

TMA4245 – Parallell 1

Geir-Arne Fuglstad, IMF

24. februar 2025

Plan

- Oppsummere og koble alt det nye vokabularet og den nye tenkemåten til to konkrete eksempler.
- Vise eksempel på sannsynlighetsmaksimeringsestimator

Eksempel 1: Selvrapportert høyde og vekt

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

Mål:

- Estimere gjennomsnittshøyde og varians for **alle** mannlige studenter
- Estimere gjennomsnittshøyde og varians for **alle** kvinnlige studenter

Populasjoner og utvalg

Populasjon (generelt):

En (endelig eller uendelig) mengde av alle mulige verdier (eller individer) som vi kan observere.

Populasjon (dette kurset):

En populasjon er en uendelig stor mengde av tall og kan beskrives av en sannsynlighetsfordeling $f(x)$. Vi kaller populasjonen en $f(x)$ -populasjon.

Utvalg (generelt):

En endelig delmengde av de mulige verdiene (eller individene) som vi kan observere.

Utvalg (dette kurset):

En delmengde av n tall $x_1, x_2, \dots, x_n$ fra populasjonen.

Tilfeldig utvalg (generelt):

Et utvalg som er valgt slik at alle mulige utvalg hadde samme “sannsynlighet”.

Tilfeldig utvalg (dette kurset):

For en $f(x)$ -populasjon er et tilfeldig utvalg n stokastiske variabler $X_1, X_2, \dots, X_n \sim_{\text{iid}} f(x)$.

Parametere og observatorer

Parametere:

Disse er **uobserverbare, fikserte** størrelser i populasjonen.

F.eks.: Gjennomsnittshøyden av alle mannlige studenter, og varians til alle mannlige studenter.

Observatorer:

Disse er observerbare størrelser som regnes ut basert på verdiene i utvalget.

F.eks.: Gjennomsnittet av de valgte studentene og empirisk varians av de valgte studentene.

Utvalgsfordeling:

Fordelingen til observatoren under gjentatte tilfeldige utvalg.

Estimatorer og deres egenskaper

Estimator:

En observator Θ som brukes for å anslå en parameter θ (i populasjonen).

Ønskelige egenskaper:

1. Forventningsrett, $E[\Theta] = \theta$.
2. Mest effisient; lavest mulig varians $\text{Var}[\Theta]$.

NB: I oppgaver prioriterer dere 1. først og så 2..

Monte Carlo

Utgangspunkt:

Vi har $X \sim f(x)$ og ønsker å bestemme $E[g(X)]$.

Tidligere terminologi:

Vi simulerer verdier $x_1, x_2, \dots, x_n$ uavhengig av hverandre fra $f(x)$ og tilnærmer verdien av $E[g(X)]$ med

$$E[g(X)] \approx \frac{1}{n} \sum_{i=1}^n g(x_i) = \overline{g(x)}.$$

Mer presis terminologi:

Vi gjør et tilfeldig utvalg $X_1, X_2, \dots, X_n$ fra en $f(x)$ -populasjon og bruker

$$E[\widehat{g(X)}] = \frac{1}{n} \sum_{i=1}^n g(X_i) = \overline{g(X)}$$

som en estimator for $E[g(X)]$.

Eksempel 2: Estimere vaksinasjonsdekning i Nigeria

Bakgrunn:

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

Mål:

- Bestem 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.



Sannsynlighetsmaksimeringsestimatoren

Det er vanlig med forkortelsen SME eller MLE (maximum likelihood estimator).

Rimelighetsfunksjon:

For stokastiske variabler $X_1, \dots, X_n$ trukket fra simultanfordeling $f(x_1, \dots, x_n; \theta)$ der θ er en parameter av interesse, er rimelighetsfunksjonen gitt ved

$$L(\theta; x_1, \dots, x_n) = f(x_1, \dots, x_n; \theta).$$

NB: L er en funksjon av θ med $x_1, \dots, x_n$ fikserte, og f er en funksjon av $x_1, \dots, x_n$ med θ fiksert.

SME:

For stokastiske variabler $X_1, \dots, X_n$ trukket fra simultanfordeling $f(x_1, \dots, x_n; \theta)$ der θ er en parameter av interesse, er estimatet gitt ved

$$\hat{\theta} = u(x_1, \dots, x_n) = \arg \max_{\theta} \{L(\theta; x_1, \dots, x_n)\}.$$

Estimatoren er gitt ved

$$\hat{\Theta} = u(X_1, \dots, X_n) = \arg \max_{\theta} \{L(\theta; X_1, \dots, X_n)\}.$$

Ekstremt viktig i praksis!