02402: Introduktion til Statistik

Uge 1: Introduktion og R

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Agenda

1. Praktiske informationer
2. Introduction to Statistics - a primer
3. Statistik og ingeniører
4. Deskriptiv statistik
   - Middelværdi og median
   - Varians og standardafvigelse
   - Fraktiler
   - Kovarians og Korrelation
5. Software: R & RStudio

Overview

Praktiske informationer

- Undervisning
  - Forelæsninger: Tirsdag 8-10 i 303A, Aud 42-43
    - Vi kører en Kahoot til forelæsningen (vær med!)
    - Feedback med *Wyblo* appen.
  - Øvelser: Tirsdag 10-12.
    - Bygning 303A, the foyer.
    - Bygning 324, lokale 020, 030, 040, 050, 060 + the foyer.
- Eksamen
  - Lørdag den 17. december 2022
  - 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: 2 timer, ugens pensum
  - Øvelser: 2 timer, øvelser og online-quizzer
  - Efter undervisningen: Online "exam quiz" (test hvor godt du har styr på materialet!)
    - Area9: Ny, eksperimentiel undervisningsplatform.
- Undervisningsmateriale
  - Tilgængeligt under *Material* på kursushjemmesiden (all in English)
  - Forelæsningsslides og R kode opdateres før hver forelæsning.

Praktiske informationer

- Kursushjemmeside: 02402.compute.dtu.dk
  - Online bog
  - Pensum
  - Undervisningsplan / agenda
  - Øvelser & løsninger (engelsk)
  - Slides, dansk og engelsk
  - Tidligere års forelæsninger givet som podcasts af Per Brockhoff (English and Danish)
  - Quizzer
- DTU Learn.
  - DTU Learn: https://learn.inside.dtu.dk/d2l/home/125821
  - Beskeder og forum
  - Projekter - formulering og aflevering

Special for E2022

This semester 02323 will have lectures in English (given by M.S. Khalid)

- 02323 lectures: Fridays 8-10, Building 306, auditorium 33.

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

- 02402 forelæsninger: Tirsdage 8-10, Bygning 303A, auditorium 42-43.

NOTE: small differences in content btw 02323 and 02402 (more content in 02402!) – please check https://02323.compute.dtu.dk/ and https://02402.compute.dtu.dk/

Overview

Introduction to Statistics - a primer

New England Journal of Medicine:

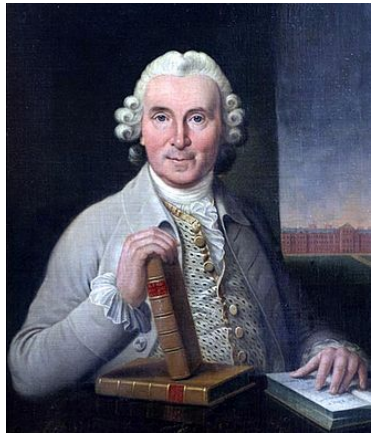
EDITORIAL: Looking Back on the Millennium in Medicine,
N Engl J Med, 342:42-49, January 6, 2000.
`http://www.nejm.org/doi/full/10.1056/NEJM200001063420108`

Millennium list

- Elucidation of human anatomy and physiology
- Discovery of cells and their substructures
- Elucidation of the chemistry of life
- **Application of statistics to medicine**
- Development of anesthesia
- Discovery of the relation of microbes to disease
- Elucidation of inheritance and genetics
- Knowledge of the immune system
- Development of body imaging
- Discovery of antimicrobial agents
- Development of molecular pharmacotherapy

James Lind

"*One of the earliest clinical trials took place in 1747, when James Lind treated 12 scorbutic ship passengers with cider, an elixir of vitriol, vinegar, sea water, oranges and lemons, or an electuary recommended by the ship's surgeon. The success of the citrus-containing treatment eventually led the British Admiralty to mandate the provision of lime juice to all sailors, thereby eliminating scurvy from the navy.*"
(Se også `http://en.wikipedia.org/wiki/James_Lind`).



John Snow

" The origin of modern epidemiology is often traced to 1854, when John Snow demonstrated the transmission of cholera from contaminated water by analyzing disease rates among citizens served by the Broad Street Pump in London's Golden Square. He arrested the further spread of the disease by removing the pump handle from the polluted well."
(Se også `http://en.wikipedia.org/wiki/John_Snow_(physician)`).

