

Sandsynlighedsteori og Statistik: Ugeseddel 12

Levi van Boekel

Uge 47, 2024

Opgave 1

1.1

95-pct. konfidensintervallet er definatorisk givet ved:

$$\theta_0 \in \left[\hat{\theta}_n - 1.96 \cdot se(\hat{\theta}_n), \hat{\theta}_n + 1.96 \cdot se(\hat{\theta}_n) \right] = \left[\theta(\hat{\theta}), \bar{\theta}(\hat{\theta}) \right] \quad (1)$$

Og angiver, et lukket interval, hvorom det gælder, at 95 pct. af sandsynlighedsmassen ligger for den stokastiske variabel, $\frac{\hat{\theta}_n - \theta_0}{se(\hat{\theta}_n)} \stackrel{a}{\sim} N(0, 1)$. Ved indsættelse af opgavens værdier fås:

$$\theta_0 \in [10.31, 14.51] \quad (2)$$

Dvs. at 95 pct. af udfaldene af parameteren (hvor udfaldene er standard normalfordelt) ligger i ovennævnte interval.

1.2

Vi definerer nul hypotesen som den hypotese, der beskriver, at udsagnet er sandt, dvs.

$$\mathcal{H}_0 : \theta_0 = 10 \quad \Theta_0 = \{10\} \quad (3)$$

Og den alternative hypotese, som beskriver, at udsagnet ikke er sandt:

$$\mathcal{H}_A : \theta_0 \neq 10 \quad \Theta_A = \{\theta_0 \in \mathbb{R} \mid \theta_0 \neq 10\} \quad (4)$$

1.3

Vi ved, at $Z_n(\theta_0 = a) \xrightarrow{d} N(0, 1)$ hvis $\mathcal{H}_0$ er sand, $\theta_0 = a$, mens grænseværdierne divergerer mod $-\infty$ og ∞ , hvis $\mathcal{H}_0$ ikke er sand.

Vi definerer derfor *Wald statistikken* og bestemmer herved, om det er sandsynligt, at udfaldet kommer fra en standard normalfordeling:

$$Z_n(\theta_0 = 10) = \frac{\hat{\theta}_n - \theta_0}{se(\hat{\theta}_n)} = \frac{12.41 - 10}{1.07} \approx 2.25 \quad (5)$$

De kritiske værdier, med $\alpha = 0.05$, er:

$$\Phi^{-1}(0.025) \approx -1.96$$

$$\Phi^{-1}(0.975) \approx 1.96$$

Og vi forkaster som bekendt hypotesen, hvis $Z_n(\theta_0 = 10) < \Phi^{-1}(\frac{\alpha}{2})$ eller $Z_n(\theta_0 = 10) > \Phi^{-1}(\frac{1-\alpha}{2})$, hvor det sidste er opfyldt her. P-værdien måler sandsynligheden for at få noget mere ekstremt end end givne realisation og er:

$$p_n(\theta_0 = 10) = 2 \cdot (1 - \Phi(|z_n(\theta_0 = 10)|)) = 2 \cdot (1 - \Phi(2.25)) = 0.0244 \quad (6)$$

Som altså tilsiger, at ssh. for at få noget mere ekstremt end $|z_n(\theta_0 = 10)|$ er 0.0244, og da dette er mindre end kravet om $\alpha = 0.05$, afvises hypotesen.

1.4

Vi definerer den kvadrede Wald statistisk som:

$$w_n(\theta_0 = a) \equiv h_n(\theta_0 = a)^2 = 2.25^2 = 5.0625 \quad (7)$$

Hvis null hypotesen er sand, og vi kun betragter en parameter, der restrikeres, vil teststørrelsen være asymptotisk standard χ^2 fordelt;

$$w_n(\theta_0 = a) \overset{a}{\sim} \chi_1^2 \quad (8)$$

Den kritiske værdi er her: $\text{crit}_{\chi_1^2, 1-0.05/2} = 3.84$, som da igen afviser hypotesen, da $2.25^2 > 3.84$. P-værdien er:

$$p_n(\theta_0 = 10) = (1 - F_{\chi_1^2}(|w_n(\theta_0 = 10)|)) = 0.0244 \quad (9)$$

Og dermed sammenfaldende, hvorfor vi igen (på baggrund af $\alpha = 0.05$) afviser null hypotesen.

