

# 02323 Introduktion til statistik

## Uge 1: Introduktion og R

Nicolai Siim Larsen  
DTU Compute  
Danmarks Tekniske Universitet  
2800 Kgs. Lyngby

- 1 Praktiske informationer
- 2 Introduktion og motivation
- 3 Deskriptiv Statistik
  - Middelværdi og median (centralitetsmål)
  - Varians og standardafvigelse
  - Fraktiler
  - Kovarians og korrelation
- 4 Software: R & RStudio

# Dagsorden

- 1 Praktiske informationer
- 2 Introduktion og motivation
- 3 Deskriptiv Statistik
  - Middelværdi og median (centralitetsmål)
  - Varians og standardafvigelse
  - Fraktiler
  - Kovarians og korrelation
- 4 Software: R & RStudio

# Praktiske informationer

## • Undervisning

- Forelæsninger: Fredag 8-10
  - Bygning 306, Aud. 33.
- Øvelser: Fredag 10-12
  - Bygning 324, stueetagen (Foyer øvelsesområder: 003, 004, 005 og 008. Lokaler: 020, 030, 040, 050 og 070).

## • Eksamen

- Mandag den 27. maj 2024.
- 4 timers multiple choice-prøve.

## • Obligatoriske projekter

- 2 projekter, som skal bestås for at kunne gå til eksamen.
- For hvert projekt vælges et af fire emner.
- De som tidligere har bestået behøver ikke at lave projekterne igen.

# Praktiske informationer

## • Generel ugeseddel

- Før undervisningen: Læs de relevante kapitler/afsnit i bogen/e-noten.
- Forelæsninger: Gennemgang af ugens pensum.
- Øvelser: Opgaveregning og online quizzet.
- Efter undervisningen: Area9 og "eksamensquizzet".

## • Undervisningsmateriale

- Tilgængeligt under *Material* på kursushjemmesiden (på engelsk).
- Forelæsningsdialekter og R-kode opdateres før hver forelæsning.

# Praktiske informationer

- Kursushjemmeside: [02323.compute.dtu.dk](https://02323.compute.dtu.dk)
  - Bog
  - Pensum
  - Undervisningsplan (agenda)
  - Øvelser og løsninger (engelsk)
  - Dias (dansk og engelsk)
  - Tidligere års forelæsninger (dansk og engelsk)
  - Quizzer
- DTU Learn
  - Beskeder
  - Projekter - formulering og aflevering
- Ed Discussion
  - Spørgsmål og diskussioner

# Special for F24

This semester 02402 will have lectures in English (given by M.S. Khalid).

- 02402 lectures: Tuesday 13-15.

Omvendt kan studerende, der følger 02402, komme til danske forelæsninger i 02323.

- 02323 forelæsninger: Fredag 8-10.

NOTE: There are small differences between the courses in weeks 6 and 12.

# Dagsorden

- 1 Praktiske informationer
- 2 Introduktion og motivation
- 3 Deskriptiv Statistik
  - Middelværdi og median (centralitetsmål)
  - Varians og standardafvigelse
  - Fraktiler
  - Kovarians og korrelation
- 4 Software: R & RStudio



# Indledning

Statistik er grundlæggende en matematisk videnskab om indsamling, beskrivelse, analyse og fortolkning af data.

*Man vil uddrage viden og lære fra observerede data.*

Sandsynlighedsregning er en gren af matematik, der beskæftiger sig med beskrivelse og analyse af tilfældighed.

*Man vil udlede viden og lære fra en teoretisk model.*

Felterne er svære at adskille, og metoder fra begge felter bruges almindeligvis sammen i ingeniørarbejde.

Et fælles mål: Beskrive og forstå tilfældig variation og usikkerheder kvantitativt!

