

# STK1100 våren 2018

## Estimering

Svarer til avsnitt 7.1 i læreboka

Ørnulf Borgan  
Matematisk institutt  
Universitetet i Oslo

1

## Måling av lungefunksjon

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



I en studie i Hordaland på 1990-tallet ble FEV1 målt for 1642 ikke-røykende, friske menn i alder 30-34 år

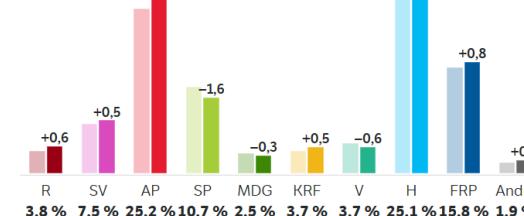
Hvordan kan vi bruke informasjonen til å fastlegge et «normalområde» for FEV1 for menn i alder 30-34 år?  
(I praksis vil en også ta hensyn til høyde og BMI, men det ser vi bort fra her)

3

## Politisk meningsmåling

Partibarometer april 2018

Norstat for NRK. Periode 3/4–9/4. 925 intervjuer. Feilmarginer fra 1,2–3,2 pp.



Spør et tilfeldig utvalg på 925 personer hva de ville ha stemt hvis det hadde vært valg. 233 ville ha stemt Ap

Ap's oppslutning er  $\frac{233}{925} = 0.252$

Hvor sikkert er dette anslaget?

2

## The German tank problem



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

Hvordan kan serienumrene gi denne informasjonen?

4

## Statistiske modeller

Felles for de tre situasjonene er at vi har data  $x_1, x_2, \dots, x_n$  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

5

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 framkommer fra populasjonen

Vi vil anta at  $x_1, x_2, \dots, x_n$  er observerte verdier av stokastiske variabler  $X_1, X_2, \dots, X_n$  og at vi kjenner fordelingen til de stokastiske variablene (med unntak av en eller flere parametere)

6

- For meningsmålingen vil vi anta at  $X_1, X_2, \dots, X_n$  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  $X_1, X_2, \dots, X_n$  er uavhengige og  $N(\mu, \sigma^2)$ -fordelte
- For tanksene vil vi anta at  $X_1, X_2, \dots, X_n$  er et tilfeldig utvalg (uten tilbakelegging) blant tallene  $1, 2, \dots, N$

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

7

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

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

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

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

8

## Bernoulli variabler og Binomisk fordeling

Anta at  $X_1, X_2, \dots, X_n$  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  $n = 925$  og vi observerte  $y = 233$

Vi får dermed **estimatet**  $\hat{p} = \frac{y}{n} = \frac{233}{925} = 0.252$

9

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

For meningsmålingen gir denne estimatet

$$p^* = \frac{y+2}{n+4} = \frac{235}{929} = 0.253$$

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?

10

## Mean square error (MSE)

Vi ser på den generelle situasjonen der  $X_1, X_2, \dots, X_n$  har en fordeling som avhenger av en parameter  $\theta$

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

Konkret ønsker vi at

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

skal være så liten som mulig

Merk at (detaljer på forelesningen)

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

Se på situasjonen der  $Y \sim \text{binomisk}(n, p)$

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

Her er (detaljer på forelesningen)

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

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

12

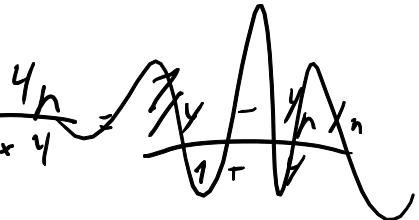
$$MSE(\hat{p}) = V(\hat{p}) + \text{Schierekt}^2$$

A (ternativer) estimator

$$E(\hat{p}) = E\left(\frac{Y+2}{n+y}\right) = \frac{E(Y)+2}{n+y} = \frac{np+2}{n+y}$$

