

Kursus 02403

Introduktion til Statistik

Forelæsning 1: Intro og beskrivende statistik

Jan Kloppenborg Møller

DTU Informatik

Bygning 305 - rum 210

Danmarks Tekniske Universitet

2800 Lyngby – Danmark

e-mail: jkm@imm.dtu.dk

- 1 Praktisk Information
- 2 Introduktion til Statistik
- 3 Software: R
- 4 Beskrivende statistik: Nøgletal
- 5 Grafisk fremstilling

Oversigt

- 1 Praktisk Information
- 2 Introduktion til Statistik
- 3 Software: R
- 4 Beskrivende statistik: Nøgletal
- 5 Grafisk fremstilling

Praktisk Information

- Undervisning: Hver dag 9.00-13.00
 - 2 timer forelæsning (dagens pensum)
 - 2 timer regneøvelser (gerne vha. R)
- Selvstudie
 - R træningsopgaver
 - Ekstraopgaver
 - Læse forud (bog + Rnote)
- Eksamen: 4 timer multiple choice (fredag 22/6)

Praktisk Information

- Hjemmeside: www.imm.dtu.dk/courses/02403
 - Forelæsningsoversigt inkl. slides (Oploades før hver forelæsning):
www.imm.dtu.dk/courses/02403/pensumliste.html
 - Opgaver & løsninger:
www.imm.dtu.dk/courses/02403/opgaver.html
- Campusnet: www.campusnet.dtu.dk
 - Meddelelser og konference
 - Fildeling

Oversigt

- 1 Praktisk Information
- 2 Introduktion til Statistik
- 3 Software: R
- 4 Beskrivende statistik: Nøgletal
- 5 Grafisk fremstilling

Introduktion til Statistik - kursusoversigt

- Simpel dataanalyse (idag)
- Fordelinger (diskrete og kontinuerte)
- Inferens og hypotesprøvning
- Simulation

Introduktion til Statistik

- Hvordan behandle (eller analysere) data?
- Hvad er tilfældig variation?
- Statistik er et værktøj til at træffe beslutninger:
 - Hvor mange computere har vi solgt det sidste år?
 - Hvad er forventet pris af en aktie?
 - Er maskine A mere effektiv end maskine B ?
- Statistik er et metodefag, der kan anvendes inden for de fleste fagområder, og er derfor et meget vigtigt værktøj

Statistik og Ingeniører

- Statistik er et vigtigt værktøj i problemløsning
- Analyse af data
- Kvalitetforbedring
- Forsøgsplanlægning
- Forudsigelse af fremtidige værdier
- .. og meget mere!

Statistik

- Moderne statistik har baggrund i sandsynlighedsregning og beskrivende statistik

Statistik

- Statistik handler ofte om at analysere en *stikprøve*, der er taget fra en *population*
- Baseret på stikprøven, prøver vi at generalisere (eller udtale os) om populationen
- Det er derfor vigtigt, at stikprøven er *repræsentativ* for populationen

Oversigt

- 1 Praktisk Information
- 2 Introduktion til Statistik
- 3 Software: R
- 4 Beskrivende statistik: Nøgletal
- 5 Grafisk fremstilling

Software: R

- www.r-project.org
- Appendix C i bogen (7. og 8. udg.): Beskrivelse af R.
- R Commander: en grafisk/menustyret grænseflade.
- R-øvelse i dag.
- R på Gbaren kan køres hjemmefra.(Thinlinc)
- R kan installeres på egen computer.(Se R-note)

Oversigt

- 1 Praktisk Information
- 2 Introduktion til Statistik
- 3 Software: R
- 4 Beskrivende statistik: Nøgletal
- 5 Grafisk fremstilling

Kapitel 2: Nøgletal

- Vi anvender en række *nøgletal* for at opsummere og beskrive data (stokastiske variable)
- Gennemsnit $\bar{x}$
- Median $\tilde{x}$
- Varians s^2
- Standardafvigelse s
- Fraktiler Q_τ

Eksempel

I en politisk debat om indførelsen af en fattigdoms grænse i Danmark anføre en politiker at medianindkomsten er et dårligt mål fordi den vil flytte sig meget (nedad) hvis Mærsk vælger at flytte til udlandet.
Er udsagnet sandt? Hvorfor/hvorfor ikke