## Forskellige aspekter

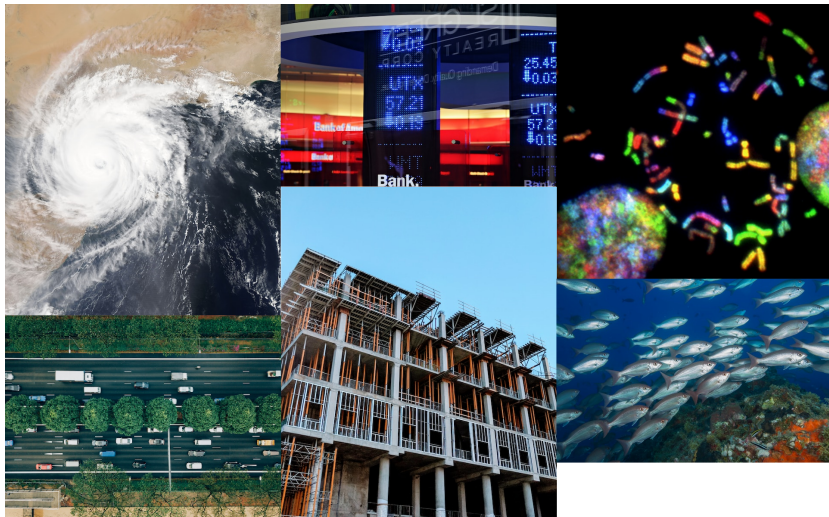
Der er mange spændende forskningsområder inden for både anvendt og teoretisk statistik.

Statistik og sandsynlighedsregning er forbundet til flere områder, f.eks.:

- Matematisk analyse
- Numerisk optimering
- Operationsanalyse
- Kontrolteori

De danner grundlaget for algoritmer i kunstig intelligens, maskinlæring og computerintensiv dataanalyse. *F.eks. er Stable Diffusion og ChatGPT baseret på avancerede statistiske modeller.*

# Anvendelse



## Intro case-historier:

### IBM big data, Novo Nordisk small data, Skive fjord

- Præsentation af Senior Scientist Hanne Refsgaard, Novo Nordisk A/S
- *IBM Social Media* podcast af Henrik H. Eliassen, IBM
- *Skive Fjord* podcasts, af Jan K. Møller, DTU

# I hverdagen

Statistik eller elementer fra faget forekommer mange steder i hverdagen, herunder:

- Nyheder
- Politik
- Reklamer
- Sport
- Arbejde

Statistik bruges ofte som beslutningsstøtte! Statistik kan bruges til at bestemme, hvad man skal undersøge nærmere.

# Almindelige fejlslutninger og bias

Statistik kan være kontraintuitivt, og vores hjerner skal trænes i statistisk tænkning for ikke at lave en række almindelige fejlslutninger. *Selvveluddannede, professionelle statistikere begår simple fejl.*

Nogle typiske biases (systematiske skævvridninger) i statistik er:

- Overlevelseshbias
- Udvælgelsesbias
- OVB (Omitted-variable bias)

Den sidste bias er tæt knyttet til koncepterne p-hacking og konfunderende variable.

# Kursets overordnede mål og afgrænsning

Kurset skal bl.a. gøre jer bedre til at:

- Behandle og analysere data hensigtsmæssigt
- Beskrive og forstå tilfældig variation og usikkerheder
- Tænke kritisk over statistiske udsagn
- Forstå mulighederne og begrænsningerne af statistik

Kurset skal også forberede jer til videregående kurser inden for bl.a. forsøgsplanlægning, tidsrækkeanalyse, kvalitetskontrol, sandsynlighedsregning, statistisk modellering, dataanalyse, maskinlæring og kunstig intelligens.

# Kursets indhold i store træk

En stor del af kurset omhandler:

- 1 Formulering af modeller
- 2 Udregning af konfidensintervaller
- 3 Udførelse af hypotesetest

i forskellige kontekster og setups.

Sandsynlighedsregningen bliver vores primære værktøj.