Denest

$$\text{Schierekt} = E(p^*) - p = \frac{np+2}{n+y} - p = \frac{2-p}{n+y}$$



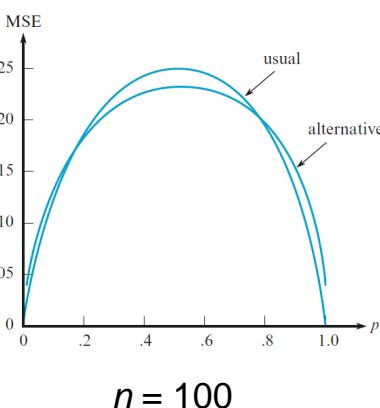
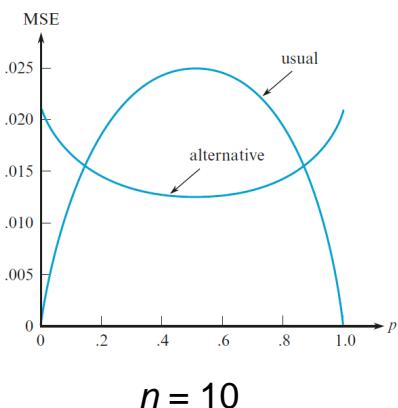
For variation

$$\begin{aligned} V(p^*) &= V\left(\frac{Y+2}{n+y}\right) = \frac{1}{(n+y)^2} V(Y) = \frac{np(1-p)}{(n+y)^2} \\ &= \frac{p(1-p)}{n+8+16/n} \end{aligned}$$

Denest

$$MSE(p^*) = V(p^*) + \text{Schierekt} = \underbrace{\frac{p(1-p)}{n+8+16/n}}_{\dots} + \dots$$

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



13

## Forventningsrette estimatorer

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

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

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

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

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

14

## Uavhengige og identisk fordelte variabler

Anta at  $X_1, X_2, \dots, X_n$  er uavhengige og identisk fordelte (u.i.f.) med forventning  $\mu$  og varians  $\sigma^2$

Da er

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

en forventningsrett estimator for  $\mu$  og

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

15

Videre er (detaljer på forelesningen)

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

en forventningsrett estimator for  $\sigma^2$

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, frikle menn er (ca.) normalfordelt med forventningsverdi 4.48 liter og standardavvik 0.60 liter

16

$x_1, x_2, \dots, x_n$  märk. sene fördelning

og  $E(x_i) = \mu$  og  $V(x_i) = \sigma^2$ ?

Hör att

$$\hat{\mu} = \bar{x}$$

estimator för  $\mu$ . Föreställer med

Hurden estimerar  $\sigma^2$ ?

Hvis  $\mu$  kända katt hittat, kvarna vi estimerar  $\sigma^2$   
med

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2} \quad \text{X}$$

För vi att

$$E(\hat{\sigma}^2) = \frac{1}{n} \sum_{i=1}^n E\{(x_i - \mu)^2\} \stackrel{V(x_i) = \sigma^2}{=} \frac{1}{n} (\sigma^2 + \sigma^2 + \dots + \sigma^2)$$
$$= \frac{1}{n} n \sigma^2 = \sigma^2$$

Hör nu räkt att  $\mu$  är okänd. Prova med estimationen

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{X}$$

Här

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

Merkt at

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n\bar{x}^2 \\ &= \sum_{i=1}^n x_i^2 - n\bar{x}^2\end{aligned}$$

Merkt osai at for enkvar Y har vi at

$$V(Y) = E(Y^2) - (EY)^2$$

...

