

# Course 02402/02323 Introducerende Statistik

## Forelæsning 1: Intro, R og beskrivende statistik

Klaus K. Andersen og Per Bruun Brockhoff

DTU Compute, Statistik og Dataanalyse  
Danmarks Tekniske Universitet  
2800 Lyngby – Danmark  
e-mail: klaus@cancer.dk

# Oversigt

- 1 Praktisk Information
- 2 Introduction to Statistics - a primer
- 3 Intro Case historier: IBM Big data, Novo Nordisk small data, Skive fjord
- 4 Introduktion til Statistik
- 5 Beskrivende statistik: Nøgletal
  - Gennemsnit
  - Median
  - Spredning
  - Fraktiler
  - Kovarians og Korrelation
  - Plots/figurer
- 6 Software: R

# Oversigt

- 1 Praktisk Information
- 2 Introduction to Statistics - a primer
- 3 Intro Case historier: IBM Big data, Novo Nordisk small data, Skive fjord
- 4 Introduktion til Statistik
- 5 Beskrivende statistik: Nøgletal
  - Gennemsnit
  - Median
  - Spredning
  - Fraktiler
  - Kovarians og Korrelation
  - Plots/figurer
- 6 Software: R

# Praktisk Information

- Undervisning: Tirsdage 8-12/fredage 8-12
- Generel daglig agenda:
  - FØR undervisningsmodulet: læs det annoncerede i eNoten!
  - 2x45 minutters forelæsning (ugens pensum)
  - 2 timers øvelser (Enote Exercises (eNe) og små quiz-spørgsmål)
  - EFTER undervisningsmodulet: Test dig selv med online eksamens-quiz
- Skriftlig eksamen: Tirsdag 13/12 2016.
- OBLIGATORISKE projekter: 2 stk - skal godkendes for at kunne gå til eksamen.

# Praktisk Information

- Hjemmeside: [introstat.compute.dtu.dk](http://introstat.compute.dtu.dk)
  - Læsemateriale: eNote
  - Forelæsningsplan
  - Øvelser & besvarelser
  - Slides
  - Podcasts af nye og gl. forelæsninger (På dansk OG engelsk)
  - Quizzer
- Campusnet: [www.campusnet.dtu.dk](http://www.campusnet.dtu.dk)
  - Meddelelser, visse (få) dokumenter
  - Links til interessante historier
  - Projekter - beskrivelse OG aflevering

# Oversigt

- 1 Praktisk Information
- 2 Introduction to Statistics - a primer
- 3 Intro Case historier: IBM Big data, Novo Nordisk small data, Skive fjord
- 4 Introduktion til Statistik
- 5 Beskrivende statistik: Nøgletal
  - Gennemsnit
  - Median
  - Spredning
  - Fraktiler
  - Kovarians og Korrelation
  - Plots/figurer
- 6 Software: R

# 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](http://www.nejm.org/doi/full/10.1056/NEJM200001063420108)

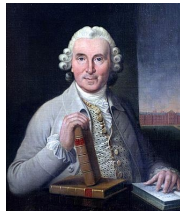
# Millennium list

- 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." (See also [http://en.wikipedia.org/wiki/James\\_Lind](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." (See also [http://en.wikipedia.org/wiki/John\\_Snow\\_\(physician\)](http://en.wikipedia.org/wiki/John_Snow_(physician))).*



## Google - *Big Data*

A quote from New York Times, 5. August 2009, from the article titled "For Today's Graduate, Just One Word: Statistics" is:

*"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."*

(Og Politiken, 12/2 2014 - se links i CampusNet)



# 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."*



# Oversigt

- 1 Praktisk Information
- 2 Introduction to Statistics - a primer
- 3 Intro Case historier: IBM Big data, Novo Nordisk small data, Skive fjord
- 4 Introduktion til Statistik
- 5 Beskrivende statistik: Nøgletal
  - Gennemsnit
  - Median
  - Spredning
  - Fraktiler
  - Kovarians og Korrelation
  - Plots/figurer
- 6 Software: R

