UNIVERSITETET I OSLO
Matematisk Institutt

EKSAMEN I:	**STK 4020 – Bayesiansk statistikk**
	Del I av to deler
VED:	**Nils Lid Hjort**
TID FOR EKSAMEN:	**Del I: 9.xii. 2004 kl. 12:00 – 21.xii. s.å. kl. 14:55;**
	Del II: 17.i. n.å., muntlig høring

Oppgavene til **Del I** *deles ut torsdag 9. desember kl. 12:00, ved Nils Lid Hjort, 8. etasje, Matematisk Institutt. Oppgavesettet blir også lagt ut på kursets hjemmeside. Skriftlig løsning,* **i to eksemplarer**, *helst tekstbehandlet, skal leveres til Nils Lid Hjort, senest tirsdag 21. desember s.å. kl. 14:55. Relevante figurer kan gjerne inngå i besvarelsen. Også kopier av programmer som er benyttet skal legges ved. Kandidatene skal arbeide uavhengig av hverandre.*

Del II *er en muntlig eksamen, og avholdes mandag 48. desember (også kalt 17. januar); nærmere opplysninger om dette blir gitt senere. Dette er altså oppgavesettet til* **Del I**. *Det inneholder fire oppgaver og er på tilsammen ni sider, inkludert et to-sides appendiks med noen nyttige opplysninger, også om* `R`*-detaljer.*

Oppgave 1

VI SKAL UT PÅ TUR i en homogen Poisson-skog. Trærne (av den bestemte typen *Tsuga Canadensis*) er fordelt slik at (a) antall trær $N(A)$ innenfor et område A med areal $\mathrm{ar}(A)$ (målt i kvadratmeter) er Poisson-fordelt med parameter $\lambda\,\mathrm{ar}(A)$; (b) antall trær $N(A)$ og $N(B)$ innenfor ikke-overlappende områder A og B er stokastisk uavhengige. Her er λ en ukjent parameter som altså svarer til den gjennomsnittlige tretettheten pr. kvadratmeter.

(a) Fra et vilkårlig utgangspunkt i skogen, la A være en sirkel med radius y. Hva er sjansen for at det ikke skal være noen trær innenfor A? Bruk dette til å vise at sannsynlighetsfordelingen til Y, distansen fra dette utgangspunktet til nærmeste tre, har sannsynlighetstetthet

$$f(y,\lambda) = e^{-\lambda\pi y^2} 2\lambda\pi y \quad \text{for } y > 0.$$

For denne fordelingen viser et par integrasjonsøvelser at

$$\mathrm{E}Y = \frac{\frac{1}{2}}{\sqrt{\lambda}} \quad \text{og} \quad \mathrm{Var}\,Y = \frac{1}{\lambda}\Big(\frac{1}{\pi} - \frac{1}{4}\Big).$$

– Og hvor mange trær (av denne bestemte typen) er det i skogen? Man velger via et kart n forskjellige startsteder i skogen, alle nøyaktig spesifiserte, og fra hver av disse måler man distansen til det nærmeste tre. Dette gir målingene $Y_1,\ldots,Y_n$. I det følgende skal du anta at $Y_1,\ldots,Y_n$ er uavhengige med den samme sannsynlighetstetthet, nemlig $f(y,\lambda)$ fra forrige punkt.

(b) Diskuter kort om disse antagelsene synes rimelige eller ikke. Vis at ML-estimatoren (maximum likelihood-estimatoren, som kanskje heter rimelighetsfunksjonsmaksimeringsestimatoren på norsk) blir

$$\widehat{\lambda} = \frac{1}{\pi}\frac{1}{W_n}, \quad \text{der } W_n = \frac{1}{n}\sum_{i=1}^{n} Y_i^2.$$

Hva er dennes approksimative normalfordeling?

(c) Du er statistikeren, og trenger en apriorifordeling for λ, som blir konstruerert på følgende måte. Samtaler med skogforskere oppsummeres ved de to opplysningene (i) i en $10\,m \times 10\,m$-rute i skogen forventer man å finne i gjennomsnitt to trær; (ii) sjansen for å finne flere enn ti trær (elleve eller flere) i denne $10\,m \times 10\,m$-ruten er 0.01. Du bruker en Gamma-fordeling som matcher disse opplysningene. Vis at dette leder til å bruke $\lambda \sim \mathcal{G}(c\,0.02, c)$, der $c = 0.134$.

