Statistikk

Jo Eidsvik

Matematiske fag, NTNU

Parameterestimering

Situasjon

- Data $X_1, \ldots, X_n$, oftest antatt uavhengige.
- Vi antar at mekanismen (fordelingen) er kjent $X_i \sim f(x; \theta)$
- Parameteren θ er imidlertid ukjent og estimeres.

Estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$.

$$E(\hat{\theta}) = \theta, \qquad Var(\hat{\theta}) \to 0, \tag{1}$$

når n blir stor (mot uendelig).

Maximum likelihood estimering (MLE)

- MLE = Maximum likelihood estimator = Sannsynlighetsmaksimeringsestimatoren
- Ide: maksimer simultantetthet til data, med hensyn på parameter.

$$l(\theta) = \ln[f(x_1;\theta)\dots f(x_n;\theta)] \tag{2}$$

$$\hat{\theta} = \text{argmax}_\theta l(\theta) \tag{3}$$

Pivot, Observator (Statistic)

- En pivot er som en observator (funksjon av data), men er et uttrykk som inneholder parameteren θ, og som har en kjent fordeling.
- Eksempel $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, $T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$, $Y = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.
- Man kan ofte finne transformasjoner av MLE som inneholder parameteren og har en kjent fordeling.
- Med en observator kan man lage et konfidensintervall.

Eksempel - levetider

X_i = tid til en parkeringsvakt kommer.
$X_1, \ldots, X_n$ er uavhengige eksponensialfordelte med parameter λ. Tetthet er $f(x;\lambda) = \lambda e^{-\lambda x}$, $x > 0$.

$$\hat{\lambda} = \frac{n}{(X_1 + \ldots + X_n)}$$

Her er $n = 10$, $\sum_{i=1}^{10} x_i = 42.5$, $\hat{\lambda} = 0.24$.
Hvordan finne et konfidensintervall for λ?

1. Finn en observator (pivot) med kjent fordeling. (Eksakt.)
2. Bruk sentralgrenseteoremet. (Approksimasjon, asymptotisk riktig, $n \to \infty$).

Eksempel - levetider

- $Y_1 = 2\lambda X_1$, $f(y_1) = \lambda e^{-\lambda(y_1/(2\lambda))}\frac{1}{2\lambda} = \frac{1}{2}e^{-y_1/2}$. Dette er en kji-kvadrat fordeling med 2 frihetsgrader.
- Summen av kji-kvadrat fordelte variable er kji-kvadrat fordelt, med frihetsgradene summert: $\sum_{i=1}^{n} 2\lambda X_i \sim \chi^2_{2n}$.

95 % konfidensintervall:

$$P(\chi^2_{2n,0.025} < 2\lambda \sum_{i=1}^{n} X_i < \chi^2_{2n,0.975}) = 0.95$$

$$P\left(\frac{\chi^2_{2n,0.075}}{2\sum_{i=1}^{n} X_i} < \lambda < \frac{\chi^2_{2n,0.975}}{2\sum_{i=1}^{n} X_i}\right) = 0.95$$

Eksempel - levetider

$\chi^2_{20,0.025} = 9.6$, $\chi^2_{20,0.975} = 32.4$.

$$P\left(\frac{\chi^2_{20,0.025}}{2\sum_{i=1}^{n} X_i} < \lambda < \frac{\chi^2_{20,0.975}}{2\sum_{i=1}^{n} X_i}\right) = 0.95$$

Konfidensintervall blir $(0.11, 0.40)$.