## 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.

# Oversigt

- 1 Praktisk Information
- 2 Introduction to Statistics - a primer
- 3 Intro Case historier: IBM Big data, Novo Nordisk small data, Skive fjord
- 4 Introduktion til Statistik**
- 5 Beskrivende statistik: Nøgletal
  - Gennemsnit
  - Median
  - Spredning
  - Fraktiler
  - Kovarians og Korrelation
  - Plots/figurer
- 6 Software: R

# 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!

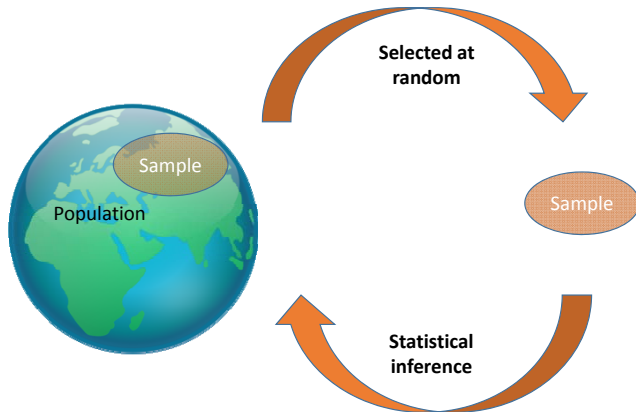
# Statistik

- 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

# Statistik



# Oversigt

- 1 Praktisk Information
- 2 Introduction to Statistics - a primer
- 3 Intro Case historier: IBM Big data, Novo Nordisk small data, Skive fjord
- 4 Introduktion til Statistik
- 5 **Beskrivende statistik: Nøgletal**
  - Gennemsnit
  - Median
  - Spredning
  - Fraktiler
  - Kovarians og Korrelation
  - Plots/figurer
- 6 Software: R

## Kapitel 2: Nøgletal

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

# 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 tyngdepunkt eller 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)



# Eksempel: Højder af unge mænd:

- Stikprøve (Sample):

```
x <- c(185, 184, 194, 180, 182)
```

n=5

## Eksempel: Højder af unge mænd:

- Stikprøve (Sample):

```
x <- c(185, 184, 194, 180, 182)
```

n=5

- Gennemsnit:

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

## Eksempel: Højder af unge mænd:

- Stikprøve (Sample):

```
x <- c(185, 184, 194, 180, 182)
```

n=5

- Gennemsnit:

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

- Median, først ordn data: 180 182 184 185 194

## Eksempel: Højder af unge mænd:

- Stikprøve (Sample):

```
x <- c(185, 184, 194, 180, 182)
```

n=5

- Gennemsnit:

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

- **Median**, først ordn data: 180 182 184 185 194  
Og så vælg det midterste (idet n er ulige)(3'te) tal: 184

## Eksempel: Højder af unge mænd:

- Stikprøve (Sample):

```
x <- c(185, 184, 194, 180, 182)
```

n=5

- Gennemsnit:

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

- **Median**, først ordn data: 180 182 184 185 194  
Og så vælg det midterste (idet n er ulige)(3'te) tal: 184
- Hvad nu hvis en person på 235cm tilføjes til data:  
Median = 184.5  
Mean = 193.33

# Varians og standardafvigelse

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

- Varians

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

- Standardafvigelse (spredning) (=standard deviation)

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

## Eksempel: Højder af unge mænd:

- **Data**  $n=5$ :  
185 184 194 180 182

## Eksempel: Højder af unge mænd:

- **Data**  $n=5$ :  
185 184 194 180 182
- **Varians**,  $s^2 =$

$$\begin{aligned} \frac{1}{4} & ((185 - 185)^2 + (184 - 185)^2 + (194 - 185)^2 + (180 - 185)^2 \\ & + (182 - 185)^2) \\ & = 29 \end{aligned}$$



## Eksempel: Højder af unge mænd:

- **Data**  $n=5$ :  
185 184 194 180 182
- **Varsians**,  $s^2 =$

$$\begin{aligned} \frac{1}{4} & ((185 - 185)^2 + (184 - 185)^2 + (194 - 185)^2 + (180 - 185)^2 \\ & + (182 - 185)^2) \\ & = 29 \end{aligned}$$

- **Standardafvigelse**,  $s = \sqrt{s^2} =$

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

# 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(=percentiles=quantiles)

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.

