

Oblig 1 - STK 1110

Kevin Alexander Aslesen

Opgave 1

- a) Tettuets funksjonen for en log-normal fordeling er:

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

Gitt uavhengighet er da joint tettuets funksjonen lik produktet av disse:

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

Dette er det vi kaller likelihood-funksjonen. Men

presist er likelihood-funksjonen for μ og σ^2 like:

$$L(\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i=1}^n \left(\frac{1}{x_i}\right) e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (\ln(x_i) - \mu)^2}$$

Tar logaritmen av denne for å få log-likelihood funksjonen:

$$l(\mu, \sigma^2) = \ln(L(\mu, \sigma^2))$$

$$= \ln \left(\frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i=1}^n \left(\frac{1}{x_i}\right) e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (\ln(x_i) - \mu)^2} \right)$$

$$= -\ln((2\pi\sigma^2)^{n/2}) + \ln\left(\prod_{i=1}^n \frac{1}{x_i}\right) + \ln(e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (\ln(x_i) - \mu)^2})$$

$$= -\frac{n}{2} \ln(2\pi\sigma^2) + \sum_{i=1}^n \ln\left(\frac{1}{x_i}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^n (\ln(x_i) - \mu)^2$$

$$= -\frac{n}{2} \ln(\pi) - \frac{n}{2} \ln(\sigma^2) - \sum_{i=1}^n \ln(x_i) - \frac{1}{2\sigma^2} \sum_{i=1}^n (\ln(x_i) - \mu)^2$$

Deriverer log-likelihood funksjonen for μ :

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= -0 - 0 - 0 - \frac{\partial}{\partial \mu} \left(\frac{1}{2\sigma^2} \sum_{i=1}^n (\ln(x_i) - \mu)^2 \right) \\ &= -\frac{1}{2\sigma^2} \sum_{i=1}^n (-1) 2(\ln(x_i) - \mu) \end{aligned}$$

product + negelen

$$\begin{aligned}\frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^n (\ln(x_i) - \mu) \\ &= \frac{1}{\sigma^2} \left(\sum_{i=1}^n \ln(x_i) - n\mu \right)\end{aligned}$$

$\left(\sum_{i=1}^n \mu = n\mu \right)$

Setter den deriverte lik null for å finne $\hat{\mu}_{MLE}$:

$$\frac{\partial l}{\partial \mu} = 0$$

$$\frac{1}{\sigma^2} \left(\sum_{i=1}^n \ln(x_i) - n\mu \right) = 0$$

$$\sum_{i=1}^n \ln(x_i) - n\mu = 0$$

$$n\mu = \sum_{i=1}^n \ln(x_i)$$

$$\mu = \frac{1}{n} \sum_{i=1}^n \ln(x_i)$$

altså $\hat{\mu}_{MLE} = \underline{\underline{\frac{1}{n} \sum_{i=1}^n \ln(x_i)}}$.

Derivere si log-likelihod funksjonen for $\theta = \sigma^2$:

$$\begin{aligned}\frac{\partial l}{\partial \theta} &= -\frac{n}{2\theta} + \frac{1}{2\theta^2} \sum_{i=1}^n (\ln(x_i) - \mu)^2 \\ &= -\frac{n}{2\theta} + \frac{1}{2\theta^2} \sum_{i=1}^n (\ln(x_i) - \hat{\mu})^2\end{aligned}$$

Setter den deriverte lik null for å finne $\hat{\sigma}_{MLE}^2$:

$$\frac{\partial \ell}{\partial \theta} = 0$$

$$-\frac{n}{2\theta} + \frac{1}{2\theta^2} \sum_{i=1}^n (\ln(x_i) - \mu)^2 = 0$$

$$\frac{1}{\theta} \sum_{i=1}^n (\ln(x_i) - \mu)^2 = n$$

$$\theta = \frac{1}{n} \sum_{i=1}^n (\ln(x_i) - \mu)^2$$

Med $\theta = \sigma^2$ så har vi at $\hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} \sum_{i=1}^n (\ln(x_i) - \mu)^2$.

b) Vi vet at $y_i \sim N(\mu, \sigma^2)$ og at maksimum likelihood

estimatorene for μ og σ^2 i en normal fordeling er gitt

ved $\hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n y_i$ og $\hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \mu)^2$. Men siden

