Matematisk Statistik — Første Obligatoriske Aflevering

Mads Heller — fpg820
Sunniva Martinsen Abildgaard — bzg977
Christian Friis — snx635
Victor Z. Nygaard — nfq499
Andreas Hove Rising — fbr426

20. april 2020

Opgave a)

Det vises her først, at $\tilde{\theta}_n$ er central. Per definition 4.6 ses der, at der skal findes middelværdien:

$$\begin{aligned} E\left[\tilde{\theta}_n\right] &= E\left[\frac{S}{n}\right] \\ &= E\left[\frac{\sum_i X_i}{n}\right] \\ &= \frac{\sum_i E\left[X_i\right]}{n} \\ &= \frac{\sum_i \theta}{n} \\ &= \frac{n \cdot \theta}{n} \\ &= \theta \end{aligned}$$

Hvilket viser, at den er central.

Det er givet, at $\tilde{\theta}_n = \frac{S}{n} = \frac{\sum_i X_i}{n} = \sum_i(\frac{X_i}{n})$. Hvis vi lader $Y_i = \frac{X_i}{n}$, ses det, at Y_i er normalfordelt. Det er en lineær transformation af en normalfordelt stokastisk variabel (korollar 18.29 MI). Dermed er $\tilde{\theta}_n = \sum_i(\frac{X_i}{n}) = \sum_i(Y_i)$ en sum af normalfordelte stokastiske variable, som derfor også er normalfordelt (korollar 18.29 MI).

For at vise at den er konstistent, anvendes der store tals lov, som kendes fra tidligere kursus MI. Det kan gøres, da X_i'erne er uafhængige og identisk normalfordelt.
Bemærk: $\tilde{\theta}$ er defineret som $\frac{1}{n}\sum_{i=1}^{n} X_i$:

$$\begin{aligned} \tilde{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \theta \\ \Rightarrow \tilde{\theta}_n \xrightarrow{P} \theta \end{aligned}$$

Der skal så findes variansen af $\tilde{\theta}_n$. Der genkaldes følgende formel fra MI:

$$V\left[a + \beta X\right] = \beta^2 V\left[X\right] \text{ for } \alpha, \beta \in \mathbb{R}.$$

Deraf kan variansen findes:

$$\begin{aligned} V\left[\tilde{\theta}_n\right] &= V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n^2}\cdot V\left[\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n^2}\underset{i=1}{} V\left[X_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\theta^2 \\ &= \frac{n\cdot\theta^2}{n^2} \\ &= \frac{\theta^2}{n} \end{aligned}$$

Hvoraf variansen af $\tilde{\theta}_n$ er fundet.

Opgave b)

Der skal her findes den asymptotiske fordeling til $\check{\theta}_n$ via deltametoden. Deltametoden findes ved BMS theorem A.12.

Det er givet, at:

$$\check{\theta}_n^2 \overset{as}{\sim} \mathcal{N}\left(\theta^2, \frac{1}{n}\theta^4\right)$$

Der dannes funktionen:

$$f(x) = \sqrt{x}$$

Det er kendt at $f : U \longrightarrow \mathbb{R}$ er kontinuert for $U \subset (0,\infty)$ og derfor også målelig[1]. Derudover er den differentiabel i alle $x \in (0,\infty)$ og vil derfor også være det i θ^2. Det vil sige at antagelserne for deltametoden er opfyldt. Per deltametode fås så:

$$\begin{aligned} \check{\theta}_n = f\left(\check{\theta}_n^2\right) &\overset{as}{\sim} \mathcal{N}\left(f\left(\theta^2\right), \frac{1}{n}\cdot Df\left(\theta^2\right)\cdot 2\theta^4\cdot Df\left(\theta^2\right)^T\right) \\ &= \mathcal{N}\left(\theta, \frac{1}{n}\cdot\frac{1}{2\theta}\cdot 2\theta^4\cdot\frac{1}{2\theta}\right) \\ &= \mathcal{N}\left(\theta, \frac{1}{2n}\theta^2\right) \end{aligned}$$

Da Df er Jacobimatricen for en 1×1 matrice i dette tilfælde. Altså er det den afledte i 1 dimension. Heraf er den asymptotiske fordeling til $\check{\theta}_n$ fundet via deltametoden.

