Obligatorisk oppgave 1 i STK1110 – Høsten 2022

Oppgave 1

a) Sannsynlighetstettheten til log-normalfordelingen er:

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2\sigma^2}(\log(x)-\mu)^2}.$$

Det gir likelihood-funksjonen:

$$\begin{aligned} f(x_1,\ldots,x_n;\mu,\sigma^2) =& \prod_{i=1}^{n} f(x_i;\mu,\sigma^2) \\ =& \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma x_i} e^{-\frac{1}{2\sigma^2}(\log(x_i)-\mu)^2} \\ =& (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} \prod_{i=1}^{n} x_i^{-1} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\log(x_i)-\mu)^2}. \end{aligned}$$

Vi tar logaritmen og får:

$$\begin{aligned} l(\mu,\sigma^2;x_1,\ldots,x_n) =& \log(f(x_1,\ldots,x_n;\mu,\sigma^2)) \\ =& -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\log(x_i) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\log(x_i)-\mu)^2. \end{aligned}$$

Vi deriverer med hensyn på μ og σ^2 og setter lik 0:

$$\begin{aligned} \frac{\partial}{\partial\mu} l(\mu,\sigma^2;x_1,\ldots,x_n) =& \frac{1}{\sigma^2}\sum_{i=1}^{n}(\log(x_i)-\mu) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}\log(x_i) - n\mu\right) = 0 \\ \rightarrow \mu =& \frac{1}{n}\sum_{i=1}^{n}\log(x_i) \\ \frac{\partial}{\partial\sigma^2} l(\mu,\sigma^2;x_1,\ldots,x_n) =& -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(\log(x_i)-\mu)^2 = 0 \\ \rightarrow \sigma^2 =& \frac{1}{n}\sum_{i=1}^{n}(\log(x_i)-\mu)^2 \\ =& \frac{1}{n}\sum_{i=1}^{n}\left(\log(x_i) - \frac{1}{n}\sum_{i=1}^{n}\log(x_i)\right)^2. \end{aligned}$$

Det betyr at MLE for μ og σ^2 er

$$\hat{\mu}_{mle} = \frac{1}{n}\sum_{i=1}^{n}\log(X_i) \text{ og } \widehat{\sigma^2}_{mle} = \frac{1}{n}\sum_{i=1}^{n}\left(\log(X_i) - \frac{1}{n}\sum_{i=1}^{n}\log(X_i)\right)^2.$$

Vi setter inn forsikringskravene, og får estimatene $\hat{\mu}_{mle} = 2.78$ og $\widehat{\sigma^2}_{mle} = 0.77$. Vi ser at disse ikke er så forskjellige fra momentestimatene på tross av at uttrykkene er vidt forskjellige.

b) Vi vet at $\log(X_i) = Y_i \overset{uif}{\sim} N(\mu, \sigma^2)$. Videre vet vi at MLE for forventning og varians i normalfordelingen er henholdsvis $\bar{Y}$ og $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Vi setter inn, og får:

$$\hat{\mu}_{mle} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n}\log(X_i)$$

$$\widehat{\sigma^2}_{mle} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \frac{1}{n}\sum_{i=1}^{n}Y_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\log(X_i) - \frac{1}{n}\sum_{i=1}^{n}\log(X_i)\right)^2,$$

som er nøyaktig samme uttrykk som vi fikk i a).

c) Vi har:

$$\log(f(x;\mu,\sigma)) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \log(x) - \frac{1}{2\sigma^2}(x-\mu)^2.$$