$y_i = \log(x_i) = \ln(x_i)$, så kan vi skrive disse om

som:

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n \ln(x_i)$$

$$\hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} \sum_{i=1}^n (\ln(x_i) - \mu)^2$$

Dette var mye enklere enn metoden i a) ...

c) For én enkelt observasjon av log-likelihood lik:

$$\ln(f(x; \mu, \theta)) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\theta) - \ln(x_1) - \frac{1}{2\theta}(\ln(x) - \mu)^2$$

hvor $\theta = \sigma^2$. Finne de forstegjøllige deriverte av denne

mhlp. μ og θ :

$$\begin{aligned} \frac{\partial}{\partial \mu} \ln(f(x; \mu, \theta)) &= -\frac{\partial}{\partial \mu} \left(\frac{1}{2\theta} (\ln(x) - \mu)^2 \right) \\ &= -\frac{1}{2\theta} (-1) 2 (\ln(x) - \mu) \\ &= \frac{1}{\theta} (\ln(x) - \mu) \end{aligned}$$

product
regel

$$\begin{aligned} \frac{\partial}{\partial \theta} \ln(f(x; \mu, \theta)) &= -\frac{\partial}{\partial \theta} \left(\frac{1}{2} \ln(\theta) \right) - \frac{\partial}{\partial \theta} \left(\frac{1}{2\theta} (\ln(x) - \mu)^2 \right) \\ &= -\frac{1}{2\theta} + \frac{1}{2\theta^2} (\ln(x) - \mu)^2 \end{aligned}$$

$$\begin{aligned} \frac{\partial^2}{\partial \mu^2} \ln(f(x; \mu, \theta)) &= \frac{\partial}{\partial \mu} \left(\frac{1}{\theta} (\ln(x) - \mu) \right) \\ &= -\frac{1}{\theta} \end{aligned}$$

$$\begin{aligned} \frac{\partial^2}{\partial \theta^2} \ln(f(x; \mu, \theta)) &= \frac{\partial}{\partial \theta} \left(-\frac{1}{2\theta} + \frac{1}{2\theta^2} (\ln(x) - \mu)^2 \right) \\ &= \frac{1}{2\theta^2} - \frac{1}{\theta^3} (\ln(x) - \mu)^2 \end{aligned}$$

$$\begin{aligned}
 \frac{\partial^2}{\partial \mu \partial \theta} \ln(f(x; \mu, \theta)) &= \frac{\partial}{\partial \mu} \left(\frac{\partial}{\partial \theta} \ln(f(x; \mu, \theta)) \right) \\
 &= \frac{\partial}{\partial \theta} \left(\frac{1}{\theta} (\ln(x) - \mu) \right) \\
 &= -\frac{1}{\theta^2} (\ln(x) - \mu)
 \end{aligned}$$

Observasjons matrisen for én observasjon er gitt ved:

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

$$\begin{aligned}
 \text{hvor } I_{11}(\mu, \theta) &= -E \left(\frac{\partial^2}{\partial \mu^2} \ln(f(x; \mu, \theta)) \right) \\
 &= -E \left(-\frac{1}{\theta} \right) \\
 &= \frac{1}{\theta}
 \end{aligned}$$

$$\begin{aligned}
 I_{12}(\mu, \theta) &= I_{21}(\mu, \theta) = -E \left(\frac{\partial^2}{\partial \mu \partial \theta} \ln(f(x; \mu, \theta)) \right) \\
 &= -E \left(-\frac{1}{\theta^2} (\ln(x) - \mu) \right) \\
 &= \frac{E(\ln(x) - \mu)}{\theta^2} \\
 &= \frac{E(\ln(x)) - \mu}{\theta^2} \\
 &= \frac{E(y) - \mu}{\theta^2} \\
 &= \frac{\mu - \mu}{\theta^2} = 0
 \end{aligned}$$

$$\begin{aligned}
I_{zz} &= -E \left(\frac{\partial^2}{\partial \theta^2} \ln f(x_i | \mu, \sigma^2) \right) \\
&= -E \left(\frac{1}{z\theta^2} - \frac{1}{2\theta^3} (\ln(x) - \mu)^2 \right) \\
&= -\frac{1}{\theta^2} + \frac{1}{2\theta^3} E((\ln(x) - \mu)^2) \\
&= -\frac{1}{\theta^2} + \frac{1}{2\theta^3} E((y - \mu)^2) \\
&= -\frac{1}{\theta^2} + \frac{1}{2\theta^3} V(y) \quad \text{hvor } V(y) = \sigma^2 = \theta \\
&= -\frac{1}{\theta^2} + \frac{1}{2\theta^2} \\
&= \frac{1}{2\theta^2}
\end{aligned}$$

