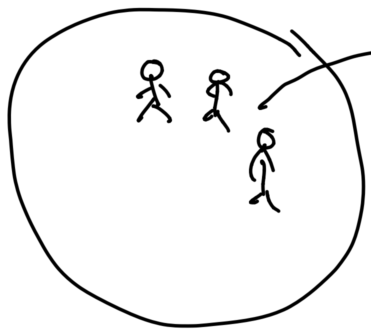
TMA4240 H2017 L15 09.10.2017

Del 1: Sannsynlighetsregning (1-7)



POPULASJON

stokastisk variabel

kjente parametre

$$X \sim N(\mu, \sigma^2)$$

høyden til en tilfeldig valgt student

$E(X)$

$Var(X)$

$SD(X) = \sqrt{Var(X)}$

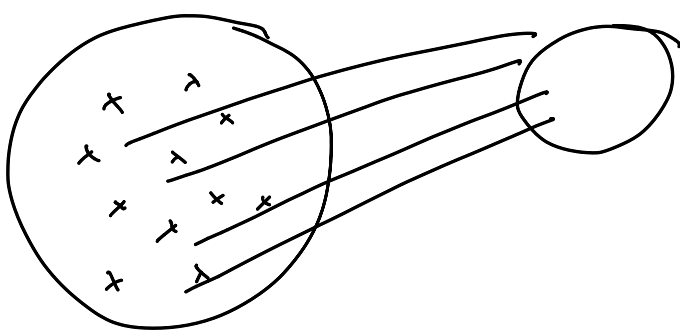
$P(X > 185)$?

Del 2: Statistisk inferens

finn ut: trekke konklusjoner om egenskaper til en populasjon

knytte til parametre

Fra populasjon til utvalg [8.1]



POPULASJON = alle enheter/individ vi ønsker å studere

UTVALG (sample) av størrelse n = delmengde av populasjonen

bør være representative for hele populasjonen

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

Formelt: $X_1, X_2, X_3, \ldots, X_n$ SV

stokastiske variable

med sannsynlighetsfordeling $f(x)$:

uavhengighet

$$f(x_1, x_2, \ldots, x_n) = f(x_1) \cdot f(x_2) \cdot f(x_3) \cdots f(x_n)$$

simultanfordeling produkt av marginalfordelinger

$X_1, X_2, \ldots, X_n$ kalles

uavhengige	(u)	i
identisk	(i)	i
fordelte	(f)	d

u.i.f observasjoner
(engelsk i.i.d)

Fremover (del 2): vi trekker et tilfeldig utvalg for å lære om populasjoner:

- anslå (estimere) ukjente parametere i $f(x)$

  f.eks: μ, σ^2, p

- teste en hypotese.

Observatorer og utvalgsfordelinger [8.2-8.3]

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

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

De to viktigste observatorene er

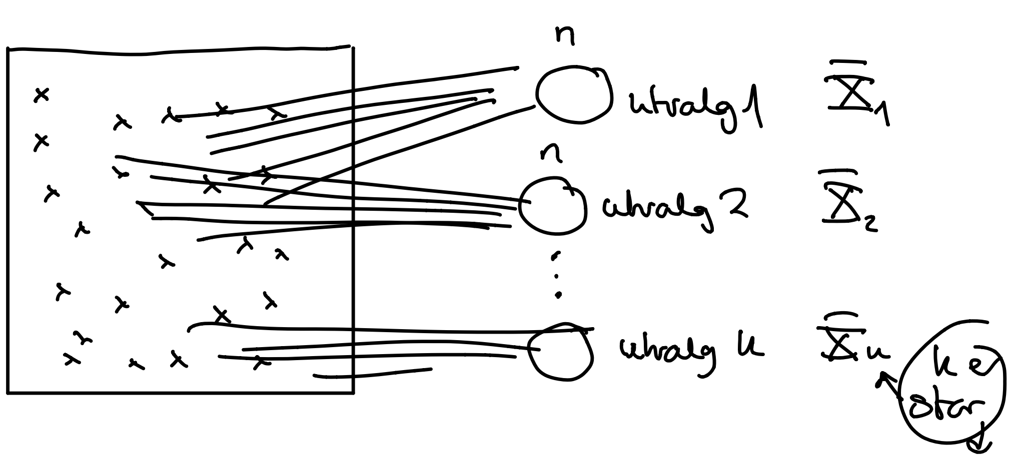
$$\boxed{\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