# Grundlæggende om statistik

Statistik kan generelt opdeles i to dele:

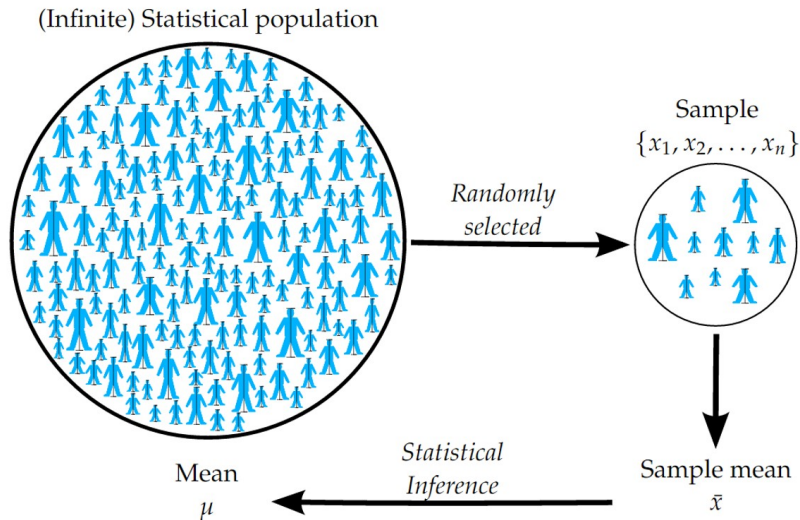
- Beskrivende statistik (deskriptiv statistik)
- Konkluderende statistik (statistisk inferens)

Statistik handler typisk om at analysere en *stikprøve*, taget ud af en *population*.

Ud fra stikprøven, udtaler vi os generelt om populationen.

Det er derfor vigtigt at stikprøven er *repræsentativ* for stikprøven. *I langt det meste af kurset vil vi bare antage, at stikprøverne er repræsentative.*

# Populationen og stikprøven



# Dagsorden

- 1 Praktiske informationer
- 2 Introduktion og motivation
- 3 Deskriptiv Statistik
  - Middelværdi og median (centralitetsmål)
  - Varians og standardafvigelse
  - Fraktiler
  - Kovarians og korrelation
- 4 Software: R & RStudio

## Det generelle setup

Der er en underliggende population, hvorfra der er udtaget en repræsentativ stikprøve med  $n$  observationer.

Stikprøven bliver almindeligvis repræsenteret med en vektor

$$x = (x_1, x_2, \dots, x_n).$$

Den sorterede stikprøve er så

$$(x_{(1)}, x_{(2)}, \dots, x_{(n)}),$$

hvor  $x_{(1)}$  angiver den mindste observation og  $x_{(n)}$  angiver den største observation.

# Nøgletal (Summary statistics)

*Nøgletal* bruges til at opsummere og beskrive data.

- *Positionsmål*
  - f.eks.: gennemsnit, median og fraktiler
- *Spredningsmål*
  - f.eks.: varians og standardafvigelse
- *Sammenhængsmål*
  - f.eks.: kovarians og korrelation

Husk at skelne mellem nøgletal for populationen og stikprøven!

## Gennemsnit, definition 1.4

**Gennemsnittet** er et nøgletal, der angiver tyngdepunktet for data.

**Middelværdien af en stikprøve (Stikprøvegennemsnittet):**

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

Vi siger, at  $\bar{x}$  er et *estimat* for populationens middelværdi.

## Median, Definition 1.5

**Medianen** er et også nøgletal, der angiver centreringen for data.

I nogle tilfælde, f.eks. hvis man har ekstreme observationer, er medianen at foretrække frem for gennemsnittet.

### **Medianen af en stikprøve (stikprøvemedianen):**

Den midterste observation (af de sorterede data) eller gennemsnittet af de to midterste observationer (af de sorterede data) afhængigt af, om stikprøven har et lige eller ulige antal observationer.