Opgave 2

2.1

Log likelihood funktionen bliver:

$$\log f_{Y_i}(\mu | i) = \sum_{i=1}^n \left[-\log(\mu) - \frac{y_i}{\mu} \right]$$

Med score funktionen givet ved;

$$\begin{aligned} \frac{\delta \log f_{Y_i}(y | \mu)}{\delta \mu} &= \sum_{i=1}^n \left[-\frac{1}{\mu} + \frac{y_i}{\mu^2} \right] \\ &= -\frac{n}{\mu} + \frac{\sum_{i=1}^n y_i}{\mu^2} \end{aligned}$$

FOC er:

$$\begin{aligned} \frac{n}{\mu} + \frac{\sum_{i=1}^n y_i}{\mu^2} &= 0 \\ \frac{\sum_{i=1}^n y_i}{\mu^2} &= -\frac{n}{\mu} \\ \hat{\mu} &= -\frac{1}{n} \sum_{i=1}^n y_i \end{aligned}$$

Vi tjekker at ovenstående er et minima ved at se, at den andenafledte er:

$$\begin{aligned} &\sum_{i=1}^n \left[\frac{1}{\mu^2} - \frac{2y_i}{\mu^3} \right] \\ &\left[\frac{n}{\mu^2} - \frac{2 \sum_{i=1}^n y_i}{\mu^3} \right] \\ &\left[\frac{n}{\mu^2} - \frac{2n\mu}{\mu^3} \right] \\ &\left[-\frac{n}{\mu^2} \right] < 0 \end{aligned}$$

Hvorfor vores estimatorer altså er et minima. Vi bemærker desuden, at estimatoren nu er det inverse af hvad vi fandt tidligere og nu blot er det aritmetiske gennemsnit.

2.2

For at vise, at estimatoren er unbiased tager vi den forventede værdi:

$$\mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(Y_i) = \frac{1}{n} \cdot n \cdot \mu_0 = \mu_0$$

2.3

Variansen af estimatoren er:

$$\text{Var}\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) = \frac{1}{n^2} = \frac{\text{Var}(Y_i)}{n} = \frac{\mu_0^2}{n}$$

2.4

Vi ved fra CLT, at:

$$\sqrt{n} \cdot (\mu_n - \mu_0) \sim N(0, \mu_0^2) \quad (10)$$

Og:

$$\mu_n \overset{a}{\sim} N\left(\mu_0, \frac{\mu_0^2}{n}\right) \quad (11)$$

Hvor iid. antagelserne er tilstrækkelige, dvs. alle stokastiske variable kommer fra den samme (kendte fordeling), er identisk fordelt heraf, samt at de er uafhængige.

Opgave 3

3.1

Log-likelihood bidraget i en enkelt observation er givet ved:

$$\log f_{Y_i}(\mu \mid y_i) = \log \ell(y_i \mid \mu) = -\log(\mu) - \frac{y_i}{\mu}$$

3.2

Den anden ordens afledte, i et enkelt punkt, er (se tidligere opgave for udledning):

$$H_i(\mu_0) = \frac{1}{\mu_0^2} - \frac{2y_i}{\mu_0^3}$$

3.3

Informationen findes:

$$\begin{aligned} I(\mu_0) &= -\mathbb{E}(H_i(\mu_0)) \\ &= -\left[\frac{1}{\mu_0^2} - \frac{2}{\mu_0^3} \cdot \mathbb{E}(Y_i) \right] \\ &= -\left[\frac{1}{\mu_0^2} - \frac{2}{\mu_0^2} \right] \\ &= \frac{1}{\mu_0^2} \end{aligned}$$

Hvilket evalueret i estimatoren bliver:

$$I(\hat{\mu}_n) = \frac{1}{\left[\frac{1}{n} \sum_{i=1}^n y_i\right]^2} = \frac{1}{\hat{\mu}_n^2}$$

3.4

Variansen på estimatoren er:

$$V(\hat{\mu}_n) = I(\hat{\mu}_n)^{-1} \cdot \frac{1}{n} = \frac{\hat{\mu}_n^2}{n}$$

Som er nøjagtig det samme resultat som i 2.3.

Opgave 4

4.1

Vi må igen ty til en Bernoulli fordeling, idet variablerne er bineære og der dermed ikke er andre mulige fordelinger. Vi har altså:

$$y_i \sim \text{Bernoulli}(\theta_0) \quad (12)$$

For at gøre brug af MLE analysen skal vi gøre brug af iid. antagelserne. Den første antagelse tilsiger, at alle de stokastiske variable kommer fra en Bernoulli fordeling, hvilket her forekommer rimeligt, idet der ikke er andre muligheder. Den anden antagelse tilsiger, at de stokastiske variable alle kommer fra den samme fordeling, hvilket ikke virker rimeligt, idet man hermed indirekte antager, at sandsynligheden for at overleve ikke afhænger af f.eks. hvilken klasse man rejser på (som vi jo vidste ikke var tilfældet i tidligere opgaver). Den tredje antagelse om uafhængighed virker heller ikke spor plausibel, da en persons overlevelse afhænger af hvor mange redningsbåde der er, hvilket påvirkes af en anden persons overlevelse.

