

TMA4245 – Parallel 1

Geir-Arne Fuglstad, IMF
3. mars 2025

Motivasjon – Konfidensintervall basert på tilfeldig utvalg

1. Vi har en $f(x; \theta)$ -populasjon der θ er en **ukjent** parameter vi ønsker å estimere.
2. Vi gjør et tilfeldig utvalg $X_1, X_2, \dots, X_n \sim_{\text{iid}} f(x; \theta)$.
3. Vi lager et intervallestimat $\hat{\theta}_L(x_1, \dots, x_n) \leq \theta \leq \hat{\theta}_U(x_1, \dots, x_n)$.
4. Utfordringen er å kvantifisere hvor sikker vi er på at ulikhetene holder!

Hvordan kan vi kvantifisere hvor sikker vi er på et intervallestimat $1.32 \leq \theta \leq 1.45$?

Konfidensintervall for tilfeldig utvalg

Løsning:

Vi kan bare garantere hvor sannsynlig det er at ulikhetene skal holde før vi gjør utvalget!

Definisjon:

Anta $X_1, X_2, \dots, X_n \sim_{\text{iid}} f(x; \theta)$ og $0 < \alpha < 1$. Da er

$$[\hat{\theta}_L(x_1, \dots, x_n), \hat{\theta}_U(x_1, \dots, x_n)]$$

et $100(1 - \alpha)\%$ -konfidensintervall for θ hvis

$$P\left(\hat{\theta}_L(X_1, \dots, X_n) \leq \theta \leq \hat{\theta}_U(X_1, \dots, X_n)\right) = 1 - \alpha.$$

NB: Vi kan også skrive at $\hat{\theta}_L(x_1, \dots, x_n) \leq \theta \leq \hat{\theta}_U(x_1, \dots, x_n)$ er et $100(1 - \alpha)\%$ -konfidensintervall for θ .

Tolkning:

Det var $100(1 - \alpha)\%$ sannsynlighet for at intervallet skulle inneholde θ før vi gjorde det tilfeldige utvalget. Med andre ord vil $100(1 - \alpha)\%$ av intervaller vi lager på denne måten inneholde θ .

Eksempel 1: Selvrapportert høyde og vekt

- Vi har 225 studenter fra USA
 - 99 menn
 - 126 kvinner
- Fra disse har vi selvrapportert
 - Høyde

Mål:

- Bestem et **95%-konfidensintervall** for gjennomsnittshøyde av **alle** mannlige studenter (og for alle kvinnelige)
- Bestem et **95%-konfidensintervall** for variansen av høydene til **alle** mannlige studenter (og for alle kvinnelige)

Eksempel 2: Estimere vaksinasjonsdekning i Nigeria

Bakgrunn:

- Folketall: 211 millioner (2021)
- Areal: 0.9 millioner km²
- Ingen fullstendige registre over vaksinerte

Mål:

- Bestem et **95%-konfidensintervall** for andel med første meslingvaksine blant **alle** 12—23 måneder gamle barn i 2017

Observert i 2017:

- 6059 barn 12—23 måneder
- 3227 av disse vaksinerte

NB: Vi skal late som dataene er et tilfeldig utvalg selv om det er mer komplisert enn dette i virkeligheten.



Konfidensintervall – mer generelt

Definisjon:

Anta $X_1, X_2, \dots, X_n$ er trukket fra simultanfordelingen $f(x_1, \dots, x_n; \theta)$ og $0 < \alpha < 1$. Da er

$$[\hat{\theta}_L(x_1, \dots, x_n), \hat{\theta}_U(x_1, \dots, x_n)]$$

et $100(1 - \alpha)\%$ -konfidensintervall for θ hvis

$$P\left(\hat{\theta}_L(X_1, \dots, X_n) \leq \theta \leq \hat{\theta}_U(X_1, \dots, X_n)\right) = 1 - \alpha.$$

NB: Vi kan også skrive at $\hat{\theta}_L(x_1, \dots, x_n) \leq \theta \leq \hat{\theta}_U(x_1, \dots, x_n)$ er et $100(1 - \alpha)\%$ -konfidensintervall for θ .