$$E(Y^2) = V(Y) + (EY)^2$$

För denna at

$$E(\hat{\sigma}^2) = E \left\{ \frac{1}{n} (\sum x_i^2 - n\bar{x}^2) \right\}$$

$$= \frac{1}{n} \sum_{i=1}^n \left( E(x_i^2) - nE(\bar{x}^2) \right)$$

$$= \frac{1}{n} \left\{ \sum_{i=1}^n \left( \underbrace{V(x_i)}_{\hat{\sigma}^2} + \underbrace{E(x_i)}_{\mu^2}^2 \right) - n \left( \underbrace{V(\bar{x})}_{\frac{\sigma^2}{n}} + \underbrace{(E\bar{x})^2}_{\mu^2} \right) \right\}$$

$$= \frac{1}{n} (n\sigma^2 - r^2) = \frac{n-1}{n} \sigma^2$$

der  $\hat{\sigma}^2$  er ikke forventningsrett

En forventningsrett estimator for  $\sigma^2$  er

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

## Uniform diskret fordeling

Anta at  $X_1, X_2, \dots, X_n$  er et tilfeldig utvalg uten tilbakelegging blant tallene  $1, 2, \dots, N$ , 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

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

17

Ved å bruke denne formelen (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		Monthly Production Speer Ministry
	Serial Number Estimate	Munitions Record 10 Aug. 42	
June, 1940	169	1000	122
June, 1941	244	1550	271
August, 1942	327	1550	342

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

Vi vil estimere  $\theta$

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

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

Vi kan få en bedre estimator ved å ta utgangspunkt i

$$Y_n = \max X_i$$

19

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  $\theta$  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  $n > 1$

20

## Konsistente estimatorer

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

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  $k = \varepsilon / \sqrt{V(\hat{\theta})}$

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

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

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

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

21

Anta at  $Y \sim \text{binomisk}(n, p)$

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

Vi har at

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

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

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

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

22

## Standardfeil

Når vi rapporterer resultatet av en undersøkelse, bør vi ikke nøye oss med å oppgi et estimat. 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 parametere. 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.)

23

For meningsmålingen fikk vi estimatet

$$\hat{p} = \frac{y}{n} = \frac{233}{925} = 0.252$$

Her er

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

og den estimerte standardfeilen blir

$$s_{\hat{p}} = \sqrt{\frac{0.252 \cdot 0.748}{925}} = 0.0143$$

24

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

Her er

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

og den estimerte standardfeilen blir

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

25

## 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

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

Denne metoden kalles vanligvis for (parametrisk) **bootstrap**

26

Anta at  $x_1, x_2, \dots, x_n$  er observerte verdier av stokastiske variable  $X_1, X_2, \dots, X_n$  og at  $X_i$ -ene har en fordeling  $f(x_1, x_2, \dots, x_n; \theta)$  som avhenger av en parameter  $\theta$  (eventuelt flere parametere)

Ut fra de observerte  $x_i$ -ene får vi estimatet  $\hat{\theta}$

For  $b = 1, 2, \dots, B$  gjør vi nå følgende:

- Genererer et bootstrap-utvalg  $x_1^*, x_2^*, \dots, x_n^*$  fra fordelingen  $f(x_1, x_2, \dots, x_n; \hat{\theta})$
- Beregner estimatet  $\hat{\theta}_b^*$  ut fra bootstrap-utvalget (på samme måte som  $\hat{\theta}$  ble beregnet ut fra de opprinnelige observasjonene)

27

Bootstrap estimatet for standardfeilen er da det empiriske standardavviket for  $\hat{\theta}_b^*$ -ene, dvs

$$s_{\hat{\theta}} = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}_b^* - \bar{\theta}^*)^2}$$

Det er litt komplisert å bestemme standardfeilen for estimatet i «the German tank problem»

Men med bootstrap-metoden går det greit (se kommandoer på neste side)

28

Anta for illustrasjon at  $n = 16$  og at det største serienummeret vi observerer, er 160. Da får vi estimatet

$$\hat{N} = \frac{16+1}{16} 160 - 1 = 169$$

Kommandoene nedenfor gir et bootstrap estimat for standardfeilen

```
n=16;
B=1000;
Nhat=zeros(1,B);
for b=1:B
    Xmax=max(randsample(169,n));
    Nhat(b)=((n+1)/n)*Xmax-1;
end
std(Nhat)
ans =
10.0980
```