STK1100 våren 2023

Estimering

Svarer til avsnitt 7.1 i læreboka

Matematisk institutt
Universitetet i Oslo

Politisk meningsmåling

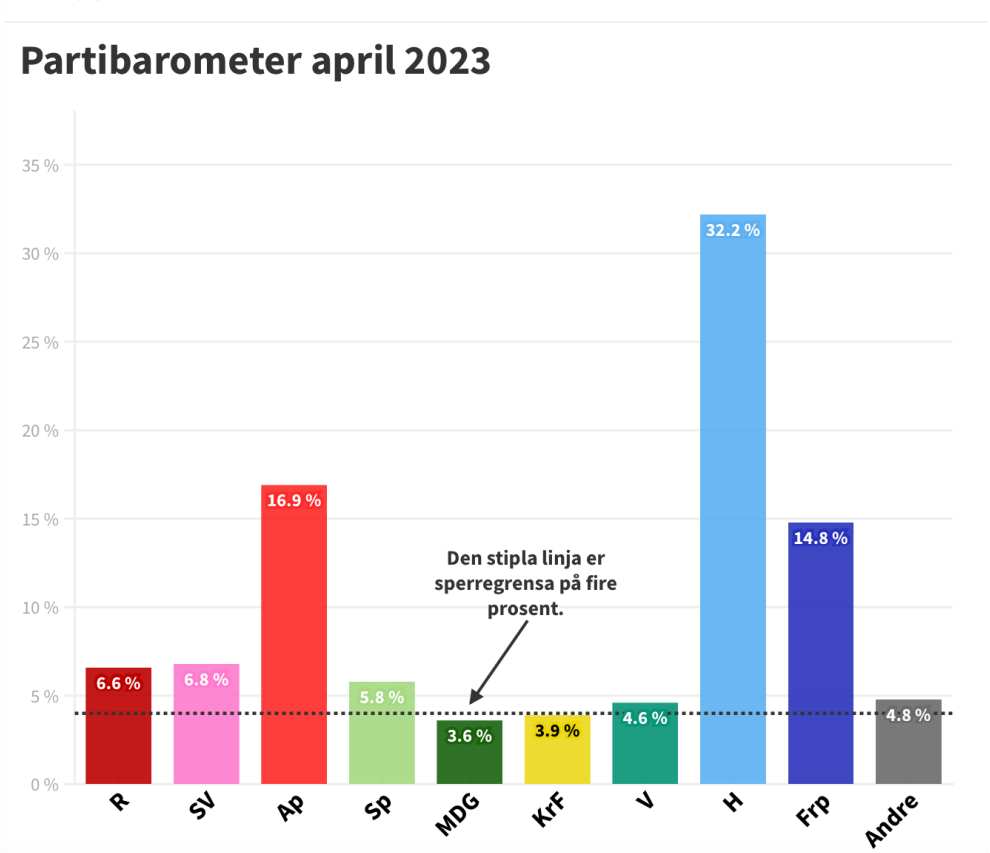
680 personer har svart på hva de ville ha stemt hvis det hadde vært stortingsvalg (april 23)

115 ville ha stemt Ap

Ap's oppslutning er

$$\frac{115}{680} = 0.169$$

Hvor sikkert er dette anslaget?



Laget av Opinion, 3.-11. april 2023

Måling av lungefunksjon

Et mål på lungefunksjon er FEV1 (forced expiratory volume in 1 second)

I en studie utført i Bergen ble FEV1 målt for 1642 ikke-røykende, friske menn i alder 30-34 år

Hvordan kan vi bruke informasjonen til å fastlegge et «normalområde» for FEV1 for menn i alder 30-34 år?
(I praksis vil en også ta hensyn til høyde og BMI, men det ser vi bort fra her)

The German tank problem

Under andre verdenskrig brukte de allierte (blant annet) serienumrene for tyske tanks til å anslå hvor mange tanks tyskerene hadde, og hvor stor produksjonen var for ulike måneder.

