Introduktion til statistik

Forelæsning 1: Intro, R og beskrivende statistik

Peder Bacher

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 009
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: pbac@dtu.dk

Forår 2017

Kapitel 1: Simple plots og deskriptiv statistik

Teknikker til at "se" på data! (deskriptiv statistik)

Opsummerende størrelser for stikprøve

- Gennemsnittet ($\bar{x}$)
- Standard afvigelse (s)
- Empirisk varians (s^2)
- Fraktiler og percentiler *(f.eks. 15% af data ligger under 0.15 fraktil)*
- Median, øvre- og nedre kvartiler
- Empirisk korrelation (r) *(mellem to stikprøver)*

Simple plots

- Scatter plot *(xy plot)*
- Histogram *(empirisk tæthed)*
- Kumulativ fordeling *(empirisk fordeling)*
- Boxplots, søjlediagram, cirkeldiagram (lagkagediagram)

Chapter 1: Simple Graphics and Summary Statistics

Look at data as it is! (descriptive statistics)

Summary statistics

- Sample mean: $\bar{x}$
- Sample standard deviation: s
- Sample variance: s^2
- Quantiles and percentiles *(e.g. 15% of data is below 0.15 quantile)*
- Median, upper- and lower quartiles
- Sample correlation (r) *(between two samples)*

Simple graphics

- Scatter plot *(xy plot)*
- Histogram *(empirical density)*
- Cumulative distribution *(empirical distribution)*
- Boxplots, Bar charts, Pie charts

Oversigt

Praktisk Information

- Undervisning: Hver fredag 8-10
- Generel daglig agenda:
  - FØR undervisningsmodulet: læs det annoncerede i bogen!
  - 2x45 minutters forelæsning (ugens pensum)
  - 2 timers øvelser: Excersises i bogen (HUSK PEN OG PAPIR)
- Skriftlig eksamen: Søndag 28. maj
- OBLIGATORISKE projekter: 2 stk - skal godkendes for at kunne gå til eksamen
- *Installer lige socrative app på dit device*

Praktisk Information

- Campusnet: `www.campusnet.dtu.dk`
  - Meddelelser
  - Links til interessante historier
  - Projekter - download og aflevering
- Hjemmeside: `02323.compute.dtu.dk`
  - Forelæsningsplan
  - Læsemateriale: Introduction to Statistics at DTU
  - Øvelser & besvarelser
  - Slides og R-scripts
  - Podcasts af forelæsninger (02402) (nu engelsk, forrige på dansk)
  - Quizzer
  - Projekt info

Eksempel med terning

- Hvordan kan man teste om en terning er fair?
- F.eks. givet en terning, svar på: Er der 1/6 sandsynlighed for at slå en sekser?
  - Stort set umuligt at beskrive med fysik
  - Derfor:
    - *Kast med terningen, observer og derefter udregn statistik*
    - Afgør om der er 1/6 $\pm$ *fejlmargin* sandsynlighed for at slå sekser med terningen

Der er altid en hvis sandsynlighed for at tage fejl! men den kan styres til at matche risikoen man vil tage

Hvor mange gange skal jeg slå med terningen?

- Hvor mange gange skal jeg slå med terningen for at afgøre om terningen slår seksere med 1/6 $\pm$ *fejlmargin* sandsynlighed?
- Det kan I nemt beregne om 13 uger :)
- Beregn det med R:

```
alpha <- 0.05
## Fejlmargen vi vil tillade (kaldet præcisionen)
ME <- 0.01
## Beregn antal gange vi skal slå med terningen
p * (1-p) * (qnorm(1-alpha/2)/ME)^2
```

Statistikken historie og anvendelse i medicin

New England Journal of medicine:

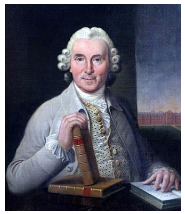
EDITORIAL: Looking Back on the Millennium in Medicine, *N Engl J Med*, 342:42-49, January 6, 2000.
`http://www.nejm.org/doi/full/10.1056/NEJM200001063420108`

Millennium list (10 vigtigste bidrag til udvikling af medicin)

- 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

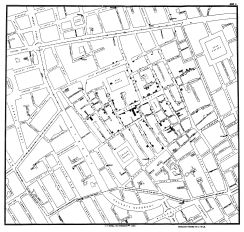
"*One of the earliest clinical trials took place in 1747, when James Lind treated 12 scorbutic ship passengers with cider, an elixir of vitriol, vinegar, sea water, oranges and lemons, or an electuary recommended by the ship's surgeon. The success of the citrus-containing treatment eventually led the 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`).



