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.

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

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

Algoritme for MLE

1. Sett opp log likelihood $l(\theta)$.
2. Deriver log-likelihood $l'(\theta) = dl/d\theta$.
3. Løs $l'(\hat{\theta}) = 0$ mhp $\hat{\theta}$.

Eksempel binomisk

- $I_1, \ldots, I_n$ er n uavhengige forsøk.
- $I_i \in \{0, 1\}$.
- Konstant suksess-sannsynlighet $p = P(I_i = 1)$.
- $X = \sum_{i=1}^{n} I_i$ = antall suksesser.

Ukjent parameter er her $\theta = p$

Eksempel binomisk

$$l(p) = \ln[p^{I_1}(1-p)^{1-I_1} \dots p^{I_n}(1-p)^{1-I_n}] = \ln[p^X(1-p)^{n-X}]$$

$$l'(p) = \frac{d(X \ln p + (n-X)\ln(1-p)}{dp} = X/p - (n-X)/(1-p)$$

$$l'(\hat{p}) = 0, \quad X - \hat{p}X = n\hat{p} - X\hat{p}, \quad \hat{p} = X/n.$$

Sentralgrenseteoremet

$$\hat{p} \to N(p, p(1-p)/n), \quad n \to \infty.$$

$$\hat{p} \to N(p, \hat{p}(1-\hat{p})/n), \quad n \to \infty.$$

Dette kan brukes til konfidensintervall for p.

Konfidensintervall

$$P(\hat{p} - z_{1-\alpha/2}\sqrt{p(1-p)/n} < p < \hat{p} - z_{\alpha/2}\sqrt{p(1-p)/n}) = 1-\alpha, \qquad n \to \infty.$$

Konfidensintervall:

$$\hat{p} \pm z_{\alpha/2}\sqrt{p(1-p)/n}, \quad n \to \infty.$$

EU valg 1994, spørreundersøkelse

$n = 6592$, $\hat{p} = 0.513$
Konfidensintervall ($1 - \alpha$=95 %):

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{p(1-p)/n} = (0.501, 0.525).$$

Bredde konfidensintervall

Bredde er $B = 2z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$.
Med 95 % er $z_{1-\alpha/2} = 1.96$. Kan anta her at $\hat{p} \approx 0.5$.
Ønsket bredde konfidensintervall $B = 0.05$:

$$n = \frac{2^2 1.96^2 \hat{p}(1-\hat{p})}{B^2} \approx 1600$$

Ønsket bredde konfidensintervall $B = 0.01$:

$$n = \frac{2^2 1.96^2 \hat{p}(1-\hat{p})}{B^2} \approx 40000$$