STK1110
Statistiske metoder og dataanalyse

OBLIG 1

Egil Furnes
Student: 693784

Problem 1

a)

Vi skal utlede moment-estimatorene for α og γ.

Vi har for $X \sim \text{Gamma}(\alpha, \gamma)$:

$$\mathbb{E}[X] = \frac{\alpha}{\gamma} \quad V[X] = \frac{\alpha}{\gamma^2}$$

Setter lik $\overline{x}$ og s^2:

$$\hat{\gamma} = \frac{\overline{x}}{s^2} \quad \hat{\alpha} = \frac{\overline{x}^2}{s^2}$$

Vi kan da bruke en parameterisering med skala $\beta = 1/\gamma$:

$$\hat{\alpha} = \frac{\overline{x}}{s^2} \quad \hat{\beta} = \frac{s^2}{\overline{x}}$$

```r
x <- scan("data/forsikringskrav.txt")
xb <- mean(x)
s2 <- mean((x-xb)^2)
alpha <- xb^2/s2
gamma <- xb/s2
beta <- s2/xb
```

```
> xb
[1] 24.13847
> s2
[1] 836.3836
> alpha
[1] 0.696649
> gamma
[1] 0.02886053
> beta
[1] 34.6494
```

Følgende har vi funnet:

$$\overline{x} = 24.14 \quad s^2 = 836.38 \quad \hat{\alpha} = 0.69 \quad \hat{\gamma} = 0.0289 \quad \hat{\beta} = 34.65$$

b)

For $i.i.d.$ $x_1, \ldots, x_n \sim \text{Gamma}(\alpha, \gamma)$ med tetthet $f(x; \alpha, \gamma) = \frac{\gamma^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\gamma x} \quad \alpha, \gamma > 0$ er log-likelihood-funksjonen:

$$l(\alpha, \gamma) = n\alpha \log \gamma - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \gamma \sum_{i=1}^{n} x_i$$

Vi har også $\hat{\alpha} = \overline{x}^2/s^2$ og $\hat{\gamma} = \overline{x}/s^2$ som gir:

```
n <- length(x)
logL <- n*alpha*log(gamma) - n*lgamma(alpha) + (alpha-1)*sum(log(x))
    - gamma*sum(x)
```

```
> n
[1] 6377
> logL
[1] -27264.51
```

Med dette har vi funnet at log-likelihood verdien er:

$$\log \mathrm{L} = -27264.51$$

c)

Fra tidligere har vi uttrykket for log L

$$l(\alpha, \gamma) = n\alpha \log \gamma - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \gamma \sum_{i=1}^{n} x_i$$

Nå kan vi derivere med hensyn på γ og sette lik 0 for å finne maksimum likelihood.

$$\frac{\partial l}{\partial \gamma} = \frac{n\alpha}{\gamma} - \sum_{i=1}^{n} x_i$$

$$\Rightarrow \hat{\gamma}(\alpha) = \frac{n\alpha}{\sum_i x_i} = \frac{\alpha}{x}$$

For å sammenligne med moment-estimatoren for γ for gitt α:

$$\overline{x} = \frac{\alpha}{\gamma} \quad \Rightarrow \quad \gamma = \frac{\alpha}{\overline{x}}$$

Som er samme uttrykk som

$$\hat{\gamma}_{MLE}(\alpha) = \hat{\gamma}_{MOM}(\alpha) = \frac{\alpha}{\overline{x}}$$

Kan også sjekke at det er et maksimum med

$$\frac{\partial^2 l}{\partial \gamma^2} = -\frac{n\alpha}{\gamma^2} < 0 \quad \text{for} \quad \alpha, \gamma > 0$$

d)

For denne funksjonen bør vi bruke $\log(\alpha)$ er for å sikre at $\alpha > 0$ uten å måtte begrense optimieringen på noe vis.

```r
alpha0 <- xb^2/s2
gamma0 <- xb/s2

negloglikgamma <- function(logalpha, x = x)
{
  n <- length(x)
  alpha <- exp(logalpha)
  gamma <- alpha / mean(x)  # MLE for gamma gitt alpha
  logL <- n*alpha*log(gamma) - n*lgamma(alpha) +
    (alpha - 1)*sum(log(x)) - gamma*sum(x)
  -logL                     # returnér negativ logL
}

fit.ml <- optim(log(alpha0), negloglikgamma, x = x, method = "BFGS")

alpha_mle <- exp(fit.ml$par)
gamma_mle <- alpha_mle / xb
logL_mle  <- -fit.ml$value
```

```
> alpha_mle
[1] 1.388346
> gamma_mle
[1] 0.05751589
> logL_mle
[1] -26488.28
> logL
[1] -27264.51
```