Det gir:

$$\begin{aligned}
\frac{\partial \log(f(x;\mu,\sigma))}{\partial \mu} &= \frac{1}{\sigma^2}(x-\mu) \\
\frac{\partial \log(f(x;\mu,\sigma))}{\partial (\sigma^2)} &= -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(\log(x)-\mu)^2 \\
\frac{\partial^2 \log(f(x;\mu,\sigma))}{\partial \mu^2} &= -\frac{1}{\sigma^2} \\
\frac{\partial^2 \log(f(x;\mu,\sigma))}{\partial \mu \partial (\sigma^2)} &= -\frac{1}{\sigma^4}(\log(x)-\mu) \\
\frac{\partial^2 \log(f(x;\mu,\sigma))}{\partial (\sigma^2)^2} &= \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(\log(x)-\mu)^2.
\end{aligned}$$

Da blir informasjonsmatrisa for én observasjon gitt ved:

$$I(\mu,\sigma^2) = \begin{pmatrix} I_{11}(\mu,\sigma^2) & I_{12}(\mu,\sigma^2) \\ I_{21}(\mu,\sigma^2) & I_{22}(\mu,\sigma^2) \end{pmatrix}$$

med

$$
\begin{aligned}
I_{11}(\mu,\sigma^2) &= -\,\mathrm{E}\left(\frac{\partial^2 \log(f(X;\mu,\sigma^2))}{\partial\mu^2}\right) = -\mathrm{E}\left(-\frac{1}{\sigma^2}\right) = \frac{1}{\sigma^2}\\
I_{12}(\mu,\sigma^2) &= -\,\mathrm{E}\left(\frac{\partial^2 \log(f(X;\mu,\sigma^2))}{\partial\mu\partial(\sigma^2)}\right) = -\mathrm{E}\left(-\frac{1}{\sigma^4}(\log(X)-\mu)\right)\\
&= \frac{1}{\sigma^4}(\mathrm{E}(\log(X)) - \mu)\\
&= \frac{1}{\sigma^4}(\mathrm{E}(Y) - \mu) = 0\\
I_{22}(\mu,\sigma^2) &= -\,\mathrm{E}\left(\frac{\partial^2 \log(f(X;\mu,\sigma^2))}{\partial(\sigma^2)^2}\right) = -\mathrm{E}\left(\frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(\log(X)-\mu)^2\right)\\
&= -\,\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}\mathrm{E}((Y-\mu)^2)\\
&= -\,\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}\mathrm{V}(Y)\\
&= -\,\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}\sigma^2 = \frac{1}{2\sigma^4},
\end{aligned}
$$

slik at

$$
I(\mu,\sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.
$$

Den inverse av informasjonsmatrisa er da

$$
I(\mu,\sigma^2)^{-1} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}.
$$

Vi vet da at for store n er

$$
\hat{\mu}_{mle} \overset{tiln.}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \text{ og } \widehat{\sigma^2}_{mle} \overset{tiln.}{\sim} N\left(\mu, \frac{2\sigma^4}{n}\right).
$$

Standardfeilen til de to estimatorene er da tilnærmet lik (for store n)

$$
\sigma_{\hat{\mu}_{mle}} \approx \frac{\sigma}{\sqrt{n}} \text{ og } \sigma_{\widehat{\sigma^2}_{mle}} \approx \sigma^2\sqrt{\frac{2}{n}}
$$

som kan estimeres med

$$
s_{\hat{\mu}_{mle}} = \sqrt{\frac{\widehat{\sigma^2}_{mle}}{n}} \text{ og } s_{\widehat{\sigma^2}_{mle}} = \widehat{\sigma^2}_{mle}\sqrt{\frac{2}{n}}.
$$

Når vi setter in forsikringsdataene får vi $s_{\hat{\mu}_{mle}} = 0.011$ og $s_{\widehat{\sigma^2}_{mle}} = 0.014$, som er svært lave på grunn av at n er såpass stor.
d) Vi bootstrapper $\hat{\mu}_{mle}$ og $\widehat{\sigma}^2_{mle}$ som følger:
For $b = 1, \ldots, B$

- Trekk $x^*_{b1}, \ldots, x^*_{bn}$ fra $x_1, \ldots, x_n$ med tilbakelegging.
- Beregn $\hat{mu}^*_{mle} = \frac{1}{n}\sum_{i=1}^n Y_i = \frac{1}{n}\sum_{i=1}^n \log(x^*_i)$ og $\widehat{\sigma}^{2,*}_{mle} = \frac{1}{n}\sum_{i=1}^n \left(\log(x^*_i) - \frac{1}{n}\sum_{i=1}^n \log(x^*_i)\right)^2$.