## Eksempel: Højde på studerende

- **Stikprøve:** Studerendes højde i cm,  $n = 5$ .

$$(x_1, x_2, x_3, x_4, x_5) = (185, 184, 194, 180, 182)$$

- **Gennemsnit:**

$$\bar{x} = \frac{1}{5}(185 + 184 + 194 + 180 + 182) = 185$$

- **Median:**

- Først sorteres data: (180, 182, 184, 185, 194).
- Da  $n$  er ulige, vælges det midterste tal: 184.
- Hvis vi tilføjer en 235 cm høj person til stikprøven:
  - *Gennemsnit:* 193
  - *Median:* 184.5



# Stikprøvevariens (sample variance) og -standardafvigelse (sample standard deviation), Definition 1.10

Stikprøvevariansen indikerer, hvor meget observationerne er spredt:

- Stikprøvevariens

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

- Stikprøvestandardafvigelse

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

# Eksempel med spredning: Højde på studerende

- **Stikprøve:** Studerendes højde i cm,  $n = 5$ .

$$(x_1, x_2, x_3, x_4, x_5) = (185, 184, 194, 180, 182)$$

- **Stikprøvevarians:**

$$s^2 = \frac{1}{4}((185 - 185)^2 + (184 - 185)^2 + \dots + (182 - 185)^2) = 29$$

- **Stikprøvestandardafvigelse:**

$$s = \sqrt{29} = 5.385$$

## Variationskoefficienten, Definition 1.12

Standardafvigelsen og variansen er de primære nøgletal til at beskrive variationen i data.

Nogle gange ønsker man at sammenligne variationen mellem forskellige datasæt; da kan det være en god ide at se på et forholdsmæssigt tal:

**Variationskoefficient:**

$$CV = \frac{s}{\bar{x}}$$

# Fraktiler (percentiles eller quantiles)

Medianen beregnes som det punkt, der deler data ind i to halvdele.

Mere generelt kan vi beregne *fraktiler*. Ofte beregner man:

- 0%, 25%, 50%, 75%, 100%-fraktilerne

Bemærk:

- Medianen er 50%-fraktilen.
- 25%, 50%, 75%-fraktilerne kaldes hhv. *første*, *anden* og *tredje* kvartil, betegnet med hhv.  $Q_1$ ,  $Q_2$  og  $Q_3$ .
- Dette giver anledning til spredningsmålet *den interkvartile variationsbredde* (*Inter Quartile Range* eller IQR):  $Q_3 - Q_1$

# Fraktiler, Definition 1.7

$p$ -fraktilen,  $q_p$ , kan defineres ud fra følgende procedure:

- 1 Sorter de  $n$  observationer fra mindst til størst:  $(x_{(1)}, x_{(2)}, \dots, x_{(n)})$ .
- 2 Beregn  $pn$ .
- 3 Hvis  $pn$  er et heltal: Tag gennemsnittet af den  $pn$ 'te og den  $(pn + 1)$ 'te ordnede observation:

$$q_p = (x_{(np)} + x_{(np+1)}) / 2$$

- 4 Hvis  $pn$  ikke er et heltal:

$$q_p = x_{(\lceil np \rceil)}$$

hvor  $\lceil np \rceil$  er *ceiling* ("loftet") af  $np$ , dvs. det mindste heltal større en  $np$ . Man afrunder altså  $np$  op til nærmeste heltal.

## Eksempel: Høyde på studerende

- **Sorteret stikprøve:** Studerendes høyde i cm,  $n = 5$ .

$$(x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(5)}) = (180, 182, 184, 185, 194)$$

- **Nedre kvartil, Q1:**

- Her er  $p = 0.25$  og  $n = 5$ , hvorfor  $np = 1.25$ .
- Det mindste heltal større end  $np$  er 2.
- $Q1 = q_{0.25} = x_{(\lceil 1.25 \rceil)} = x_{(2)} = 182$ .