Google - *Big Data*

Citat fra New York Times-artiklen *For Today's Graduate, Just One Word: Statistics* (5 August 2009)
`http://www.nytimes.com/2009/08/06/technology/06stats.html`

I keep saying that the sexy job in the next 10 years will be statisticians, said Hal Varian, chief economist at Google. And I'm not kidding.

IBM - *Big Data*

" The key is to let computers do what they are good at, which is trawling these massive data sets for something that is mathematically odd, said Daniel Gruhl, an I.B.M. researcher whose recent work includes mining medical data to improve treatment. And that makes it easier for humans to do what they are good at - explain those anomalies."

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.

Overview

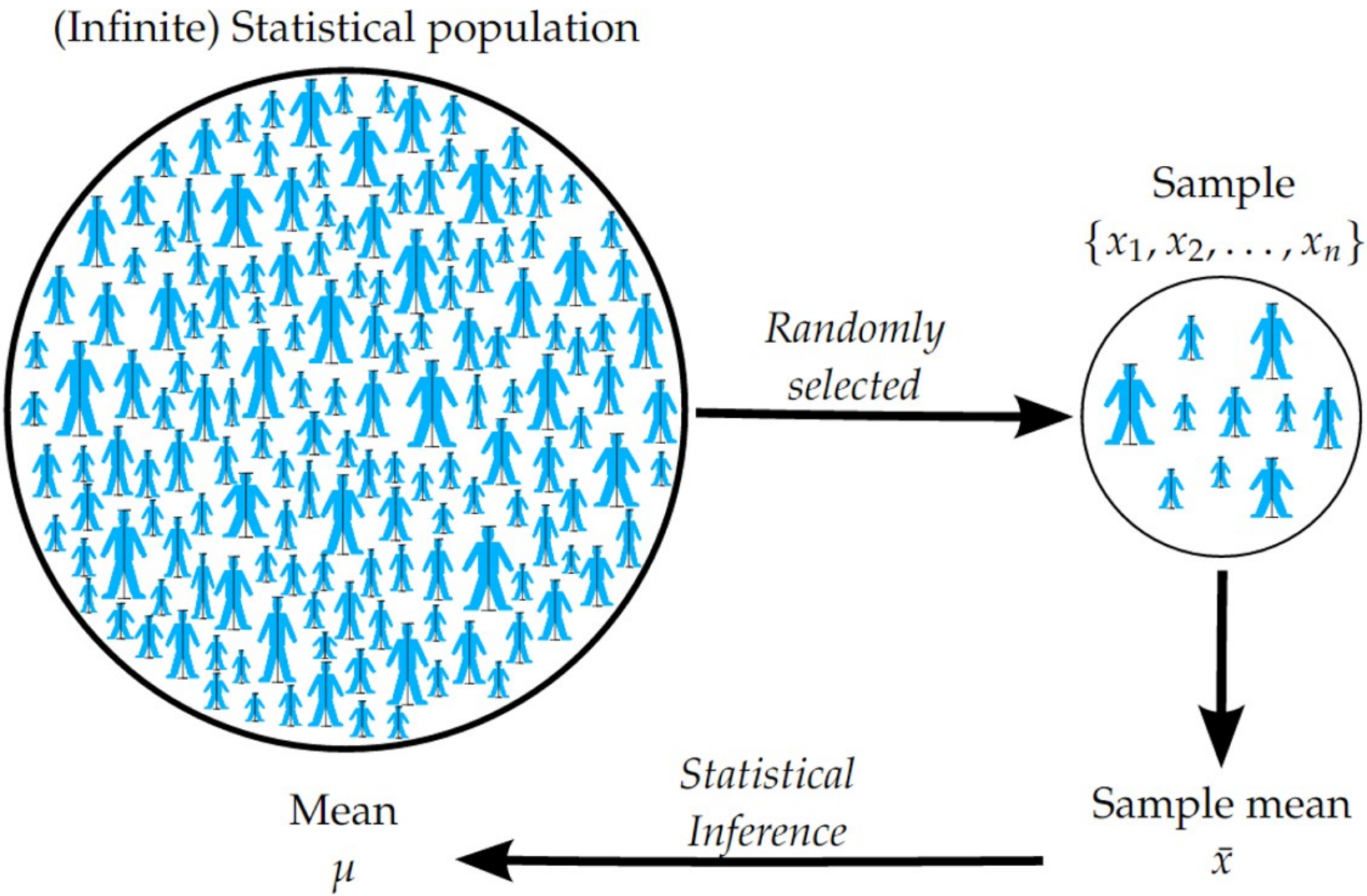
Statistik og ingeniører

- Analyse af data (både "small & big"data)
- Forstå tilfældig variation
- Forstå fordelene (og begrænsningerne) af statistik ifm. problemløsning
- Kvalitetskontrol / kvalitetsforbedring
- Forsøgsplanlægning
- Forudsigelse af fremtidige værdier
- ... og meget mere!