Hvordan kan serienumrene gi denne informasjonen?

Statistiske modeller

Felles for de tre situasjonene er at vi har data $x_1, x_2, ..., x_n$ for n enheter:

- For meningsmålingen er $x_i = 1$ hvis person nummer i ville ha stemt Ap, $x_i = 0$ ellers

- For målingene av lungefunksjon er x_i FEV1-målingen for person nummer i

- For tanksene er x_i serienummeret for den i-te tanksen de allierte fikk informasjon om

På grunnlag av utvalget, dvs. de observerte x_i-ene, ønsker vi å få kunnskap om den populasjonen observasjonene kommer fra

Vi må da ha en modell som angir hvordan de observerte x_i-ene fremkommer fra populasjonen

Vi vil anta at $x_1, x_2, ..., x_n$ er observerte verdier av stokastiske variabler $X_1, X_2, ..., X_n$ og at vi kjenner fordelingen til de stokastiske variablene - med unntak av en eller flere parametre, som er ukjente

- For meningsmålingen vil vi anta at $X_1, X_2, ..., X_n$ er uavhengige og Bernoulli-fordelte, dvs $P(X_i = 1) = p$ og $P(X_i = 0) = 1 - p$

- For målingene av lungefunksjon vil vi anta at $X_1, X_2, ..., X_n$ er uavhengige og $N(\mu, \sigma^2)$- fordelte

- For tanksene vil vi se på en forenklet situasjon og anta at $X_1, X_2, ..., X_n$ er et tilfeldig utvalg (uten tilbakelegging) blant tallene $1, 2, ..., N$

I alle de tre tilfellene ønsker vi å anslå verdien av en eller flere ukjente parametere, dvs. p, μ, σ^2 og N

Generelt vil vi anta at $x_1, x_2, ..., x_n$ er observerte verdier av stokastiske variabler $X_1, X_2, ..., X_n$ og at X_i-ene har en fordeling som avhenger av en parameter θ (det kan være flere parametere, men vi vil fokusere på én om gangen nå)

Vi vil anslå verdien til θ (eller estimere verdien som det heter på «statistikerspråket») på grunnlag av observasjonene våre

Til det bruker vi en estimator $\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_n)$

På grunnlag av de observerte x_i-ene , får vi estimatet $\hat{\theta} = \hat{\theta}(x_1, x_2, ..., x_n)$

Bernoulli variabler og binomisk fordeling

Anta at $X_1, X_2, ..., X_n$ er uavhengige og Bernoulli-fordelte, dvs $P(X_i = 1) = p$ og $P(X_i = 0) = 1 - p$

Da er $$Y = \sum_{i=1}^{n} X_i \sim \text{binomisk}(n, p)$$

En naturlig estimator for p er $$\hat{p} = \frac{Y}{n}$$

For meningsmålingen har vi $n = 680$ og vi observerte $y = 115$

Vi får dermed estimatet $$\hat{p} = \frac{y}{n} = \frac{115}{680} = 0.169$$

En annen estimator for p er $p^* = \frac{Y+2}{n+4}$

For meningsmålingen gir denne estimatet

$$p^* = \frac{y+2}{n+4} = \frac{117}{684} = 0.171$$

Det er liten forskjell på $\hat{p}$ og p^* her

Forskjellen er større når n er mindre og y er nær 0 eller n

Hvordan kan vi avgjøre hvilken av estimatorene $\hat{p}$ og p^* som er best?