Man kan altså undersøge fænomener man ikke forstår og derefter begynde at forstå dem!

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)`).



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.*"

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.*"

Optagelse af gæsteforedrag af Henrik H. Eliassen IBM på introstat hjemmesiden

Introduktion til Statistik

- Hvordan behandles (eller analyseres) data?
- Hvordan beskrives *tilfældig variation*?
- Statistik er et værktøj til at træffe beslutninger
- Meget vigtigt værktøj i ingeniørens værktøjskasse:
  - Analyse af data
  - Forsøgsplanlægning
  - Forudsigelse af fremtidige værdier
  - ... og meget mere!

Anvendelser på DTU (mest Compute)

- Energisystemer:
  - Prognoser af sol- og vindkraft
  - Modellering af energilagring, menneskers adfærd, spildevandsanlæg
- Styring:
  - Mekaniske systemer (e.g. robotter, biler, skibe, vindmøller, ...)
- Medicin (Compute):
  - Statistik på medicinforsøg, Kunstig bugspytkirtel
- Billedanalyse:
  - Billeder er observeret data!
  - Røntgenbilleder, 3D skanninger, Video, ...

Anvendelser på DTU (mest Compute)

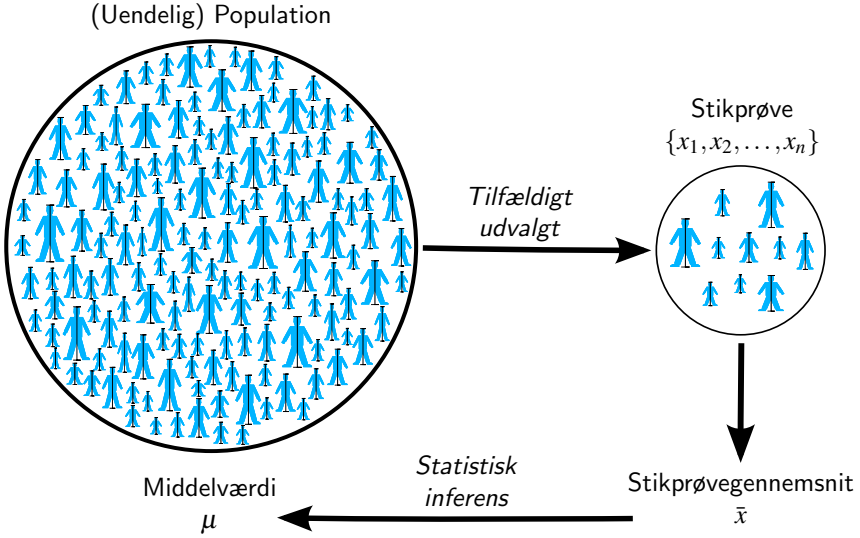
- Signalbehandling:
  - Elektriske systemer (filtre, forstærkere, ...)
- Computer science:
  - Internet data (trafik, Google, Facebook, osv.)
  - Tekstgenkendelse, Sikkerhed: Server angreb etc.
- Byg:
  - Tests af materialeegenskaber og konstruktioner
  - Energisystemer og indeklima
- Management:
  - Finans, spørgeskema undersøgelser, ...
- Kemi, fysik, miljø, ...