Statistik

- Deskriptiv statistik vs. 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.

Statistik


(Infinite) Statistical population
Randomly selected
Sample
$\{x_1, x_2, \ldots, x_n\}$
Sample mean
$\bar{x}$
Statistical Inference
Mean
μ

Overview

Nøgletal (Summary statistics)

Nøgletal bruges til at opsummere og beskrive data.

- Mål for *centralitet*
  - e.g.: gennemsnit ($\bar{x}$) og median
- Mål for *afvigelse*
  - e.g.: varians (s^2) og standardafvigelse (s)
- Mål for *sammenhæng*
  - e.g.: kovarians og korrelation

Bemærk forskellen mellem, f.eks., (stikprøvens) gennemsnit $\bar{x}$ og (populationens) gennemsnit μ.

Gennemsnit, definition 1.4

Gennemsnittet er et nøgletal, der angiver tyngdepunkt eller centrering for data.

Middelværdien af en stikprøve (gennemsnittet):

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

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

Median, Definition 1.5

Medianen er et også nøgletal, der angiver centrering

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

Median (stikprøvemedian):
Den midterste observation (af de sorterede data).

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

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

Stikprøvevarians indikerer hvor meget observationerne er spredt:

- Stikprøvevarians

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

- Stikprøvestandardafvigelse

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

Eksempel med spredning: Højder af unge mænd

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

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

- **Stikprøvevarians:**

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

- **Stikprøvestandardafvigelse:**

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

Variationskoefficienten, Definition 1.12

Standardavigelsen 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 idé at se på et forholdsmæssigt tal:

Coefficient of variation:

$$V = \frac{s}{\bar{x}} \tag{1}$$

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:

- $0, 25, 50, 75, 100$ % fraktiler

Bemærk:

- Medianen er lig 50%-fraktilen.
- 25, 50, 75 % fraktilerne kaldes ofte hhv. *første*, *anden* og *tredje* kvartil, betegnet med $Q1$, $Q2$, og $Q3$.
- *Inter Quartile Range* (IQR): $Q3 - Q1$

Fraktiler, Definition 1.7

Den p'te *fraktil* (quantile), kan defineres ud fra følgende procedure:

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

$$\text{Den } p\text{'te fraktil} = \left(x_{(np)} + x_{(np+1)}\right)/2$$

4. Hvis pn er et ikke-helt tal, tag den "næste" i den sorterede liste

$$\text{Den } p\text{'te fraktil} = x_{(\lceil np \rceil)}$$

hvor $\lceil np \rceil$ er *ceiling*("loftet") of np, p, dvs. det mindste heltal større en np.

Eksempel: Højde på studerende

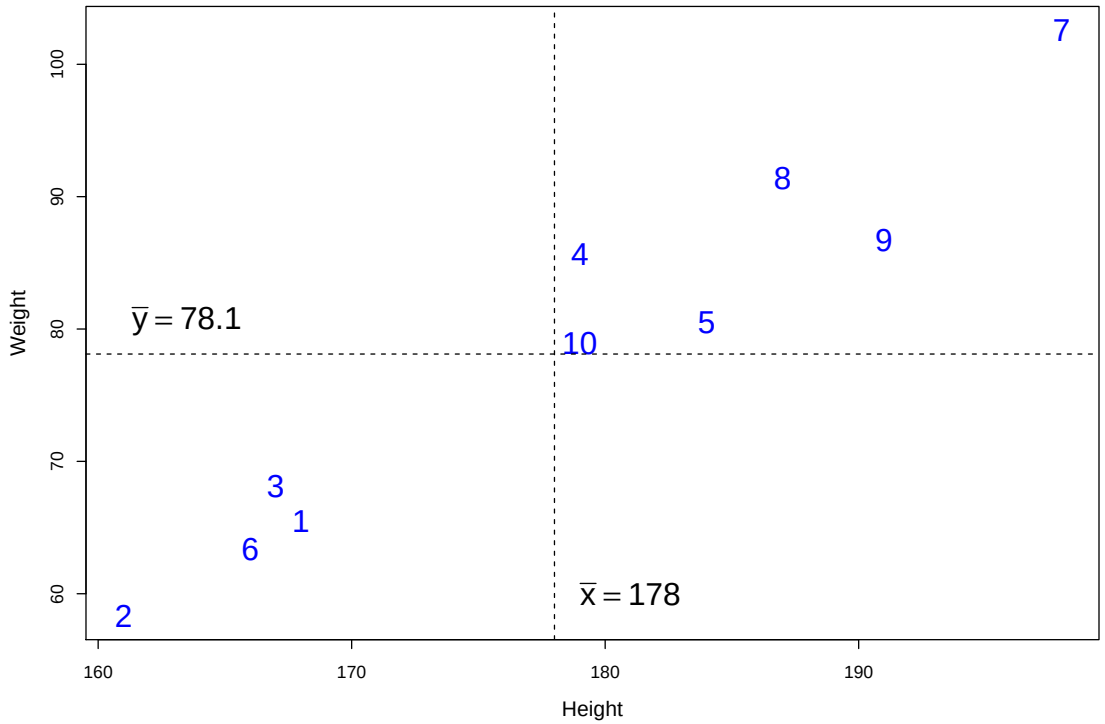
- **Stikprøve:** *Sorterede* studenterhøjder i cm.

