STK1110 - Statistiske metoder og dataanalyse høsten 2019
Løsningsforslag til ekstraoppgave for uke 40

Vinnie Ko & Ørnulf Borgan

October 3, 2019

Ekstraoppgave 8

a) Vi har likelihooden

$$\mathcal{L}(\lambda) = \prod_{i=1}^{n} f(x_i;\lambda) = \prod_{i=1}^{n} \left(\lambda e^{-\lambda x_i}\right)$$

og log-likelihooden

$$\ell(\mu) = \frac{\partial}{\partial\lambda} \ln \mathcal{L}(\lambda) = \sum_{i=1}^{n} \ln f(x_i;\lambda) = n \ln\lambda - \lambda \sum_{i=1}^{n} x_i$$

Score funksjon er dermed

$$S(\lambda) = \frac{\partial}{\partial\lambda}\ell(\lambda) = \frac{\partial}{\partial\lambda}\left(n\ln\lambda - \lambda\sum_{i=1}^{n} x_i\right) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

Maximum likelihood estimatet $\widehat{\lambda}$ er gitt ved $S(\widehat{\lambda}) = 0$. Dermed får vi $\widehat{\lambda} = 1/\overline{x}$.

b) Den kumulative fordelingsfunksjonen for eksponensialfordeling er $F(x) = 1 - e^{-\lambda x}$ når $x \geq 0$. Hvis vi setter $Y_i = 2\lambda X_i$, har vi for $y \geq 0$ at

$$F(y) = P(Y_i \leq y) = P\left(2\lambda X_i \leq y\right) = P\left(X_i \leq \frac{y}{2\lambda}\right) = 1 - e^{-\lambda\frac{y}{2\lambda}} = 1 - e^{-\frac{y}{2}}$$

Hermed får vi at sannsynlighetstetthet til Y_i er gitt ved (for $y \geq 0$)

$$f(y) = F'(y) = \frac{1}{2}e^{-\frac{1}{2}y}.$$

Sannsynlighetstettheten til kji-kvadratfordeling med ν frihetsgrader er (for $x \geq 0$)

$$f(x) = \frac{1}{2^{\nu/2}\Gamma\left(\nu/2\right)}\, x^{(\nu/2)-1} e^{-\frac{x}{2}}.$$

Når $\nu = 2$, blir dette (for $x \geq 0$)

$$f(x) = \frac{1}{2}e^{-\frac{1}{2}x}.$$

Vi har derfor at $Y_i = 2\lambda X_i \sim \chi^2_2$

c) Vi vet hvis $U_1 \sim \chi^2_{\nu_1}$ og $U_2 \sim \chi^2_{\nu_2}$ er uavhengige, så er $U_1 + U_2 \sim \chi^2_{\nu_1+\nu_2}$.

Nå har vi at $Y_i = 2\lambda X_i \overset{u.i.f}{\sim} \chi^2_2$.

Derfor er

$$2\lambda \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \sim \chi^2_{2n}.$$

d) For eksponential-fordelingen er forventningen $1/\lambda$. Dermed er $E(\overline{X}) = 1/\lambda$. Så hvis $\overline{x}$ er liten, tyder det på at λ er stor. Og hvis $\overline{x}$ er stor, tyder det på at λ er liten. Det er derfor rimelig å forkaste nullhypotesen for små eller store verdier av $\overline{x}$, eller ekvivalent for små eller store verdier av $2n\lambda_0\overline{x}$. Det gir en begrunnelse for "fasongen" av forkastningsområdet for testen gitt i oppgaven.

Av resultatet i punkt c har vi nå at

$$2n\lambda_0\overline{X} = 2\lambda \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \sim \chi^2_{2n}$$

når $H_0 : \lambda = \lambda_=$ er sann.

Vi har dermed at

$$\begin{aligned}
&P(\text{forkast } H_0 \,|\, H_0 \text{ er sann}) \\
&= P(2n\lambda_0\overline{X} \le \chi^2_{1-\alpha/2,2n} \text{ eller } 2n\lambda_0\overline{X} \ge \chi^2_{\alpha/2,2n} \,|\, \lambda = \lambda_0) \\
&= P(2n\lambda_0\overline{X} \le \chi^2_{1-\alpha/2,2n} \,|\, \lambda = \lambda_0) + P(2n\lambda_0\overline{X} \ge \chi^2_{\alpha/2,2n} \,|\, \lambda = \lambda_0) \\
&= \alpha/2 + \alpha/2 \\
&= \alpha
\end{aligned}$$

så testen har signifikansnivå α.

e) Fra punkt a har vi at $\widehat{\lambda} = 1/\overline{x}$. Så

$$\begin{aligned}
-2\ln(LR) &= -2\ln\left(\frac{\mathcal{L}(\lambda_0)}{\mathcal{L}(\widehat{\lambda})}\right) \\
&= 2\ln\mathcal{L}(\widehat{\lambda}) - 2\ln\mathcal{L}(\lambda_0) \\
&= 2\ell(\widehat{\lambda}) - 2\ell(\lambda_0) \\
&= 2\left(n\ln\widehat{\lambda} - \widehat{\lambda}\sum_{i=1}^{n} X_i\right) - 2\left(n\ln\lambda_0 - \lambda_0\sum_{i=1}^{n} X_i\right) \\
&= 2n\ln\widehat{\lambda} - 2n\ln\lambda_0 - 2\left(\widehat{\lambda} - \lambda_0\right)\sum_{i=1}^{n} x_i.
\end{aligned}$$

f) Vi vil teste $H_0 : \lambda = 0.35$ mot $H_a : \lambda \neq 0.35$.