Hver gang man har målinger skal man sådan set bruge statistik

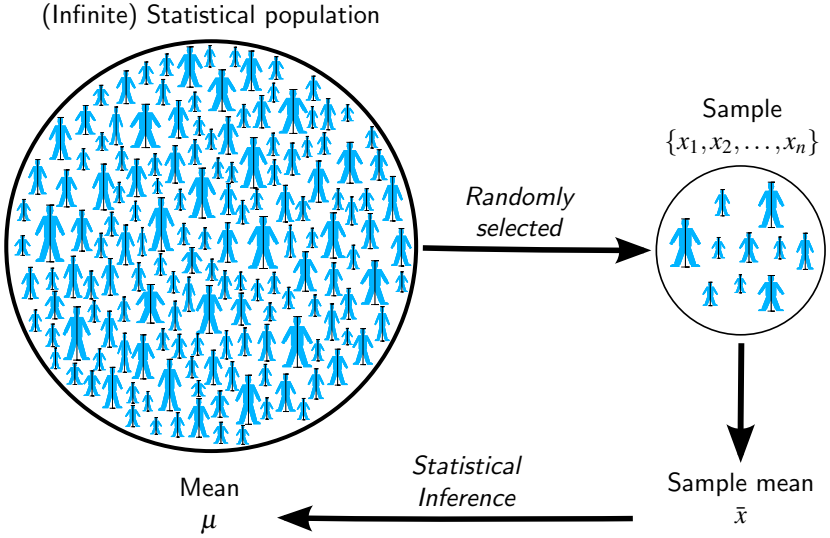
Population og stikprøve

- Statistik handler ofte om at analysere en *stikprøve* (sample), der er taget fra en *population* (population)
- Baseret på stikprøven, vil vi generalisere om populationen (dvs. beskrive noget om hele populationen)
- Det er derfor vigtigt, at stikprøven er *repræsentativ* for populationen

Population og stikprøve



Population og stikprøve



Meningsmålinger Socrative.com, room: PBAC

Hvilken af følgende er en typisk årsag til at meningsmålinger ved valg ofte fejler?

A: Stikprøvetagning er nemt, men beregningerne er for simple

B: Visse grupper er svære at få til at deltage i meningsmålinger

C: Vælgerne har ikke bestemt sig for hvad de vil stemme før valget

D: Ved ikke

B: Hvis nogle grupper er svære at få til at svare på meningsmålinger, så bliver de underrepræsenterede i stikprøven. Det er svært at korrigere for! (det var bare et eksempel, der er mange årsager til *biased* meningsmålinger)

Nøgletal (summary statistics)

Vi anvender en række *nøgletal* (eller statistikker) for at opsummere og beskrive data (en stikprøve):

- **Gennemsnit:** Tyngdepunkt eller centrering
- **Median:** Tyngdepunkt eller centrering
- **Varians:** Variation
- **Spredning:** Variation (samme enhed som stikprøve)
- **Fraktiler og kvartiler:** Siger noget om fordelingen af stikprøve
- **Variations koefficient:** Variationen i stikprøve (enhedsløs)
- **Kovarians:** Samvariation mellem datasæt
- **Korrelation:** Samvariation mellem datasæt (enhedsløs)

(Stikprøve)Gennemsnit (sample mean)

- Gennemsnittet er et nøgletal, der angiver tyngdepunkt eller centrering
- **Stikprøvegennemsnit**

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

- Vi siger, at $\bar{x}$ er et *estimat* af middelværdien for populationen (populationsgennemsnittet)

Median

- **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
- (Stikprøve)**Median**:
  Den midterste observation i den sorterede rækkefølge
  (tallet hvor der er lige mange observationer under og over)

Eksempel: Højder af unge mænd

Stikprøve (sample) $x = [185, 184, 194, 180, 182]$

$n = 5$

- **Gennemsnit**

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

- **Median**

Først sorter data: 180 182 184 185 194
Vælg så det midterste (idet n er ulige)(3'te) tal: 184

Hvis en person på 235cm tilføjes til stikprøven, hvilken bliver mest påvirket? (socrative.com eller app. Room:PBAC)

A: Gennemsnittet B: Medianen C: Påvirkes lige meget D: Ved ikke

Svar) A: Gennemsnittet. Det stiger meget mere end medianen (nyt gennemsnit 193.3 og ny median 184.5)

