Løsningsforslag øving 10, ST1301

Oppgave 1 Øyvind går to skiturer i påsken. På den en skituren viste klokken 12:32 når han startet og 12:59 når han kom tilbake på hytta. Den andre dagen viste klokken 12:32 når han startet og 12:58 når han vendte tilbake. Klokken hennes viser bare timer og minutter og ikke sekunder. Hva er sannsynligheten for at skituren den andre dagen var lenger (målt i tid) enn skituren den første dagen?

Hint: Når klokken viser f.eks. 12:32 betyr det at tiden T_1 var et sted mellom 12:32:00 og 12:33:00 ved avlesning av starttidspunkt første dag. Bruk uniforme sannsynlighetsfordelinger for å modellere hva vi vet om T_1 og de tre andre tidspunktene som inngår i problemet. La f.eks. T_1 være uniform på intervallet fra 32 til 33.

Finn sannsynligheten for hendelsen over (lengst skitur andre dag) ved å simulere 10000 realisasjoner av $T_1, T_2, \ldots, T_4$ og tell opp antall ganger hendelsen inntreffer f.eks. ved bruk av tellevariabel og if-setning.

```
skitur <- function(nsim=1000) {
  nlenger <- 0
  for (i in 1:nsim) {
     T1 <- runif(1,32,33) # Start første dag
     T2 <- runif(1,59,60) # Slutt første dag
     T3 <- runif(1,32,33) # Start andre dag
     T4 <- runif(1,58,59) # Slutt andre dag
     if (T4-T3>T2-T1) # Lengde andre skitur større enn første
        nlenger <- nlenger+1
  }
  return(nlenger/nsim)
}
> skitur(nsim=100000)
[1] 0.04156
```

Vi ser at sannsynligheten blir lik 0.041. Det kan vises at den eksakte sannsynligheten blir 1/24 (en løsning er her: `http://groups.google.com/groups?hl=en&lr=&ie=UTF-8&oe=UTF-8&threadm=alt260pbk38tvcm9kvbjjelqsmguupab2s%404ax.com&rnum=1&prev=/groups%3Fhl%3Den%26lr%3D%26ie%3DUTF-8%26oe%3DUTF-8%26selm%3Dalt260pbk38tvcm9kvbjjelqsmguupab2s%25404ax.com`).

Oppgave 2 *Programmering av likelihood funksjon. Beregning av SME.*

I denne oppgaven skal vi estimere når og hvor lange brunstperioden er i en elgpopulasjon. Dette vil naturlig nok ikke være direkte observertbart men kan estimeres fra relevante data. Her skal vi bruke data gjort tilgjengelige av Erling Solberg ved NINA i Trondheim. Last ned dataene `ovul2.dat` fra hjemmesiden, og last disse inn i R ved å bruke `read.table` og `attach`. Disse dataene består av tre variable. $t_1, t_2, \ldots, t_n$ (har navnet `tid` i data.framen) er antall dager siden nyttår jaktdag i fant sted. Variablen n_i er antall elgkyr fellt på dag i, og variabelen x_i er antall ut av disse som hadde ovulert (hatt eggløsning) på fellingstidspunktet. (Dette ble bestemt laboratorium ved å undersøke om en spesiell strukturer (en eller flere corpus luteum.[1] var tildstede i ovariene.). Hos elg inntreffer ovulasjon en (eller flere ganger hvis først ovulasjon ikke fører til befruktning) i løpet av høsten for kjønnsmodne individer. Dermed vil andelen individer som har ovulert øke utover høsten og forløpet på denne sammenhengen vil kunne fortelle oss om gjennomsnittet og standardavviket til fordelingen av ovulasjonstidpunktene i populasjonen som vi her ønsker å estimere.

Lag først et spredningsplot av andelen (x_i/n_i) som har ovulert på ulike tidpunkt t_i. La størrelsen på hvert punkt i plottet avhenge av n_i slik at det fremkommer av plottet hvor mange observasjoner hvert punkt representer. Dette kan gjøres via argumentet `cex`. (Se `?plot.default`.)