Mean square error (MSE)

Vi ser på den generelle situasjonen der $X_1, X_2, ..., X_n$ har en fordeling som avhenger av en parameter θ

Vi ønsker at estimatoren $\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_n)$ skal være nær θ

Konkret ønsker vi at

$$\mathrm{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

skal være så liten som mulig

Merk at (detaljer på forelesningen)

$$\mathrm{MSE}(\hat{\theta}) = V(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2 = \text{varians} + (\text{skjevhet})^2$$

Se på situasjonen der Y : binomisk(n, p)

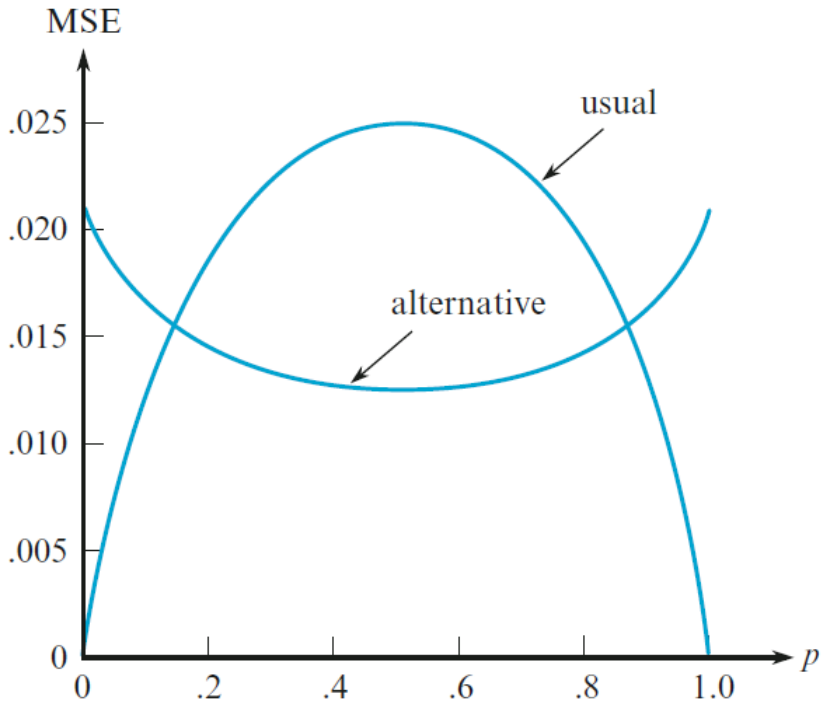
Vi har estimatorene $\hat{p} = \frac{Y}{n}$ og $p^* = \frac{Y+2}{n+4}$

Her er (detaljer på forelesningen)

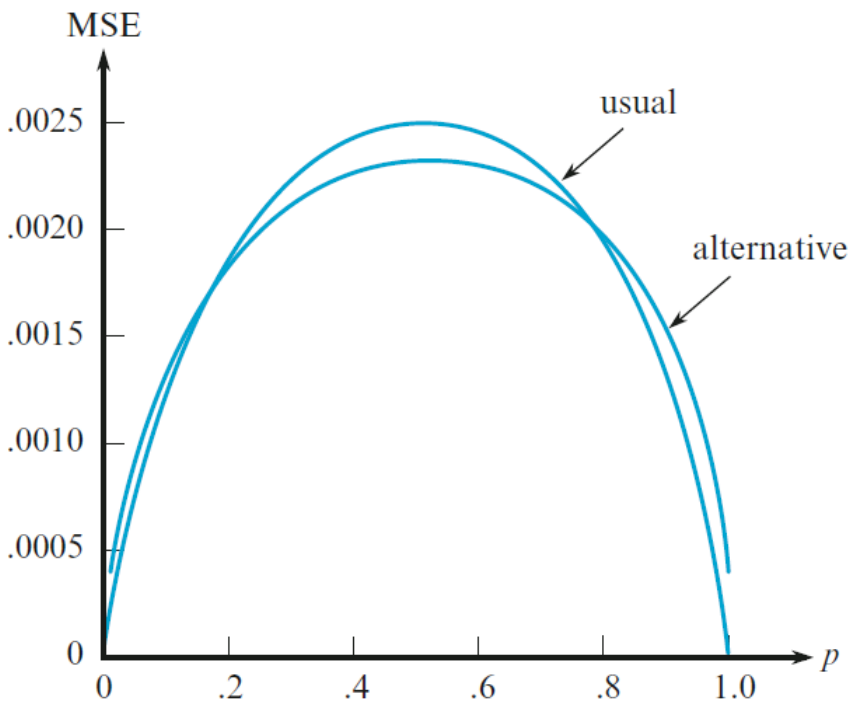
$$\mathrm{MSE}(\hat{p}) = \frac{p(1-p)}{n}$$

$$\mathrm{MSE}(p^*) = \frac{p(1-p)}{n+8+16/n} + \left(\frac{2/n - 4p/n}{1+4/n} \right)^2$$

