

# Introduktion til statistik

## Forelæsning 1: Intro, R og beskrivende statistik

Peder Bacher

DTU Compute, Dynamiske Systemer  
Bygning 303B, Rum 010  
Danmarks Tekniske Universitet  
2800 Lyngby – Danmark  
e-mail: pbac@dtu.dk

Forår 2021

# Kapitel 1: Simple plots og deskriptiv statistik

Tag en *stikprøve*: Brug deskriptiv statistik til at "se" på den!

## Opsummerende størrelser for stikprøve

- Gennemsnittet ( $\bar{x}$ )
- Standardafvigelse ( $s$ )
- Empirisk varians ( $s^2$ )
- Fraktiler og percentiler (*f.eks. 15% af data ligger under 0.15 fraktilen*)
- Median, øvre- og nedre kvartiler
- Empririsk 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

Take a *sample*: Use descriptive statistics to “look” at it!

## 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

- 1 Praktisk Information
- 2 Introduction to Statistics - a primer
- 3 Population og stikprøve
- 4 Beskrivende statistik: Nøgletal
  - Gennemsnit
  - Median
  - Spredning
  - Fraktiler
  - Kovarians og Korrelation
- 5 Beskrivende statistik: Grafisk fremstilling
- 6 Software: R
- 7 Projekter

# Praktisk Information

- Generel daglig agenda:
  - FØR undervisningsmodulet: læs det annoncerede i bogen!
  - 2x45 minutters forelæsning
  - 2 timers øvelser: Excercises i bogen (HUSK PEN OG PAPIR)
- Mere:
  - Area9: Adaptive learning
  - Test quizzes: Gamle eksamensspørgsmål
  - Zoom kanal til online TA hjælp 'Introstat\_02323': Se meddelelse
- Skriftlig eksamen: Dato står på hjemmesiden
- OBLIGATORISKE projekter: 2 stk - skal godkendes for at kunne gå til eksamen
- *Installer lige Socrative app på dit device*