Oftest beregner man fraktilerne

- 0, 25, 50, 75, 100 % fraktiler
- Bemærk: 50% fraktilen svarer til medianen

## Fraktiler(=percentiles=quantiles), Definition 1.6

Den  $p$ 'te fraktil, også kaldet *quantile*, kan defineres ud fra følgende procedure:

- 1 Ordne de  $n$  observationer fra mindst til størst:  $x_{(1)}, \dots, x_{(n)}$ .

# Fraktiler(=percentiles=quantiles), Definition 1.6

Den  $p$ 'te fraktil, også kaldet *quantile*, kan defineres ud fra følgende procedure:

- 1 Ordne de  $n$  observationer fra mindst til størst:  $x_{(1)}, \dots, x_{(n)}$ .
- 2 Beregn  $pn$ .

## Fraktiler(=percentiles=quantiles), Definition 1.6

Den  $p$ 'te fraktil, også kaldet *quantile*, kan defineres ud fra følgende procedure:

- 1 Ordne de  $n$  observationer fra mindst til størst:  $x_{(1)}, \dots, x_{(n)}$ .
- 2 Beregn  $pn$ .
- 3 Hvis  $pn$  er et helt tal: Midl den  $pn$ 'te og  $(pn + 1)$ 'te ordnede observationer

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

## Fraktiler(=percentiles=quantiles), Definition 1.6

Den  $p$ 'te fraktil, også kaldet *quantile*, kan defineres ud fra følgende procedure:

- 1 Ordne de  $n$  observationer fra mindst til størst:  $x_{(1)}, \dots, x_{(n)}$ .
- 2 Beregn  $pn$ .
- 3 Hvis  $pn$  er et helt tal: Midl den  $pn$ 'te og  $(pn + 1)$ 'te ordnede observationer

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

- 4 Hvis  $pn$  er et ikke-helt tal: tag den "næste" i den ordnede liste:

$$\text{Den } p\text{'te fraktil} = x_{(\lceil np \rceil)} \quad (2)$$

hvor  $\lceil np \rceil$  er *ceiling*("loftet") af  $np$ , dvs. det mindste hele tal større end  $np$ .

## Eksempel: Højder af unge mænd:

- **Data,  $n=5$ :**  
185 184 194 180 182



## Eksempel: Højder af unge mænd:

- **Data**,  $n=5$ :  
185 184 194 180 182
- **Nedre kvartil**,  $Q_1$ , først ordn data: 180 182 184 185 194

## Eksempel: Højder af unge mænd:

- **Data**,  $n=5$ :  
185 184 194 180 182
- **Nedre kvartil**,  $Q_1$ , først ordn data: 180 182 184 185 194  
Og så vælg det rigtige baseret på  $np = 1.25$ :  
 $Q_1 = 182$

## Eksempel: Højder af unge mænd:

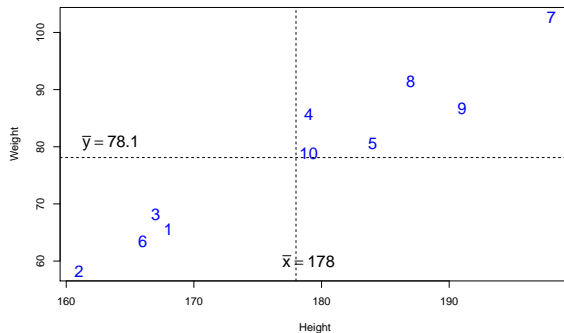
- **Data**,  $n=5$ :  
185 184 194 180 182
- **Nedre kvartil**,  $Q_1$ , først ordn data: 180 182 184 185 194  
Og så vælg det rigtige baseret på  $np = 1.25$ :  
 $Q_1 = 182$
- **Øvre kvartil**,  $Q_3$ , først ordn data: 180 182 184 185 194

