

Introduksjon til parameterestimering

TMA4240/TMA4245 Statistikk

Håkon Tjelmeland

Institutt for matematiske fag

Norges teknisk-naturvitenskapelige universitet

Sannsynlighetsregning og statistisk inferens

- ★ Frem til nå har vi sett på situasjoner hvor vi
 - har en eller flere stokastiske variabler
 - kjenner fordelingen til disse, også verdien til parametrene
 - regner så ut sannsynligheter
- ★ Dette kalles sannsynlighetsregning
- ★ Ofte er situasjonen annerledes! Et eksempel:
 - interessert i andelen av defekte artikler, p , i et (stort) vareparti
 - undersøker n tilfeldig valgte artikler
 - la X være antall av de n undersøkte artiklene som var defekte
 - forstår at $X \sim \text{binomisk}(n, p)$
 - verdien til p er ukjent
 - ønsker å bruke observert verdi for X til å si noe om verdien til p
 - hva er vårt beste «gjett» på verdien til p ?
 - kan vi angi et intervall som verdien til p «sannsynligvis» er innenfor?
- ★ Et annet eksempel:
 - er interessert i et fysisk størrelse μ
 - gjør gjentatte (n) målinger av μ , MEN med målefeil!
 - for måling i : anta $X_i \sim N(\mu, \sigma^2)$
 - verdien til μ er ukjent, kanskje er også verdien til σ^2 ukjent
 - ønsker å bruke observerte verdier for $X_1, X_2, \dots, X_n$ til å si noe om μ
 - hva er vårt beste «gjett» på verdien til μ ?
 - kan vi angi et intervall som verdien til μ «sannsynligvis» er innenfor?
- ★ Dette kalles statistisk inferens
 - ut fra det vi har observert ønsker vi å si noe om hele populasjonen

Populasjon og utvalg

Definisjon (Populasjon og utvalg)

En populasjon består av alle mulige observasjoner man kan gjøre. Et utvalg er en delmengde av en populasjon.

- ★ Ønsker å bruke hva vi har observert på utvalget til å få informasjon/si noe om hele populasjonen
- ★ Et utvalg må være representativt
- ★ Dersom elementene i en populasjon har fordeling $f(x)$ sier vi at vi har en $f(x)$ -populasjon

Definisjon (Tilfeldig utvalg)

La $X_1, X_2, \dots, X_n$ være n uavhengige stokastiske variabler, alle med samme fordeling $f(x)$. Vi sier da at vi har et tilfeldig utvalg fra $f(x)$ -populasjonen.

- ★ Merk: Hvis vi har et tilfeldig utvalg fra $f(x)$ -populasjonen blir simultanfordelingen til $X_1, X_2, \dots, X_n$ lik

$$f(x_1, x_2, \dots, x_n) = f(x_1) \cdot f(x_2) \cdot \dots \cdot f(x_n) = \prod_{i=1}^n f(x_i)$$

- ★ Når observert verdier for et utvalg, må starte med å «se på» de observerte verdiene
 - regne ut tall som oppsummerer de viktigste egenskaper
 - gjennomsnitt, median, empirisk varians, empirisk standardavvik, kvartildifferanse
 - lage plott som visualiserer viktige egenskaper
 - histogram, boksplott, normalfordelingsplott, spredningsplott

Mål på lokasjon

Definisjon (Gjennomsnitt)

Anta at vi har observert et utvalg bestående av n observasjoner eller målinger, som vi betegner med $x_1, x_2, \dots, x_n$. Gjennomsnittet er da

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

- ★ Gjennomsnitt er det mest brukte målet på lokasjon
- ★ Gjennomsnitt er sensitivt for ekstreme verdier

Definisjon (Median)

Anta at vi har observert et utvalg bestående av n observasjoner eller målinger, som vi betegner med $x_1, x_2, \dots, x_n$. Sorter de observerte verdiene fra minst til størst, og la $x_{(i)}$ være den i -te minste verdien slik at $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. Medianen til datasettet er da

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{hvis } n \text{ er odde,} \\ \frac{1}{2} \left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)} \right) & \text{hvis } n \text{ er like.} \end{cases}$$

- ★ Median er ikke sensitivt for ekstreme verdier
- ★ Andre mål på lokasjon finnes: Hva bruker man ved bedømming av stil i skihopp?

Mål på variasjon

Definisjon (Empirisk varians og standardavvik)

Anta at vi har observert et utvalg bestående av n observasjoner eller målinger, som vi betegner med $x_1, x_2, \dots, x_n$. Empirisk varians er da

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

der $\bar{x}$ er gjennomsnittet. Empirisk standardavvik er

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

- ★ Empirisk standardavvik har samme dimensjon som x_i -ene
- ★ Empirisk varians og standardavvik er de mest brukte målene på variasjon
- ★ Empirisk varians og standardavvik er sensitive for ekstreme verdier

Definisjon (Empirisk kvantil og kvartildifferanse)

Anta at vi har observert et utvalg bestående av n observasjoner eller målinger, som vi betegner med $x_1, x_2, \dots, x_n$. Sorter de observerte verdiene fra minst til størst, og la $x_{(i)}$ være den i -te minste verdien slik at $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. For $\alpha = \frac{i-1}{n}$, $i \in \{1, 2, \dots, n\}$ er da den empiriske α -kvantilen gitt ved

$$q_\alpha(x_1, x_2, \dots, x_n) = x_{(i)}.$$

Kvartildifferansen (inter quartile range (IQR) på engelsk) er

$$IQR(x_1, x_2, \dots, x_n) = q_{0.75}(x_1, x_2, \dots, x_n) - q_{0.25}(x_1, x_2, \dots, x_n).$$

- ★ Kvartildifferansen er ikke sensitiv for ekstreme verdier

Oppsummering

- ★ Har definert begrepene
 - populasjon
 - utvalg
 - tilfeldig utvalg
 - gjennomsnitt
 - median
 - empirisk varians og standardavvik
 - kvartildifferanse
- ★ Har diskutert at vi ønsker å bruke verdier observert på et utvalg til å få informasjon om hele populasjonen
- ★ Starter med å «se på» de observerte verdiene:
 - mål på lokasjon
 - mål på variasjon
 - ulike typer plott (diskuterer i egen video)