Her finner vi likelihood-verdiene som ser rimelige ut i forhold til den vi fant i b).

$$\log \mathrm{L} = -27264.51 \quad \log \mathrm{L}_{MLE} = -26488.28$$

e)

```r
set.seed(1705)
B <- 2000
boot <- replicate(B, {
  xb_ <- mean(xb <- sample(x, n, TRUE))
  s2_ <- mean((xb - xb_)^2)
  a0  <- xb_^2 / s2_
  a   <- exp(optim(log(a0), negloglikgamma, x = xb, method = "BFGS")
     $par)
  c(alpha = a, gamma = a / xb_)
})

alpha_se <- sd(boot["alpha", ])
gamma_se <- sd(boot["gamma", ])
alpha_ci <- quantile(boot["alpha", ], c(.025, .975))
gamma_ci <- quantile(boot["gamma", ], c(.025, .975))
```

```
> alpha_se
[1] 0.03080641
> gamma_se
[1] 0.002002218
> alpha_ci
    2.5%    97.5%
1.331434 1.449643
> gamma_ci
      2.5%      97.5%
0.05383500 0.06154386
```

Med dette finner vi følgende 95% konfidens-intervaller for α og γ

$$\begin{aligned} \hat{\alpha}_{CI} &\in [1.331,\ 1.450] \\ \hat{\gamma}_{CI} &\in [0.0538,\ 0.0615] \end{aligned}$$

f)

```r
set.seed(1705)
B <- 2000
mu_hat <- mean(x)
mu_boot <- replicate(B, mean(sample(x, n, TRUE)))
mu_ci <- quantile(mu_boot, c(.025, .975))

mu_hat
mu_ci

# decisions via CI inclusion of 25
rej_05 <- !(25 >= mu_ci[1] & 25 <= mu_ci[2])
rej_01 <- !(25 >= quantile(mu_boot, c(.005, .995))[1] &
              25 <= quantile(mu_boot, c(.005, .995))[2])
rej_05
rej_01
```

```
> rej_05
2.5%
TRUE
> rej_01
 0.5%
FALSE
```

Vi ser det at `rej_05` gir $TRUE$ men `rej_01` gir $FALSE$, og dermed er det naturlig å tro at P-veriden for testen er et sted $0.01 < p < 0.05$.

Problem 2

a)

Vi finner et 95% konfidensintervall

$$\overline{X} \pm t_{0.975,14} \frac{S}{\sqrt{15}}$$

```r
x <- c(525,587,547,558,591,531,571,551,566,622,561,502,556,565,562)
n <- length(x)
xb <- mean(x)
s <- sd(x)
tcrit <- qt(.975, df = n-1)
lo <- xb - tcrit * s / sqrt(n)
hi <- xb + tcrit * s / sqrt(n)
```

```
> xb
[1] 559.6667
> s
[1] 28.55988
> lo
[1] 543.8507
> hi
[1] 575.4826
```

Med dette finner vi

$$\bar{x} \approx 559.67, \quad S \approx 28.56, \quad \mu \in [543.85,\ 575.48]$$

b)

I denne oppgaven generer jeg 1000 datasett med størrelse $n = 15$ bestående av stokastiske variabler $X_1, \ldots, X_{15} \overset{\text{uif}}{\sim} N(558, 30^2)$ med `rnorm()`.

```r
set.seed(1705)
B <- 10000
n <- 15
mu <- 558
sigma <- 30

cov <- replicate(B, {
  y <- rnorm(n, mu, sigma)
  xb <- mean(y); s <- sd(y)
  tcrit <- qt(.975, df = n - 1)
  lo <- xb - tcrit * s / sqrt(n); hi <- xb + tcrit * s / sqrt(n)
  (mu >= lo) & (mu <= hi)
})

```

```r
cover <- mean(cov)
se_mc <- sqrt(cover * (1 - cover) / B)
```

Fra utskriften nedenfor finner vi at simuleringen ga en dekning som nærmer seg veldig det teoretiske konfidensintervallet $0.9534 \approx 0.95$. Dette er å forvente i en Monte Carlo simulering, ettersom vi har introdusert ikke-reduserbart støy med en standardfeil på omtrent 0.0021.

```
> cover
[1] 0.9534
> se_mc
[1] 0.002107805
```

Andelen $\approx 95\%$ av disse simuleringene inneholder den sanne verdien $\mu = 558$ som stemmer overens med forventet antocyaninnhold for blåbær.

c)