Standardfeilen til $\hat{\mu}_{mle}$ og $\widehat{\sigma}^2_{mle}$ kan da estimeres med

$$s_{\hat{\mu}_{mle}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (\hat{\mu}^*_{mle} - \bar{\hat{\mu}}^*_{mle})^2} \text{ og } s_{\widehat{\sigma^2}_{mle}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (\widehat{\sigma^2}^*_{mle} - \overline{\widehat{\sigma^2}}^*_{mle})^2},$$

der $\bar{\hat{\mu}}^*_{mle} = \frac{1}{n}\sum_{i=1}^n \hat{\mu}^*_{mle}$ og $\overline{\widehat{\sigma^2}}^*_{mle} = \frac{1}{n}\sum_{i=1}^n \widehat{\sigma^2}^*_{mle}$.

Vi lar $B = 1000$, setter inn forsikringskravene og får standardfeilene 0.011 og 0.014, som er det samme som vi fikk ved å bruke den tilnærmede fordelingen til maksimum likelihood-estimatoren for store utvalg i c). Det skyldes at n her er såpass stor at det er en god tilnærming.

e) Når utvalgsstørrelsen n er nokså stor, slik som for forsikringskravene, vil, ifølge sentralgrenseteoremet,

$$Z = \frac{\bar{X} - \phi}{S/\sqrt{n}} \overset{tiln.}{\sim} N(0,1),$$

der $\phi = \mathrm{E}(X_i)$ og $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$. Da har vi:

$$\begin{aligned} \mathrm{P}\left(-1.96 \leq \frac{\bar{X} - \phi}{S/\sqrt{n}} \leq 1.96\right) &\approx 0.95 \\ \rightarrow \mathrm{P}\left(\bar{X} - 1.96\frac{S}{\sqrt{n}} \leq \phi \leq \bar{X} + 1.96\frac{S}{\sqrt{n}}\right) &\approx 0.95, \end{aligned}$$

slik at et tilnærmet 95% konfidensintervall for $\mathrm{E}(X_i)$ er gitt ved

$$\bar{x} \pm 1.96\frac{s}{\sqrt{n}},$$

der $\bar{x}$ og s er observerte verdier av $\bar{X}$ og S. Vi setter inn for forsikringskravene og får konfidensintervallet $(23.4, 24.8)$ for $\mathrm{E}(X_i)$.

f) Dersom $X_1, \ldots, X_n$ uavhengige og indentisk normalfordelt, så er

$$\frac{(n-1)S^2}{\mathrm{V}(X_i)} \sim \chi^2_{n-1}.$$

Da er

$$\begin{aligned} \mathrm{P}\left(\chi^2_{0.975,n-1} \leq \frac{(n-1)S^2}{\mathrm{V}(X_i)} \leq \chi^2_{0.025,n-1}\right) &= 0.95 \\ \rightarrow \mathrm{P}\left(\frac{(n-1)S^2}{\chi^2_{0.025,n-1}} \leq \mathrm{V}(X_i) \leq \frac{(n-1)S^2}{\chi^2_{0.975,n-1}}\right) &= 0.95, \end{aligned}$$

slik at et 95% konfidensintervall for $\mathrm{V}(X_i)$ er gitt ved

$$\left(\frac{(n-1)s^2}{\chi^2_{0.025,n-1}}, \frac{(n-1)s^2}{\chi^2_{0.975,n-1}} \right).$$