Opgave c)

Den simultane fordeling af $(X_1, X_2, \ldots, X_n)$ har tæthed f mht. Lebesguemålet, givet ved

$$f(x_1,x_2,\ldots,x_n) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\theta^2}}e^{\frac{-(x_i-\theta)^2}{2\theta^2}} = \left(\frac{1}{2\pi\theta}\right)^{\frac{n}{2}} e^{-\frac{1}{2\theta^2}\sum_{i=1}^{n}(x_i-\theta)^2}$$

Dermed bliver loglikelihoodfunktionen

$$\begin{aligned} \ell_x(\theta) &= -\log(f(x_1,x_2,\ldots,x_n)) \\ &= -\frac{n}{2}\log(1) + \frac{n}{2}\log(2\pi\theta) + \frac{1}{2\theta^2}\sum_{i=1}^{n}(x_i-\theta)^2 \\ &= \frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\theta^2) + \frac{SS}{2\theta^2} + \frac{n}{2} - \frac{S}{\theta} \end{aligned}$$

[1] Se evt. Øvelse 8.7 i Schilling

Vi kan nu finde score -og informationsfunktionen ved at differentiere loglikelihoodfunktionen hhv. en og to gange mht. θ

$$S(x,\theta) = 2\theta\frac{n}{2\theta^2} + (-2)\frac{SS}{2\theta^3} - (-1)\frac{S}{\theta^2} = \frac{n}{\theta} - \frac{SS}{\theta^3} + \frac{S}{\theta^2}$$
$$I(x,\theta) = 3\frac{SS}{\theta^4} - 2\frac{S}{\theta^3} - \frac{n}{\theta^2}$$

Ved at tage middelværdien af $I(x,\theta)$, og bruge linearitet, finder vi Fisherinformationen.

$$\begin{aligned} i(\theta) := E_\theta(I(x,\theta)) &= \frac{3}{\theta^4}\sum_{i=1}^{n} EX_i^2 - \frac{2}{\theta^3}\sum_{i=1}^{n} EX_i - E\frac{n}{\theta^2} \\ &= \frac{6n}{\theta^2} - \frac{2n}{\theta^2} - \frac{n}{\theta^2} \\ &= \frac{3n}{\theta^2} \end{aligned}$$

hvor vi har brugt at $EX_i^2 = 2\theta^2$ da

$$VX_i = EX_i^2 - (EX_i)^2 \Leftrightarrow \theta^2 = EX_i^2 - (\theta)^2 \Leftrightarrow EX_i^2 = 2\theta^2$$

Opgave d)

For at finde MLE for θ løser vi scoreligningen $S(x,\theta) = 0$.

$$S(x,\theta) = 0 \Leftrightarrow \frac{n}{\theta} - \frac{SS}{\theta^3} + \frac{S}{\theta^2} = 0 \Leftrightarrow n\theta^2 + S\theta - SS = 0$$

Det genkender vi som en andengradsligning, som bekendt har løsninger

$$\theta = \frac{-S \pm \sqrt{S^2 + 4\cdot n\cdot SS}}{2n}$$

Da $\theta > 0$ skal ovenstående være større end 0 for at være en løsning. Vi ser at kvadratroden og nævneren altid er positiv. Endvidere kan vi lave vurderingen

$$4\cdot n\cdot SS \geq 0 \Leftrightarrow S^2 + 4\cdot n\cdot SS \geq S^2 \Leftrightarrow \sqrt{S^2 + 4\cdot n\cdot SS} \geq S$$

Hvorfor tælleren kun er ikke-negativ med plusløsningen.
Altså vil løsningen til score ligningen være givet ved;

$$\theta = \frac{-S + \sqrt{S^2 + 4\cdot n\cdot SS}}{2n}.$$

Vi ved at $\log f$ har præcis et staitonært punkt, som er maksimum, derfor må det entydigt fundne stationære punkt for $-\log f$ nødvendigvis være et minimumspunkt (EH, indledende matematisk analyse).