Gennemsnit

- Gennemsnittet er et nøgletal, der angiver tyngdepunkt eller centrering af data
- Gennemsnit:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- Vi siger, at $\bar{x}$ er et *estimat* af middelværdien

Median

Medianen er et også nøgletal, der angiver centrering af data. I nogle tilfælde, f.eks. hvis man har ekstreme værdier, er medianen at foretrække frem for middelværdien

- Median: Den midterste observation, i den sorterede rækkefølge $(x_{(1)} \leq \dots \leq x_{(n)})$

$$\tilde{x} = \begin{cases} x_{(k)} & , n = 2k - 1 \\ \frac{1}{2}(x_{(k)} + x_{(k+1)}) & , n = 2k \end{cases}$$

Varians og standardafvigelse

Variansen (eller standardafvigelsen) siger noget om hvor meget data spredter:

- Varians

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

- Standardafvigelse (spredning)

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

Variationskoefficient

Standardafvigelse og variansen er nøgletal for den absolutte variation. Hvis man gerne vil sammenligne variationen mellem forskellige datasæt, er det en god idé at anvende et relativt nøgletal, nemlig variationskoefficienten:

$$V = \frac{s}{\bar{x}} \cdot 100$$

Fraktiler

Medianen beregnes som det punkt, der deler data ind i to halvdele. Man kan naturligvis finde andre punkter, der deler data ind i andre dele, og det man kalder fraktiler.

Ofte beregner man fraktilerne

- 0, 25, 50, 75, 100 % fraktiler og/eller
- 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 % fraktiler
- Bemærk: 50% fraktilen svarer til medianen

Beregning af Fraktiler

Lad $(x_{(1)} \leq \dots \leq x_{(n)})$ betegne den sorterede rækkefølge, da kan p -fraktilen beregnes som

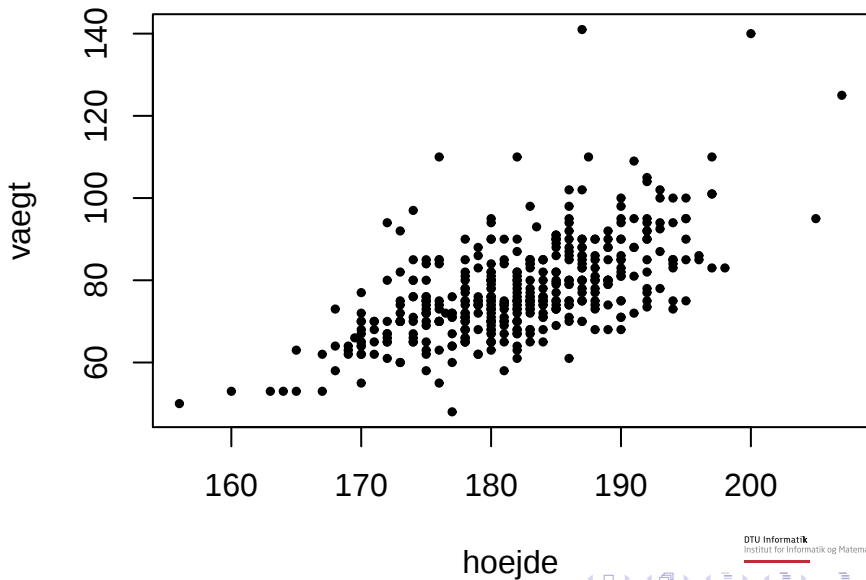
$$Q_p = \begin{cases} x_{(\lceil np \rceil)} & ; \text{ hvis } np \text{ ikke er et heltal} \\ \frac{1}{2}(x_{(np)} + x_{(np+1)}) & ; \text{ hvis } np \text{ er et heltal} \end{cases}$$