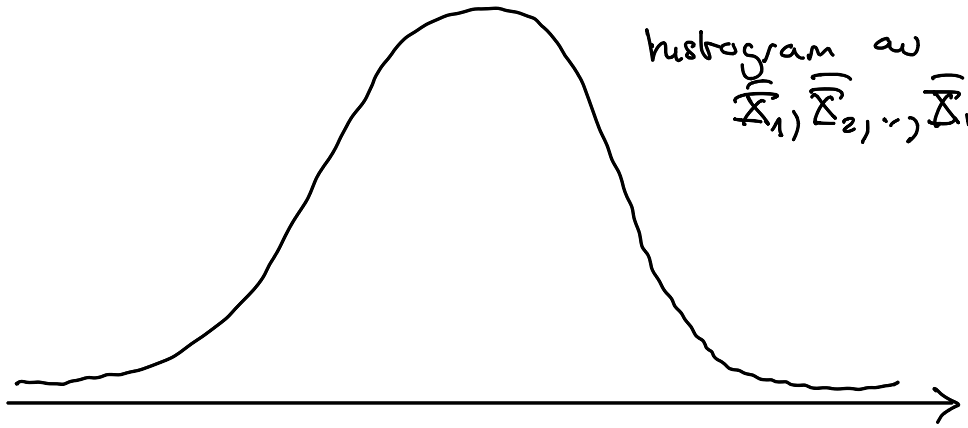
utvalgsgjennomsnitt — utvalgsvarians

lokasjon, forventning i populasjon $\mu = E(X)$

spredning, si noe om $Var(X) = E[(X-\mu)^2]$

Fordelingen til gjennomsnittet $\bar{X}$





histogram av $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_n$

$\Rightarrow$ Vi har "vist" at $\bar{X}$ er en SV 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 $Var(\bar{X})$.

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

$$\left[\begin{array}{l} E(aX) = aE(X) \\ E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) \end{array}\right]$$

$$E(\bar{X}) = \frac{1}{n}\left(\underbrace{\mu + \mu + \ldots + \mu}_{n \text{ ganger}}\right) = \frac{1}{n} \cdot n \cdot \mu = \underline{\underline{\mu}}$$

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

$$\left[\quad Var(a \cdot X) = a^2 Var(X) \qquad E((X-\mu)^2) = Var(X) \qquad Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) \quad\right]$$

X_i ene er uavhengige

X_i ene er uavhengige

$$Var(\bar{X}) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \underbrace{Var(X_i)}_{\sigma^2} = \left(\frac{1}{n}\right)^2 (\underbrace{\sigma^2 + \sigma^2 + \dots + \sigma^2}_{n \text{ ganger}})$$

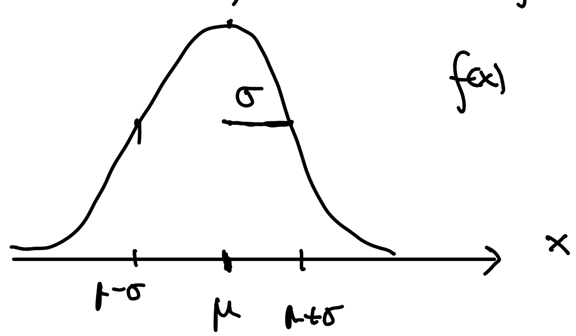
$$= \left(\frac{1}{n}\right)^2 n\sigma^2 = \underline{\underline{\frac{\sigma^2}{n}}} \qquad SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Gjelder uansett hva $f(x)$ er.

2) Hvis $X_1, \dots, X_n$ er u.i.f normal(μ, σ) $N(\mu, \sigma^2)$

sd

varians



$\bar{X}$ vil være normalfordelt såen $X_1, \dots, X_n$ er det

en sum av N vil også være N

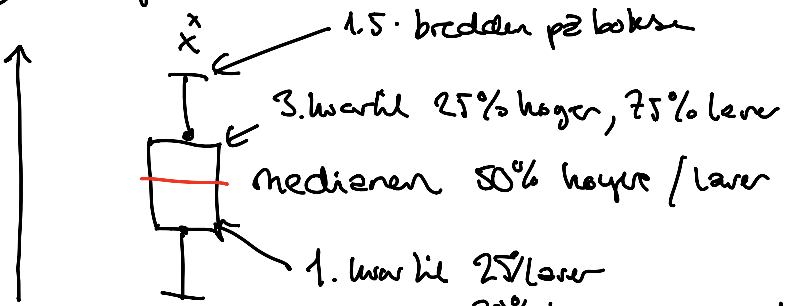
med $E(\bar{X}) = \mu$ og $Var(\bar{X}) = \frac{\sigma^2}{n}$, $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

