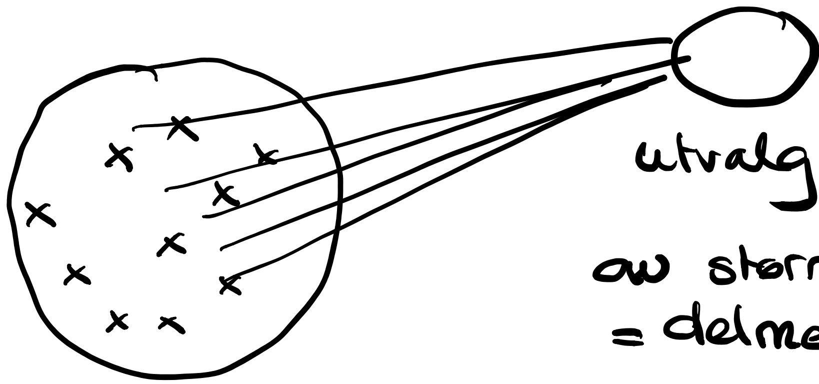
TMA 4240
F15, kap 8

Bolk 2:
STATISTISK INFERENS

Fra populasjon til utvalg [8.1]



utvalg (sample) av størrelse n
= delmengde av populasjonen

populasjon = alle enheter/individer vi ønsker å studere

↑ bør være representativt for hele populasjonen

Tilfeldig utvalg (random sample)
= enhetene som trekkes fra populasjonen velges tilfeldig og uavhengig av hverandre.

Formelt: $X_1, X_2, \ldots, X_n$ SV
stok. var

med sannsynlighetsfordeling $f(x)$,

trukket tilfeldig.

$$f(x_1, x_2, \dots, x_n) = f(x_1) \cdot f(x_2) \cdots f(x_n)$$

simultanfordeling

marginalford.

uavhengighet

$X_1, X_2, \dots, X_n$ kalles uavhengige (u)

identisk (i)

fordelte (f)