Opgave e)

Vi ser at det ikke er muligt at bruge T2.20 i BMS til at slutte at MLE er asymptotisk normalfordelt, idet der forekommer problemer med regularitetsbetingelsen. Bemærk at vi fra BMS D1.3 har at tætheden for en sandsynlighedsfordeling i en ekspotentiel familie skal være på form;

$$\frac{1}{c(\theta_k)}e^{\langle\theta_k, t(x)\rangle}, \tag{1}$$

med hensyn til $d\mu(x)$ for en parameter $\theta_k \in \Theta$. Dog har vi i vores case at vores X vil have tæthed g mht. Lebesguemålet på $\mathbb{R}^2$;

$$g(x) \propto e^{\frac{-(x-\theta)^2}{2\theta^2}} = e^{\frac{-x^2-\theta^2+2x\theta}{2\theta^2}} = e^{(x,x^2)^T(\theta^{-1},-(2\theta)^{-2})-\frac{1}{2}}$$

Her identificerer vi en stikprøvefunktion $t : \mathbb{R} \to \mathbb{R}^2$ som $t(x) = (x, x^2)$. Hvis observationerne kom fra en eksponentiel familie ville parameteren være 2-dimensional ifølge definition 2.13(2), men θ er en positiv konstant, så den må komme fra et parameterrum af dimension 1, hvilket er i modstrid med definitionen og dermed kan T2.20 ikke benyttes.

Opgave f)

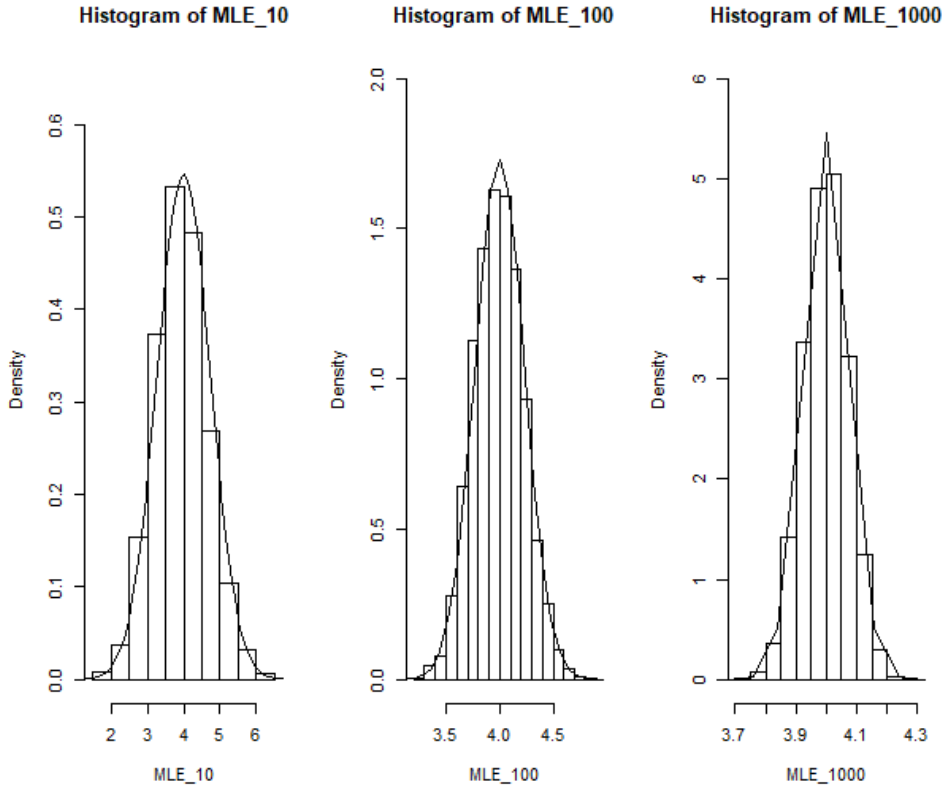
Vi simulerer MLE $\hat{\theta}_n$, som vi har et udtryk for fra d). Vi simulerer med $n = 10$, $n = 100$, $n = 1000$. For $\theta = 4$ sammenligner vi middelværdien af MLE $\hat{\theta}_n$ og middelværdien fra normalfordelingen givet i opgaven. Simuleringen ses herunder

```
# Vaelger parametre
set.seed(205)
theta <- 4
### Simulerer MLE fra opgave d) med 5000 datasaet
MLEMidFunc <- function(tilf, n, middel, stdaf) {
    MLE_n <- c() # Initialisering
    for (i in 1:tilf) {
        x <- rnorm(n,middel, stdaf)
        MLE_n[i] <- (-sum(x) + sqrt(sum(x)^2 + 4*n*sum(x^2)))/(2*n) # Scoreligning
    }
    # Sammenligner med parametrene af den i opgaven angivet normalfordeling
    c(mean(MLE_n), middel) # (MLE middelvaerdi, Teoretisk middelvaerdi)
}
MLEMidFunc(5000, 10^1, theta, theta) # 3.92068 4.00000
MLEMidFunc(5000, 10^2, theta, theta) # 3.987991 4.000000
MLEMidFunc(5000, 10^3, theta, theta) # 3.997768 4.000000
```