La oss nå anta at ovulasjonstidspunktet T til et gitt individ i populasjonen er normalfordelt med forventning μ og standardavvik σ. Ikke alle kyrne ovulerer fordi ikke alle har nådd kjønnsmodning som 3-åringer. La oss derfor anta at hvert individ ovulerer en gang i løpet av høsten med sannsynlighet q. Disse antakalsene utgjør tilsammen vår modell. Vi ønsker nå å sette opp likelihoodfunksjonen for de observerte dataene for denne modellen. Vi trenger altså å finne et uttrykk for sannsynligheten for dataene gitt parameterne μ, σ, og q.

La O være hendelsen at et gitt individ ovulerer en eller annen gang i løpet av høsten. Sannsynligheten for at ovulasjon har funnet sted ved tidpunkt t_i, altså at $T < t_i$ blir da

$$\begin{aligned} P(T \leq t_i | O_i) & \\ &= P(\frac{T-\mu}{\sigma} \leq \frac{t_i - \mu}{\sigma}) \\ &= P(Z \leq \frac{t_i - \mu}{\sigma}) \\ &= G(\frac{t_i - \mu}{\sigma}) \end{aligned} \tag{1}$$

hvor G er kumulativ tetthet til en standard normalfordelt variabel.

[1] Se `http://en.wikipedia.org/wiki/Corpus_luteum`.

I likelihoodfunksjonen vår trenger vi de ubetingede sannsynligheten for hendelsen $T < t_i$. La oss kalle denne sannsynligheten for p_i. Bruker vi lov om totalsannsynlighet får vi at

$$\begin{aligned} p_i &= P(T < t_i) \\ &= P(T < t_i|O_i)P(O_i) + P(T < t_i|\bar{O}_i)P(\bar{O}_i) \\ &= G(\frac{t_i - \mu}{\sigma})q + 0 \cdot (1 - q) \end{aligned} \tag{2}$$

i og med at hendelsen $T < t_i$ har sannsynlighet 0 hvis individ i ikke ovulerer på noe tidspunkt i løpet av høsten (hendelsen $\bar{O}_i$).

Ser vi på antall individer som har ovulert X_i på et bestemt dag i, hver med sannsynlighet p_i, for vi at X_i må være binomisk fordelt med parameter p_i og n_i. Dermed blir likelihoodfunksjonen (sannsynligheten for alle de observerte dataene)

$$L(\mu, \sigma, q) = \prod_{i=1}^{n} \binom{n_i}{x_i} p_i^{x_i}(1 - p_i)^{n_i - x_i}, \tag{3}$$

hvor hver p_i er en funksjon av μ, σ og q gitt ved ved ligning (2).

Lag et plot av p_i som funksjon av t_i for $\mu = 260$, $\sigma = 10$ og $q = 0.8$ (altså funksjonen gitt ved (2)). Hvordan passer funksjonen med spredningsplottet av x_i/n_i og t_i?

Programmer så en funksjon som beregner verdien av minus log likelihoodfunksjonen for gitte parameterverdier. For å fungere sammen med `optim` må likelihoodfunksjonen må ta i mot parameterverdiene som første argument i form av en vektor, samt vektorer `t` og `x` som inneholder de observerte dataene. Tips: Beregn p_i'ene og ta vare på dette som et mellomregningsresultatet (bruk `pnorm`). Beregn så log-likelihoodverdien ved bruk av elementvise operasjoner på vektorer, funksjonen `sum`, og eventuellt og `dbinom` med tilleggsargumentet `log=T`.

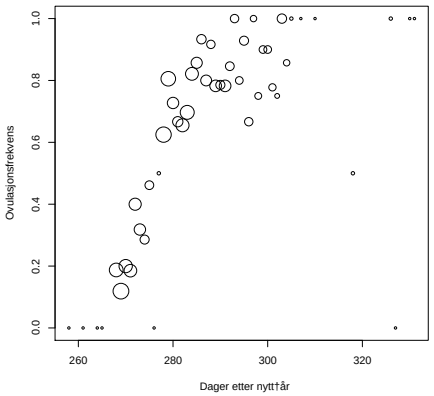
Beregn så sannsynlighetsmaksimeringsestimatene av μ, σ, og q ved bruk av `optim`. Det kan være viktig å velge gode startverdier for parameterne. Kontroller at `optim` finner samme maksimum om du endrer på startverdiene.

Hva blir SMEene av parameterne i modellen? Legg inn et plot av funksjonen gitt ved (`??`) i spredningsplottet for t_i og x_i ved å bruke `lines`. Hvordan passer den tilpassede modellen med dataene?

