Mot statistisk inferens

Vi ønsker å kartlegge norske stemmeberettiges holdning til EU.
Anta at andelen mot norsk medlemskap i EU er lik p.

Tenk at vi trekker 10 enkle utvalg på 1000 personer.
Da kan vi beregne de empiriske andelene $\hat{p}_1,\ldots,\hat{p}_{10}$.

Trolig er alle forskjellige, men i nærheten av p.

Å si noe om hvordan $\hat{p}_1,\ldots,\hat{p}_{10}$, eller $\hat{p}_1$ hvis vi bare tar ett utvalg, forholder seg til p er et eksempel på det som kalles **statistisk inferens**.

Følgende skille er fundamentalt i denne sammenhengen:

Parameter: Detter er en teoretisk størrelse som beskriver populasjonen. I eksemplet foran var p parameteren.

Statistikk: Dette er en empirisk størrelse som kan beregnes i utvalget. (I *norsk* statistisk ordbruk er det vanlig å bruke uttrykket observator for statistikk). I eksemplet var statistikken $\hat{p}$.

At $\hat{p}_1,\ldots,\hat{p}_{10}$ varierer skyldes den tilfeldige mekanismen utvalget er konstruert i henhold til.

Her er noen viktige egenskaper og begreper i forbindelse med utvalgsfordelinger.

Skjevhet dreier seg om sentret i utvalgsfordelingen. En estimator er **forventningsrett** hvis forventningen i utvalgsfordelingen er lik verdien til parameteren som skal estimeres.

Hvor variabel eller usikker en statistikk er, beskrives av spredningen i utvalgsfordelingen. Spredningen bestemmes av utvalget og størrelsen n på utvalget. Spredningen synker med utvalgsstørrelsen.

Neste figur illustrerer de to egenskapene ved en utvalgsfordeling.

Formell statistisk inferens

- To viktige metoder
  - *Konfidensintervall*
  - *Signifikanstester*
- Basert på *fordelingen* til en statistikk (observator)
- Krever *sannsynlighetsmodell* for dataene
- Statistisk inferens baserer seg på at dataene kommer fra et tilfeldig utvalg eller et randomisert eksperiment
  - Viktig å huske på!

Konfidensintervall intuitivt

- SATM-poeng:
  - Dersom peng for individene i populasjonen er N(μ,σ)-fordelt, vet vi at gjennomsnittet $\bar{x}$ er N(μ,σ/√n)-fordelt
  - Antar at vi vet at σ=100. For n=500 er da σ/√n=4.5
  - 68-95-99.7-regelen: $\bar{x}$ er i [μ-2σ/√n,μ+2σ/√n] = [μ-9,μ+9] med ca 95% sannsynlighet
  - Å si at $\bar{x}$ er 9 poeng mindre eller større enn μ er det samme som å si at μ er 9 poeng fra $\bar{x}$
  - Altså vil den sanne verdien av μ i 95% av utvalg vil ligge i intervallet:

Konfidensnivå for forventning

- Normalfordelte data: $\bar{x}$ er eksakt N(μ,σ/√n)-fordelt
- Sentralgrenseteorem for store utvalg: $\bar{x}$ er tilnærmet N(μ,σ/√n)-fordelt
- Vi så at vi kunne finne et omtrentlig konfidensintervall for μ ved å bruke 68-95-99.7-regelen
- Skal nå se hvordan vi lager mer presise konfidensintervall for μ
- Må starte med å finne feilmarginene for et bestemt konfidensnivå C
- Går veien om standard normalfordeling: Da kan vi finne generelle feilmarginer som alltid kan brukes for konfidensnivå C når gjennomsnittet $\bar{x}$ er normalfordelt