(d) Man foretok $n = 25$ slike målinger, og fant disse tallene, i meter:

9.55	4.92	7.85	1.27	2.91	3.77	2.94	3.39	
4.29	4.14	6.13	0.16	4.94	5.59	2.53	9.16	
6.71	4.42	2.64	7.66	0.85	2.21	1.38	0.94	2.70

For disse målingene finner man (blant annet)

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = 4.122, \quad \frac{1}{n}\sum_{i=1}^{n} Y_i^2 = 23.458, \quad \text{empirisk standardavvik } 2.596.$$

Finn den eksakte aposteriorifordelingen.

(e) Plott den eksakte aposterioritettheten for λ sammen med «den fattige Bayesianers approksimasjon», som bruker normalapproksimasjonen via ML-analysen. Kommenter resultatet.

(f) Man er interessert i sannsynligheten p for at det i en $10\,m \times 10\,m$-rute skal befinne seg minst fem trær. Simuler 10,000 p-verdier fra den relevante aposteriorifordelingen, vis frem et histogram over disse, og beregn Bayes-estimat, samt et 90% troverdighetsintervall, for p.

Oppgave 2

Visse oppfatninger er blandede, og vi skal bevege oss mot blandings-priors, der apriorifordelingen for de ukjente parametre er satt sammen av ulike komponenter. Men vi starter forsiktig nok med en enklere øvelse. For å komme gjennom matematikken på en oversiktlig nok måte skriver vi

$$\phi(x, \sigma) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\Big(-\tfrac{1}{2}\frac{x^2}{\sigma^2}\Big)$$

for normaltettheten med standardavvik σ.

(a) Anta at målinger $y_1, \ldots, y_n$ for gitt θ er uavhengige $\mathrm{N}(\theta, 1)$, og at θ har apriorifordelingen $\mathrm{N}(\theta_0, \sigma_0^2)$. Innfør

$$\widehat{\theta} = \frac{(1/\sigma_0^2)\theta_0 + (1/n^{-1})\theta_0}{(1/\sigma_0^2) + (1/n^{-1})} = \frac{\sigma_0^2 \bar{y} + (1/n)\theta_0}{\sigma_0^2 + 1/n},$$
$$\widehat{\sigma}^2 = \frac{1}{(1/\sigma_0^2) + (1/n^{-1})} = \frac{\sigma_0^2}{n\sigma_0^2 + 1}.$$

Vis at simultanfordelingen for (θ, data) kan skrives

$$\begin{aligned}\pi(\theta) L_n(\theta) &= \phi(\theta - \theta_0, \sigma_0) \prod_{i=1}^n \phi(y_i - \theta, 1) \\ &= \phi(\theta - \widehat{\theta}, \widehat{\sigma}) \phi(\bar{y} - \theta_0, (\sigma_0^2 + 1/n)^{1/2}) h(y_1, \ldots, y_n),\end{aligned}$$

der

$$h(y_1, \ldots, y_n) = \frac{1}{\sqrt{n}} \Big(\frac{1}{\sqrt{2\pi}}\Big)^{n-1} \exp(-\tfrac{1}{2} n s^2), \quad \text{med } s^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2.$$

Dette kan vises på flere måter, og gjerne ved algebraiske omkalfatringer; du kan for eksempel arbeide med «konstanter» og «eksponentialledd» hver for seg, og vise at de matcher for venstre og høyre side av den formel som skal vises.

(b) Vi har ti målinger av typen beskrevet over:

`0.486 -0.288 3.427 0.651 2.055 3.118 1.327 0.891 1.872 1.468`

med gjennomsnitt $\bar{y} = 1.5007$. Anta først at en innledende introspeksjon og prior-elicitasjon manifesterer seg som apriorifordelingen $\mathrm{N}(0, \sigma_0^2)$ for θ, der $\sigma_0 = 0.25$. Hva blir aposteriorifordelingen for θ? Plott apriori- og aposterioritetthet i samme diagram.