# Praktisk Information

- DTU Learn: [learn.inside.dtu.dk/d2l/home/44519](https://learn.inside.dtu.dk/d2l/home/44519)
  - Meddelelser (slå email notifications til)
  - Projekter - download og aflevering
- Hjemmeside: [02323.compute.dtu.dk](https://02323.compute.dtu.dk)
  - Forelæsningsplan
  - Læsemateriale: Introduction to Statistics at DTU
  - Øvelser & besvarelser
  - Slides og R-scripts
  - Podcasts af forelæsninger (02402) (engelsk, gamle 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  $1/6 \pm \text{fejlmargin}$  sandsynlighed?
- Det kan I nemt beregne om 13 uger :)
- Beregn det med R:

```
alpha <- 0.05
## Fejlmargen vi vil tillade (kaldet præcisionen, margin of error)
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."*  
(se [http://en.wikipedia.org/wiki/James\\_Lind](http://en.wikipedia.org/wiki/James_Lind)).

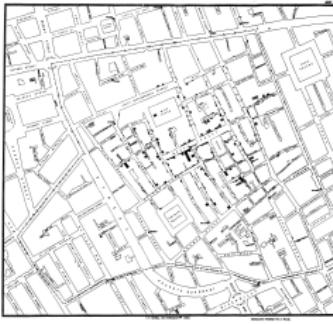


*Man kan altså undersøge fænomener man ikke forstår og derefter begynde at forstå dem!*

# 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."*

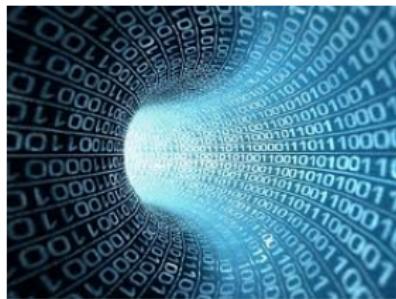
(se [http://en.wikipedia.org/wiki/John\\_Snow\\_\(physician\)](http://en.wikipedia.org/wiki/John_Snow_(physician))).



Google - *Big Data*

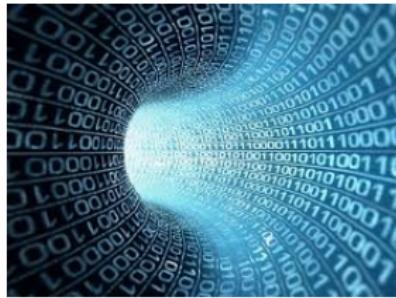
A quote from New York Times, 5. August 2009, from the article titled "For Today's Graduate, Just One Word: Statistics" is:

*"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."*



# 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 hjemmesiden.

# Hvad er Statistik?

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

# Spørgsmål Socrative.com, room: PBAC

Tror du, at din karakter til eksamen ligger blandt de 50% højeste, dvs. ligger du i den bedste halvdel til eksamen?

- TRUE: Ja
- FALSE: Nej

# Anvendelser på DTU (mest Compute)

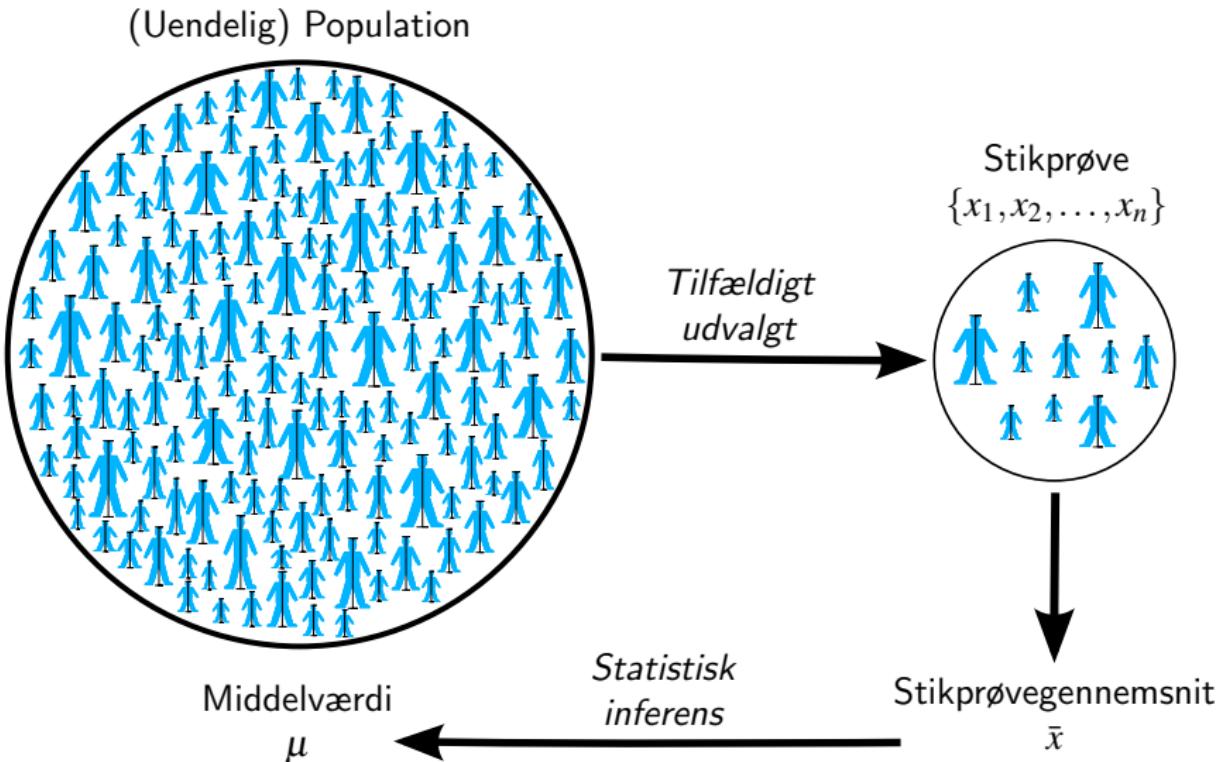
- Energisystemer:
  - Prognoser af sol- og vindkraft
  - Modellering af energilagring, menneskers adfærd, spildevandsanlæg
- Styring:
  - Mekaniske systemer (e.g. robotter, biler, skibe, vindmøller, ...)
- Medicin (Compute):
  - Statistik på medicinforsøg, Kunstig bugspytkirtel
- Billedanalyse:
  - Billeder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

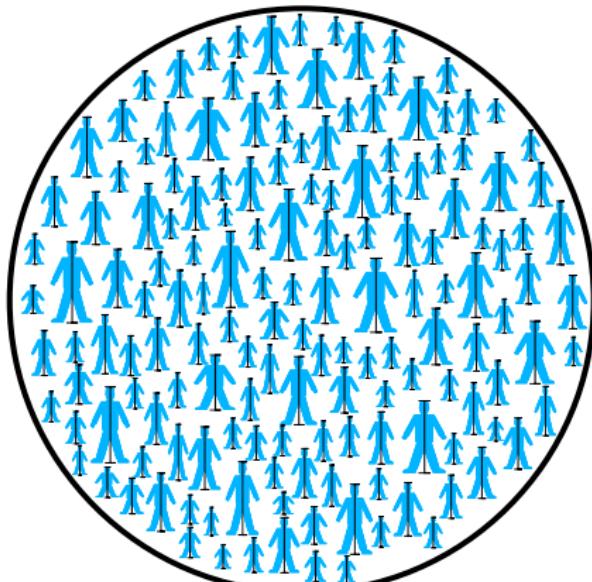
- Statistik handler ofte om at analysere en *stikprøve* (sample), der er taget fra en *population* (population)
- Baseret på stikprøven, vil vi generalisere om populationen (dvs. beskrive noget om hele populationen)
- Det er derfor vigtigt, at stikprøven er *repræsentativ* for populationen

# Population og stikprøve



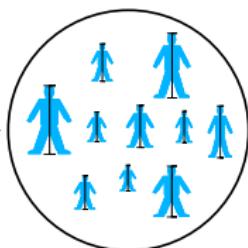
# Population og stikprøve

(Infinite) Statistical population



*Randomly selected*

Sample  
 $\{x_1, x_2, \dots, x_n\}$



Sample mean  
 $\bar{x}$

*Statistical Inference*

# Meningsmålinger

Socrative.com, room: PBAC