Stikprøvevarians (sample variance) og -standardafvigelse (sample standard deviation)

Stikprøvevarians siger noget om 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 (sample): $x = [185, 184, 194, 180, 182]$

$n = 5$

- **Stikprøvevarians** (sample variance)

$$s^2 = \frac{1}{4}\left((185 - \bar{x})^2 + (184 - \bar{x})^2 + (194 - \bar{x})^2 + (180 - \bar{x})^2 + (182 - \bar{x})^2\right) = 29$$

- **Standardafvigelse** (sample standard deviation)

$$s = \sqrt{s^2} = \sqrt{29} = 5.385$$

Fraktiler (percentiles eller quantiles)

- Medianen beregnes som det punkt, der deler data ind i to halvdele
- Man kan naturligvis finde punkter som deler i andre dele:
  De punkter kaldes **fraktiler**
- Ofte beregner man
  - $0, 25, 50, 75, 100\%$ fraktilerne kaldes **kvartilerne** (quartiles)
  - 50% fraktilen er altså medianen

- Eksempel: 10% fraktilen er punktet (estimat) hvor 10% af observationerne ligger under

Fraktiler (percentiles eller quantiles, Definition 1.7)

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

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

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

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

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

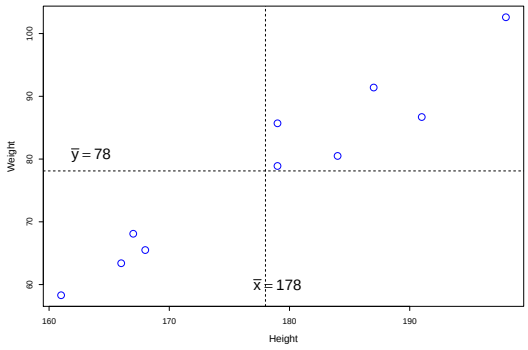
hvor $\lceil np \rceil$ er *ceiling*("loftet") af np, dvs. det mindste hele tal større end np

Eksempel på fraktiler: Højder af unge mænd

- **Data,** n=10: 168 161 167 179 184 166 198 187 191 179
- Sorteret: 161 166 167 168 179 179 184 187 191 198
- **Nedre kvartil (25% fraktil),** Q_1:
  Sorter og så vælg det rigtige baseret på $np = 2.5$:
  $Q_1 = 167$
- **Øvre kvartil (75% fraktil),** Q_3:
  Sorter og så vælg det rigtige baseret på $np = 7.5$:
  $Q_3 = 187$

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

Kovariansen

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

Korrelationskoefficient

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

hvor s_x og s_y er standard afvigelsen for henholdsvis x og y

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

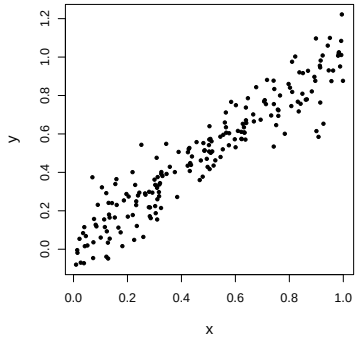
Korrelation - egenskaber

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

Korrelation Socrative.com, room: PBAC

Hvad er korrelationen mellem x og y?

A: ca. -0.95 B: ca. 0 C: ca. 0.95

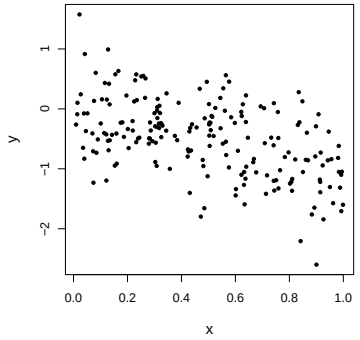


Svar) C: ca. 0.95

Korrelation

Hvad er korrelationen mellem x og y?

A: ca. 0 B: ca. -0.5 C: ca. -0.95

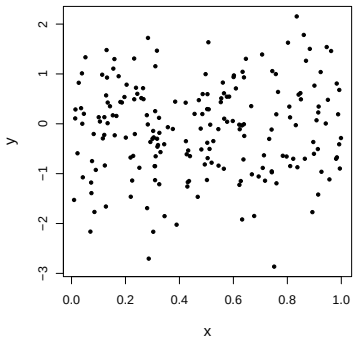


Svar) B: ca. -0.5

Korrelation

Hvad er korrelationen mellem x og y?