Observasjonsmatrisen er da:

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

setter vi inn σ^2 for θ så har vi matrisen:

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

Siden n er stor og X_1, \dots, X_n ikke er avhengig av μ og σ^2 , så betyr det at maksimum likelihood estimatorene $\hat{\mu}$ og $\hat{\sigma}^2$ har tilnærmet normalfordeling med henholdsvis forventning μ og σ^2 , samt henholdsvis

varians $\frac{1}{n} I_{11}(\mu, \sigma^2)$ og $\frac{1}{n} I_{22}(\mu, \sigma^2)$.

Et estimat på standardfeilen til $\hat{\mu}_{MLE}$ og $\hat{\sigma}_{MLE}^2$ er da:

$$\begin{aligned}\sigma_{\hat{\mu}_{MLE}} &= \sqrt{\frac{1}{n} I_{11}(\mu, \sigma^2)} \\ &= \sqrt{\frac{1}{n} \left(\frac{1}{\sigma^2} \right)} \\ &= \underline{\underline{\frac{\sigma}{\sqrt{n}}}}\end{aligned}$$

$$\begin{aligned}\sigma_{\hat{\sigma}_{MLE}^2} &= \sqrt{\frac{1}{n} I_{22}(\mu, \sigma^2)} \\ &= \sqrt{\frac{1}{n} \left(\frac{1}{2\sigma^4} \right)} \\ &= \underline{\underline{\frac{\sigma^2}{\sqrt{n}}}}\end{aligned}$$

d) Leser inn datuen om forsikringskravene:

```
url = "https://www.uio.no/studier/emner/matnat/math/STK1110/data/forsikringskrav.txt"
data_table = read.table(url, header = FALSE)
data = data_table$V1
n = length(data)
```

henger en bootstrap fordeling for forventninga μ for å estimera standardfeilen i $\hat{\mu}_{MLE}$. Sammenliknar dette med det vi filede i c).

```

B = 1000
data_mean_boot = c()
for (i in 1:B){
  data_boot = c()
  for (i in 1:n){
    krav = sample(data, 1)
    data_boot = append(data_boot, krav)
  }
  data_mean_boot = append(data_mean_boot, mean(data_boot))
}
print(mean(exp(data_mean_boot)))
print(sd(exp(data_mean_boot)))
print(mean(data))
print(sd(data)/sqrt(n))

```

Før at standardfeilen for $\hat{\mu}_{MLE}$ ved bootstrap er $S_{\hat{\mu}_{MLE}} = 0.366$,

mens for estimeringen fra c) er $\sigma_{\hat{\mu}_{MLE}} = 0.362$. Her

stemmer altså teorien med målingene.

Gjør tilsvarende for $\hat{\sigma}^2_{MLE}$:

```

B = 1000
data_sd_boot = c()
for (i in 1:B){
  data_boot = c()
  for (i in 1:length(data)){ # nolint
    krav = sample(data, 1)
    data_boot = append(data_boot, krav)
  }
  data_sd_boot = append(data_sd_boot, sd(data_boot))
}
print(sd(data_sd_boot))
print(sqrt(2/length(data)) * sd(data)^2)

```

Her får vi at $S_{\hat{\sigma}^2_{MLE}} = 1.56$, mens for estimeringen

fra c) er $\sigma_{\hat{\sigma}^2_{MLE}} = 14.81$. Dette var veldig stor forskjell,

som betyr at teorien ikke stemmer med målingene.

Vil tro det er utykket fra c) det er noe feil med siden den er så stor.

e) Lager et 95 % konfidensintervall for $E(x_i)$.

