ell p_{k-1}(i_k,j_k)} \\
&= \frac{p_{k-1}(i_k,j_k) \cdot (1-\ell)}{1 - p_{k-1}(i_k,j_k)\ell}.
\end{aligned}
$$

For resten av cellene, altså de med $(i,j) \neq (i_k, j_k)$, får vi på liknende vis

$$
\begin{aligned}
p_k(i,j) &= P(X_{ij} = 1 | Y_{1,1} = \ldots = Y_{i_{k-1},j_{k-1}} = Y_{i_k,j_k} = 0) \\
&= \frac{P(X_{ij} = 1, Y_{1,1} = \ldots = Y_{i_{k-1},j_{k-1}} = Y_{i_k,j_k} = 0)}{P(Y_{1,1} = \ldots = Y_{i_{k-1},j_{k-1}} = Y_{i_k,j_k} = 0)} \\
&= \frac{P(X_{ij} = 1, Y_{1,1} = \ldots = Y_{i_{k-1},j_{k-1}} = 0)}{P(Y_{1,1} = \ldots = Y_{i_{k-1},j_{k-1}} = 0)} \\
&\quad \cdot \frac{P(Y_{i_k,j_k} = 0 | X_{ij} = 1, Y_{1,1} = \ldots = Y_{i_{k-1},j_{k-1}} = Y_{i_k,j_k} = 0)}{P((Y_{i_k,j_k} = 0 | Y_{1,1} = \ldots = Y_{i_{k-1},j_{k-1}} = 0)} \\
&= p_{k-1}(i,j) \cdot P(Y_{i_k,j_k} = 0 | X_{i,j} = 1) / \\
&\quad (P(Y_{i_k,j_k} = 0 | X_{i_k,j_k} = 0) P(X_{i_k,j_k} = 0 | Y_{1,1} = \ldots = Y_{i_{k-1},j_{k-1}} = 0) \\
&\quad + P(Y_{i_k,j_k} = 0 | X_{i_k,j_k} = 1) P(X_{i_k,j_k} = 1 | Y_{1,1} = \ldots = Y_{i_{k-1},j_{k-1}} = 0)) \\
&= p_{k-1}(i,j) \cdot \frac{1}{1 \cdot (1 - p_{k-1}(i_k,j_k)) + (1-\ell) \cdot p_{k-1}(i_k,j_k)} \\
&= \frac{p_{k-1}(i,j)}{1 - p_{k-1}(i_k,j_k)\ell}.
\end{aligned}
$$

d) Posteriorifordelingen kan f.eks. beregnes som følger.

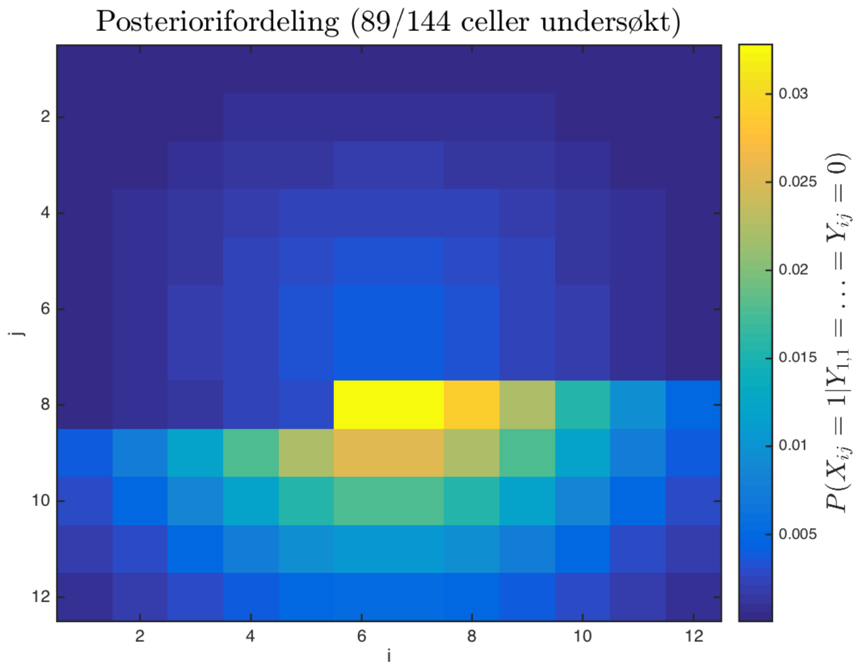
```
lik = 0.9;
po = NaN(nx,ny);
K = 144;

if K == 0
    po = pr;
```

```
end
for k = 1:K
   [ik,jk] = ind2sub([nx,ny],k);
   mlik = 1 - lik*pr(ik,jk);
   for i = 1:nx
       for j = 1:ny
           if nx*(i-1)+j~=k
               po(i,j) = pr(i,j)./mlik;
           else
               po(i,j) = pr(i,j).*(1-lik)./mlik;
           end
       end
   end
   pr = po;
end
```