observasjoner

u. i. f

engelsk i. i. d

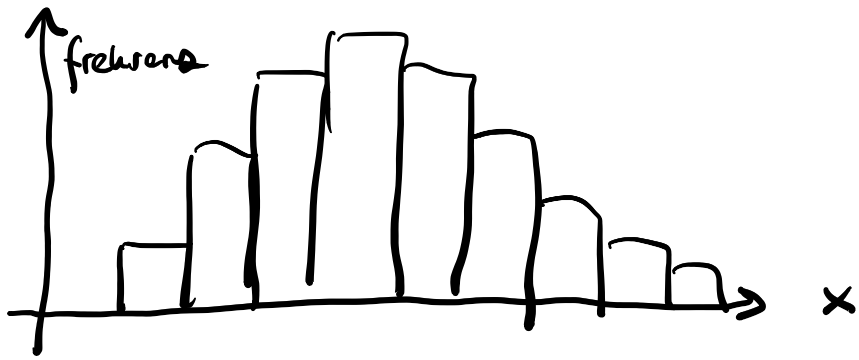
distributed

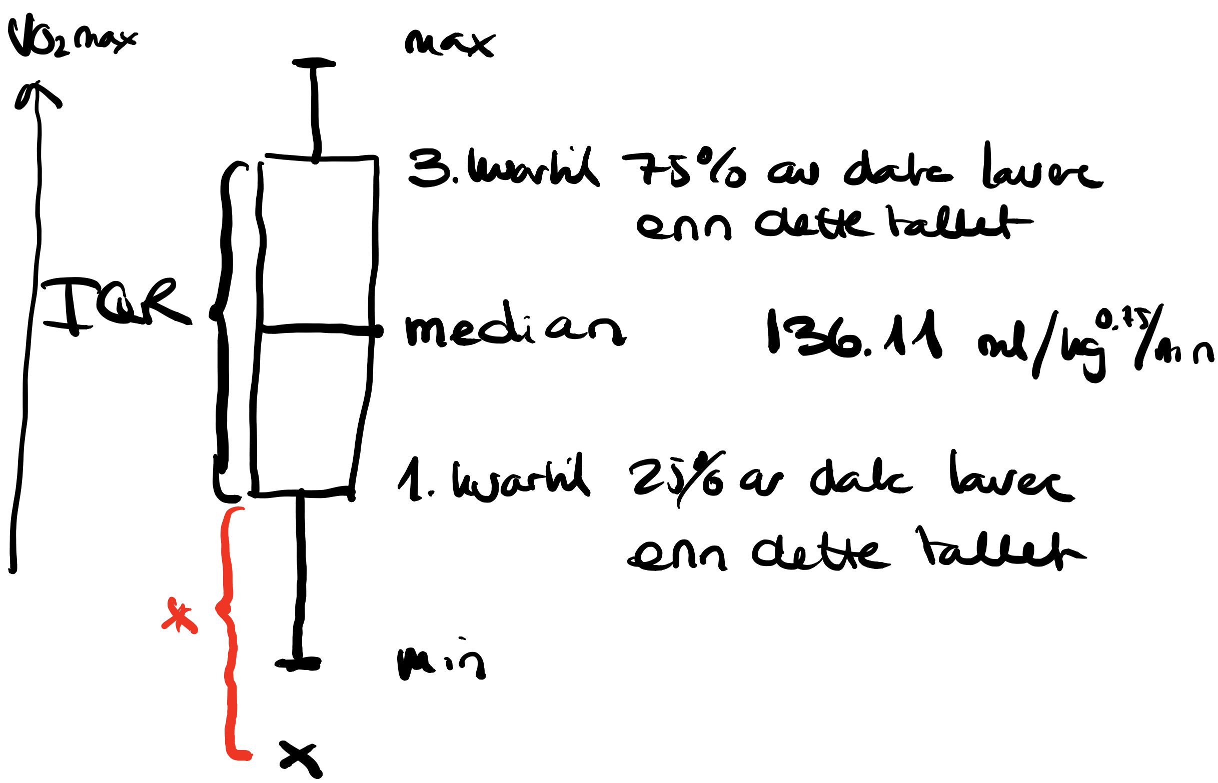
independent

identically

EKS: VO_2max data

$n = 1477$ obs av oksygenopptak





IQR = interquartile range

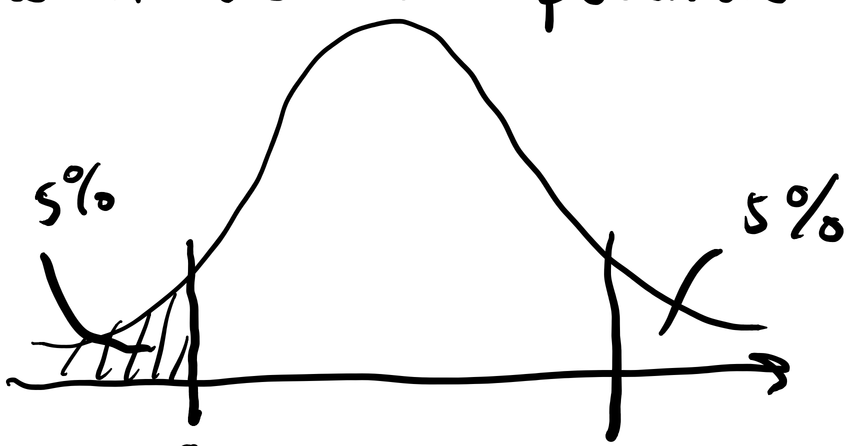
* > 1.5 · IQR ⇒ tegnes enkeltobs

Ordbruk: kvartil, kvantil, persentil & kritisk verdi

Utvalg: sorter og del inn i 4 deler (kvartil)

1. kvartil: 25% data er ≤ dette tallet
2. — " — 50% : median
3. — " — 75%
4. — " — 100%

Deler inn i 100 deler → persentiler



↑ 5 prosentil

kritisk verdi ↑ for ulike valg av prosentil

Kvantil: generelt tall.

Vi trekker et tilfeldig utvalg for å lære om populasjonen

- anslå (estimere) ukjente parameter i $f(x)$
  μ, σ, p, λ

- teste en hypotese

[8.2-8.3]

Observatorer og utvalgsfordelinger

Observator (eng: statistic) er en funksjon av SV i et tilfeldig utvalg.
SV: stokastiske variabler

Fordelingen til en observator heter en utvalgsfordeling (sampling distribution).