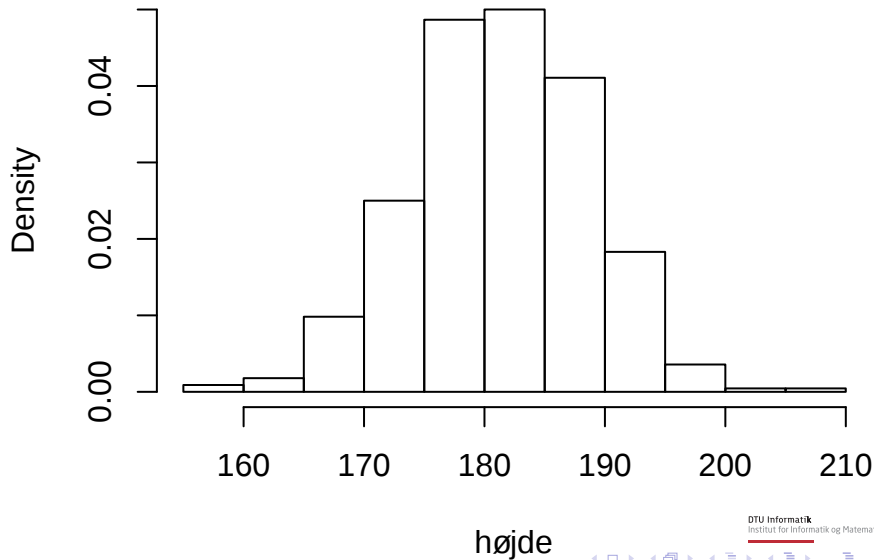
Oversigt

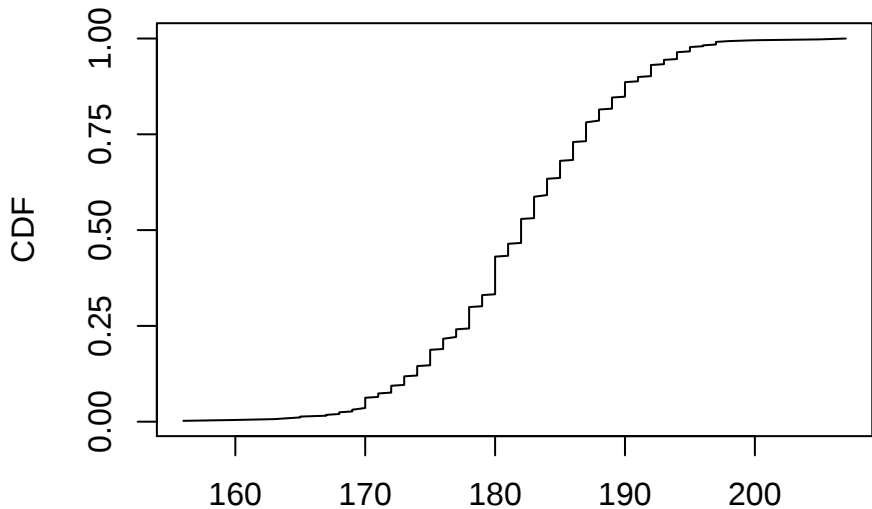
- 1 Praktisk Information
- 2 Introduktion til Statistik
- 3 Software: R
- 4 Beskrivende statistik: Nøgletal
- 5 Grafisk fremstilling

Figurer/Tabeller

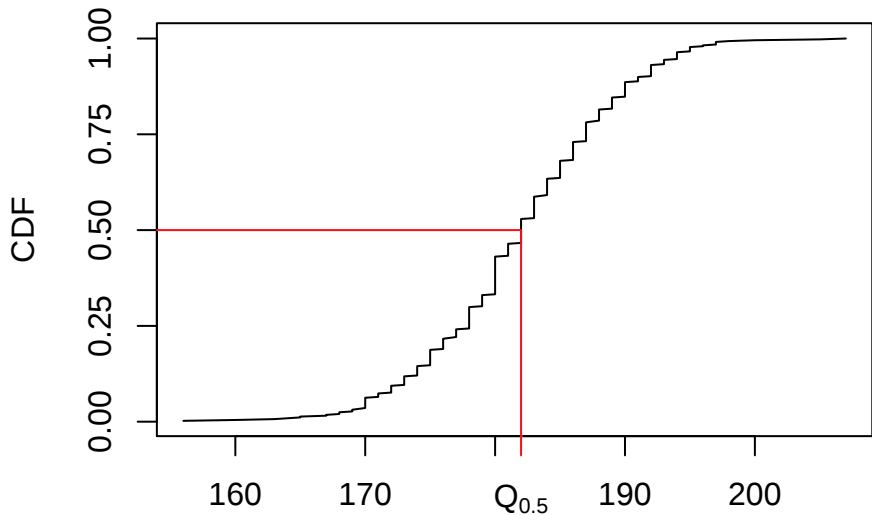
- Kvantitative data:
 - Scatter plot (xy plot)
 - Histogram
 - Kumulativ fordeling
 - Boxplots
- Antalsdata:
 - Bar charts (pareto diagram)
 - Pie charts