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

- **Nedre kvartil, Q1:**
  - Eftervis at $np = 1.25$, da $p = 0.25$ og $n = 5$.
  - Det mindste heltal større end np er 2.
  - $Q1 = x_{(2)} = 182$.
- **Øvre kvartil, Q3:**
  - Eftervis at $np = 3.75$, da $p = 0.75$ og $n = 5$.
  - Det mindste heltal større end np er 4.
  - $Q3 = x_{(4)} = 185$.
- **IQR:**
  - $Q3 - Q1 = 3$

Kovarians og Korrelation - mål for sammenhæng

Height (x_i)	168	161	167	179	184	166	198	187	191	179
Weight (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9



Kovarians og Korrelation - Def 1.18 og 1.19

Kovariansen er defineret ved

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

Korrelationskoefficienten 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 standardfvigelserne for hhv. x og y.

Kovarians og korrelation

Student	1	2	3	4	5	6	7	8	9	10
Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (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

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

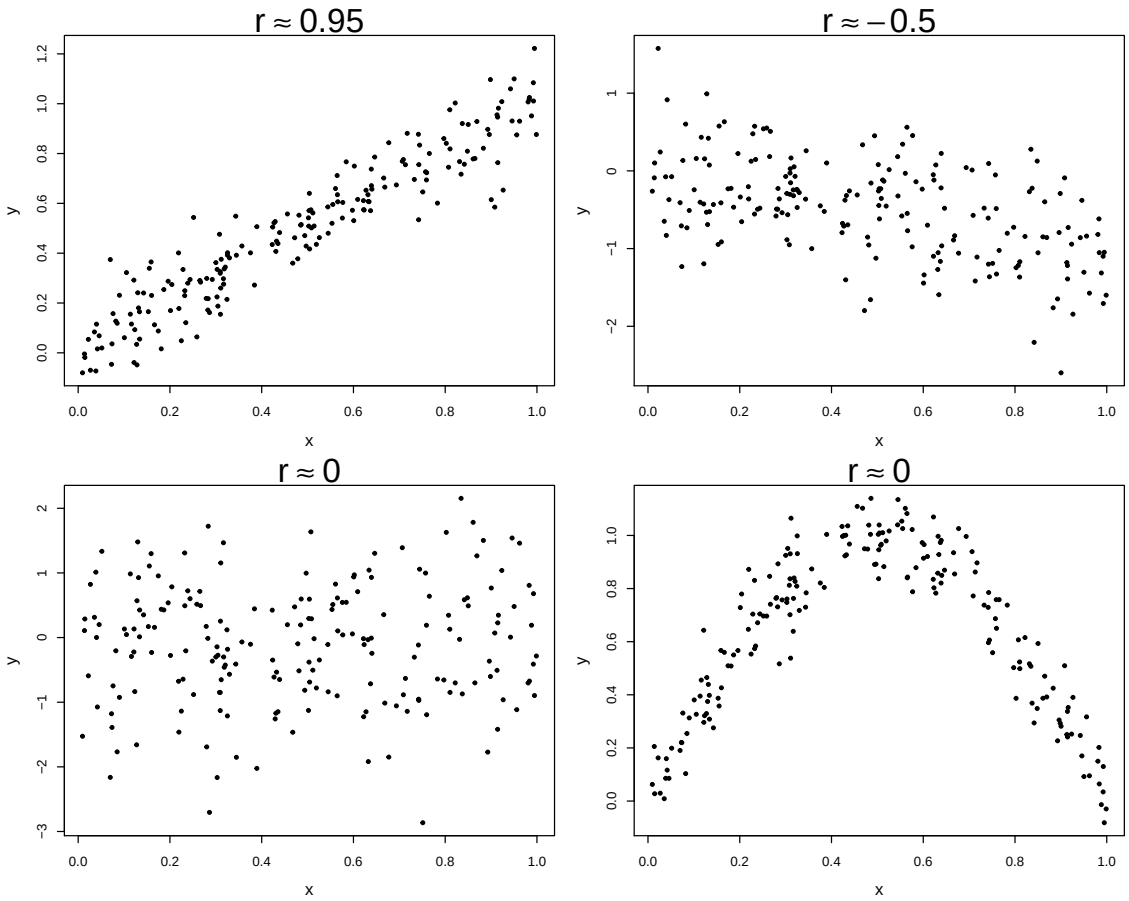
$$s_x = 12.21, \quad \text{og} \quad s_y = 14.07$$