(c) Vi tenker oss nå at en grundigere og mer detaljert forhåndsanalyse avdekker en noe mer nyansert apriorifordeling, som tar høyde for at θ kan komme annetsteds fra en bare fra $\mathrm{N}(0, \sigma_0^2)$, riktignok med lav sannsynlighet. Apriorifordelingen er nå

$$\pi = p_1\pi_1 + p_2\pi_2 + p_3\pi_3 = 0.01\,\mathrm{N}(-1, \sigma_0^2) + 0.98\,\mathrm{N}(0, \sigma_0^2) + 0.01\,\mathrm{N}(1, \sigma_0^2),$$

der fortsatt $\sigma_0 = 0.25$. Vis at aposteriorifordelingen tar formen

$$\pi(\theta \,|\, \mathrm{data}) = \widetilde{p}_1 \pi_1(\theta \,|\, \mathrm{data}) + \widetilde{p}_2 \pi_2(\theta \,|\, \mathrm{data}) + \widetilde{p}_3 \pi_3(\theta \,|\, \mathrm{data}),$$

en kombinasjon av de tre aposteriorifordelinger man ville fått om apriorifordelingene var henholdsvis $\mathrm{N}(-1, \sigma_0^2)$, $\mathrm{N}(0, \sigma_0^2)$, $\mathrm{N}(1, \sigma_0^2)$.

(d) Finn uttrykk for og beregnede verdier for $\widetilde{p}_1, \widetilde{p}_2, \widetilde{p}_3$ her, og kommenter i hvilken grad komponentvektene 0.01, 0.98, 0.01 har flyttet seg i lys av data. Plott aposteriori- og aprioritetthet i samme diagram. Sammenlign med resultatet fra den enklere apriorifordelingen, og kommenter.

(e) Analysen over bygget på at standardavviket σ_0 i hver av apriori-komponentene var et kjent tall. Hvordan vil du gjøre analysen, hvis essens altså er å komme frem til en aposteriorifordeling for θ, dersom σ_0 var ukjent?

Oppgave 3

Sicut cervus desiderat ad fontes aquarum – og hvor mange hjorter er det i skogen? De lar seg slett ikke telle direkte, og lar seg kun observere på filmrullene når de kommer til de observasjonsposter feltbiologene har rigget opp i nærheten av vannhullet. Feltbiologene har tellet antall hjorter ved ti ulike anledninger, og funnet

`36 33 33 47 32 43 40 37 45 40`

Vi tenker oss at disse $m = 10$ observasjonene $y_1, \ldots, y_m$ er uavhengige binomiske observasjoner, $y_i \sim \text{Bin}(N, \theta)$, der N er det ukjente antall hjorter i skogen og θ er deres oppdagelsessannsynlighet. Oppgaven består i å anslå N, med et supplerende troverdighetsintervall.

(a) Vis at modellens likelihood blir

$$L_m(N,\theta) = \Big[\prod_{i=1}^m \binom{N}{y_i}\Big]\theta^{m\bar{y}}(1-\theta)^{m(N-\bar{y})} \quad \text{for } N \geq y_{\max} \text{ og } \theta \in (0,1),$$

der $\bar{y} = (1/m)\sum_{i=1}^m y_i$ og $y_{\max} = \max\{y_1, \ldots, y_m\}$.

– Andre apriorifordelinger kan absolutt tenkes anvendt her, men vi skal bruke

$$\pi(N,\theta) = \text{const.}\, N^{-1} \quad \text{for } 1 \leq N \leq N_{\max} \text{ og } \theta \in (0,1),$$

for en passende høy $N_{\max}$; den eksakte verdien for $N_{\max}$ er ikke meget avgjørende, så lenge den er høy. Det ligger også i antagelsen at N og θ er uavhengige, i sin apriorifordeling, og at θ er uniform over $(0,1)$.

(b) Sett opp et uttrykk for aposteriorifordelingen $\pi(N,\theta\,|\,\text{data})$. Vis at

$$N\,|\,\text{data} \sim \text{const.}\, N^{-1}\Big[\prod_{i=1}^m \binom{N}{y_i}\Big]\frac{(m\bar{y})!\,(mN-m\bar{y})!}{(mN+1)!} \quad \text{for } y_{\max} \leq N \leq N_{\max}.$$

Beregn disse sannsynlighetene og plott dem.