- **Øvre kvartil, Q3:**

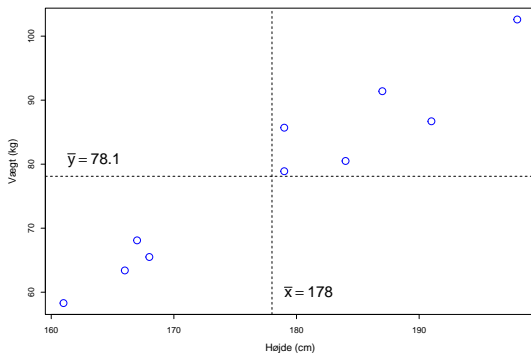
- Her er  $p = 0.75$  og  $n = 5$ , hvorfor  $np = 3.75$ .
- Det mindste heltal større end  $np$  er 4.
- $Q3 = q_{0.75} = x_{(\lceil 3.75 \rceil)} = x_{(4)} = 185$ .

- **IQR:**

- $Q3 - Q1 = 185 - 182 = 3$ .

# Kovarians og korrelation - Sammenhængsmål

Højde (cm) - ( $x_i$ )	168	161	167	179	184	166	198	187	191	179
Vægt (kg) - ( $y_i$ )	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9



# Stikprøvekovarians og -korrelation - Def 1.18 og 1.19

Stikprøvekovariansen er defineret ved

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Stikprøvekorrelationskoefficienten er defineret ved

$$r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) = \frac{s_{xy}}{s_x \cdot s_y}$$

hvor  $s_x$  og  $s_y$  standardfvgelserne for hhv.  $x$  og  $y$ .



# Stikprøvekovarians og -korrelation

Studerende (ID)	1	2	3	4	5	6	7	8	9	10
Højde (cm) - ( $x_i$ )	168	161	167	179	184	166	198	187	191	179
Wægt (kg) - ( $y_i$ )	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9
$(x_i - \bar{x})$	-10	-17	-11	1	6	-12	20	9	13	1
$(y_i - \bar{y})$	-12.6	-19.8	-10	7.6	2.4	-14.7	24.5	13.3	8.6	0.8
$(x_i - \bar{x})(y_i - \bar{y})$	126.1	336.8	110.1	7.6	14.3	176.5	489.8	119.6	111.7	0.8

$$\begin{aligned}
 s_{xy} &= \frac{1}{9}(126.1 + 336.8 + 110.1 + 7.6 + 14.3 + 176.5 + 489.8 \\
 &\quad + 119.6 + 111.7 + 0.8) \\
 &= \frac{1}{9} \cdot 1493.3 \\
 &= 165.9
 \end{aligned}$$

$$s_x = 12.21 \quad \text{og} \quad s_y = 14.07$$

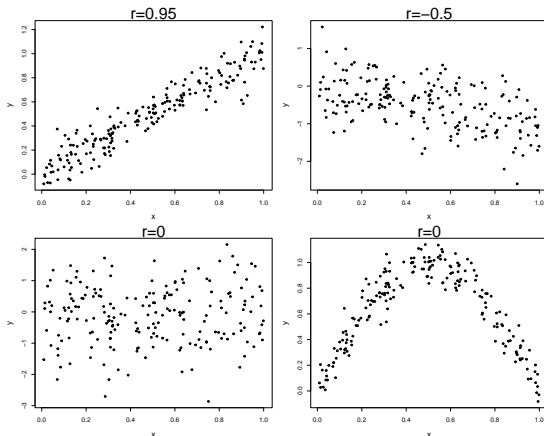
$$r = \frac{165.9}{12.21 \cdot 14.07} = 0.97$$

# Egenskaber for korrelationskoefficienten

De vigtigste egenskaber for korrelationskoefficienten er:

- $r$  er altid mellem  $-1$  og  $1$ :  $-1 \leq r \leq 1$
- $r$  er et mål for lineær sammenhæng mellem  $x$  og  $y$
- $r = \pm 1$  hvis og kun hvis punkterne ligger på en ret linie
- $r > 0$  hvis den generelle trend i scatterplottet er positiv
- $r < 0$  hvis den generelle trend i scatterplottet er negativ

# Korrelation



# Figurer/Tabeller

- Numeriske data

- Scatterplot (xy plot)
- Histogram
- Kumuleret fordeling
- Boxplot

- Tælledata

- Søjlediagram (bar chart)
- Cirkeldiagram (pie chart)

# Dagsorden

- 1 Praktiske informationer
- 2 Introduktion og motivation
- 3 Deskriptiv Statistik
  - Middelværdi og median (centralitetsmål)
  - Varians og standardafvigelse
  - Fraktiler
  - Kovarians og korrelation
- 4 **Software: R & RStudio**

# Software: R & RStudio

- R: Software/programmeringssprog for statistisk analyse og datavisualisering.
- R & RStudio: Gratis, open source, virker på alle platforme.
- Mange ekstrapakker i R til alskens dataanalyse.
- Introduceres i bogen.
- Integreret del af kurset.
- Learn by doing. Og: brug Google!

# Software: R

```
> # Addition  
> 2 + 3  
  
## [1] 5
```

```
> # Definer en variabel som en værdi  
> x <- 3  
> x  
  
## [1] 3
```

```
> # Definer en variabel som en vektor  
> x <- c(1, 4, 6, 2); x  
  
## [1] 1 4 6 2
```

```
> # Definer x som en vektor med heltallene fra 1 til 10  
> x <- 1:10
```

# Software: R

```
# Højde data  
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
```

```
# Stikprøvegennemsnit  
mean(x)
```

```
## [1] 178
```

```
# Stikprøvemedian  
median(x)
```

```
## [1] 179
```

```
# Stikprøvevarians  
var(x)
```

```
## [1] 149.1
```



# Software: R

```
# Stikprøvestandardafvigelse  
sd(x)
```

```
## [1] 12.21
```

```
# Stikprøvefraktil  
quantile(x, type = 2)
```

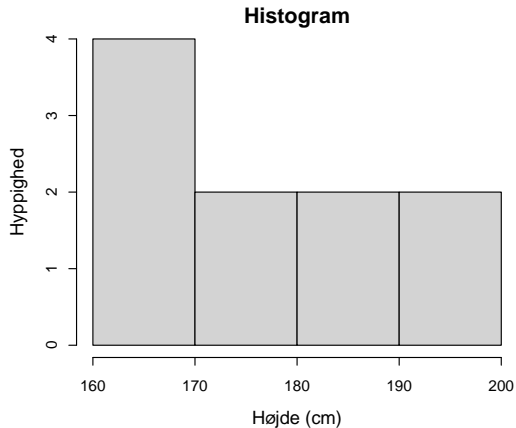
```
## 0% 25% 50% 75% 100%  
## 161 167 179 187 198
```

```
# Stikprøvefraktiler 0%, 10%, .., 90%, 100%  
quantile(x, probs = seq(0, 1, by = 0.10), type = 2)
```

```
## 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%  
## 161.0 163.5 166.5 168.0 173.5 179.0 184.0 187.0 189.0 194.5 198.0
```

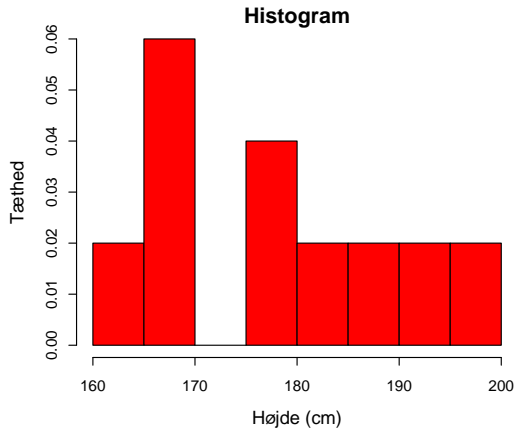
# R: Histogram

```
# Et histogram af højderne  
hist(x, main = "Histogram", ylab="Hyppighed", xlab="Højde (cm)")
```