Hvilken af disse årsager, tror du, ofte fremhæves som grunden til forkerte meningsmålinger?

- A: Stikprøvetagningen er ikke et problem, men de statistiske beregninger er meget komplikerede
- B: Det er svært at tage en repræsentativ stikprøve  
(det er bl.a. svært at få alle samfundsgrupper til at svare lige meget, samt forudse hvor meget de vil møde op og stemme)
- C: De fleste svarer, at de ved ikke hvem de vil stemme på

# 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)

# Stikprøvegennemsnit (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
- **Median (stikprøvemedian):**  
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

# 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)

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

- **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 halvdeler
- 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:  $x_{(1)}, \dots, x_{(n)}$
- 2 Beregn  $pn$
- 3 Hvis  $pn$  er et helt tal: Midl den  $pn$ 'te og  $(pn + 1)$ 'te sorterede observationer

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

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

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

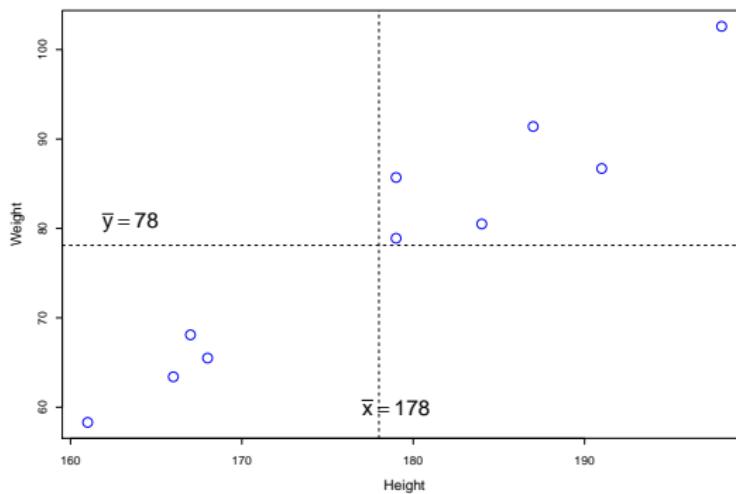
hvor  $\lceil np \rceil$  er *ceiling*("loftet") af  $np$ , dvs. det mindste hele tal større end  $np$

## 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

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

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

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

# Korrelation - egenskaber

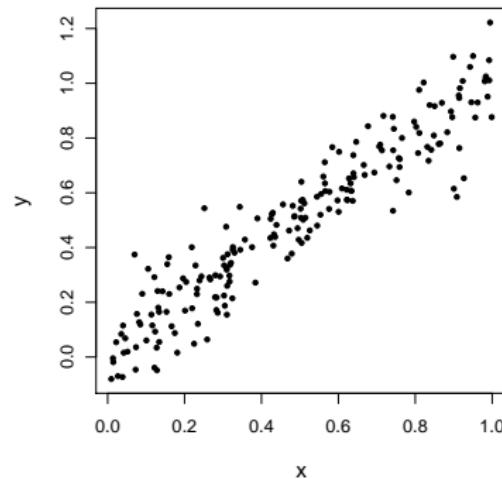
## Korrelation - egenskaber

- $r$  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 i scatter plottet er positiv
- $r < 0$  hvis den generelle trend i scatter plottet er negativ

# Korrelation Socrative.com, room: PBAC

Hvad er korrelationen mellem  $x$  og  $y$ ?

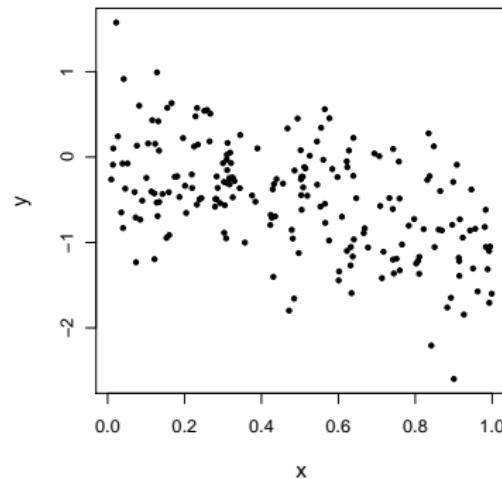
- A: ca. -0.95    B: ca. 0    C: ca. 0.95



# Korrelation

Hvad er korrelationen mellem  $x$  og  $y$ ?

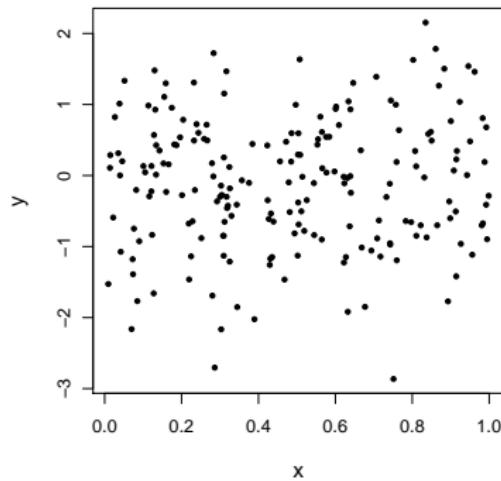
- A: ca. 0    B: ca. -0.5    C: ca. -0.95



# Korrelation

Hvad er korrelationen mellem  $x$  og  $y$ ?

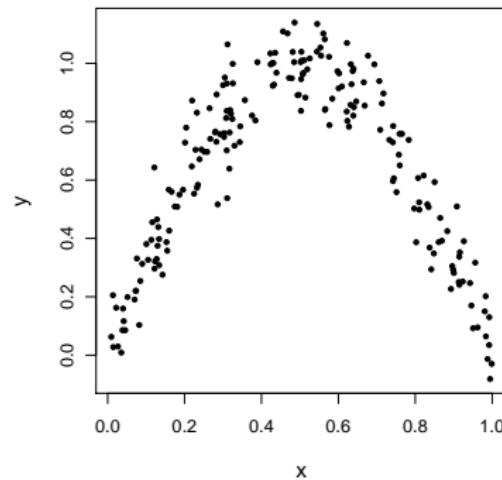
- A: ca. -0.5    B: ca. 0    C: ca. 0.5



# Korrelation

Hvad er korrelationen mellem  $x$  og  $y$ ?

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



# Beskrivende statistik (explorative statistics)

- Undersøg et data sæt
- Beskriv data ved at fremhæve de vigtige pointer: alt data, 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 open source beregningsmiljø

# 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)

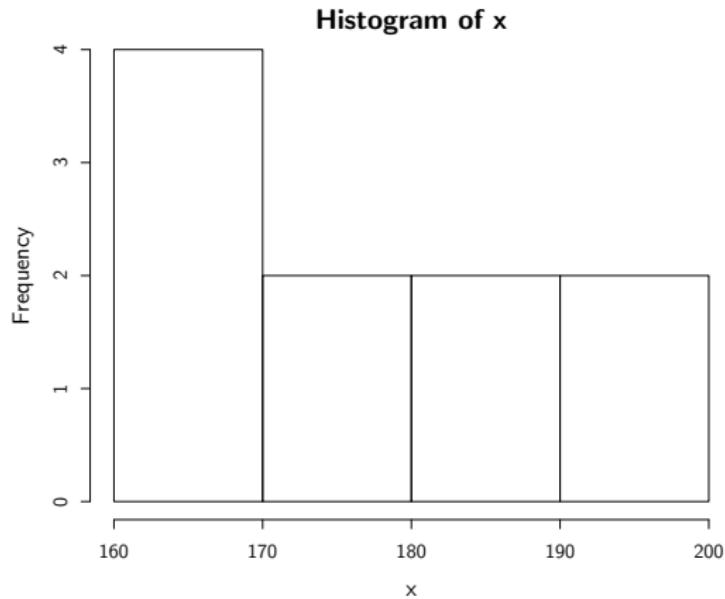
##    0%   25%   50%   75% 100%
## 161   167   179   187   198
```

```
## 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
```