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

Korrelation - egenskaber

- r er altid mellem -1 og 1: $-1 \leq r \leq 1$.
- r er et mål for den lineære 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
  - Barplot (bar chart)
  - Lagkagediagram (pie chart)

Overview

Software: R & RStudio

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

Software: R

```
> # Adding numbers in the console
> 2 + 3

## [1] 5
```

```
> # Assigning a number to a variable
> x <- 3
> x

## [1] 3
```

```
> # Assigning a vector to a variable
> x <- c(1, 4, 6, 2); x

## [1] 1 4 6 2
```

```
> # A vector of integers from 1 to 10
> ( x <- 1:10 )

##  [1]  1  2  3  4  5  6  7  8  9 10
```

Software: R

```r
# Height data from before
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
```

```r
# Sample mean
mean(x)
```

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

```r
# Sample median
median(x)
```

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

```r
# Sample variance
var(x)
```

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

Software: R

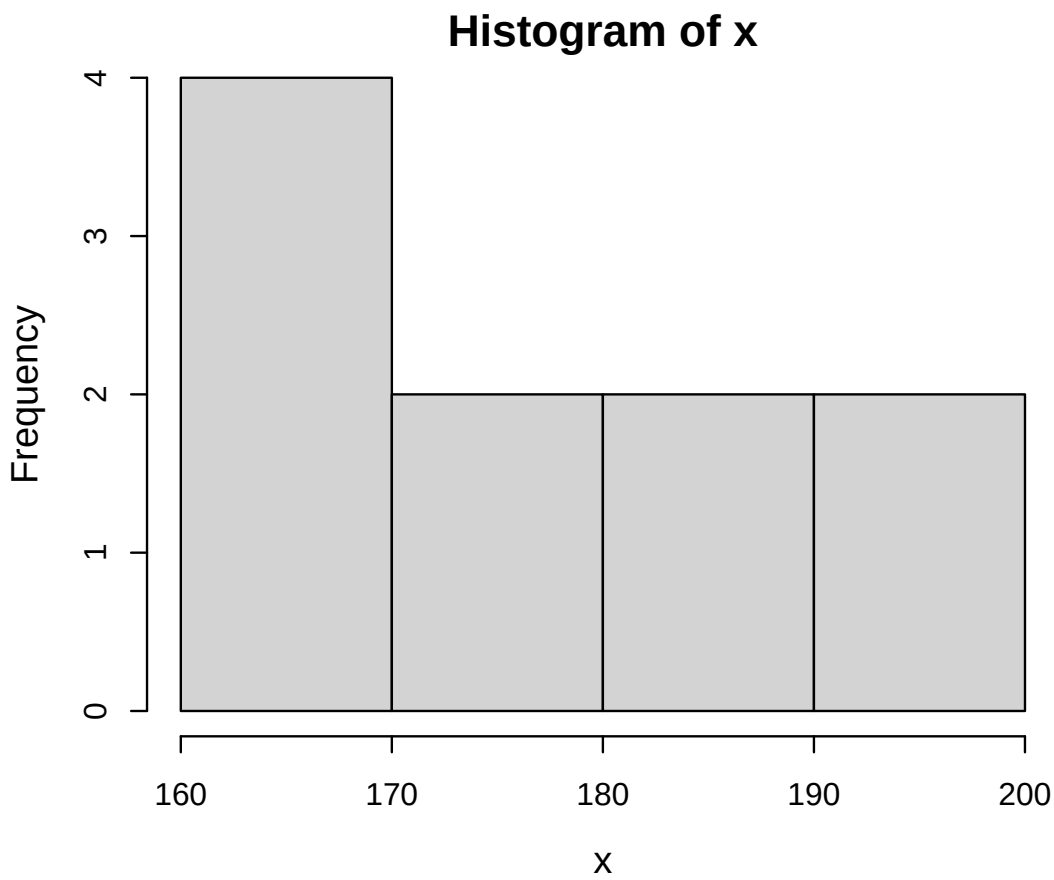
```r
  # Sample standard deviation
  sd(x)

## [1] 12.21
```

```r
# Sample quartiles
quantile(x, type = 2)

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