For varians har vi ligeledes lavet;

```
MLEVarFunc <- function(tilf, n, middel, stdaf) {
    MLE_n <- c() # Initialisering
    for (i in 1:tilf) {
        x <- rnorm(n, middel, stdaf)
        MLE_n[i] <- (-sum(x) + sqrt(sum(x)^2 + 4*n*sum(x^2)))/(2*n) # Scoreligning
    }
    c(var(MLE_n), middel^2/(3*n)) # (MLE varians, Teoretisk varians)
}
MLEVarFunc(5000, 10^1, theta, theta) # 0.5354607 0.5333333
MLEVarFunc(5000, 10^2, theta, theta) # 0.05336975 0.05333333
MLEVarFunc(5000, 10^3, theta, theta) # 0.005348861 0.005333333
```

Vi sammenligner den teoretiske tæthed og de overstående simulerede værdier for $n = 10^1, 10^2, 10^3$ opnås følgende tre plots;



Vi ser dermed at det (tilsynladende) forekommer muligt approximere MLE med en normalfordeling.

Opgave g)

Fra opgave a ved vi at

$$\tilde{\theta}_n \overset{as}{\sim} \mathcal{N}\left(\theta, \frac{\theta^2}{n}\right)$$

Fra opgave b ved vi at

$$\check{\theta}_n \overset{as}{\sim} \mathcal{N}\left(\theta, \frac{\theta^2}{2n}\right)$$

Fra opgave f ved vi at

$$\hat{\theta}_n \overset{as}{\sim} \mathcal{N}\left(\theta, \frac{\theta^2}{3n}\right)$$

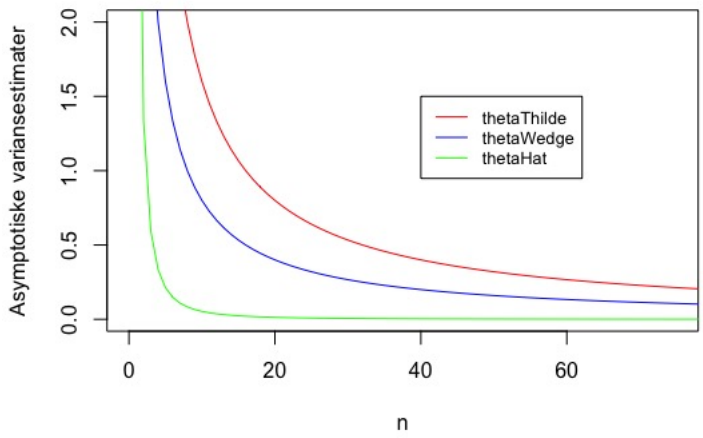
Vi ser for hver linje at variansen går hurtigere mod 0 når n vokser. Betragt for eksempel $\theta = 6$. Hvis $n = 12$ har vi at

$$V\hat{\theta}_{12} = 1$$

For at få opnå samme varians skal man bruge hhv.

$$V\tilde{\theta}_n = 1 \Leftrightarrow \frac{1}{n} = \frac{1}{36} \Leftrightarrow n = 36$$
$$V\check{\theta}_n = 1 \Leftrightarrow \frac{1}{2n} = \frac{1}{36} \Leftrightarrow n = 18$$

Her skal vi altså bruge 3 og 1.5 gange så mange observationer for samme præcision. Nedenfor ses hvor hurtigt de asymptotiske varianser aftager ift. hinanden.