Men, hvordan kan jeg sjekke at $X_1, .., X_n$ er et tilfeldig utvalg fra normalfordelingen [8.8]

---

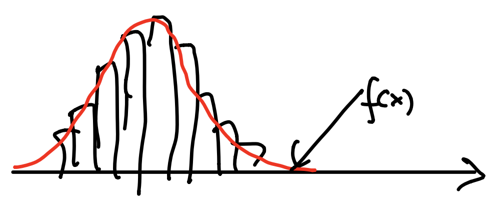
a) boxplott

IQR = inter quartile range

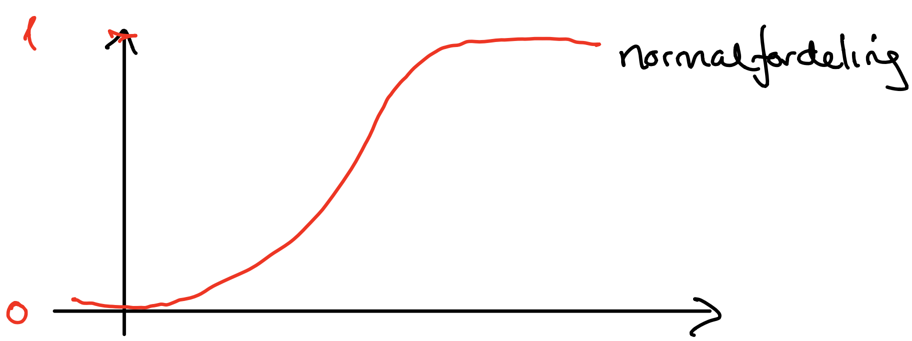


N-fordeling er symmetrisk – da ligger medianen midt i boksen

b) histogram ← $f(x)$



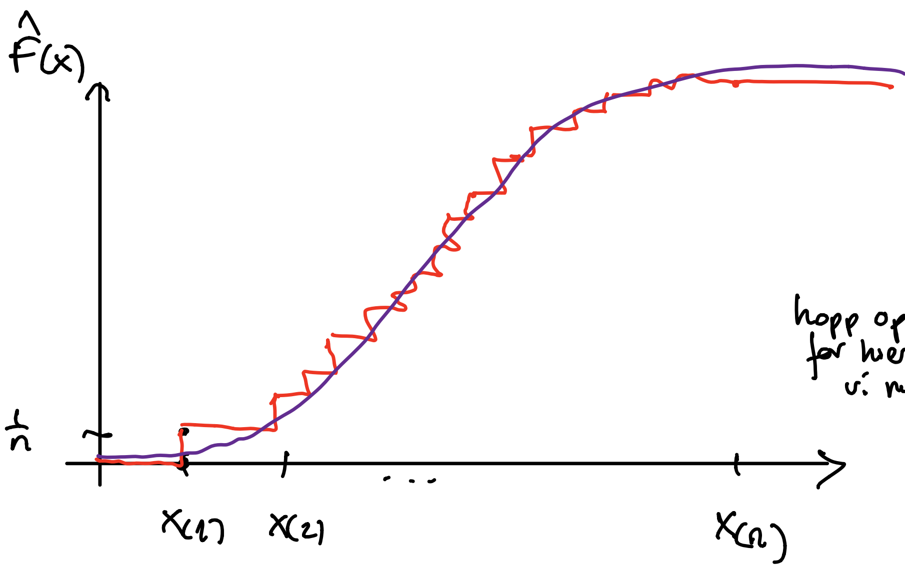
c) $F(x) = P(X \le x)$ (cdf)



Empirisk $\hat{F}(x) = \dfrac{\# X_i \le x}{n} \xrightarrow{n \to \infty} F(x)$

n = # observasjon

= antall



hopp opp med $\frac{1}{n}$ for hver ny obs vi møter

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

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

Φ er cdf i $N(0,1)$

kumulativ ford. funksjon

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

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

Matlab: normplot.

3) Hva hvis X ikke er normalfordelt?
Hva med $\bar{X}$ da?

Når n er stor ($n \geq 30$) bruker vi sentralgrense-teoremet

$$Z = \frac{\bar{X} - E(\bar{X})}{SD(\bar{X})} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{n \to \infty} N(0,1)$$

Selv om X ikke er normalfordelt vil $\bar{X}$ når n er stor være normalfordelt.

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

Se slides for hvor raskt dette går for noen fordelinger.