Vi setter inn forsikringskravene og får estimatet $s^2 = 837$ og konfidensintervallet $(808, 866)$ for $\mathrm{V}(X_i)$.
g) Vi bruker estimatet S^2 for $\mathrm{V}(X_i)$, og vi bootstrapper dette som følger:
For $b = 1, \ldots, B$

- Trekk $x^*_{b1}, \ldots, x^*_{bn}$ fra $x_1, \ldots, x_n$ med tilbakelegging.
- Beregn $\bar{x}^* = \frac{1}{n}\sum_{i=1}^n x^*_{bi}$ og $s^{2,*}_b = \frac{1}{n-1}\sum_{i=1}^n (x^*_{bi} - \bar{x}^*)^2$.

Vi sorterer $s^{2,*}_1, \ldots, s^{2,*}_B$ i stigende rekkefølge

$$s^{2,*}_{(1)} \leq \ldots \leq s^{2,*}_{(B)},$$

og 95% persentilintervallet for $\mathrm{V}(X_i)$ er da gitt ved

$$(s^{2,*}_{(B\cdot 0.025)}, s^{2,*}_{(B\cdot 0.975)}).$$

Vi lar $B = 1000$, setter inn forsikringskravene og får konfidensintervallet $(689, 1025)$. Gjennomsnittet av bootstrap-estimatene $s^{2,*}_1, \ldots, s^{2,*}_B$ er 839, som er ganske nær $s^2 = 837$, så det ser ikke ut til å være noen skjevhet av betydning, som vi trenger å korrigere for.

Når vi sammenligner med intervallet fra g), basert på antakelsen om normalfordeling, ser vi at persentilintervallet er mye videre. I dette tilfellet bør vi stole mest på persentilintervallet, da det ikke baserer seg på antakelser om dataene som vi vet er feil, slik som det fra g). Vi ser altså at konfidensintervallet fra g), basert på en feilaktig antakelse om normalfordelte data, her gir en drastisk underestimering av usikkerheten rundt den faktiske størrelsen på variansen til forsikringskravene.

Oppgave 2

a) Vi antar at $X_1, \ldots, X_{15} \overset{uif}{\sim} N(\mu, \sigma^2)$, og da er

$$T = \frac{\bar{X} - \mu}{S/\sqrt{15}} \sim t_{15-1}$$

Da har vi:

$$\mathrm{P}\left(-t_{0.025,14} \leq \frac{\bar{X} - \mu}{S/\sqrt{15}} \leq t_{0.025,14} \right) = 0.95$$
$$\rightarrow \mathrm{P}\left(\bar{X} - t_{0.025,14}\frac{S}{\sqrt{15}} \leq \mu \leq \bar{X} + t_{0.025,14}\frac{S}{\sqrt{15}} \right) = 0.95,$$

som gir følgende 95% konfidensintervall for μ:

$$\bar{x} \pm t_{0.025,14} \frac{s}{\sqrt{15}}. \tag{1}$$

For antocyandataene i oppgaven gir det intervallet $(544, 575)$.
b) Vi har trukket 10000 datasett fra fordelingen $N(558, 30^2)$ og beregnet konfidensintervallet (1) for hvert av disse. Deretter har vi talt opp hvor mange av intervallene som inneholder den sanne verdien $\mu = 558$. Det var 9486, som gir en andel på omtrent 0.95. Et 95% konfidensintervall er nettopp definert slik at dersom en beregner dette intervallet for mange forskjellige tilfeldige utvalg, skal omtrent 95% av dem inneholde den sanne verdien av parameteren, i dette tilfellet μ. Vi får her bekreftet at det er tilfellet.
c) I oppgave 1c) er kvantilen $t_{0.025,14} = 2.14$ fra t-fordelingen byttet ut med tilsvarende kvantil i standard normalfordeling $z_{0.025} = 1.96$. Når n er stor, kan vi gjøre dette fordi t_{n-1}-fordelingen da er tilnærmet standardnormalfordelingen, men spørsmålet er om det gjelder her? Forskjellen mellom kvantilene tyder på at det ikke er tilfellet.