I denne deloppgaven benytter vi heller det tilnærmede utvalget

$$\left(\overline{X} - 1.96\frac{S}{\sqrt{15}}, \quad \overline{X} + 1.96\frac{S}{\sqrt{15}}\right)$$

med

$$\bar{X} = \frac{1}{15}\sum_{i=1}^{15} X_i \quad S^2 = \frac{1}{15-1}\sum_{i=1}^{15}(X_i - \bar{X})^2$$

```r
set.seed(1705)
cov_z <- replicate(B, {
  y <- rnorm(n, mu, sigma)
  xb <- mean(y)
  s <- sd(y)
  lo <- xb - 1.96 * s / sqrt(n)
  hi <- xb + 1.96 * s / sqrt(n)
  (mu >= lo) & (mu <= hi)
})
cover_z <- mean(cov_z)
se_mc_z <- sqrt(cover_z * (1 - cover_z) / B)
```

Med denne fremgangsmåten finner vi at 93.34% av intervallene inneholder $\mu = 558$, i motsetning til forrige metode som ga 95.34%. Med dette kan vi konkludere at å bruke 1.96 heller enn $t_{0.975,14}$ blir et for smalt intervall.

```
> cover_z; se_mc_z
[1] 0.9334
[1] 0.00249328
```

d)

Nå ser vi heller på σ og hvor mange intervaller som inneholder $\sigma = 30$.

```
set.seed(1705)
cov_sigma <- replicate(B, {
  y <- rnorm(n, mu, sigma)
  s <- sd(y)
  lo <- sqrt((n - 1) * s^2 / qchisq(.975, df = n - 1))
  hi <- sqrt((n - 1) * s^2 / qchisq(.025, df = n - 1))
  (sigma >= lo) & (sigma <= hi)
})
cover_sigma <- mean(cov_sigma)
se_mc_sigma <- sqrt(cover_sigma * (1 - cover_sigma) / B)
```

Med denne fremgangsmåten finner vi at 94.89% av intervallene inneholder $\sigma = 30$ med et standardavvik på omtrent 0.002.

```
> cover_sigma; se_mc_sigma
[1] 0.9489
[1] 0.002202017
```

e)

Nå gjentar vi oppgave b) med de nye datasettene, med antakelse om at $Z_1, \ldots, Z_15 \overset{\text{uif}}{\sim} t_7$.

```
set.seed(1705)
cov_t7 <- replicate(B, {
  z <- rt(n, df = 7)
  y <- mu + sigma * z
  xb <- mean(y)
  s <- sd(y)
  tcrit <- qt(.975, df = n - 1)
  lo <- xb - tcrit * s / sqrt(n); hi <- xb + tcrit * s / sqrt(n)
  (mu >= lo) & (mu <= hi)
})
cover_t7 <- mean(cov_t7)
se_mc_t7 <- sqrt(cover_t7 * (1 - cover_t7) / B)
```

Med dette finner vi at dekningen er 95.31% altså tilbake på statistisk signifikansnivå 5%. Vi ser det at selv om data ikke er normalfordelte her, men heller t_7-fordelt med tyngre haler er middelverdien relativt robust allerede for $n = 15$.

```
> cover_t7; se_mc_t7
[1] 0.9531
[1] 0.002114247
```

f)

Vi lager nå et konfidensintervall for $\widetilde{\sigma}$ og undersøke andelen som inneholder denne gitt

$$V(Z_i) = \frac{7}{7-2}$$

$$\widetilde{\sigma}^2 = V(X_i) = V(\mu + \sigma Z_i) = \sigma^2 V(Z_i) = 1.4\sigma^2$$

```r
set.seed(1705)
tilde_sigma <- sqrt(1.4) * sigma

cov_t7_sigma <- replicate(B, {
  y <- mu + sigma * rt(n, df = 7)
  s <- sd(y)
  lo <- sqrt((n - 1) * s^2 / qchisq(.975, df = n - 1))
  hi <- sqrt((n - 1) * s^2 / qchisq(.025, df = n - 1))
  (tilde_sigma >= lo) & (tilde_sigma <= hi)
})

cover_t7_sigma <- mean(cov_t7_sigma)
se_mc_t7_sigma <- sqrt(cover_t7_sigma * (1 - cover_t7_sigma) / B)
```

I dette tilfellet får vi dekning for $\widetilde{\sigma}$ på $89.16\% \ll 95\%$ og samtidig standardavvik på 0.0031. Nå ser vi langt dårligere dekning enn for i d). Dette kan være siden t_7 ikke er fordelt som χ^2 og intervallene blir igjen for smale.

```
> cover_t7_sigma; se_mc_t7_sigma
[1] 0.8916
[1] 0.003108849
```