Log likelihood funktionen er:

$$f_{Y_i}(\mu | y_i) = \sum_{i=1}^n [y_i \log(\theta) + (1 - y_i) \cdot \log(1 - \theta)] \quad (13)$$

Score funktionen er:

$$\begin{aligned} \frac{\delta f_{Y_i}(\mu | y_i)}{\delta \theta} &= \sum_{i=1}^n \left[\frac{y_i}{\theta} - \frac{(1 - y_i)}{1 - \theta} \right] \\ &= \frac{\sum_{i=1}^n y_i}{\theta} - \frac{n - \sum_{i=1}^n y_i}{1 - \theta} \end{aligned}$$

Med FOC:

$$\begin{aligned} \frac{\sum_{i=1}^n y_i}{\theta} &= \frac{n - \sum_{i=1}^n y_i}{1 - \theta} \\ \sum_{i=1}^n y_i(1 - \theta) &= (n - \sum_{i=1}^n y_i)\theta \\ \sum_{i=1}^n y_i &= n\theta \\ \hat{\theta}_n &= \frac{1}{n} \sum_{i=1}^n y_i \hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n y_i \end{aligned}$$

Som blot er det aritmetiske gennemsnit. MLE estimatet bliver da:

$$\hat{\theta}(y_1, \dots, y_{2201}) \approx 0.323 \quad (14)$$

4.2

Variansen af estimatoren er:

$$\text{Var}\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n Y_i\right) = \frac{\text{Var}(Y_i)}{n} = \frac{\theta_0(1 - \theta_0)}{n}$$

Da iid. er opfyldt, ved vi fra CLT, at fordeling af estimatoren er: $\sqrt{n} \cdot (\hat{\theta}_n - \theta_0) \sim N(0, \theta_0(1 - \theta_0))$.

Vi kan også skrive den asymptotiske fordeling op som:

$$\hat{\theta}_n \overset{a}{\sim} N\left(\theta_0, \frac{\theta_0(1 - \theta_0)}{n}\right) \quad (15)$$

4.3 & 4.4

Giver samme resultat. Hvis $S(\theta) > 0$ er det et udtryk for, at vi er til venstre for toppunktet (tegn graf selv). Koden er:

```
import excel "/Users/levivanboekel/Desktop/ku/sandstat/data/Titanic.xlsx", sheet("Sheet1")
firstrow clear
mlexp(survived*log(theta) +(1-survived)*log(1-theta))
```

4.5

Log likelihood funktionen er:

$$\log L_n(\theta_i) = \sum_{i=1}^{2201} [y_i \cdot \log(\theta + \delta \cdot \text{first}_i) + (1 - y_i) \cdot \log(1 - \theta - \delta \text{first}_i)]$$

Koden kan ændres til det følgende:

```
import excel "/Users/levivanboekel/Desktop/ku/sandstat/data/Titanic.xlsx", sheet("Sheet1")
firstrow clear
g f = class==1
mat define init = J(1,2,0.2)
mlexp(survived*log(theta + delta*f) +(1-survived)*log(1-theta-delta*f)), from (init)
```

Og vi får:

$$\begin{aligned}\hat{\theta}(y_1, \dots y_{2201}) &\approx 0.2707 \\ \hat{\delta}(y_1, \dots y_{2201}) &\approx 0.3538\end{aligned}$$

Resultatet beskriver, at sandsynligheden for at overleve er ca. 7 pct. højere, hvis man er på første klasse, end hvis man ikke er (idet θ og $\theta + \delta$ er middelværdier for Bernoullifordelingen).

4.6

Det er klart, at det er et eksempel på en betinget model, idet parametrene netop afhænger af, hvorvidt indikatorvariablen er opfyldt. De betingende middelværdier er, jf. svaret i den tidligere opgave, blot MLE estimaterne, dvs $\hat{\theta}$, hvis man rejser på andet end første klasse og $\hat{\theta} + \hat{\delta}$, hvis man rejser på første klasse. Det er dog under forudsætning af, at iid. antagelserne holder.

Opgave 5

5.1

Nullhypotesen er:

$$\mathcal{H}_0 : \delta = 0 \quad \Theta_0 = \{0 < \theta < 1, \delta = 0\} \quad (16)$$

Mens den alternative hypotese er:

$$\mathcal{H}_A : \delta \neq 0 \quad \Theta_A = \{0 < \theta < 1, \delta \neq 0\} \quad (17)$$

5.2

Vi ved, at $Z_n(\delta_0 = 0) \xrightarrow{d} N(0, 1)$ hvis $\mathcal{H}_0$ er sand, $\delta_0 = 0$, mens grænseværdierne divergerer mod $-\infty$ og ∞ , hvis $\mathcal{H}_0$ ikke er sand.