(c) Fra denne aposteriorifordelingen for N skal du finne to Bayes-estimater samt et 90% troverdighetsintervall. Det første Bayes-estimatet skal anvende kvadratisk tapsfunksjon mens det andre skal bruke tapsfunksjonen

$$L(N,\widehat{N}) = \begin{cases} 1 & \text{hvis } \widehat{N} \neq N, \\ 0 & \text{hvis } \widehat{N} = N. \end{cases}$$

(d) Finn fordelingen for θ gitt N, i aposteriorifordelingen. Bruk dette, gjerne via `sample` i `R`, til å simulere 100 eller 1000 par (N,θ) fra aposteriorifordelingen. Plott disse, og kommenter det du finner. Vis også frem aposteriorifordelingen for θ, og gi et estimat og et 90% troverdighetsintervall.

(e) Resultatene over ble nådd under den antagelse at θ har er uniform apriorifordeling. Anta nå at dyreforskerne enes om, for denne hjortepopulasjonen, i denne skogen, og med forskernes observasjonsplan, at θ istedet bør modelleres som en Beta-fordeling med forventning 0.30 og standardavvik 0.15. Undersøk hvordan aposteriorifordelingen for N, og for θ, forandrer seg.

(f) Det viser seg ved nærmere analyse av dyreforskernes observasjonsplan at de fem første observasjonene (36, 33, 33, 47, 32) stammer fra tellinger foretatt under litt andre forutsetninger enn de fem siste (43, 40, 37, 45, 40). Mer konkret anses det rimelig at den hjorteoppdagelsessannsynlighet θ_2 som ligger til grunn for de siste fem tellingene er noe høyere enn den θ_1 som vedrører de første fem tellingene. Uten nødvendigvis å fullføre alle dine tankerekker, gjør rede for hvordan en slik type tilleggsinnsikt kan bli brukt til å finne et forhåpentligvis skarpere anslag for N.

Oppgave 4

HVOR FORSKJELLIGE ER DE SKJULTE SANNSYNLIGHETER? Man har observert

`5 3 2 6 3 7 3 8 2 1`

fra ti uavhengige binomiske forsøk, hver med tallparameter $n = 12$. Dette svarer altså til data $y_i \sim \text{Bin}(12, p_i)$, men der p_i-ene ikke nødvendigvis er like. Denne oppgaven går ut på å estimere disse p_i-ene og å kvantifisere nettopp hvor ulike, eller hvor like, de ser ut til å være.

(a) Vi starter med en enkel «kjenn din prior»-øvelse. Anta at p kommer fra fordelingen $\text{Beta}(kp_0, k(1-p_0))$, der tettheten altså er

$$\text{be}(p \mid k, p_0) = \frac{\Gamma(k)}{\Gamma(kp_0)\Gamma(kq_0)} p^{kp_0-1}(1-p)^{kq_0-1} \quad \text{for } 0 < p < 1,$$

der vi skriver $q_0 = 1 - p_0$. Vis at

$$\text{E}\, p = p_0 \quad \text{og} \quad \text{Var}\, p = \frac{p_0 q_0}{k+1}.$$

Parameteren k styrer altså graden av konsentrasjon rundt p_0.

(b) Dernest gjennomgås en generell øvelse om hierarkiske modeller, som vi skal få bruk for under. Anta at kjellerparameteren ϕ styrer under-overflaten-parameteren θ som så styrer data y; mer konkret antas at (i) ϕ har apriorifordelingen $\pi(\phi)$; (ii) θ gitt ϕ har fordeling $f(\theta \mid \phi)$; (iii) y gitt (ϕ, θ) har fordeling $g(y \mid \theta)$. Den simultane fordelingen er dermed gitt:

$$(\phi, \theta, y) \sim \pi(\phi) f(\theta \mid \phi) g(y \mid \theta).$$

I den situasjon vi om et øyeblikk skal vende tilbake til er $\phi = (k, p_0)$, $\theta = (p_1, \ldots, p_{10})$, og $y = (y_1, \ldots, y_{10})$.

Din oppgave nå er å sette opp formler for
- marginalfordelingen for data y;
- aposteriorifordelingen for ϕ;
- fordelingen for θ gitt (ϕ, y).