De to viktigste observatorene er

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

utvalgsgjsn utvalgsvarians

lokasjon

si noe om

$\mu = E(X)$

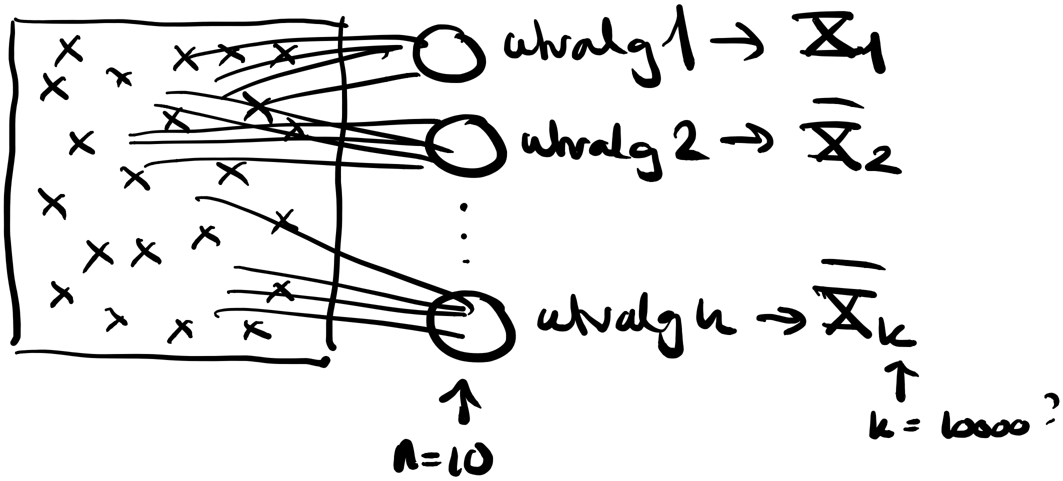
spredning

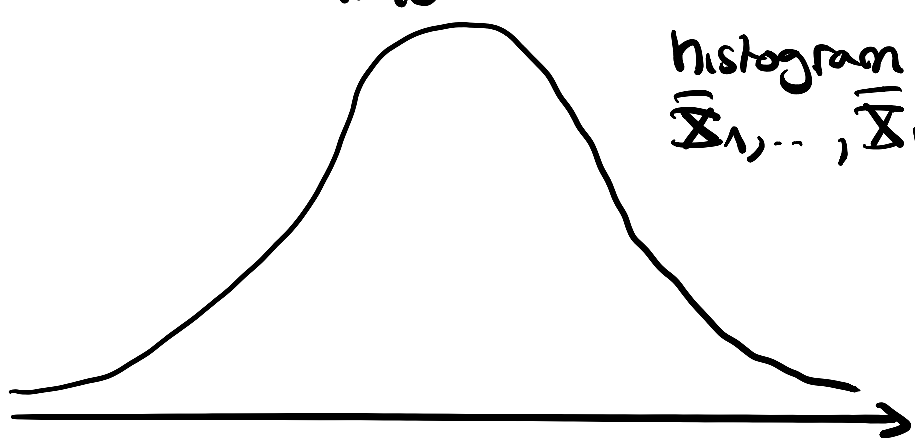
si noe om

$\sigma^2 = Var(X)$
$= E[(X-\mu)^2]$

[8.4]

Fordelingen til utvalgsgjennomsnittet $\bar{X}$





histogram av $\bar{X}_1, \dots, \bar{X}_k$

Obs: $\bar{X}$ er en stokastisk variabel og har derfor en fordeling.

1) $X_1, X_2, \ldots, X_n$ u.i.f $f(x)$

med $E(X_i) = \mu$ og $Var(X_i) = \sigma^2$

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Finn $E(\bar{X})$ og $Var(\bar{X})$.