A: ca. -0.5 B: ca. 0 C: ca. 0.5

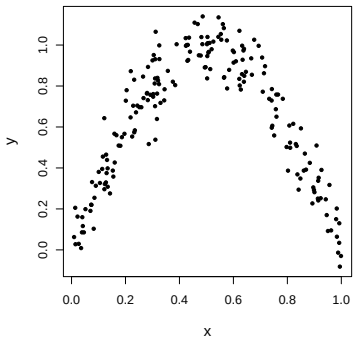


Svar) B: ca. 0

Korrelation

Hvad er korrelationen mellem x og y?

A: ca. -0.5 B: ca. 0 C: ca. 0.5



Svar) B: ca. 0

Beskrivende statistik (explorative statistics)

- Undersøg et data sæt
- Beskriv data ved at fremhæve de vigtige pointer så andre hurtigt kan se indhold, trends og synlige sammenhænge
- Præsenter data for andre, som ikke kender det
- Grafisk fremstilling med forskellige plots:
  - Histogram (empirisk tæthedsfunktion)
  - Empirisk kumulativ tæthedsfunktion
  - Boxplot
  - Scatterplot

Software: R

- Installer R og Rstudio på egen computer
- Introduceres i bogen (kap. 1.5)
- Er integreret i mange ting i kurset vi gør
- Globalt og hurtigt voksende open source beregningsmiljø
- ADVARSEL: R kan IKKE erstatte vores hjerner!!!! (Læs sektion 1.5.4!)

Software: R

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

```
> ## Define a vector
> x <- c(1, 4, 6, 2)
> x
```

```
> ## A sequence from 1 to 10
> x <- 1:10
> x
```


Software: R

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

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

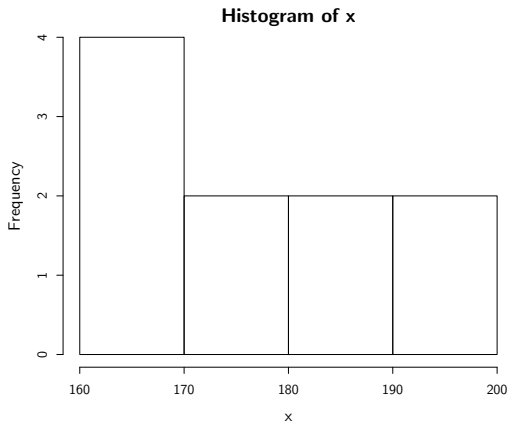

Software: R

```
## Sample quartiles
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  164  166  168  174  179  184  187  189  194  198
```

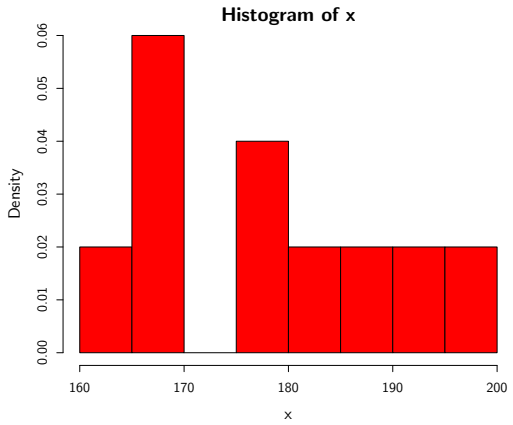

Histogram

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



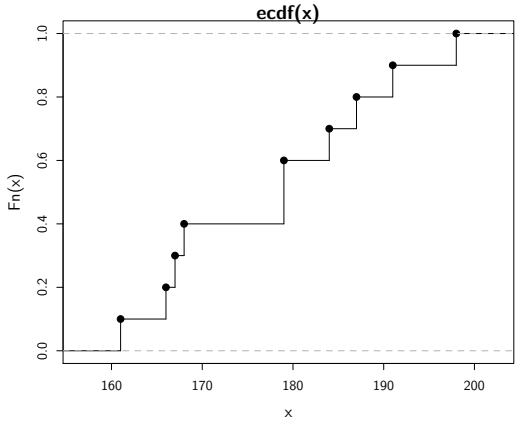
Empirisk tæthed (empirical density plottes med density histogram)