For testen i punkt d kan vi bruke følgende R-kode.

```
> HP.data = c(2, 1.3, 6, 1.9, 5.1, 0.4, 1, 5.3, 15.7, 0.7, 4.8, 0.9, 12.2, 5.3, 0.6)
> n = length(HP.data)
>
> # Maximum likelihood estimatet
> lambda.hat = 1/mean(HP.data)
> print(lambda.hat)
[1] 0.2373418
>
> # Nullhypotese og testobservator
> lambda.0 = 0.35
> test.statistic = 2*n*lambda.0*mean(HP.data)
> print(test.statistic)
[1] 44.24
>
```

```
> # i) Kritiske verdier når signifikansnivået er 5%
> alpha = 0.05
> critical.value = c(qchisq(alpha/2, df=2*n), qchisq(1-alpha/2, df=2*n))
> print(critical.value)
[1] 16.79077 46.97924
>
> # ii) Kritiske verdier når signifikansnivået er 10%
> alpha = 0.10
> critical.value = c(qchisq(alpha/2, df=2*n), qchisq(1-alpha/2, df=2*n))
> print(critical.value)
[1] 18.49266 43.77297
```

i) Når $\alpha = 0.05$ ligger verdien av testobservatoren (44.24) mellom de to kritiske verdiene, dvs. i intervallet $(16.79, 46.98)$, så vi forkaster ikke H_0.

ii) Når $\alpha = 0.10$ er verdien av testobservatoren (44.24) større enn den øvre kritiske verdien (43.77), så vi forkaster H_0.

P-verdien er det minste signifikansnivået vi kan bruke og få forkastning av H_0 (se side 458 i læreboka). Her forkaster vi H_0 når $\alpha = 0.10$, så P-verdi ≤ 0.10. Men vi forkaster ikke H_0 når $\alpha = 0.05$, så så P-verdi > 0.05. Altså har vi $0.05 < P$-verdi ≤ 0.10.

g) Den minste α vi kan bruke for testen i punkt d og få forkastning av H_0 er gitt ved at $\chi^2_{\alpha/2,2n} = 44.24$. Derfor er P-verdi $= 2\{1 - F_{2n}(44.24)\}$ der F_{2n} er den kumulative kji-kvadrat fordelingen med $2n$ frihetsgrader. R-koden nedenfor gir at P-verdien er 9.1%.

```
> p.value = 2*(1 - pchisq(test.statistic, df=2*n))
> print(p.value)
[1] 0.09071112
```

h) R-koden nednefor gir at $-2\ln(LR) = 2.59$.

i) Når $\alpha = 0.05$ er verdien av testobservatoren (2.59) mindre enn den kritiske verdien (3.84), så vi forkaster ikke H_0.

i) Når $\alpha = 0.10$ er verdien av testobservatoren (2.59) mindre enn den kritiske verdien (2.71), så vi forkaster ikke H_0.

```
> LR.test.statistic = 2*n*log(lambda.hat) - 2*n*log(lambda.0) - 2*(lambda.hat-lambda.0)*
    sum(HP.data)
> print(LR.test.statistic)
[1] 2.587041
>
> # Kritisk verdi når signifikansnivået er 5%
> alpha = 0.05
> critical.value = qchisq(p=1-alpha, df=1)
> print(critical.value)
[1] 3.841459
>
> # Kritisk verdi når signifikansnivået er 10%
> alpha = 0.1
> critical.value = qchisq(p=1-alpha, df=1)
> print(critical.value)
[1] 2.705543
```

i) P-verdien er her gitt ved P-verdi $= 1 - F_1(2.59)$ der F_1 er den kumulative kji-kvadrat fordelingen med 1 frihetsgrad. R-koden nedenfor gir at P-verdien er 10.8%.

```
> p.value = 1 - pchisq(LR.test.statistic, df=1)
> print(p.value)
[1] 0.1077415
```

j) De to testene er begge tester for nullhypotesen $H_0 : \lambda = 0.35$ mot den alternative hypotesen $H_a : \lambda \neq 0.35$. For testen i punkt d har vi den faktiske fordelingen til testobservatoren. Så for denne testen er signifikansnivå og P-verdi beregnet eksakt. For testen i punkt e har vi bare en tilnærming til den faktisk fordelingen for testobservatoren. Så her har vi tilnærmete verdier for signifikansnivå og P-verdi.