## Eksempel: Højder af unge mænd:

- **Data**,  $n=5$ :  
185 184 194 180 182
- **Nedre kvartil**,  $Q_1$ , først ordn data: 180 182 184 185 194  
Og så vælg det rigtige baseret på  $np = 1.25$ :  
 $Q_1 = 182$
- **Øvre kvartil**,  $Q_3$ , først ordn data: 180 182 184 185 194  
Og så vælg det rigtige baseret på  $np = 3.75$ :  
 $Q_3 = 185$

# Kovarians og Korrelation - mål for sammenhæng

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



# Kovarians og Korrelation - Def. 1.17 og 1.18

The sample covariance is given by

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (3)$$

The sample correlation coefficient is given by

$$r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) = \frac{s_{xy}}{s_x \cdot s_y} \quad (4)$$

where  $s_x$  and  $s_y$  is the sample standard deviation for  $x$  and  $y$  respectively.

# Kovarians og Korrelation - mål for sammenhæng

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

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

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

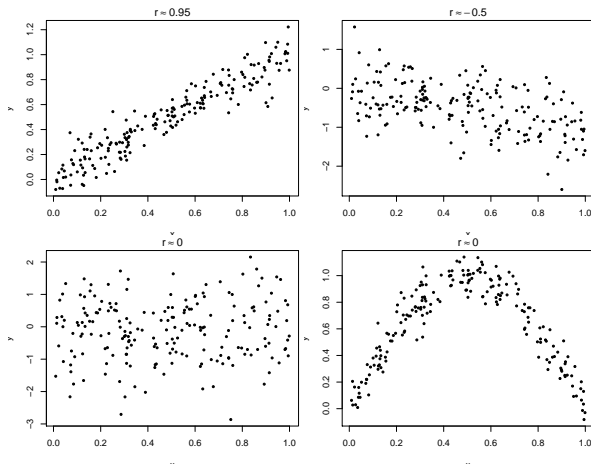
$$r = \frac{165.9}{12.21 \cdot 14.07} = 0.97$$

# Korrelation - egenskaber

- $r$  is always between  $-1$  and  $1$ :  $-1 \leq r \leq 1$
- $r$  measures the degree of linear relation between  $x$  and  $y$
- $r = \pm 1$  if and only if all points in the scatterplot are exactly on a line
- $r > 0$  if and only if the general trend in the scatterplot is positive
- $r < 0$  if and only if the general trend in the scatterplot is negative



# Korrelation



# Plots/figurer

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

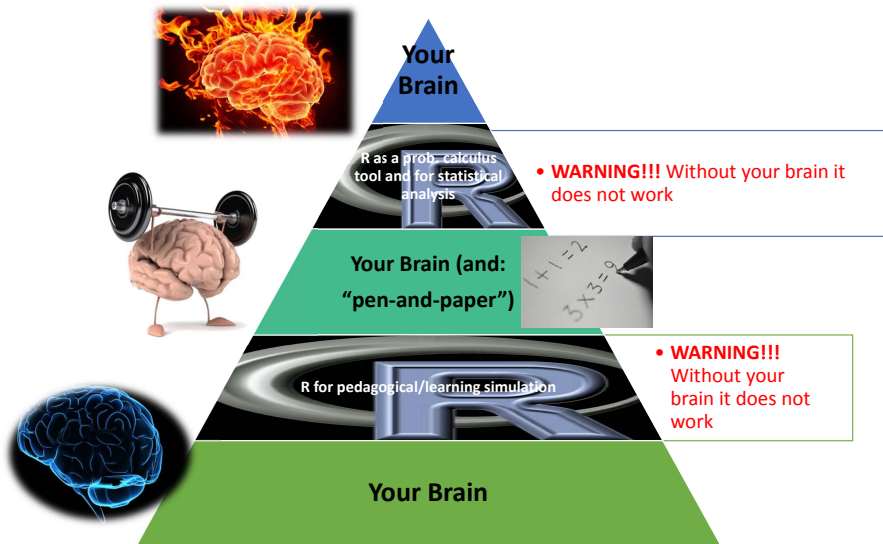
# Oversigt

- 1 Praktisk Information
- 2 Introduction to Statistics - a primer
- 3 Intro Case historier: IBM Big data, Novo Nordisk small data, Skive fjord
- 4 Introduktion til Statistik
- 5 Beskrivende statistik: Nøgletal
  - Gennemsnit
  - Median
  - Spredning
  - Fraktiler
  - Kovarians og Korrelation
  - Plots/figurer
- 6 Software: R