```r
# Sample quantiles 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
```

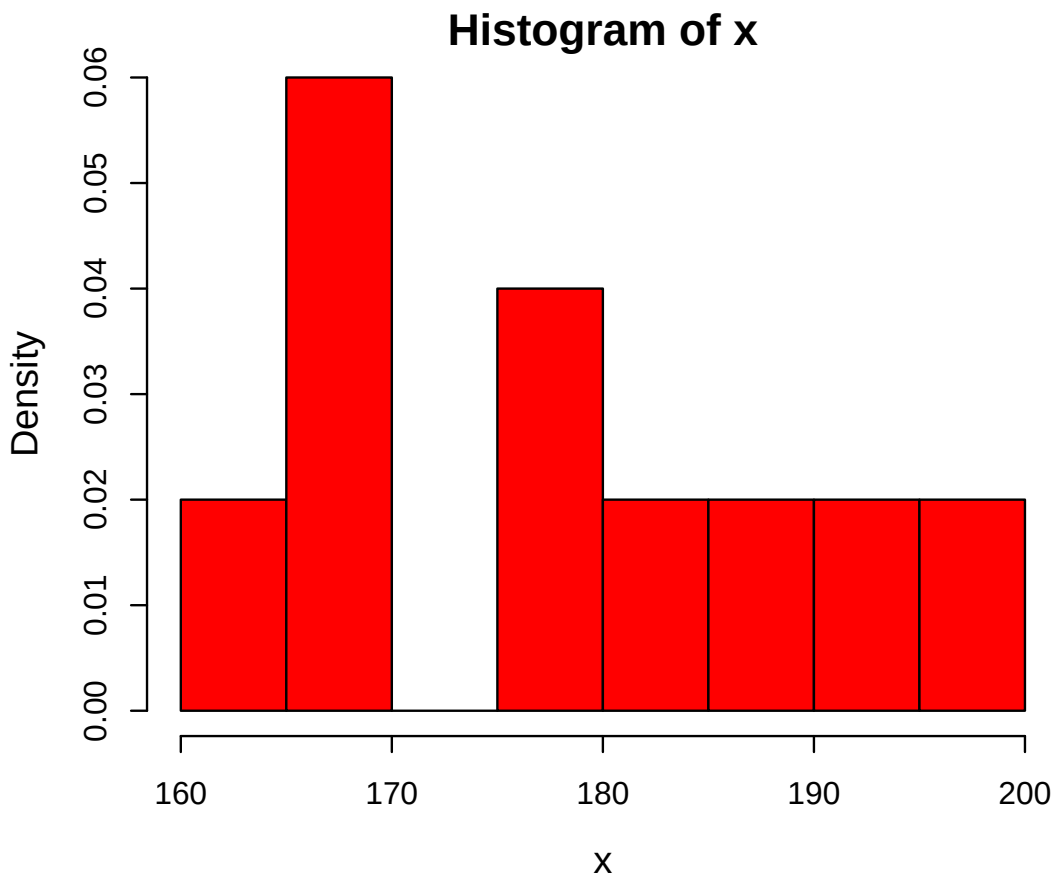
R: Histogram