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



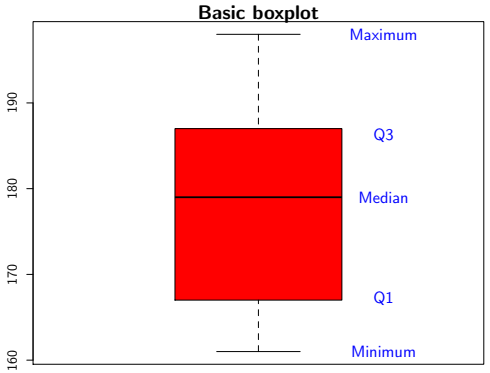
Empirisk kumuleret distribution

```
## Empirical cumulated distribution function (ecdf)
plot(ecdf(x), verticals=TRUE)
```



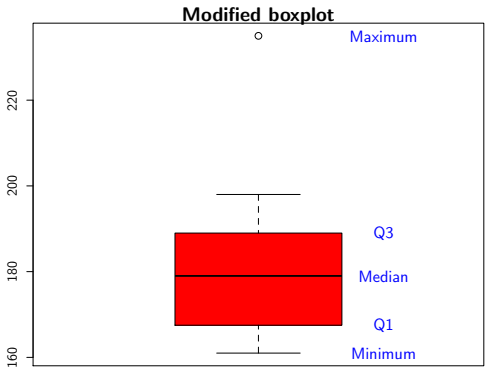
Boxplot

```
## 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")
```



Modified boxplot

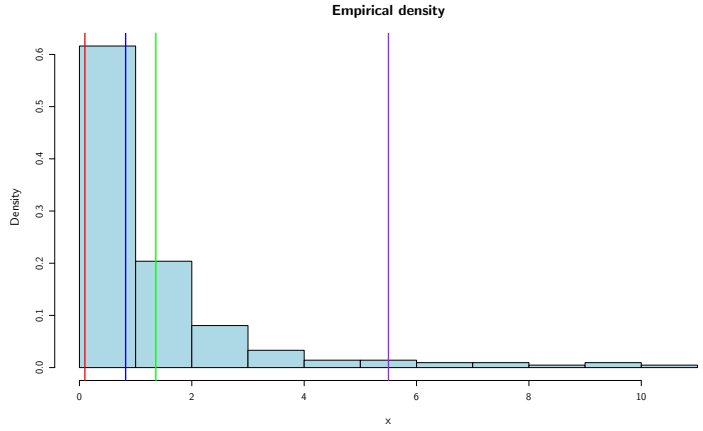
```
## A modified boxplot of the heights with an
## extreme observation, 235cm added:
## The modified version is the default
boxplot(c(x,235), col="red", main="Modified boxplot")
text(1.3, quantile(c(x,235)), c("Minimum","Q1","Median","Q3",
     "Maximum"),col="blue")
```



Spørgsmål Socrative.com, room: PBAC

Hvad er markeret med den lilla linie?

A: Gennemsnittet B: Medianen C: 10% fraktilen D: 95% fraktilen

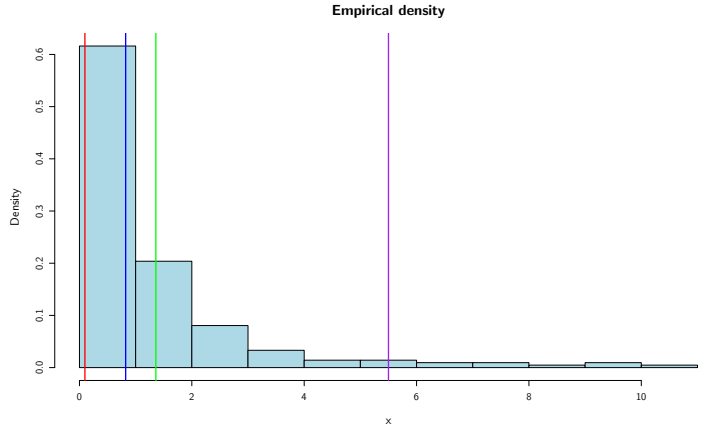


Svar) D: 95% Fraktil

Spørgsmål Socrative.com, room: PBAC

Hvad er markeret med den røde linie?