Noen forsiktighetsregler

- Data bør være fra et *enkelt randomisert utvalg* (SRS) av populasjonen
  - Viktig med *uavhengige* observasjoner fra populasjonen
- Andre, korrigerte formler for mer kompliserte design
- Følsomt for *uteliggere*
- Lite robust for små *n* (bør ha *n*>15) når data ikke er normalfordelte
- Må kjenne σ. Senere skal vi se på hva vi gjør når σ er ukjent
  - Hvis n stor, kan vi bruke [$\bar{x}$ -z*s/√n, $\bar{x}$ +z*s/√n] (som da er et *tilnærmet* konfidensintervall for μ)

6.2 Signifikanstester

- Konfidensintervaller er nyttige når vi ønsker å estimere en populasjonsparameter
- Signifikanstester er nyttige dersom vi ønsker å teste en hypotese om en parameter i en populasjon
- Bruker observerte data til å teste hypotesen om populasjonen
- Typisk prosedyre
  - Beregn sannsynlighet for observert utfall av en statistikk (eller noe mer ekstremt) gitt antatt hypotese
  - Hvis sannsynlighet liten, forkast hypotese

Teststatistikk

Baserer test på en statistikk som estimerer parameteren vi er interessert i (ofte den samme som vi ville brukt til et konfidensintervall for parameteren)

– Eks.: $\bar{x}_1$-$\bar{x}_2$ estimerer $\mu = \mu_1$-μ_2

•Verdier langt fra parameterverdi spesifisert av H_0 gir bevis mot H_0

•H_a angir hvilken retning som teller:

– Ensidig H_a: $\mu_1 > \mu_2$ angir at vi må ha stor $\bar{x}_1$-$\bar{x}_2$ som bevis mot H_0

– Ensidig H_a: $\mu_1 < \mu_2$ angir at vi må ha liten $\bar{x}_1$-$\bar{x}_2$ som bevis mot H_0

– Tosidig H_a: $\mu_1 \neq \mu_2$ angir at vi må ha stor $|\bar{x}_1$-$\bar{x}_2|$ som bevis mot H_0

Standardisert test-statistikk

For å undersøke hvor langt estimatet er fra parameterverdien spesifisert av H_0, standardiserer vi estimatet

$$z = \frac{estimat - parameterverdi\ under\ H_0}{standard\ avvik\ \ for\ estimat}$$

Signifikanstest: P-verdi

- *P-verdi*: Sannsynligheten for at et utfall er like ekstremt eller mer ekstremt enn faktisk utfall (beregnet ved å anta at parameterverdien gitt av H_0 er sann)
  - Ekstremt: Lang fra hva vi ville forvente hvis H_0 var sann. Retning på hva som regnes som ekstremt: Bestemmes av H_a og H_0
  - *Jo mindre P-verdien er, jo sterkere bevis har vi mot H_0*

Statistisk signifikans

Hvordan konkludere?

- Forkaster H_0 når *P-verdi* er liten nok

Signifikansnivå α: Grenseverdi for når vi forkaster

–Forkaster H_0 når *P-verdi* $\leq \alpha$

–Ikke grunnlag for å forkaste H_0 når P-verdi>α

–Typisk: α=0.05 (eller 0.01)

Signifikant betyr at bevisene mot nullhypotesen oppfyller standarden satt av α. Typisk utsagn er «Resultatene er signifikante ($P < 0.05$)»

- Hvis vi velger signifikansnivå α=0.05 krever vi at dataene gir bevis mot H_0 som bare vil skje i 5% av tilfellene hvis H_0 er sann
- Hvis vi velger signifikansnivå α=0.01 krever vi at dataene gir bevis mot H_0 som bare vil skje i 1% av tilfellene hvis H_0 er sann
  - *Vi krever altså sterkere bevis mot H_0 hvis vi velger α=0.01 enn hvis vi velger α=0.05*

Tester for populasjonsforventning

$H_0: \mu = \mu_0$

Data: $x_1, \ldots, x_n$

Estimator for μ: $\bar{x}$

Teststatistikk: $z = (\bar{x} - \mu_0)/\sigma_{\bar{x}} = (\bar{x} - \mu_0)/(\sigma\sqrt{n})$