# Software: R

- Installer R og Rstudio på egen computer.
- Introduceres i eNoten
- Er integreret i alt vi gør
- Globalt og hurtigt voksende open source beregningsmiljø
- ADVAAAAARRRSEEEEL: R kan IKKE erstatte vores hjerner!!!! (Læs sektion 1.5.4!)

# Use your Brain!!!



# Software: R

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

```
[1] 5
```

```
> y <- 3
```

```
> x <- c(1, 4, 6, 2)  
> x
```

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

```
> x <- 1:10  
> x
```

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

# Software: R

```
## Sample Mean and Median (data from eNote)  
x <- c(168,161,167,179,184,166,198,187,191,179)  
mean(x)
```

```
[1] 178
```

```
median(x)
```

```
[1] 179
```

```
## Sample variance and standard deviation  
var(x)
```

```
[1] 149.11
```

```
sd(x)
```

```
[1] 12.211
```

# Software: R

```
## Sample quantiles
```

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

```
##      0%   25%   50%   75%  100%
```

```
##    161   167   179   187   198
```

```
## Sample quantiles 0%, 10%,...,90%, 100%:
```

```
quantile(x,probs=seq(0, 1, by=0.10),type=2)
```

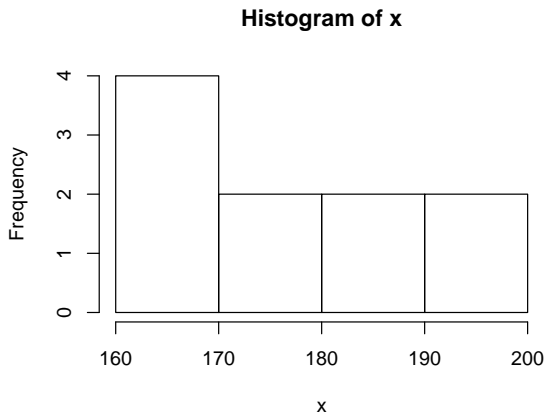
```
##      0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
```

```
## 161.0 163.5 166.5 168.0 173.5 179.0 184.0 187.0 189.0 194.5 198.0
```



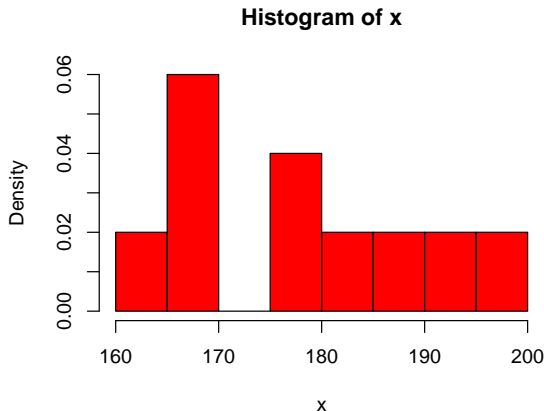
# Software: R

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



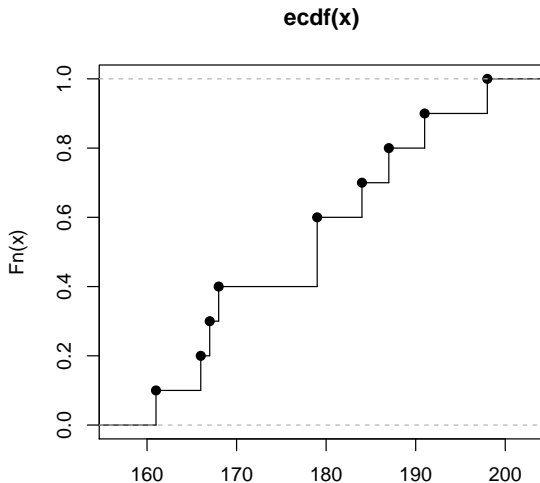
# Software: R

```
## A density histogram of the heights:  
hist(x,freq=FALSE,col="red",nclass=8)
```