```
# A histogram of the heights
hist(x)
```



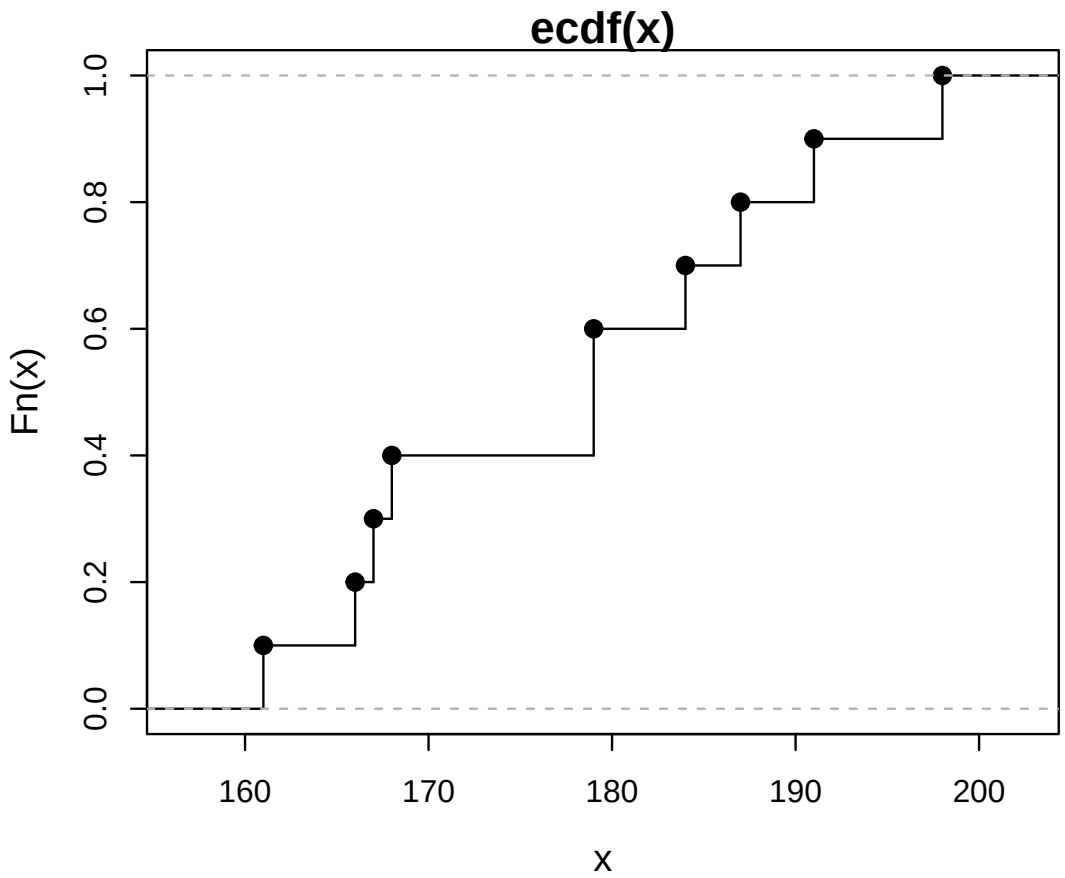
R: Empirisk tæthed

```r
# A density histogram of the heights
hist(x, prob = TRUE, col = "red", nclass = 8)
```



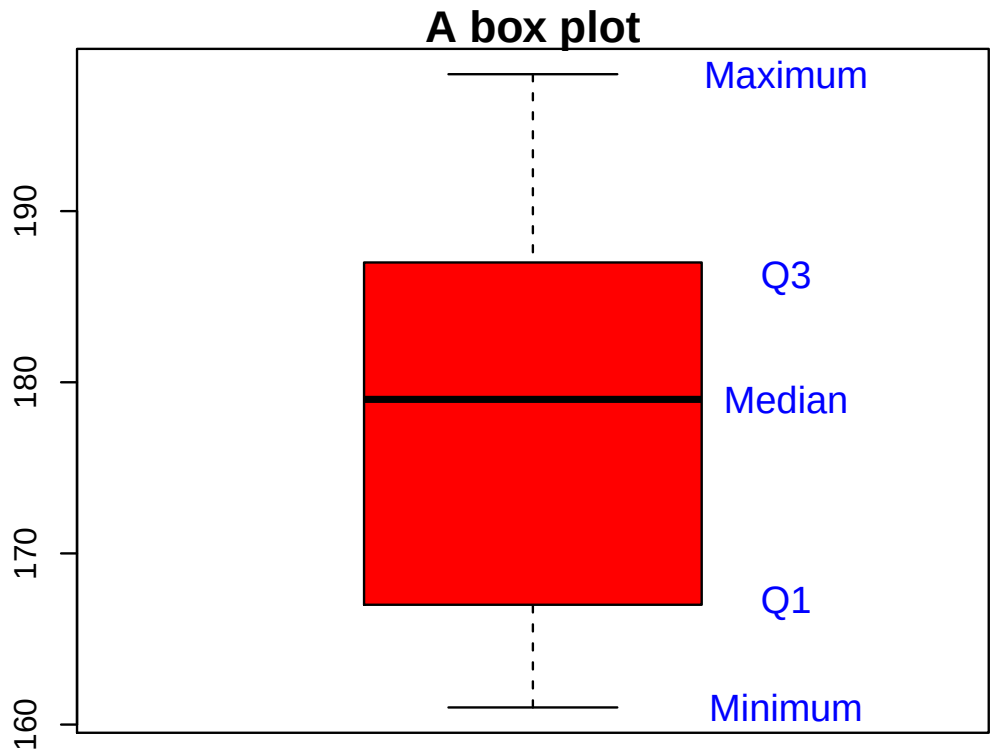
R: Empirisk kumuleret fordeling