A: Gennemsnittet B: Medianen C: 10% fraktilen D: 95% fraktilen

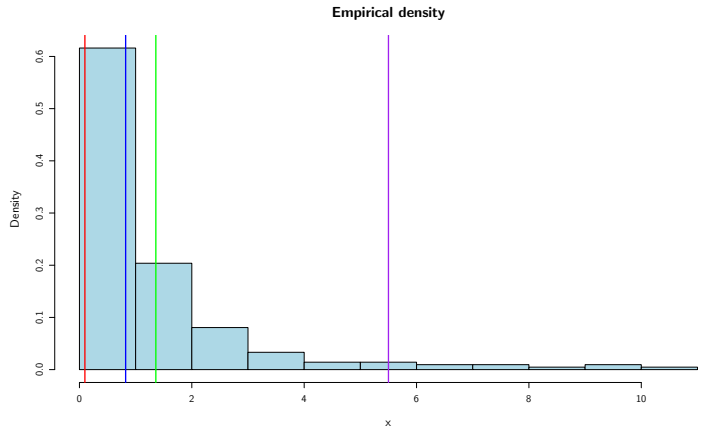


Svar) C: 10% Fraktil

Spørgsmål Socrative.com, room: PBAC

Hvad er markeret med den blå linie?

A: Gennemsnittet B: Medianen C: 10% fraktilen D: 95% fraktilen

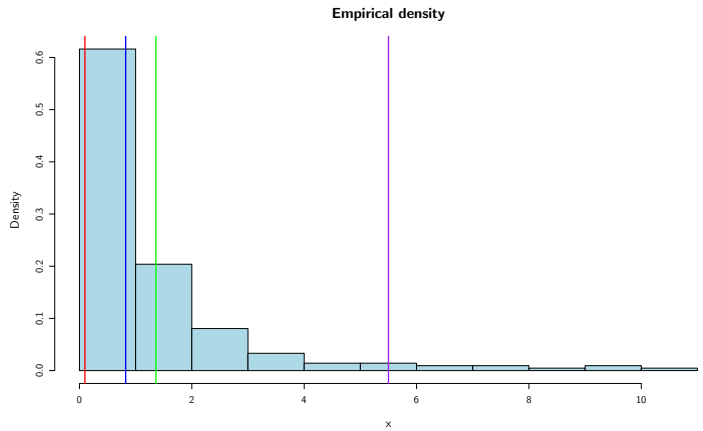


Svar) B: Medianen

Spørgsmål Socrative.com, room: PBAC

Hvad er markeret med den grønne linie?

A: Gennemsnittet B: Medianen C: 10% fraktilen D: 95% fraktilen



Svar) A: Gennemsnittet

Projekter

- Der skal laves **to projekter**
- Emne for alle projekter i 1. omgang: Beskrivende statistik, konfidensintervaller og hypotesetest
- 1. omgang er der fire projekter at vælge imellem:
  - `Handel med ETF I`
  - `Varmeforbrug i Sønderborg I`
  - `Skive fjord I`
  - `BMI I`
- Man må gerne arbejde i grupper om beregninger, men rapporterne skal skrives individuelt, se mere på `https://02323.compute.dtu.dk/Agendas/`
- Se også afleveringsdatoer på hjemmesiden
- Begynd på projekt 1 efter opgaverne i dag til grupperegning
- Bemærk: der checkes for plagiering og det bliver anmeldt!

Begge skal godkendes for at kunne gå til eksamen. Får man ikke godkendt første aflevering er der mulighed for genaflevering

Næste uge:

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