# Software: R

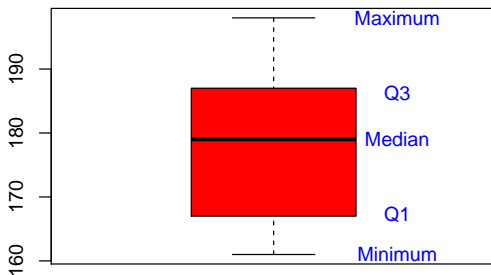
```
plot(ecdf(x), verticals=TRUE)
```



# Software: R

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

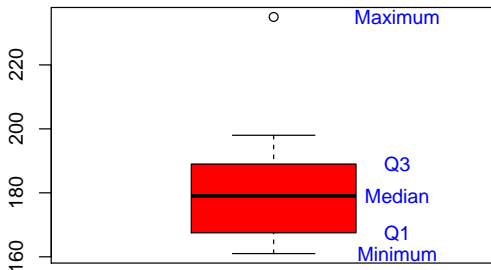
Basic boxplot



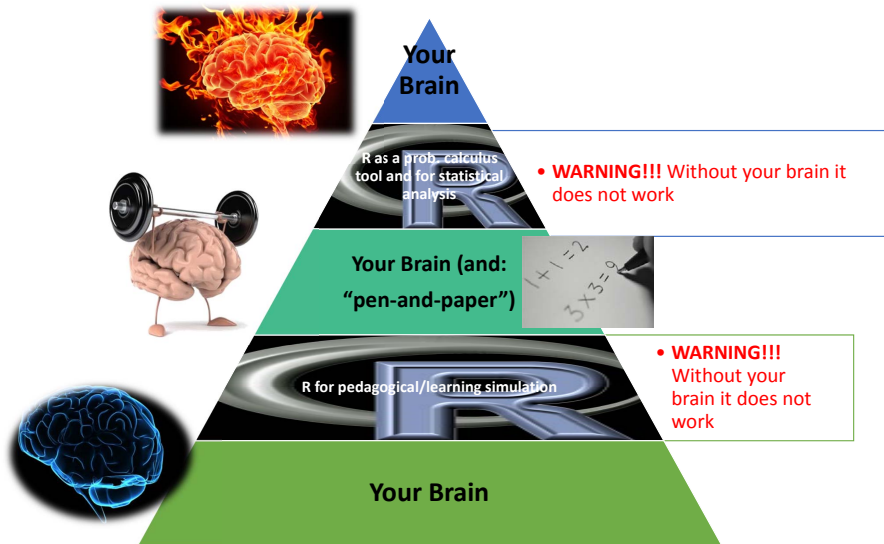
# Software: R

```
## A modified boxplot of the heights with an
## extreme observation, 235cm added:
## The modified version is the default
boxplot(c(x,235),col="red",main="Modified boxplot")
text(1.3,quantile(c(x,235)),c("Minimum","Q1","Median","Q3",
                              ,"Maximum"),col="blue")
```

**Modified boxplot**



# Use your Brain!!!



# Næste uge:

- Sandsynlighed, diskrete fordelinger - kapitel 2 i eNoten

# Oversigt

- 1 Praktisk Information
- 2 Introduction to Statistics - a primer
- 3 Intro Case historier: IBM Big data, Novo Nordisk small data, Skive fjord
- 4 Introduktion til Statistik
- 5 Beskrivende statistik: Nøgletal
  - Gennemsnit
  - Median
  - Spredning
  - Fraktiler
  - Kovarians og Korrelation
  - Plots/figurer
- 6 Software: R