```r
# Empirical cumulative distribution function of the heights
plot(ecdf(x), verticals = TRUE)
```



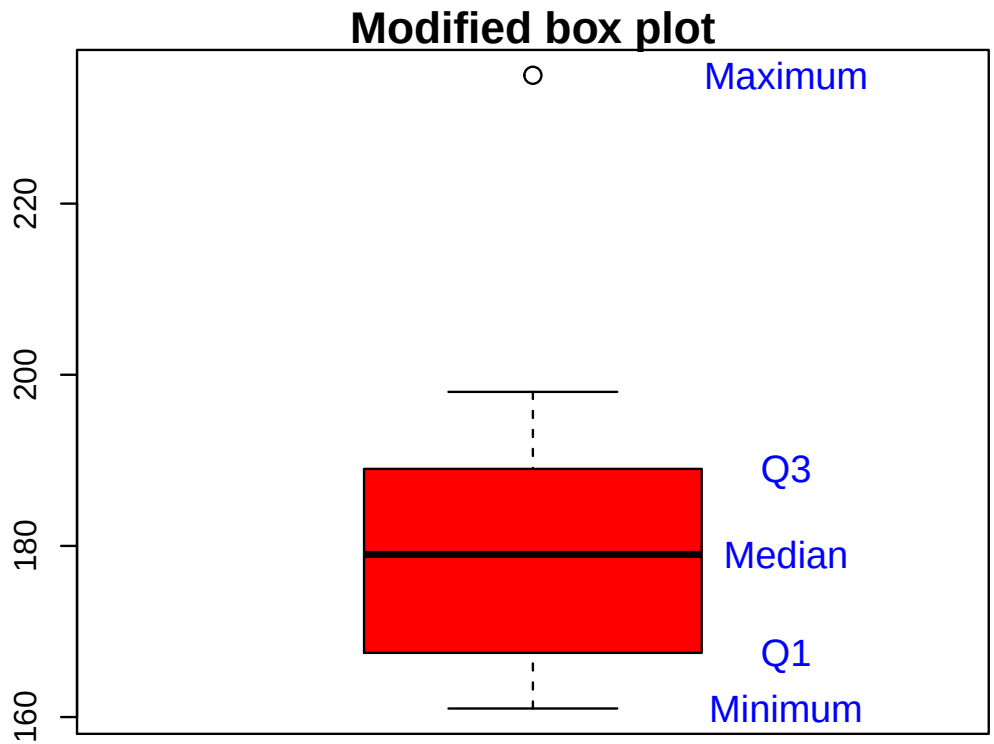
R: boxplot

```
  # Basic box plot of the heights ('range = 0' makes it "basic")
  boxplot(x, range = 0, col = "red", main = "A box plot")
text(1.3, quantile(x), c("Minimum", "Q1", "Median", "Q3", "Maximum"), col = "blue")
```



Software: R

```
  # Modified box plot of heights with an additional extreme observation (235 cm).
  # The modified version is the default.
  boxplot(c(x, 235), col = "red", main = "Modified box plot")
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.

Agenda

1. Praktiske informationer
2. Introduction to Statistics - a primer
3. Statistik og ingeniører
4. Deskriptiv statistik
   - Middelværdi og median
   - Varians og standardafvigelse
   - Fraktiler
   - Kovarians og Korrelation
5. Software: R & RStudio