Antar at $E(x_i) = \mu$ eftersom n er stor. Da vil

intervallet se slik ut:

$$\bar{x} \pm t_{0.025, n-1} \cdot \frac{s}{\sqrt{n}}$$

Den kritiske t-verdien kan tilnærmes som $t_{0.025, n-1} \xrightarrow{n \rightarrow \infty} 1.960$

siden n er stor. Fra målingene har vi $\bar{x} = 24.14$

og $s = 28.92$. Da får vi følgende:

$$24.14 \pm 0.71$$

altså intervallet er:

$$\underline{(23.43, 24.85)}$$

Oppgave 2

$$n=15$$

a) Snittet av målingene er $\bar{x} = 559.7$ og standardavviket er $s = 28.6$. Med disse verdiene og med antagelsen om at populasjonen er normalfordelt med forventning μ og varians σ^2 , så lager vi et 95% konfidens-intervall for forventning μ basert på målingene våre med:

$$\left(\bar{x} - t_{0.025, n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{0.025, n-1} \frac{s}{\sqrt{n}} \right)$$

hvor den kritiske t-verdien ved $\alpha = 0.025$ og frihetsgrader $df = n-1 = 15-1 = 14$ er $t_{0.025, 14} = 2.145$.

Settar inn alle verdiene og får intervallet:

$$\underline{(543.9, 575.5)}$$

Vi kan da si med 95% sikkerhet at den virkelige verdien for μ vil ligge mellom 543.9 mg/100g og 575.5 mg/100g.

b) Begynner først med å liste opp målingene og finne lengden som vi vet er $n = 15$:

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

Lager en $m = 10000$ datasett, hvor hvert datasett inneholder de stokastiske variablene $X_1, X_2, \dots, X_{15} \stackrel{\text{uif}}{\sim} N(588, 30^2)$ ved hjelp av funksjonen `rnorm(15, 588, 30)`.

For hvert av disse datasettene lager vi et 95%

konfidensintervall for μ som i a) og teller hvor

mange av disse intervallene som inneholder den virkelige

verdien $\mu = 558$:

```
set.seed(1)
m <- 10000
count <- 0
for (i in 1:m) {
  sim_data <- rnorm(n, 558, 30)
  sim_mean <- mean(sim_data)
  sim_sd <- sd(sim_data)
  t_0.025 <- 2.145
  conf_low <- sim_mean - t_0.025 * sim_sd / sqrt(n)
  conf_high <- sim_mean + t_0.025 * sim_sd / sqrt(n)
  if (conf_low < 558 && conf_high > 558) {
    count <- count + 1
  }
}
print(count)
```

Dette ga count = 9527.

Vi fikk også et 95.27 % av konfidensintervallene inneholdte $\mu = 558$. Dette stemmer godt med teorien bak konfidensintervall, hvor "95% sikkerhet" tilsvarer at 95 % av intervallene vi får ved mange målinger vil inneholde den virkelige verdien for μ .

c) Det tilnærmede intervallet for store utvalg er:

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

Dette er det samme som 95% intervallet, utenom at 1.96 er satt inn for t-verdien. Dette skyldes at grensverdien for t-verdien ved $\frac{\alpha}{2} = 0.025$ som går mot 1.96 når $n \rightarrow \infty$ ($t_{0.025, \infty} \stackrel{n \rightarrow \infty}{=} 1.96$). Lager det tilnærmede intervallet for de 10000 datasettene som i b) og finner antallet av dem som inneholder $\mu = 558$:

```

m <- 10000
count <- 0
for (i in 1:m) {
  sim_data <- rnorm(n, 558, 30)
  sim_mean <- mean(sim_data)
  sim_sd   <- sd(sim_data)
  conf_low <- sim_mean - 1.96 * sim_sd / sqrt(n)
  conf_high <- sim_mean + 1.96 * sim_sd / sqrt(n)
  if (conf_low < 558 && conf_high > 558) {
    count <- count + 1
  }
}
print(count)

```

Dette gir count = 9357. Altså 93.57 % av de tilnærmede intervallene for store utvalg inneholdte $\mu = 558$. Det er 1.7% lavere enn det vi fikk i b). Dette gir mening ettersom vi effektivt sett har gjort intervallene smalere med 1.96 istedenfor 2.145, som betyr at fårrer intervaller vil inneholde $\mu = 558$.

d) Siden vi antar at populasjonen er normalfordelt så finner vi et 95 % konfidensintervall for variansen σ^2 ved:

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

hvor de kritiske chi-verdiene er $\chi^2_{0.025, 14} = 26.119$
 og $\chi^2_{0.975, 14} = 5.629$. Kvadratet av dette intervallet
 gir 95 % konfidens intervall for σ . Regner dette
 for $n=10000$ datasett som i b) og teller hvor
 mange av disse intervallene som inneholder den
 virkelige verdien $\sigma = 30$:

```
m <- 10000
count <- 0
for (i in 1:m) {
  sim_data <- rnorm(n, 558, 30)
  sim_sd <- sd(sim_data)
  chi_low <- 26.119
  chi_high <- 5.692
  conf_low <- sqrt((n-1) / chi_low * sim_sd^2)
  conf_hig <- sqrt((n-1) / chi_high * sim_sd^2)
  if (conf_low < 30 && conf_hig > 30) {
    count <- count + 1
  }
}
print(count)
```

Dette gir $count = 9482$. Altså 94.82 % av
 konfidensintervallene inneholder $\sigma = 30$. Dette
 stemmer godt med det vi ønsket å få.

c) Vi kjenner at $Z_i = \frac{X_i - \mu}{\sigma}$. Frem til nå har vi antatt at $Z_i \sim N(0,1)$, men nå antar vi at $Z_i \stackrel{\text{u.f.}}{\sim} t_7$ altså Z_i er t-fordelt med 7 frihetsgrader. Med andre ord antar vi at populasjonen ikke er normalfordelt. Vi skal se hvor stor effekt dette har på antall konfidensintervaller som inneholder den virkelig verdien $\mu = 558$. Først lager vi en tilfeldig variabel t -fordeling med 7 frihetsgrader med $n = 15$ punkter ved hjelpe funksjonen $rt(n, df=7)$ og dessetter gjør den om med $X_i = \mu + \sigma \cdot Z_i = 558 + 30Z_i$.

Regner så konfidensintervaller som i b):

```
m <- 10000
count <- 0
for (i in 1:m){
  n <= 15
  tdist <- rt(n, df=7)
  xdist <- 558 + 30 * tdist
  x_mean <- mean(xdist)
  x_sd <- sd(xdist)
  t_0.025 <- 2.145
  conf_low <- x_mean - t_0.025 * x_sd / sqrt(n)
  conf_high <- x_mean + t_0.025 * x_sd / sqrt(n)
  if (conf_low < 558 && conf_high > 558) {
    count <- count + 1
  }
}
print(count)
```

Dette ga count = 9518. Altså 95.18 % av konfidensintervallene inneholder $\mu = 558$. Prosentantallet vi fikk i b) var 95.27 %. Vi ser altså at det er liten forskjell for prosentantallene mellom normal-antagelse og t₂-antagelse.

Med dette kan vi si at metoden for å lage konfidensintervall for μ med antagelse om normalfordeling er robust. Gir mening ettersom t-fordelingen har like midtpunkt som normalfordelingen.

f) laget $n=10000$ datasett som i e) og lager et 95 % konfidensintervall for standardavviket til X_i for hvert datasett som i d). Teller antall intervaller som innholder $\tilde{\sigma} = \sqrt{1.4}\sigma = 30\sqrt{1.4}$:

```
m <- 10000
count <- 0
for (i in 1:m) {
  n <= 15
  tdist <- rt(n, df=7)
  xdist <- 558 + 30 * tdist
  x_sd <- sd(xdist)
  chi_low <- 26.119
  chi_high <- 5.692
  conf_low <- sqrt((n-1) / chi_low * x_sd^2)
  conf_high <- sqrt((n-1) / chi_high * x_sd^2)
  if (conf_low < sqrt(1.4)*30 && conf_high > sqrt(1.4)*30) {
    count <- count + 1
  }
}
print(count)
```

Dette ga count = 8887. Altså 88.87% av intervallene inneholdte $\hat{\sigma} = 30\sqrt{1.41}$. Fra d) finne vi prosentandel ved normal fordeling til $\hat{\sigma}$ over 94.82 %. Dette er merkbar forskjell. Men dette gir mening ettersom t-fordelingen er en "klemt" versjon av normalfordelingen med høyere standard avvik. Dette betyr at metoden for å lage konfidensintervall for σ hvor vi antar normal fordeling ikke er robust.