Dersom vi betrakter et tidinterval slik at 95% av individene i populasjon ovulerer i dette intervallet, hvor langt må da intervallet være?

```
ovulasjonsdata <- read.table("ovul2.dat")
```

```
attach(ovulasjonsdata)
plot(tid,x/n,cex=sqrt(n)/2,
    xlab="Dager etter nyttår",
    ylab="Ovulasjonsfrekvens")
```



Likelihoodfunksjonen i oppgaveteksten kan programmeres på følgende måte. Funksjonen skal returnere verdien til likelihoodfunksjonen i et bestemt punkt (μ, σ, q) i parameterrommet. Denne vektoren er innargument i funksjonen. Det kan være hensiktsmessig (for å gjøre programkoden mer lesbar) å plukke ut parameterverdier fra denne vektoren og lagre dem i lokale variable med passende navn.

```
lnL <- function(par,x,n,t) {
  mu <- par[1]
  sigma <- par[2]
  q <- par[3]
  p <- q*pnorm((t-mu)/sigma)
  sum(-dbinom(x,size=n,prob=p,log=T))
}
```

Legg merke til hvordan beregner alle p_i'ene — forventningene til hver observasjon X_i, for parameterverdiene gitt ved innargumentet `par` før det totale log-likelihood verdien beregnes i siste linje. Dette gjøres ved hjelp av elementvise operasjoner på vektorer.

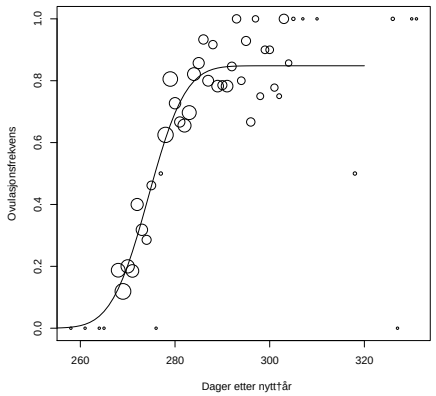
Nå som vi har likelihoodfunksjonen kan vi finne sannsynlighetsmaksimeringsestimater ved bruk av `optim`. Spredningsplottet antyder at μ ligger omkring 280, og at q ligger omkring 0.75.

```
> fit<-optim(c(280,10,.75),lnL,x=x,n=n,t=t)
```

```
There were 25 warnings (use warnings() to see them)
> fit
$par
[1] 274.386520   6.801954   0.855978

$value
[1] 355.3924

$counts
function gradient
     224       NA

$convergence
[1] 0

$message
NULL
```

Vi kan legge inn en plot av p_i som funksjon av t_i på følgende måte.

```
tt <- 240:320
lines(tt,fit$par[3]*pnorm((tt-fit$par[1])/fit$par[2]))
```



Basert på estimatet av $\sigma = 6.80$ dager følger det at et intervall som inneholder 95% av ovulasjonstidspunktene må være $2z_{0.025}\sigma = 26.65$.

Oppgave 3 Oppgave 5.4 i ISwR. Se eventuellt også notatet om binormalfordelingen fra i høst.

```
rbinorm <- function(n,rho) {
   x <- rnorm(n=n)
   y <- rho*x+rnorm(n=n,sd=sqrt(1-rho^2))
   return(cbind(x,y))
}
> plot(rbinorm(1000,.8))  # x og y verdiene kan ligge i en
                          # 2 x n matrise
> mat <- rbinorm(1000,.8)
> cor.test(mat[,1],mat[,2]) # Bare her blir estimatet blir nesten 0.8

        Pearson's product-moment correlation

data:  mat[, 1] and mat[, 2]
t = 42.26, df = 998, p-value = < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7775 0.8221
sample estimates:
   cor
0.8009

> cor.test(mat[,1],mat[,2],method="spearman")

        Spearman's rank correlation rho

data:  mat[, 1] and mat[, 2]
S = 37866128, p-value = < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
   rho
0.7728

> cor.test(mat[,1],mat[,2],method="kendall")

        Kendall's rank correlation tau

data:  mat[, 1] and mat[, 2]
z = 27.62, p-value = < 2.2e-16
alternative hypothesis: true tau is not equal to 0
```

```
sample estimates:
   tau
0.5833
```