MSE for $\hat{p}$ («usual») og p^* («alternative»)



$n = 10$



$n = 100$

Forventningsrette estimatorer

Vi ser på den generelle situasjonen med en estimator $\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_n)$ for θ

Hvis $E(\hat{\theta}) = \theta$ for alle mulige verdier av θ, sier vi at $\hat{\theta}$ er forventningsrett (engelsk: unbiased)

For en forventningsrett estimator er skjevheten $E(\hat{\theta}) - \theta$ lik 0, og det følger at

$$\mathrm{MSE}(\hat{\theta}) = V(\hat{\theta})$$

For den binomiske situasjonen er $\hat{p} = \frac{Y}{n}$ en forventningsrett estimator for p

Uavhengige og identisk fordelte variabler

Anta at $X_1, X_2, ..., X_n$ er uavhengige og identisk fordelte (u.i.f.) med forventning μ og varians σ^2

Da er

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

en forventningsrett estimator for μ og

$$V(\hat{\mu}) = V(\bar{X}) = \frac{\sigma^2}{n}$$

Videre er (detaljer på forelesningen)

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

en forventningsrett estimator for σ^2

Estimatorene $\hat{\mu} = \bar{X}$ og S^2 er spesielt aktuelle for normalfordelte data (for andre fordelingen kan det være at andre estimatorer er bedre)

For FEV1-målingene var $\bar{x} = 4.48$ og $s = \sqrt{s^2} = 0.60$

FEV1-målinger for 30-34 år gamle ikke-røykende, friske menn er (ca.) normalfordelt med forventnings-verdi 4.48 liter og standardavvik 0.60 liter

Uniform diskret fordeling

Anta at $X_1, X_2, ..., X_n$ er et tilfeldig utvalg uten tilbakelegging blant tallene $1, 2, ..., N$, der N er ukjent (jf. «the German tank problem»)

Vi vil finne en forventningsrett estimator for N

La $Y_n = \max X_i$ være det største tallet vi observerer

Da er (se eget notat)

$$E(Y_n) = \frac{n(N+1)}{n+1}$$

og en forventningsrett estimator for N er

$$\hat{N} = \frac{n+1}{n} Y_n - 1$$

Ved å bruke formler av denne typen (sammen med diverse annen informasjon) kunne de allierte anslå størrelsen av den tyske produksjonen av tanks i ulike måneder (og på ulike steder):

Date	Estimated Monthly Production	
	Serial Number Estimate	Munitions Record 10 Aug. 42
June, 1940	169	1000
June, 1941	244	1550
August, 1942	327	1550

Ruggles & Brodie (1947). An Empirical Approach to Economic Intelligence in World War II. Journal of the American Statistical Association, Vol. 42, pp.72-91

Uniform kontinuerlig fordeling

Anta at $X_1, X_2, ..., X_n$ er uavhengige og uniform fordelte over $[0, \theta]$

Vi vil estimere θ

Vi har at $E(X_i) = \frac{\theta}{2}$ og $V(X_i) = \frac{\theta^2}{12}$

Vi har at $\theta^* = 2\bar{X}$ er en forventningsrett estimator for θ og at $V(\theta^*) = \theta^2 / (3n)$
(detaljer på forelesningen)