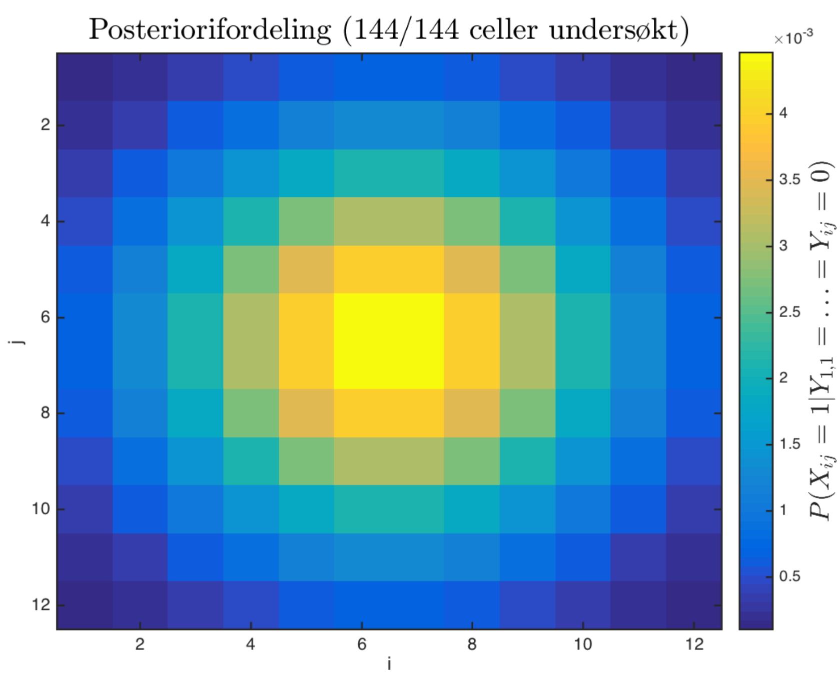
Visualisering med `imagesc` kan gjøres som vist under. For å lage alle figurene oppgaven ber om må en bruke ulike verdier av `K`.

```
figure(2);
imagesc(po);
title(sprintf('Posteriorifordeling (%i/%i celler unders{\\o}kt)',K,nx*ny), ...
'FontSize',18,'Interpreter','Latex')
cbh2 = colorbar;
xlabel('i')
ylabel('j')
ylabel(cbh2,'$P(X_{ij} = 1|Y_{1,1} = \ldots = Y_{ij} = 0)$',' ...
Interpreter','Latex','FontSize',16)
```

Figur 4 viser sannsynlighetsfordelingen underveis i søket, etter at de første 89 cellene er undersøkt uten funn. I denne situasjonen er det størst sannsynlighet for at vraket befinner seg i den nederste delen av leteområdet, som ikke er undersøkt enda. Når vi sammenlikner figur 3 og 5, ser vi at posteriorifordelingen etter at alle cellene er undersøkt uten funn er identisk med apriorifordelingen. Dersom vi leter gjennom hele området uten å finne noe, er vi altså like usikre på vrakets posisjon som vi var før vi begynte å lete. Det kan virke som om vi har mer informasjon etter å ha undersøkt 89 celler enn vi har etter å ha undersøkt alle 144, slik at vi tilsynelatende har tapt informasjon ved å undersøke resten av leteområdet. Men vi kan også tolke dette som at vi har fått mer informasjon om hvor stor spredningen i sannsynlighetsfordelingen egentlig er. Etter å ha undersøkt alle cellene vet vi med andre ord mer om hvor mye (eller lite) vi egentlig vet.



Figur 4: Posteriorifordelingen $p_{89}(i, j)$.



Figur 5: Posteriorifordelingen $p_{144}(i, j)$.