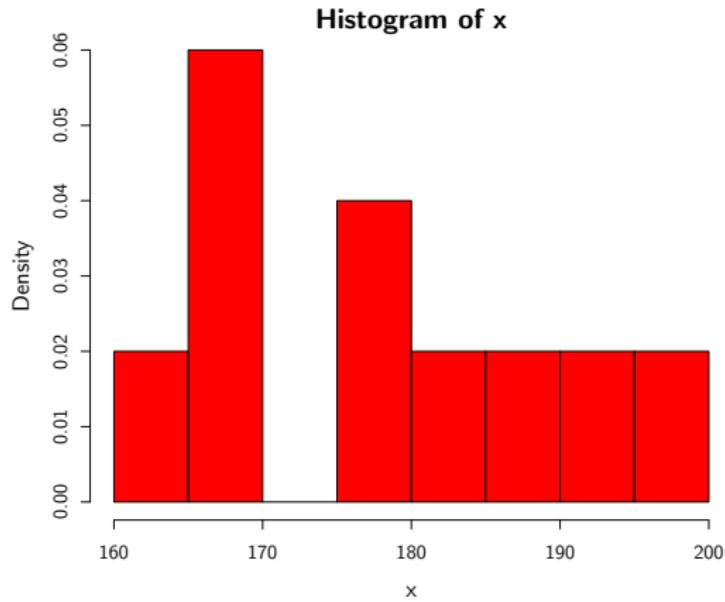
# Histogram

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



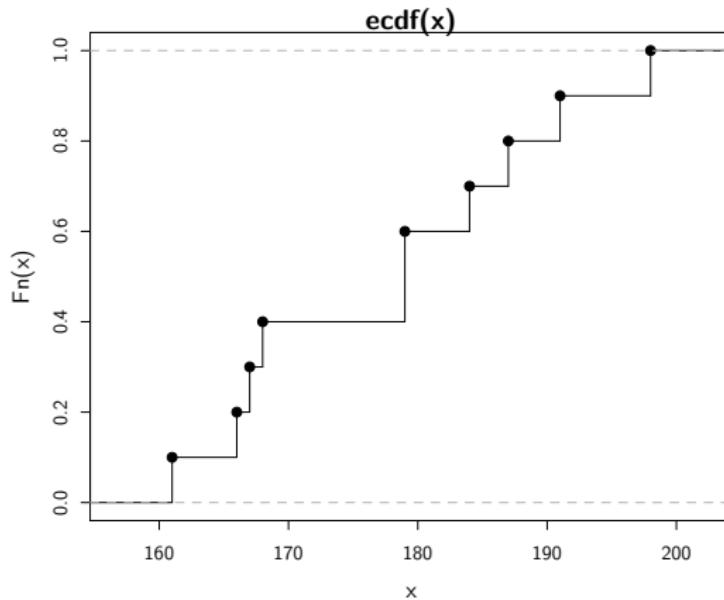
# Empirisk tæthed (empirical density plottes med density histogram)