(c) Anta at p er fordelt som over og at y gitt p er binomisk (n, p). Det klassiske Bayes-resultatet er da at $p \mid y$ blir Beta$(kp_0 + y, kq_0 + n - y)$. Vis dette, og dessuten at marginalfordelingen for y blir

$$g(y) = g(y \mid k, p_0, n) = \binom{n}{y} \frac{\Gamma(k)}{\Gamma(kp_0)\Gamma(kq_0)} \frac{\Gamma(kp_0 + y)\Gamma(kq_0 + n - y)}{\Gamma(k + n)}$$

for $y = 0, 1, \ldots, n$. Hva blir (ubetinget) forventning og varians til y? Kommenter spesialtilfellene k liten og k stor.

(d) Vis at simultanfordelingen for (p, y) kan skrives

$$\begin{aligned}(p, y) &\sim \mathrm{be}(p \mid kp_0, kq_0)\, \mathrm{bin}(y \mid n, p) \\ &= \mathrm{be}(p \mid kp_0 + y, kq_0 + n - y)\, g(y \mid k, p_0, n)\end{aligned}$$

for $p \in (0, 1)$ og $y = 0, 1, \ldots, n$.

(e) Vi vender nå tilbake til situasjonen beskrevet innledningsvis for denne oppgaven. Vår komplette modellbeskrivelse, denne gang beskrevet fra data-overflaten og nedover, er at (i) y_i-ene er betinget uavhengige Bin$(12, p_i)$; (ii) p_i-ene er betinget uavhengige fra Beta$(kp_0, k(1 - p_0))$; og (iii) (k, p_0) har en viss bakgrunnsprior. Kommenter kort disse modellantagelsene.

- For at ikke dette oppgavesettet skal bli for langt (!) skal vi nå anta at p_0 er et kjent tall, nemlig $p_0 = 0.28$. Dermed er k den eneste gjenværende bakgrunnsparameter. Full analyse for (k, p_0) kan også utføres, uten prinsipielt større problemer. Det samme gjelder situasjoner som typisk dukker opp i praksis, der det ikke er samme sample-størrelse n i hvert binomisk forsøk.

(f) Vis at likelihood-funksjonen for parameteren k blir

$$L(k) = \prod_{i=1}^{10} g(y_i \mid k, p_0, n),$$

for dataene over, og hvor altså $p_0 = 0.28$ og $n = 12$. Plott likelihood-funksjonen her, og finn spesielt ML-estimatet for k.

(g) Hvilken prior kan man bruke for k? Gjør først et forsøk på beregne, og plotte, den såkalte Jeffreys-prioren. Hvilken forbilledlig egenskap har Jeffreys-prioren? Uansett skal vi i resten av oppgaven anvende apriorifordelingen $\pi(k) = 0.1\, e^{-0.1\, k}$ for $k > 0$. Dette er også en $\mathcal{G}(1, 0.1)$-fordeling.

(h) Beregn og plott aposteriorifordelingen for k. Simuler 1000 eller 10000 kopier av vektoren $(k, p_1, \ldots, p_{10})$ fra dens aposteriorifordeling. Bruk dette til

- å estimere hver enkelt p_i;
- med et supplerende standardavviksestimat for hver;
- å gi et histogram over de samplede k, gjerne supplert med aposterioritettheten i samme diagram;
- å vise frem aposteriorifordelingene for

$$\tau = \sqrt{\frac{p_0 q_0}{k+1}} \quad \text{og} \quad \kappa = \frac{1}{10} \sum_{i=1}^{10} |p_i - \bar{p}|\,;$$

- samt eventuelle andre problemstillinger du kunne finne det interessant å belyse.

Her er τ standardavviket i den bakenforliggende fordelingen for p_i-ene. For estimering av de forskjellige p_i-ene er det av interesse, presentasjonsmessig, å plotte tre ulike estimater i samme diagram, for $i = 1, \ldots, 10$: «rå-estimatene» y_i/n (hvilken verdi av k svarer dette til?); det store felles-estimatet p^*, beregnet under den modell at alle p_i-ene er like (og hvilken verdi av k svarer dette til?); og Bayes-estimatene $\widehat{p}_i$.