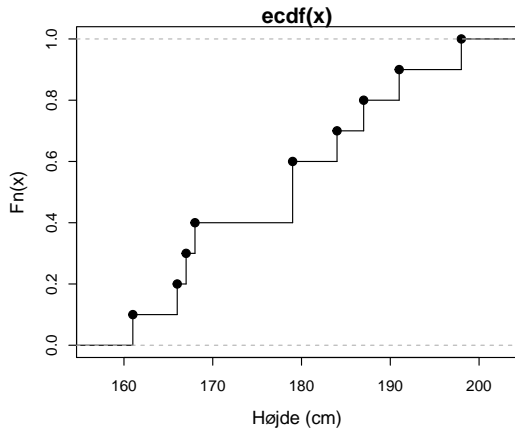
# R: Empirisk tæthed

```
# Et plot af tætheden af højderne  
hist(x, prob = TRUE, col = "red", nclass = 8, main="Histogram", ylab="Tæthed", xlab="Højde (cm)")
```



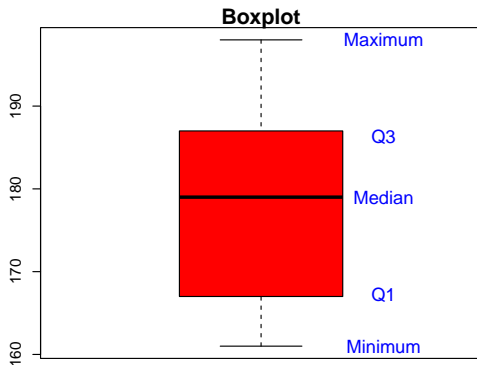
# R: Empirisk kumuleret fordeling

```
# Et plot af den empiriske fordelingsfunktion af højderne  
plot(ecdf(x), verticals = TRUE, xlab="Højde (cm)")
```



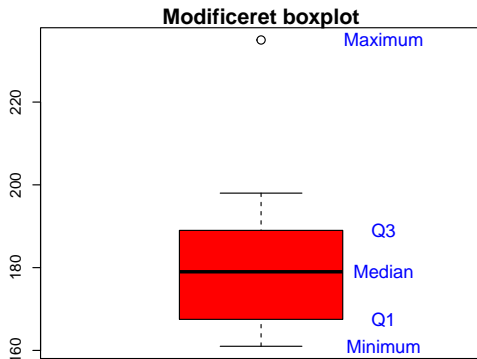
# R: boxplot

```
# Et boxplot af højderne ('range = 0' makes it "basic")  
boxplot(x, range = 0, col = "red", main = "Boxplot")  
text(1.3, quantile(x), c("Minimum", "Q1", "Median", "Q3", "Maximum"), col = "blue")
```



# Software: R

```
# Modificeret boxplot af højderne med en ekstrem observation (235 cm).
# Den modificerede version er "default".
boxplot(c(x, 235), col = "red", main = "Modificeret boxplot")
text(1.3, quantile(c(x, 235)), c("Minimum", "Q1", "Median", "Q3", "Maximum"), col = "blue")
```



## Næste uge:

- Stokastiske variable, sandsynligheder, diskrete fordelinger - kapitel 2 i bogen.

# Dagsorden

- 1 Praktiske informationer
- 2 Introduktion og motivation
- 3 Deskriptiv Statistik
  - Middelværdi og median (centralitetsmål)
  - Varians og standardafvigelse
  - Fraktiler
  - Kovarians og korrelation
- 4 Software: R & RStudio