Opgave h)

Lad z være 95%-fraktilen i en χ^2_1 fordeling. Vi opstiller først områdeestimatoren baseret på den falske Wald størrelse teoretisk ved at finde de θ der opfylder at

$$ni(\hat{\theta}_n)(\hat{\theta}_n - \theta)^2 < z$$

Vi løser

$$\begin{aligned}
(\hat{\theta}_n - \theta)^2 &< \frac{z}{ni(\hat{\theta}_n)} \\
\Leftrightarrow |\hat{\theta}_n - \theta| &< \sqrt{\frac{z}{ni(\hat{\theta}_n)}} \\
\Leftrightarrow -\sqrt{\frac{z}{ni(\hat{\theta}_n)}} < \hat{\theta}_n - \theta &< \sqrt{\frac{z}{ni(\hat{\theta}_n)}} \\
\Leftrightarrow \hat{\theta}_n + \sqrt{\frac{z}{ni(\hat{\theta}_n)}} > \theta &> \hat{\theta}_n - \sqrt{\frac{z}{ni(\hat{\theta}_n)}}
\end{aligned}$$

Bruger vi $i(\hat{\theta}_n)$ fra opgave c får vi at

$$\theta < \hat{\theta} \pm \sqrt{\frac{z}{ni(\hat{\theta}_n)}} = \hat{\theta} \pm 1.96\frac{\theta^2}{n\sqrt{3}}$$

Ved at bruge de simulerede MLE'er fra opgave f finder vi dækningsgrad og gennemsnitlig intervallængde for ovenstående konfindensinterval baseret på hhv. 10, 100 og 1.000 observationer.

```
for (i in 1:5000)
{
  KI_10[i,] <- MLE_10[i] + c(-1,1) * (sqrt(qchisq(0.95,1)) / (sqrt(3*10)/MLE_10[i]))
  KI_100[i,] <- MLE_100[i] + c(-1,1) * (sqrt(qchisq(0.95,1)) / (sqrt(3*100)/MLE_100[i]))
  KI_1000[i,] <- MLE_1000[i] + c(-1,1) * (sqrt(qchisq(0.95,1)) / (sqrt(3*1000)/MLE_1000[i]))
}

  DG_10 <- mean((theta > KI_10[,1]) * (theta < KI_10[,2])) # 0.9108
  DG_100 <- mean((theta > KI_100[,1]) * (theta < KI_100[,2])) # 0.9484
  DG_1000 <- mean((theta > KI_1000[,1]) * (theta < KI_1000[,2])) # 0.95

  LNG_10 <- mean(KI_10[,2] - KI_10[,1]) # 2.81
  LNG_100 <- mean(KI_100[,2] - KI_100[,1]) # 0.9
  LNG_1000 <- mean(KI_1000[,2] - KI_1000[,1]) # 0.29
```

Opgave i)

Vi bruger Deltametoden til at finde fordelingen af λ. Fra d ved vi at:

$$\hat{\theta}_n^2 \overset{as}{\sim} \mathcal{N}\left(\theta, \frac{1}{n}\frac{\theta^2}{3}\right)$$

Der dannes funktionen:

$$f(x) = \log x$$

Lad U være en åben delmængde af $(0,\infty)$ så er f: $U \longrightarrow \mathbb{R}$ både målelig og differentiabel ved $\theta > 0$. Vi kan derfor bruge deltametoden. Per deltametode fås så:

$$\begin{aligned}\hat{\lambda}_n = f(\hat{\theta}_n) &\overset{as}{\sim} \mathcal{N}\left(f(\theta), \frac{1}{n}\cdot Df(\theta)\cdot\frac{\theta^2}{3}\cdot Df(\theta)^T\right)\\ &= \mathcal{N}\left(log(\theta), \frac{1}{n}\cdot\frac{1}{\theta}\cdot\frac{\theta^2}{3}\cdot\frac{1}{\theta}\right)\\ &= \mathcal{N}\left(\lambda, \frac{1}{3n}\right)\end{aligned}$$

Da Df er Jacobimatricen for en 1×1 matrice i dette tilfælde. Altså er det den afledte i 1 dimension. Heraf er den asymptotiske fordeling til $\check{\lambda}_n$ fundet via deltametoden.