Vi definerer derfor *Wald statistikken* og bestemmer herved, om det er sandsynligt, at udfaldet kommet fra en standard normalfordeling:

$$Z_n(\delta = 0) = \frac{\hat{\delta}_n - \delta_0}{se(\hat{\delta}_n)} = \frac{0.3538 - 0}{0.0287525} \approx 12.31 \quad (18)$$

De kritiske værdier er:

$$\Phi^{-1}(0.025) \approx -1.96$$

$$\Phi^{-1}(0.975) \approx 1.96$$

Og idet $12.31 > 1.96$ afvises nulhypotesen. P værdien, dvs. sandsynligheden for at få noget mere ekstremt end det givne udfald bestemmes som:

$$p_n(\delta_0 = 0) = 2 \cdot (1 - \Phi(|z_n(\theta_0 = 0)|)) = 2 \cdot (1 - \Phi(|12.31|)) = 0 \quad (19)$$

Det vil sige, at sandsynligheden for at få et mere ekstremt udfald er 0 pct, hvorom det gælder, at $0.05 > 0$, hvorfor vi også på denne baggrund afviser null hypotesen.

5.3

Vi definerer den kvadrede Wald statistisk som:

$$w_n(\theta_0 = a) \equiv h_n(\theta_0 = a)^2 \approx 151.53 \quad (20)$$

Hvis null hypotesen er korrekt vil det gælde (da vi kun lægger en begrænsning over en parameter), at:

$$w_n(\theta_0 = a) \stackrel{a}{\sim} \chi_1^2 \quad (21)$$

Hvilket gælder, da $w_n(\theta_0 = a) \xrightarrow{d} Z^2$, når $n \rightarrow \infty$, hvor $Z = N(0, 1)$ og $Z^2 = \chi_1^2$. Den kritiske værdi for en χ_1^2 fordeling er $3.84 < 151.53$, hvorfor vi også her afviser null hypotesen. P-værdien er:

$$p_n(\theta_0 = 0) = (1 - F_{\chi_{(1)}^2}(151.53)) = 0 \quad (22)$$

5.4

Likelihood funktionen evalueret i den urestrikerede model er:

$$\log L_{\hat{\theta}, \hat{\delta} \neq 0} = -1310.73$$

Mens vi i den restrikerede model har:

$$\log L_{\hat{\theta}, \hat{\delta} = 0} = -1384.72$$

Derfor bliver *LR statistik*:

$$\begin{aligned} LR_n(\mathcal{H}_0) &= -2 \log \left[\frac{L_n(\tilde{\theta}_n)}{L_n(\hat{\theta}_n)} \right] \\ &= 2 \cdot \left[\log(L_n(\hat{\theta}_n)) - \log(L_n(\tilde{\theta}_n)) \right] = 2 \cdot (-1310.73 + 1384.72) \approx 147.98 \end{aligned}$$

Som er χ_1^2 fordelt, hvis null hypotesen er sand, idet vi kun restrikerer en parameter. Vi afviser dermed igen null hypotesen, idet $147.98 > 3.84$. P-værdien er også her 0.

Opgave 6

6.1

Nullhypotesen er:

$$\mathcal{H}_0 : \delta = 0, \theta = 0.35 \quad \Theta_0 = \{\theta = 0.35, \delta = 0\} \quad (23)$$

Mens den alternative hypotese er:

$$\mathcal{H}_A : \delta \neq 0, \theta \neq 0.35 \quad \Theta_A = \{0 < \theta, \delta < 1 \mid \theta \neq 0.35, \delta \neq 0\} \quad (24)$$

6.2

Likelihood funktionen evalueret i den urestrikerede model er:

$$\log L_{\hat{\theta} \neq 0.35, \hat{\delta} \neq 0} = -1310.73$$

Mens vi i den restrikerede model har:

$$\log L_{\hat{\theta}=0.35, \hat{\delta}=0} = \sum_{i=1}^{2201} [y_i \cdot \log(0.35) + (1 - y_i) \cdot \log(1 - 0.35)] = -1388.29$$

Derfor bliver LR statistikken (som nu er χ_2^2) fordelt:

$$LR_n(\mathcal{H}_0) = 2 \cdot (-1310.73 + 1388.29) = 155.11$$

Den kritiske værdi er $7.37 < 155.11$, hvorfor hypotesen afvises. P-værdien er:

$$p_n(\theta_0 = 0) = (1 - F_{\chi_{(2)}^2}(151.11)) = 0 \quad (25)$$

6.3

Nej, et Wald test kan ikke bruges her, da vi restrikerer 2 parametre samtidig, og vi dermed ikke længere kan bestemme den simple *Wald statistisk*.