Appendiks I: et par fordelinger og integraler

Vi sier at U er Gamma-fordelt med parametre (a, b), og skriver $U \sim \mathcal{G}(a, b)$, dersom dens tetthet er

$$g(u) = \frac{b^a}{\Gamma(a)} u^{a-1} e^{-bu} \quad \text{for } u > 0.$$

Dens forventning og varians er a/b og a/b^2. I R kan man kan ved hjelp av kommandoer av typen

```
rgamma(1000,a,b)
pgamma(0.10,a,b)
dgamma(0.33,a,b)
qgamma(0.88,a,b)
```

simulere Gamma-variable, finne kumulative sannsynligheter, tetthetsverdier, og kvantiler.

At integrasjonskonstanten over er som den er henger naturligvis sammen med at

$$\int_0^\infty u^{a-1} e^{-bu}\, \mathrm{d}u = \frac{b^a}{\Gamma(a)} \quad \text{for alle positive } a, b.$$

Analogt er Beta-tettheten med parametre (a, b) på formen

$$\mathrm{be}(v \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} v^{a-1}(1-v)^{b-1} \quad \text{for } v \in (0, 1)$$

fordi

$$\int_0^1 v^{a-1}(1-v)^{b-1}\, \mathrm{d}v = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \quad \text{for alle positive } a, b.$$

Spesielt er

$$\int_0^1 v^a (1-v)^b\, \mathrm{d}v = \frac{a!\, b!}{(a+b+1)!} \quad \text{for ikkenegative heltall } a, b.$$

Appendiks II: numerikk og assorterte R-grep

Produkter av mange eller store tall beregnes oftest best ved å gå veien om logaritmen. For fakulteter (factorials) bruker man $\log(100!) = \log\Gamma(101)$, osv., der $\log\Gamma(x)$ beregnes som `lgamma(x)` i R.

Man kan definere funksjoner i R, som følger:

```
ff <- function(u)
{exp(0.5*sin(u))}
```

Er funksjonen av flere variable kan man gjøre slik:

```
ff <- function(diverse)
{x <- diverse[1]
y <- diverse[2]
sqrt(abs(x) + abs(y))}
```

For å plotte funksjonene $f_1(x) = \sin(\exp(\frac{1}{2}x))$ og $f_2(x) = \cos(\sin(\cos x))$ over intervallet $x \in [0, 2\pi]$, kan man anvende

```
xval <- seq(0,2*pi,by=0.01)
f1val <- sin(exp(0.5*xval))
f2val <- cos(sin(cos(xval)))
matplot(xval, cbind(f1val, f2val), type="l")
```

Man kan innenfor `matplot`s parenteser supplere med `xlab`, `ylab`, samt med `col=c(1,1)` om man ønsker at kurvene skal plottes i samme farge (som er praktisk når resultatet skal printes ut på en vanlig sort-hvit-printer).

Av og til er funksjonene man skal beregne eller plotte såpass kompliserte at man ikke får det utført ved enkle vektorielle metoder, slik vi klarte det over med f_1 og f_2. Da kan man bruke en `for`-løkke og beregne én verdi om gangen, som vist her for funksjonen $g(a) = \int_0^1 |\cos(a\, e^x)|\, \mathrm{d}x$, her beregnet for $a \in [0, 3]$:

```
aval <- seq(0,3,by=0.01)
gval <- 0*aval
for (j in 1:length(aval))
{
a <- aval[j]
inte <- function(x)
{abs(cos(a*exp(x)))}
gval[j] <- integrate(inte, 0, 1)$value
print(c(a,gval[j])) # if I wish to check
}
```

For å sample `sim` ganger fra en gitt liste `xval` av verdier, med visse sannsynligheter `here-andnow`, anvendes

```
sample(xval, sim, replace=T, prob=hereandnow)
```

Her er det ikke engang nødvendig å skalere sannsynlighetene.

Histogrammer for observerte eller genererte verdier z får man via `hist(z)`, men det er ofte fornuftig å supplere med et par kosmetiske detaljer:

```
hist(z, probability=T, breaks=30, xlim=c(-3,3))
```

gir et histogram korrekt normalisert, med 30 vinduer, avgrenset til det bestilte intervall.

SLUTT PÅ DEL I