Tolkning:

Det var $100(1 - \alpha)\%$ sannsynlighet for at intervallet skulle inneholde θ før vi trakk de stokastiske variablene. Med andre ord vil $100(1 - \alpha)\%$ av intervaller vi lager på denne måten inneholde θ .

Monte Carlo

Velg en vilkårlig fordeling $f(x)$. Da er $\theta = E[g(X)]$ en parameter.

Hvis vi simulerer $X_1, X_2, \dots, X_n \sim_{\text{iid}} f(x)$ og n er stor gir sentralgrenseteoremet at

$$\overline{g(x)} - z_{\alpha/2} \frac{s}{\sqrt{n}} \leq E[g(X)] \leq \overline{g(x)} + z_{\alpha/2} \frac{s}{\sqrt{n}}$$

er et omtrent $100(1 - \alpha)\%$ -konfidensintervall, der

- $\overline{g(x)}$ er gjennomsnittet av simuleringene $g(x_1), \dots, g(x_n)$
- s er empirisk standardavvik av simuleringene $g(x_1), \dots, g(x_n)$
- $z_{\alpha/2}$ er kritisk verdi i standardnormalfordelingen

NB1: Konfidensnivået er litt feil på grunn av at 1) vi brukte sentralgrenseteoremet og 2) vi byttet standardavviket σ med estimatet s .

NB2: Vi kan ikke bruke t -fordelingen fordi $f(x)$ ikke nødvendigvis er en normalfordeling. I praksis er man villig til å anta $\sigma \approx s$ ved rundt $n \geq 30$.

Planlegging av undersøkelser

Hvis man ønsker en bestemt nøyaktighet i konfidensintervallet er det veldig viktig å planlegge antall forsøk n på forhånd. Dette gjelder både i virkelige undersøkelser og i Monte Carlo.

Eksempel:

$X_1, X_2, \dots, X_n \sim_{\text{iid}} N(\mu, \sigma^2)$ der μ er **ukjent** og σ^2 er **kjent**. Da vet vi at et $100(1 - \alpha)\%$ -konfidensintervall for μ er gitt ved

$$\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right].$$

Da kan man kontrollere størrelsen på intervallet ved å kontrollere n .

NB: Dere vil få andre konfidensintervall under andre antagelser og for andre estimatorer (se introduksjonsvideoer), men idéen er at ved å øke n blir konfidensintervallet smalere.

Prediksjonsintervall – Motivasjon

Konfidensintervaller kan brukes for **ukjente** parametere i en populasjon/fordeling, men vi trenger noe mer hvis vi ønsker å si noe om framtidige observasjoner basert på tidligere observasjoner.

Eksempel:

Populasjonen er personer i alder 20—40 år i Norge.

- **Konfidensintervall:** Forventet levealder, gjennomsnittslønn etter studiet, folkehelse
- **Prediksjonsintervall:** Din levealder, din lønn etter studiet, din helse

Prediksjonsintervall

Antagelser:

- $X_1, X_2, \dots, X_n$ trekkes fra simultanfordelingen $f(x_1, \dots, x_n; \theta)$
- X^* trekkes fra $g(x^*; \theta)$
- $0 < \alpha < 1$

Prediksjonsintervall:

Intervall

$$[\hat{X}_L(x_1, \dots, x_n), \hat{X}_U(x_1, \dots, x_n)]$$

er et $100(1 - \alpha)\%$ -prediksjonsintervall for X^* hvis

$$P\left(\hat{X}_L(X_1, \dots, X_n) \leq X^* \leq \hat{X}_U(X_1, \dots, X_n)\right) = 1 - \alpha.$$

Tolkning:

Det var $100(1 - \alpha)\%$ sannsynlighet for at intervallet skulle inneholde x^* før vi trakk $x_1, \dots, x_n$ **og** x^* . Med andre ord vil $100(1 - \alpha)\%$ av intervaller vi lager på denne måten inneholde x^* .