Vi har på nytt gerenert datasett som i b) og nå beregnet intervallet $\bar{x} \pm 1.96\frac{s}{\sqrt{15}}$ for hvert av dem. Nå var det 9298 som inneholdt $\mu = 558$, altså rundt 93%. Det er noe mindre enn det ønskede konfidensnivået på 95%. Det skyldes at $1.96 < t_{0.025,14} = 2.14$, slik at intervallene i denne deloppgaven blir smalere enn om vi hadde brukt t-fordelingen, som i b), noe som igjen fører til at færre av intervallene inneholder μ. Her er altså ikke $n = 15$ stor nok til at vi kan bruke konfidensintervallet for store utvalg.
d) Vi antar at $X_1, \ldots, X_{15} \overset{uif}{\sim} N(\mu, \sigma^2)$, og da vet vi fra oppgave 1g) at

$$\begin{aligned} \mathrm{P}\left(\frac{(n-1)S^2}{\chi^2_{0.025,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{0.975,n-1}} \right) &= 0.95 \\ \rightarrow \mathrm{P}\left(S\sqrt{\frac{(n-1)}{\chi^2_{0.025,n-1}}} \leq \sigma \leq S\sqrt{\frac{(n-1)}{\chi^2_{0.975,n-1}}} \right) &= 0.95 \end{aligned}$$

Vi får dermed følgende 95% konfidensintervall for σ

$$\left(s\sqrt{\frac{(n-1)}{\chi^2_{0.025,n-1}}}, s\sqrt{\frac{(n-1)}{\chi^2_{0.975,n-1}}} \right). \tag{2}$$

Vi har simulert data som i 1b), og beregnet konfidensintervallet over for hvert datasett. Det var 9476 intervaller som inneholdt $\sigma = 30$, hvilket tilsvarer en andel på rundt 95%. Det er som forventet tatt i betraktning at det er 95% konfidensintervaller og modellantakelsene stemmer.
e) Vi har nå trukket 10000 datasett som beskrevet i oppgaven, og beregnet konfidensintervallet fra (1) for hvert av dem. Nå var det 9539 som inneholdt μ, altså en andel på rundt 95%, som vi ønsker. Det tyder på at konfidensintervallet (1) er nokså robust overfor normalfordelingsantakelsen, altså at det kan gi nokså

fornuftige resultater selv om antakelsen om normalfordeling ikke stemmer helt, som hevdet på s. 406 i boka.
f) Vi har nå trukket 10000 datasett som beskrevet i oppgave 1e), og beregnet konfidensintervallet fra (2) for hvert av dem. Det er 8912 av dem som inneholder σ, altså rundt 89%. Det er betraktelig lavere enn det ønskede konfidensnivået på 95%. Det betyr at konfidensintervallet (2) ikke er særlig robust overfor normalfordelingsantakelsen, som nevnt på s. 410 i boka.

