Konfidensintervall

TMA4240/TMA4245 Statistikk

Håkon Tjelmeland
Institutt for matematiske fag
Norges teknisk-naturvitenskapelige universitet

Konfidensintervall

Observator: En observerbar funksjon av en eller flere variabler som utgjør et utvalg

- ⋆ Situasjon:
  - er interessert i en populasjon, parameter θ
  - observerer verdier for et utvalg, $X_1, X_2, \ldots, X_n$
  - har estimator for θ: $\widehat{\theta}$
  - ønsker nå et intervall som inneholder verdien til θ med en angitt sannsynlighet

Definisjon (Konfidensintervall)

Anta at vi har stokastiske variabler $X_1, X_2, \ldots, X_n$, og at sannsynlighetsfordelingen til disse inneholder en parameter θ og at verdien til θ er ukjent. La $x_1, x_2, \ldots, x_n$ være observerte verdier for de stokastiske variabler $X_1, X_2, \ldots, X_n$. Anta videre at for to observatorer $\widehat{\theta}_L(X_1, X_2, \ldots, X_n)$ og $\widehat{\theta}_U(X_1, X_2, \ldots, X_n)$ har at

$$P\left(\widehat{\theta}_L(X_1, X_2, \ldots, X_n) \leq \theta \leq \widehat{\theta}_U(X_1, X_2, \ldots, X_n)\right) = 1 - \alpha,$$

for $\alpha \in (0, 1)$. Det numeriske intervallet

$$\left[\widehat{\theta}_L(x_1, x_2, \ldots, x_n), \widehat{\theta}_U(x_1, x_2, \ldots, x_n)\right]$$

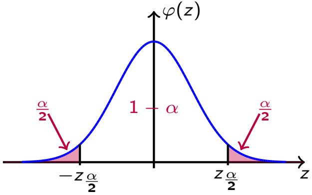
kalles da et $(1 - \alpha) \cdot 100\%$-konfidensintervall for θ

- ⋆ Hvordan finne slike observatorer $\widehat{\theta}_L(X_1, X_2, \ldots, X_n)$ og $\widehat{\theta}_U(X_1, X_2, \ldots, X_n)$?
- ⋆ Hva er tolkningen av intervallet $\left[\widehat{\theta}_L(x_1, x_2, \ldots, x_n), \widehat{\theta}_U(x_1, x_2, \ldots, x_n)\right]$?

Eksempel: Utledning av konfidensintervall

- Situasjon:
  - la $X_1, X_2, \dots, X_n$ være et tilfeldig utvalg fra $N(\mu, \sigma^2)$-populasjonen
  - verdien til μ er ukjent, ønsker et konfidensintervall for μ, antar verdien til σ^2 kjent!
- Estimator for μ: $\widehat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Vi vet at

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \Rightarrow \quad \frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$



- Dermed

$$P\left(-z_{\frac{\alpha}{2}} \le \frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \le z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

- Løser hver ulikhet hver for seg med hensyn på μ

$$\begin{aligned} -z_{\frac{\alpha}{2}} &\le \frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \\ -z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} &\le \overline{X} - \mu \\ -\overline{X} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} &\le -\mu \\ \mu &\le \overline{X} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} \end{aligned}$$

$$\begin{aligned} \frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} &\le z_{\frac{\alpha}{2}} \\ \overline{X} - \mu &\le z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} \\ -\mu &\le -\overline{X} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} \\ \overline{X} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} &\le \mu \end{aligned}$$

- Får at

$$P\left(\overline{X} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} \le \mu \le \overline{X} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}\right) = 1 - \alpha$$

- $(1-\alpha)\cdot 100\%$-konfidensintervall for μ: $\left[\overline{x} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}, \overline{x} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}\right]$

Utledning av konfidensintervall

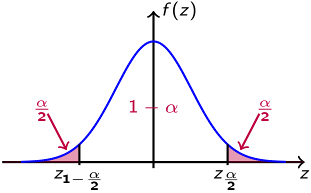
⋆ Prosedyre for å utlede et konfidensintervall for parameter θ basert på $x_1, x_2, \ldots, x_n$

1. Bestemme pivotal, $Z = h(\theta, X_1, X_2, \ldots, X_n)$
   - inneholder ingen ukjente parametre unntatt θ
   - har en kjent fordeling (som ikke avhenger av θ)

$$Z = \frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

2. Finne kvantiler $z_{1-\frac{\alpha}{2}}$ og $z_{\frac{\alpha}{2}}$ slik at

$$P\left(z_{1-\frac{\alpha}{2}} \leq h(\theta, X_1, X_2, \ldots, X_n) \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$



3. Løse hver ulikhet med hensyn på θ

4. Sette ulikhetene sammen igjen med θ i midten

$$P\left(\widehat{\theta}_L(X_1, X_2, \ldots, X_n) \leq \theta \leq \widehat{\theta}_U(X_1, X_2, \ldots, X_n)\right) = 1-\alpha$$

5. Konkludere ved å skrive opp $(1-\alpha)\cdot 100\%$-konfidensintervallet for θ

$$\left[\widehat{\theta}_L(x_1, x_2, \ldots, x_n), \widehat{\theta}_U(x_1, x_2, \ldots, x_n)\right]$$

Tolkning av konfidensintervall

$P(A)$: Andel av gangene hendelsen A skjer hvis man gjentar det stokastiske forsøket uendelig mange ganger

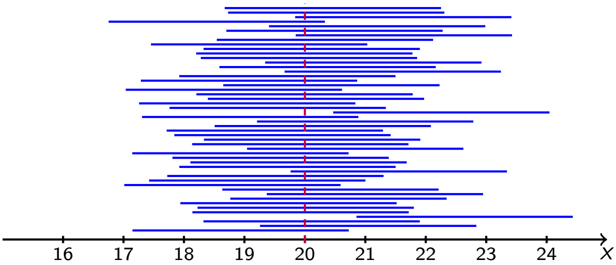
* Kan ikke tolke selve konfidensintervallet

$$\left[\widehat{\theta}_L(x_1, x_2, \ldots, x_n), \widehat{\theta}_U(x_1, x_2, \ldots, x_n)\right]$$

* Må tolke sannsynlighetsuttrykket

$$P\left(\widehat{\theta}_L(X_1, X_2, \ldots, X_n) \leq \theta \leq \widehat{\theta}_U(X_1, X_2, \ldots, X_n)\right) = 1 - \alpha$$

* Simuleringseksempel:
  - la $X_1, X_2, \ldots, X_n$ være tilfeldig utvalg fra $N(\mu, \sigma^2)$-populasjonen
  - verdien til μ er ukjent, ønsker konfidensintervall for μ, antar verdien til σ^2 kjent
  - $(1-\alpha) \cdot 100\%$-konfidensintervall for μ: $\left[\overline{x} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}, \overline{x} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}\right]$
  - Stokastisk simulering med $n = 30$, $\mu = 20$, $\sigma^2 = 5^2$ og $\alpha = 0.05$



Oppsummering

- Har definert begrepet
  - konfidensintervall
- Har diskutert
  - prosedyre for å utlede et konfidensintervall
    - tar utgangspunkt i en pivotal
  - tolkningen av konfidensintervall
    - tenker oss at vi gjentar målingene uendelig mange ganger