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



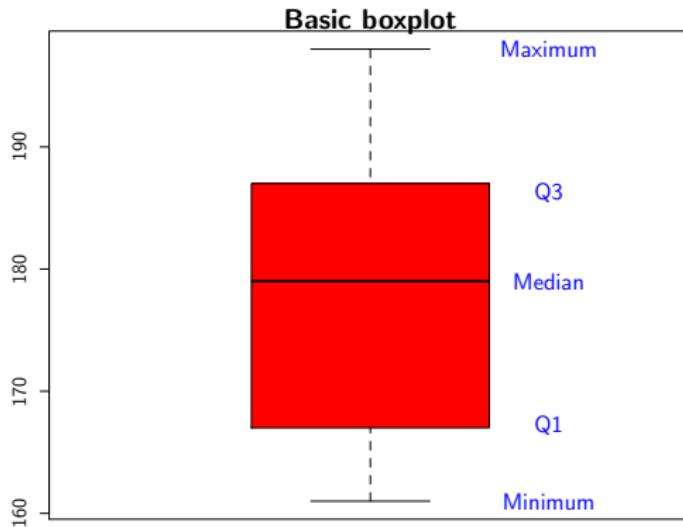
# Empirisk kumuleret distribution

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



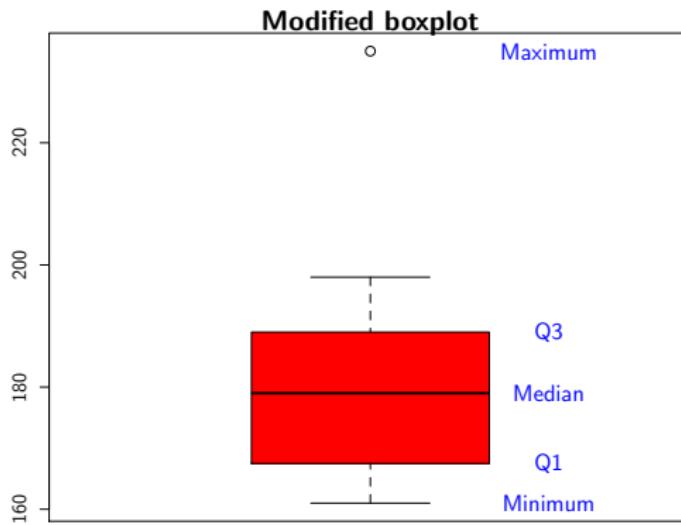
# Boxplot

```
## 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

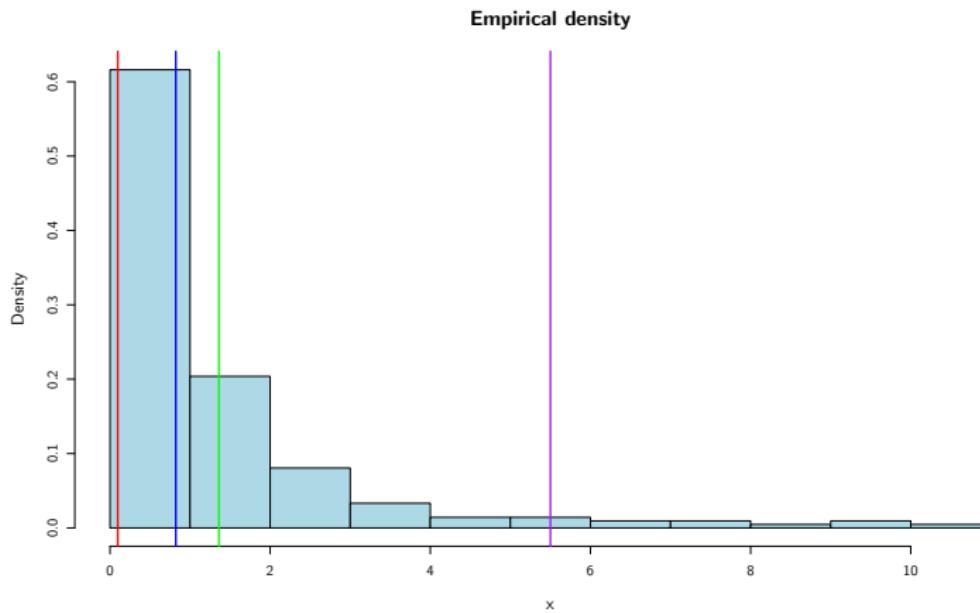
```
## A modified boxplot of the heights with an
## extreme observation, 235cm added:
## The modified version is the default
boxplot(c(x,235), col="red", main="Modified boxplot")
text(1.3, quantile(c(x,235)), c("Minimum", "Q1", "Median", "Q3",
  "Maximum"), col="blue")
```



# Spørgsmål Socrative.com, room: PBAC

Hvad er markeret med den lilla linie?

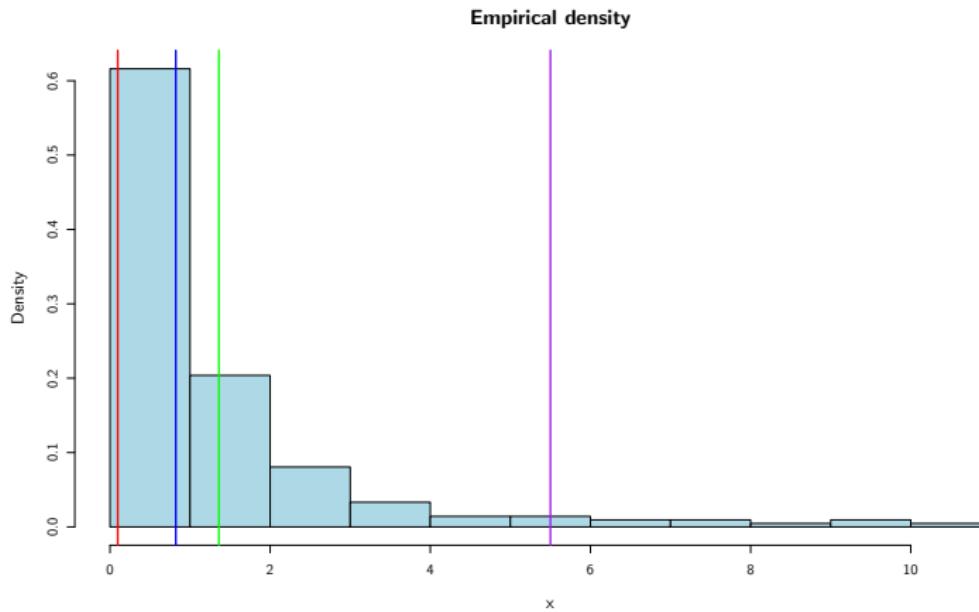
- A: Gennemsnittet   B: Medianen   C: 10% fraktilen   D: 95% fraktilen



# Spørgsmål Socrative.com, room: PBAC

Hvad er markeret med den røde linie?

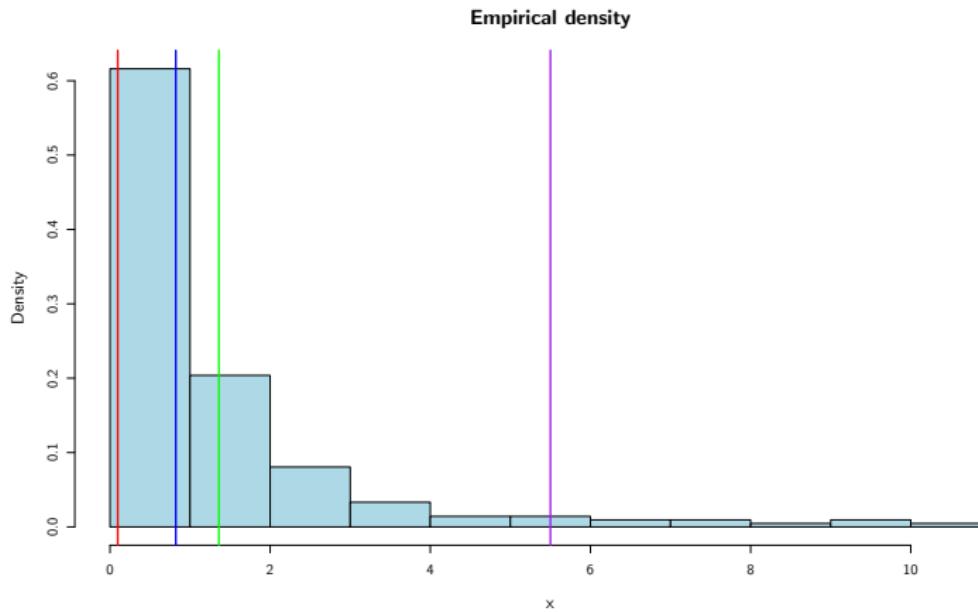
- A: Gennemsnittet   B: Medianen   C: 10% fraktilen   D: 95% fraktilen



# Spørgsmål Socrative.com, room: PBAC

Hvad er markeret med den blå linie?

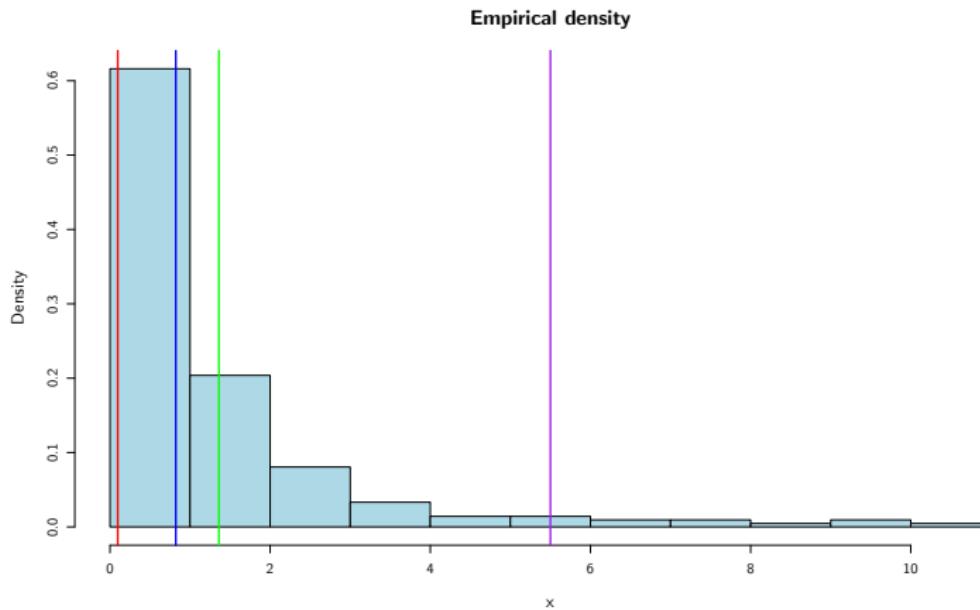
- A: Gennemsnittet   B: Medianen   C: 10% fraktilen   D: 95% fraktilen



# Spørgsmål Socrative.com, room: PBAC

Hvad er markeret med den grønne linie?

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



# 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
- Arbejde i grupper om beregninger, men rapporterne skal skrives individuelt, se mere på <https://02323.compute.dtu.dk/projects/>
- Se også afleveringsdatoer på hjemmesiden
- Begynd på projekt 1 næste gang 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
- HUSK AT BRUGE PAPIR OG BLYANT NÅR I REGNER!

# Introduktion til Statistik

## Forelæsning 2: Stokastisk variabel og diskrete fordelinger

Peder Bacher

DTU Compute, Dynamiske Systemer  
Bygning 303B, Rum 010  
Danmarks Tekniske Universitet  
2800 Lyngby – Danmark  
e-mail: pbac@dtu.dk

Forår 2021

# Kapitel 2: Diskrete fordelinger

Grundlæggende koncepter:

- Stokastisk variabel (*værdi afhængig af udfald af endnu ikke udført eksperiment*)
- Tæthedsfunktion:  $f(x) = P(X = x)$  (*pdf*)
- Fordelingsfunktion:  $F(x) = P(X \leq x)$  (*cdf*)
- Middelværdi:  $\mu = E(X)$
- Standard afvigelse:  $\sigma$
- Varians:  $\sigma^2$

Specifikke distributioner:

- Binomial (*tæl antal succes ud af n trækninger*)
- Hypergeometrisk (*trækning uden tilbagelægning*)