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{E(X_i)}_{\mu}$$

$$= \frac{1}{n} \cdot n \cdot \mu = \underline{\underline{\mu}}$$

$$Var(\bar{X}) = \underset{\substack{\uparrow \\ E\left((\bar{X}-\mu)^2\right)}}{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 Var\left(\sum_{i=1}^{n} X_i\right)$$

$$\underset{\substack{\uparrow \\ \text{uavhengig}}}{=} \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \underbrace{Var(X_i)}_{\sigma^2} = \left(\frac{1}{n}\right)^2 n \cdot \sigma^2$$

$$= \underline{\underline{\frac{\sigma^2}{n}}}$$

$\Rightarrow$ gjelder uansett hva $f(x)$ er

2) Hvis $X_1, X_2, \dots, X_n$ er u.i.f

normal (μ, σ) $\quad N(\mu, \sigma^2)$

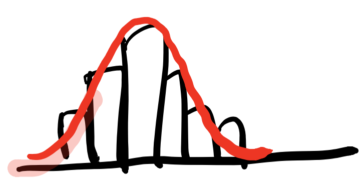
varians

to notasjoner

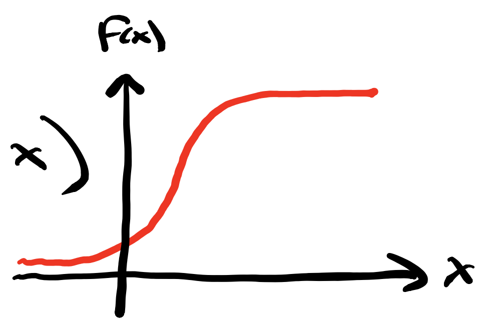
$\bar{X}$ vil også være normal fordelt med

$E(\bar{X}) = \mu$ og $Var(\bar{X}) = \frac{\sigma^2}{n}$, $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.

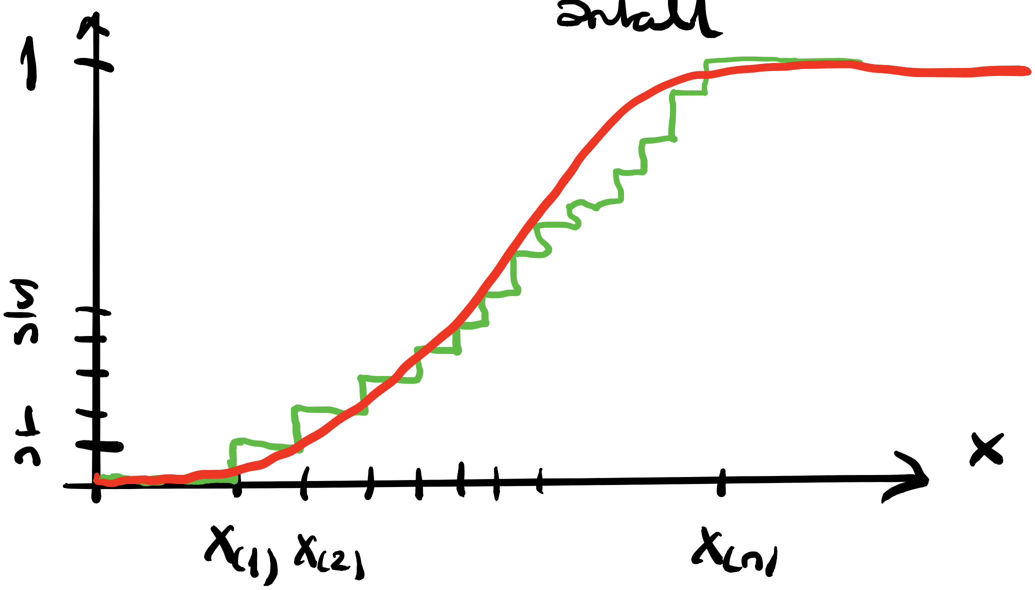
Hvordan kan sjekke at $X_1, \dots, X_n$
er et tilfeldig utvalg
fra normalfordelingen. [8.8]

a)  $f(x)$

b) $F(x) = P(X \leq x)$



Empirisk $\hat{F}(x) = \frac{\underbrace{\#}_{\text{antall}} X_i \le x}{n} \xrightarrow{n \to \infty} F(x)$



Hvis X er $N(\mu, \sigma^2)$ vil

$$\hat{F}(x) \approx \Phi\left(\frac{x-\mu}{\sigma}\right)$$

$$\Phi^{-1}(\hat{F}(x)) \approx \frac{x-\mu}{\sigma}$$

Hvis vi tar invers funksjon ender vi opp med en rett linje hvis data er normalfordelt.

$\Rightarrow$ Matlab normplot

NB: finnes også qq plot som
gjør små modifikasjoner på normplot.

$\Rightarrow$ rett linje!

3) nestegang $\rightarrow$ sentralgrenseteoremet.