R-kode

```
### Oppgave 1

## Lese data

x = scan("https://www.uio.no/studier/emner/matnat/math/STK1110/
data/forsikringskrav.txt")
y = log(x)

## a)

n = length(x)
mu.mle = mean(log(x))
sigma2.mle = mean((log(x)-mu.mle)^2)
signif(c(mu.mle,sigma2.mle),3)

## c)

s.mu.mle = sqrt(sigma2.mle/n)
s.sigma2.mle = sigma2.mle*sqrt(2/n)
c(s.mu.mle,s.sigma2.mle)

## d)

B = 1000
mu.star = rep(0,B)
sigma2.star = rep(0,B)
for(b in 1:B)
{
  x.star = sample(x,n,replace=TRUE)
  mu.star[b] = mean(log(x.star))
  sigma2.star[b] = mean((log(x.star)-mu.star[b])^2)
}
s.mu.mle.boot = sd(mu.star)
s.sigma2.mle.boot = sd(sigma2.star)
round(c(s.mu.mle.boot,s.sigma2.mle.boot))
```

```
## e)

s = sd(x)
signif(x.bar+c(-1,1)*1.96*s/sqrt(n),3)

## f)

signif(c((n-1)*s^2/qchisq(0.025,df=n-1,lower.tail=FALSE),
(n-1)*s^2/qchisq(0.975,df=n-1,lower.tail=FALSE)),3)

## g)

B = 1000
s2.star = rep(0,B)
for(b in 1:B)
{
  x.star = sample(x,n,replace=TRUE)
  s2.star[b] = var(x.star)
}
s2.star.sort=sort(s2.star)
round(c(s2.star.sort[B*0.025],s2.star.sort[B*0.975]))

c(round(s^2),round(mean(s2.star)))

### Oppgave 2

## Lese data

antocyan = c(525,587,547,558,591,531,571,551,566,622,561,502,556,
565,562)

## a)

n = length(antocyan)
x.bar = mean(antocyan)
s = sd(antocyan)
t.crit = qt(0.975,n-1)
signif(x.bar+c(-1,1)*t.crit*s/sqrt(n),3)

## b)

N = 10000
mu = 558
sigma = 30
ci.mat = matrix(0,N,2)
```

```
for(i in 1:N)
{
  x = rnorm(n,mu,sigma)
  x.bar = mean(x)
  s = sd(x)
  ci.mat[i,] = x.bar+c(-1,1)*t.crit*s/sqrt(n)
}
mu.in.int = (ci.mat[,1] <= mu)&(ci.mat[,2] >= mu)
mean(as.numeric(mu.in.int))

## c)

ci.mat.alt = matrix(0,N,2)
for(i in 1:N)
{
  x = rnorm(n,mu,sigma)
  x.bar = mean(x)
  s = sd(x)
  ci.mat.alt[i,] = x.bar+c(-1,1)*1.96*s/sqrt(n)
}
mu.in.int.alt = (ci.mat.alt[,1] <= mu)&(ci.mat.alt[,2] >= mu)
mean(as.numeric(mu.in.int.alt))

## d)

ci.mat.sigma = matrix(0,N,2)
for(i in 1:N)
{
  x = rnorm(n,mu,sigma)
  s = sd(x)
  ci.mat.sigma[i,] = s*sqrt(n-1)/sqrt(c(qchisq(0.975,n-1),qchisq(0.025,n-1)))
}
sigma.in.int = (ci.mat.sigma[,1] <= sigma)&(ci.mat.sigma[,2] >= sigma)
mean(as.numeric(sigma.in.int))

## e)

nu = 7
ci.mat.t = matrix(0,N,2)
for(i in 1:N)
{
  t = rt(n,nu)
  x = mu+sigma*t
  x.bar = mean(x)
  s = sd(x)
  ci.mat.t[i,] = x.bar+c(-1,1)*t.crit*s/sqrt(n)
```

```
}
mu.in.int.t = (ci.mat.t[,1] <= mu)&(ci.mat.t[,2] >= mu)
mean(as.numeric(mu.in.int.t))

## f)

ci.mat.sigma.t = matrix(0,N,2)
for(i in 1:N)
{
  t = rt(n,nu)
  x = mu+sigma*t
  s = sd(x)
  ci.mat.sigma.t[i,] = s*sqrt(n-1)/sqrt(c(qchisq(0.975,n-1),qchisq(0.025,n-1)))
}
sigma.in.int.t = (ci.mat.sigma.t[,1] <= sigma*sqrt(nu/(nu-2)))&
(ci.mat.sigma.t[,2] >= sigma*sqrt(nu/(nu-2)))
mean(as.numeric(sigma.in.int.t))
```