- Poisson (*antal hændelser i interval*)

# Chapter 2: Discrete Distributions

## General concepts:

- Random variable (*value is outcome of yet not carried out experiment*)
- Density function:  $f(x) = P(X = x)$  (*pdf*)
- Distribution function:  $F(x) = P(X \leq x)$  (*cdf*)
- Mean:  $\mu = E(X)$
- Standard deviation:  $\sigma$
- Variance:  $\sigma^2$

## Specific distributions:

- The binomial distribution (*dice roll*)
- The hypergeometric distribution (*draw without replacement*)
- The Poisson distribution (*number of events in interval*)

# Oversigt

- 1 Stokastisk variabel
- 2 Tæthedsfunktion (pdf)
- 3 Fordelingsfunktion (cdf)
- 4 Konkrete statistiske fordelinger
  - Binomialfordelingen
  - Hypergeometrisk fordeling
  - Eksempler
  - Poissonfordelingen
- 5 Middelværdi og varians
  - Middelværdi og varians for de diskrete fordelinger

# Praktisk information

SE PRAKTISK INFORMATION PÅ HJEMMESIDEN OG I STARTEN AF SLIDES FRA UGE 1

- Download script på <https://02323.compute.dtu.dk/material>.
- Lad os få lidt gang i den med et kampråb!

# Stokastisk variabel

En **stokastisk variabel** (random variable) tildeler en værdi til udfaldet af et eksperiment *der endnu ikke er udført*, f.eks.:

- Et terningekast
- Antallet af seksere i 10 terningekast
- Hvor stor en andel svarer ja til et spørgsmål
- km/l for en bil
- Måling af sukkerniveau i blodprøve
- ...

# Diskret eller kontinuert

- Vi skelner mellem diskret og kontinuert
- **Diskret** (kan ofte tælles):
  - Hvor mange der bruger briller herinde
  - Antal mange flyvere letter den næste time
  - ...
- **Kontinuert**:
  - Vindmåling
  - Brandstofforbrug på en køretur
  - ...
- I dag er det diskret og i næste uge er det kontinuert.

# Stokastisk variabel

- Før eksperimentet udføres: En **stokastisk variabel**

$$X_1$$

noteret med stort bogstav

- Så udføres eksperimentet: Vi har da en *realisation* eller *observation*

$$x_1$$

noteret med småt bogstav

# Stokastisk variabel og stikprøve

- Før eksperimentet udføres: **Stikprøven** som  $n$  stokastiske variable

$$X_1, X_2, \dots, X_n$$

noteret med stort bogstav

- Så udføres eksperimentet: Vi har da  $n$  realisationer (*observationer*)

$$x_1, x_2, \dots, x_n$$

noteret med små bogstav

- Dvs. vi udfører eksperimentet  $n$  gange for at lave stikprøven

# Eksempel: Simuler et terningekast

- Vælg et tal fra (1,2,3,4,5,6) med lige stor sandsynlighed for hvert udfald
- Simuler i R