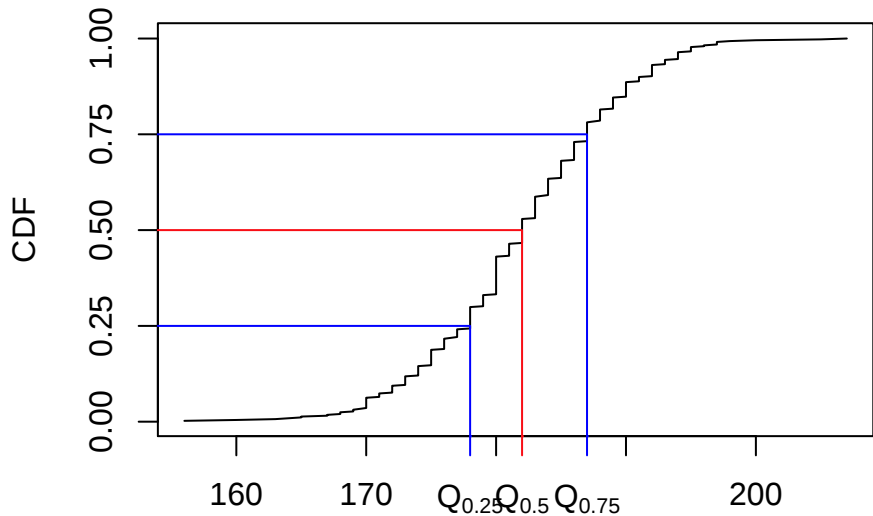


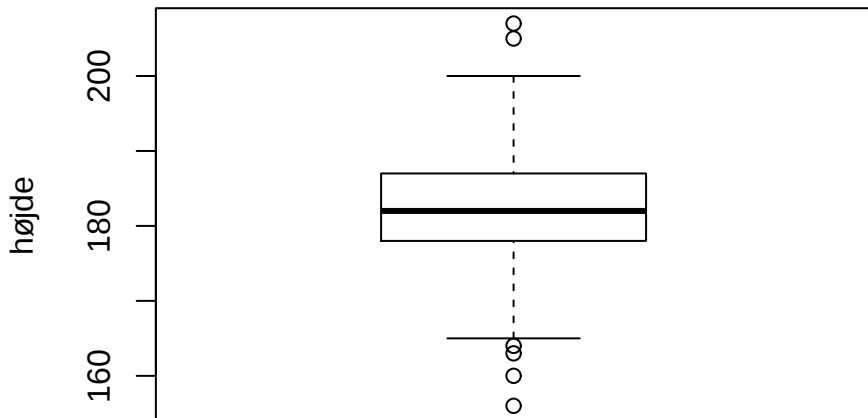


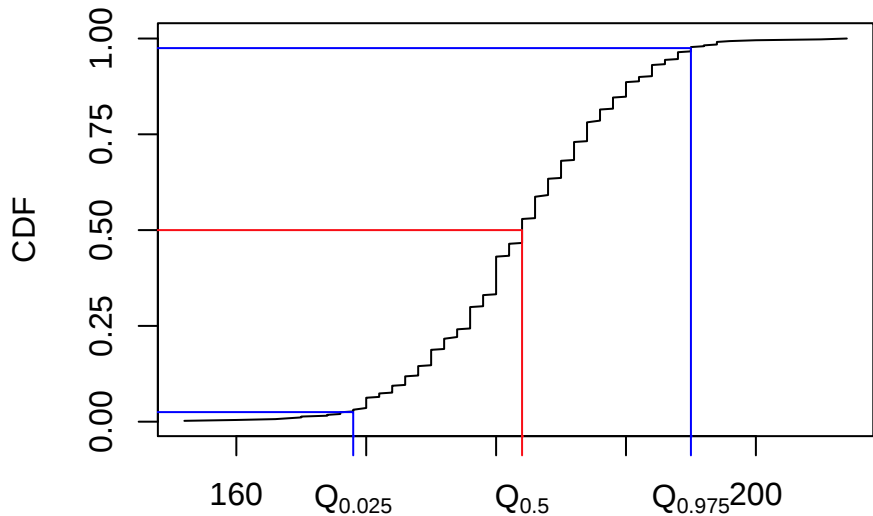
Højde



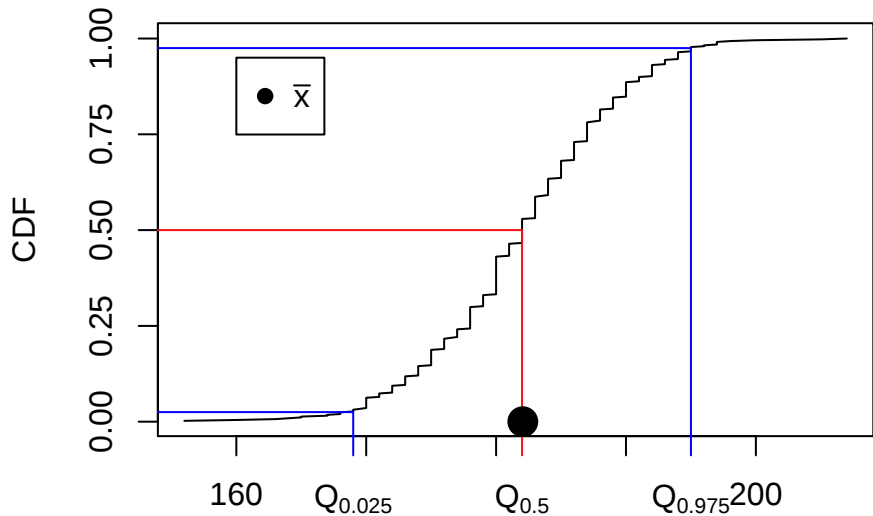
Højde



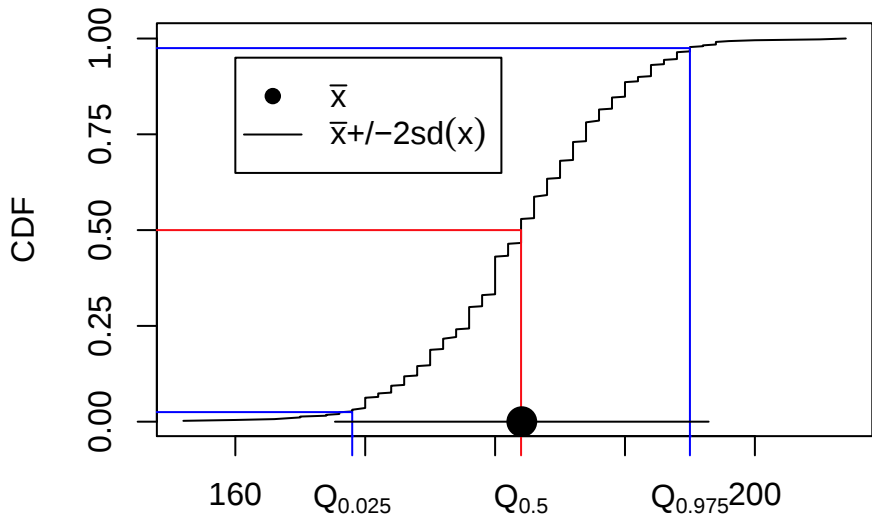




Højde



Højde



Oversigt

- 1 Praktisk Information
- 2 Introduktion til Statistik
- 3 Software: R
- 4 Beskrivende statistik: Nøgletal
- 5 Grafisk fremstilling

I morgen:

- Diskrete fordelinger - kapitel 4