Vi kan få en bedre estimator ved å ta utgangspunkt i

$$Y_n = \max X_i$$

Vi har at (detaljer på forelesningen)

$$E(Y_n) = \frac{n\theta}{n+1}$$

$$V(Y_n) = \frac{n\theta^2}{(n+2)(n+1)^2}$$

En forventningsrett estimator for θ er dermed

$$\hat{\theta} = \frac{n+1}{n} Y_n$$

Vi finner at

$$V(\hat{\theta}) = \frac{\theta^2}{n(n+2)}$$

Vi har $V(\hat{\theta}) < V(\theta^*)$ hvis $n > 1$

Konsistente estimatorer

Vi ser på den generelle situasjonen, og antar at $\hat{\theta}$ er en forventingsrett estimator for θ

Chebyshevs ulikhet gir at for alle $k > 0$ er

$$P\left(|\hat{\theta} - \theta| \geq k\sqrt{V(\hat{\theta})} \right) \leq \frac{1}{k^2}$$

Vi lar $\varepsilon > 0$ og setter $k = \varepsilon / \sqrt{V(\hat{\theta})}$

Da får vi at $P\left(|\hat{\theta} - \theta| \geq \varepsilon \right) \leq \frac{V(\hat{\theta})}{\varepsilon^2}$

Hvis $V(\hat{\theta}) \to 0$ når antall observasjoner $n \to \infty$, vil

$$P\{|\hat{\theta} - \theta| < \varepsilon\} \to 1$$

Vi sier at $\hat{\theta}$ er konsistent

Anta at Y : binomisk(n, p)

Da er $\hat{p} = \frac{Y}{n}$ en forventningsrett estimator for *p*

Vi har at

$$V(\hat{p}) = \frac{p(1-p)}{n} \quad \rightarrow 0$$

når $n \rightarrow \infty$, så $\hat{p}$ er konsistent

Anta at $X_1, X_2, ..., X_n$ er uavhengige og $N(\mu, \sigma^2)$-fordelte

Da er $\hat{\mu} = \bar{X}$ en konsistent estimator for μ

Standardfeil

Når vi rapporterer resultatet av en undersøkelse, bør vi ikke nøye oss med å oppgi et estimatet. Vi bør også si noe om hvor presist estimatet er. Det er da vanlig å oppgi (et estimat for) standardavviket

Standardavviket $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$ til en estimator $\hat{\theta}$ blir vanligvis kalt standardfeilen til estimatoren

Ofte vil $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$ avhenge av en eller flere ukjente parmetere. Hvis vi estimerer disse, får vi den estimerte standardfeilen $s_{\hat{\theta}}$

(I praktisk statistikk er det vanlig å kalle $s_{\hat{\theta}}$ for standardfeilen)

For meningsmålingen fikk vi estimatet

$$\hat{p} = \frac{y}{n} = \frac{115}{680} = 0.169$$

Her er

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

og den estimerte standardfeilen blir

$$s_{\hat{p}} = \sqrt{\frac{0.169 \cdot 0.831}{680}} = 0.0144$$

For FEV1-målingene fikk vi $\hat{\mu} = \overline{x} = 4.48$ og $s = \sqrt{s^2} = 0.60$

Her er

$$\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$$

og den estimerte standardfeilen blir

$$s_{\hat{\mu}} = \frac{0.60}{\sqrt{1642}} = 0.0148$$

Bootstrap

For enkle situasjoner kan vi finne et uttrykk for standardfeilen $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$ til en estimator

Men hvis estimatoren og/eller fordelingen til X_i-ene er komplisert, kan det være vanskeligere å finne et slikt uttrykk (som i tanks-eksempelet)

Da kan vi bruke stokastisk simulering til å finne et estimat for standardfeilen $s_{\hat{\theta}}$

Denne metoden kalles vanligvis for (parametrisk) bootstrap, som nevnes i kap. 7.1. Parametrisk og ikke-parametrisk bootstrap kommer vi tilbake til. Men først: kap. 7.2 – om metoder for å finne formler for estimatorer $\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_n)$