```
## Simuler et terningekast

## Vælg et tal fra (1,2,3,4,5,6) med lige sandsynlighed for hvert udfald
sample(1:6, size=1)

## Antal simulerede realiseringer
n <- 30
## Træk uafhængigt fra mængden (1,2,3,4,5,6) med ens sandsynlighed
sample(1:6, size=n, replace=TRUE)
```

# Tæthedsfunktion (probability density function (pdf))

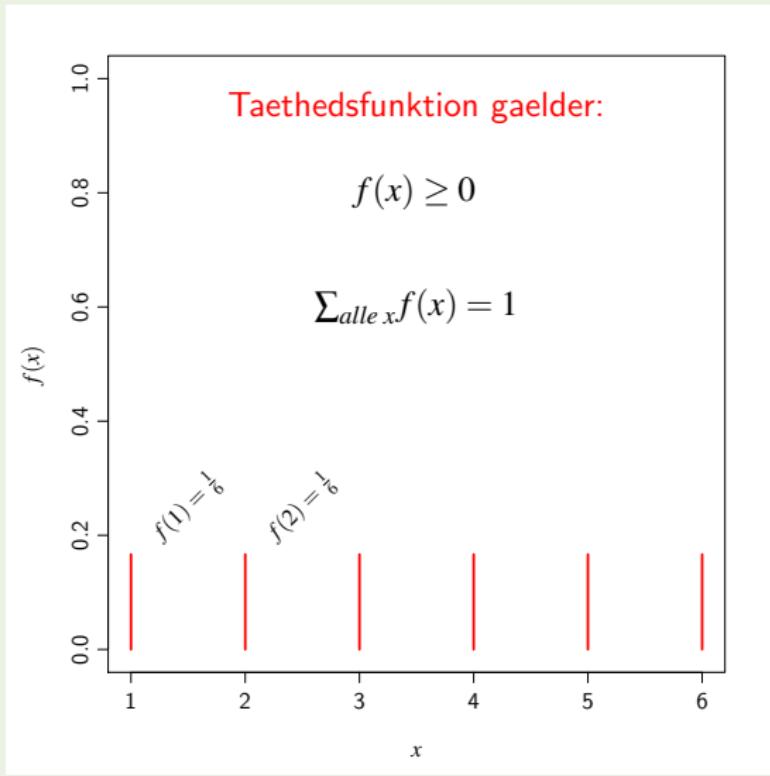
- Hvordan kan vi regne på eksperimentet før det er udført?
- Vi kender ikke værdien af variablen før eksperimentet er udført!? Løsning:  
Brug tæthedsfunktionen

Def. 2.6: En stokastisk variabel har en **tæthedsfunktion**

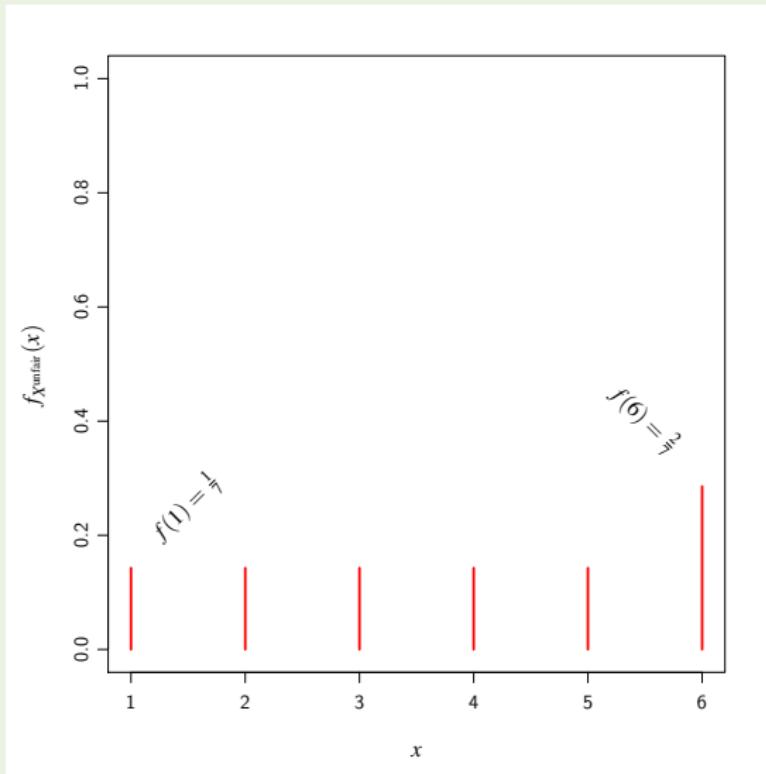
$$f(x) = P(X = x)$$

- 
- Den giver sandsynligheden for at  $X$  antager værdien  $x$  når eksperimentet udføres

# Eksempel: En fair ternings tæthedsfunktion



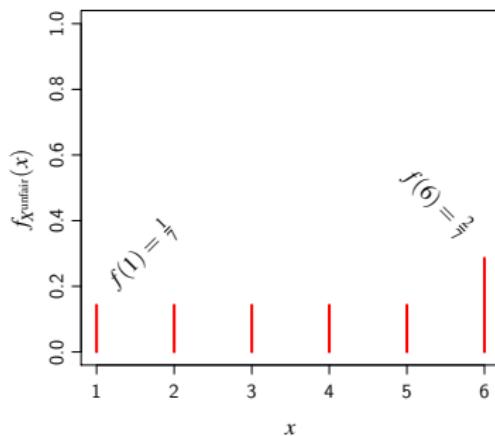
# Eksempel: En unfair ternings tæthedsfunktion



# Spørgsmål om unfair terning (socrative.com, room: PBAC)

Find nogle sandsynligheder for  $X^{\text{unFair}}$ :

- Sandsynligheden for at få en fire?
- Sandsynligheden for at få en femmer eller en sekser?
- Sandsynligheden for at få mindre end tre?



Svarmuligheder:

- A:  $\frac{3}{7}$   
 B:  $\frac{1}{6}$   
 C:  $\frac{4}{7}$   
 D:  $\frac{2}{7}$   
 E:  $\frac{1}{7}$

# Stikprøve

- Vi har en terning og vil nu undersøge om en terningen er fair.
- Hvis vi kun har en observation kan vi da se fordelingen?
- men hvis vi har  $n$  observationer, så har vi en *stikprøve* (sample)

$$\{x_1, x_2, \dots, x_n\}$$

og da kan vi begynde at “se” fordelingen.

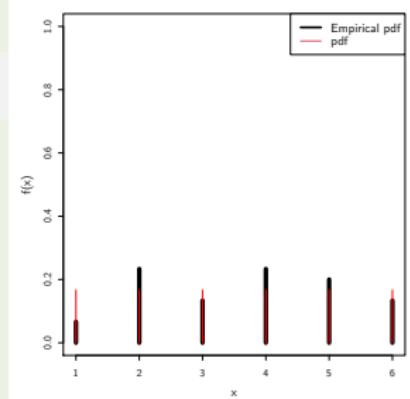
# Eksempel: Simuler $n$ kast med en fair terning