Vi vil nu gennemføre en tilsvarende undersøgelse af λ som af θ i forrige opgave og sammenligne. Først finder vi et 95% konfidensinterval.

Se først at

$$(3n)^{-1} = (n\log\theta)^{-1} \Leftrightarrow i(\hat{\lambda}_n) = 3$$

Lad igen z være 95% fraktilen fra forrige opgave.

$$\begin{aligned}ni(\hat{\lambda}_n)(\hat{\lambda}_n-\lambda)^2 = 3n(\log\hat{\theta}_n - \log\theta)^2 &< z\\ \Leftrightarrow |\log\left(\frac{\hat{\theta}_n}{\theta}\right)| &< \sqrt{\frac{z}{3n}}\\ \Leftrightarrow -\sqrt{\frac{z}{3n}} < \log\left(\frac{\hat{\theta}_n}{\theta}\right) &< \sqrt{\frac{z}{3n}}\\ \Leftrightarrow e^{-\sqrt{\frac{z}{3n}}} < \frac{\hat{\theta}_n}{\theta} &< e^{\sqrt{\frac{z}{3n}}}\\ \Leftrightarrow \hat{\theta}_n e^{\sqrt{\frac{z}{3n}}} > \theta &> \hat{\theta}_n e^{-\sqrt{\frac{z}{3n}}}.\end{aligned}$$

Vi laver en tilsvarende simulering;

```
# Initialisering
lambda <- log(theta)
lambdaKI_10 <- matrix(0,5000,2)
lambdaKI_100 <- matrix(0,5000,2)
lambdaKI_1000 <- matrix(0,5000,2)

# Simulation
for (i in 1:5000)
{
  lambdaKI_10[i,] <- c(MLE_10[i]
  * exp(-sqrt(qchisq(0.95,1)/(3*10))), MLE_10[i] * exp(sqrt(qchisq(0.95,1)/(3*10))))
  lambdaKI_100[i,] <- c(MLE_100[i] *
  exp(-sqrt(qchisq(0.95,1)/(3*100))), MLE_100[i] * exp(sqrt(qchisq(0.95,1)/(3*100))))
  lambdaKI_1000[i,] <- c(MLE_1000[i]
  * exp(-sqrt(qchisq(0.95,1)/(3*1000))), MLE_1000[i] * exp(sqrt(qchisq(0.95,1)/(3*1000))))
}
```

```
# Daekningsgrader

lambdaDG_10 <- mean((theta > lambdaKI_10[,1]) * (theta < lambdaKI_10[,2]))
lambdaDG_100 <- mean((theta > lambdaKI_100[,1]) * (theta < lambdaKI_100[,2]))
lambdaDG_1000 <- mean((theta > lambdaKI_1000[,1]) * (theta < lambdaKI_1000[,2]))

c(lambdaDG_10, lambdaDG_100, lambdaDG_1000) # 0.9320 0.9504 0.9498

# Intervallaengder

lambdaLNG_10 <- mean(lambdaKI_10[,2] - lambdaKI_10[,1])
lambdaLNG_100 <- mean(lambdaKI_100[,2] - lambdaKI_100[,1])
lambdaLNG_1000 <- mean(lambdaKI_1000[,2] - lambdaKI_1000[,1])

c(lambdaLNG_10,lambdaLNG_100,lambdaLNG_1000) # 2.8662107 0.9044782 0.2861724
```

Vi ser fra overstående at dækningsgraden for den logtransformerede variabel konvergerer hurtigere mod 95% og at det ikke lader til at have en indvirkning på den gennemsnitlige intervallængde.