```
## Simuler en fair terning

## Antal simulerede realiseringer
n <- 30
## Træk uafhængigt fra mængden (1,2,3,4,5,6) med ens sandsynlighed
xFair <- sample(1:6, size=n, replace=TRUE)
## Tæl antallet af hvert udfald
table(xFair)

## Plot den empiriske tæthedsfunktion (pdf), altså et density histogram
plot(table(xFair)/n, ylim=c(0,1), lwd=10, xlab="x", ylab="f(x)")
## Tilføj den rigtige tæthedsfunktion til plottet
lines(rep(1/6,6), type="h", lwd=3, col="red")
## legend
legend("topright", c("Empirical pdf","pdf"), lty=1, col=c(1,2), lwd=c(5,2))
```

```
## Eller bare med
hist(xFair, breaks=seq(0.5,6.5,by=1), prob=TRUE)
```

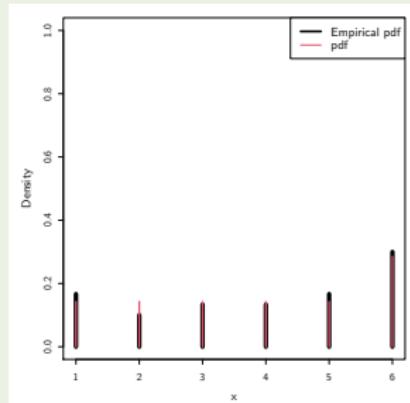


# Eksempel: Simuler $n$ kast med en ikke-fair terning

```
## Simuler en ikke-fair terning

## Antal simulerede realiseringer
n <- 30
## Træk uafhændigt fra mængden (1,2,3,4,5,6) med højere sandsynlighed for en sekser
xUnfair <- sample(1:6, size=n, replace=TRUE, prob=c(rep(1/7,5),2/7))
## Tæl antallet af hvert udfald
table(xUnfair)

## Plot den empiriske tæthedsfunktion
plot(table(xUnfair)/n, lwd=10, ylim=c(0,1), xlab="x", ylab="Density")
## Tilføj den rigtige tæthedsfunktion
lines(c(rep(1/7,5),2/7), lwd=4, type="h", col=2)
## En legend
legend("topright", c("Empirical pdf","pdf"), lty=1, col=c(1,2), lwd=c(5,2))
```



# Fordelingsfunktion (distribution function eller cumulative density function (cdf))

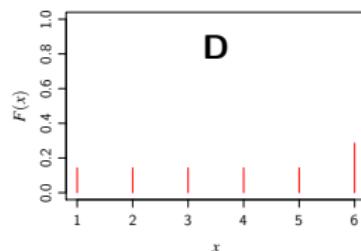
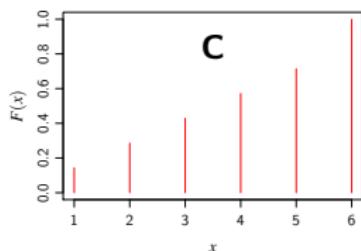
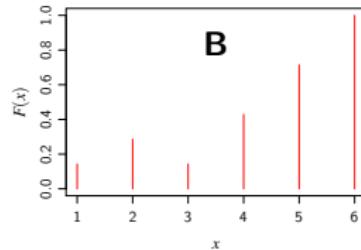
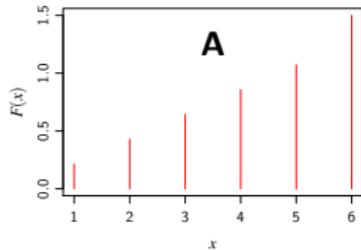
Def. 2.9: **Fordelingsfunktionen** (cdf) er tæthedsfunktionen akkumuleret

$$F(x) = P(X \leq x) = \sum_{j \text{ hvor } x_j \leq x} f(x_j)$$

Der gælder for en fordelingsfunktion (cdf):

- Den er en 'ikke-aftagende' funktion
- Den akkumuleres (assymtotisk) til 1 når  $x \rightarrow \infty$

## Spørgsmål: Fordelingsfunktion (cdf) (socrative.com, room: PBAC)



Hvilket et af ovenstående plots kan være en fordelingsfunktion (akkumuleret tæthedsfunktion, cdf)?

A, B, C eller D?

## Eksempel: Fair terning

- Lad  $X$  repræsentere værdien af et kast med en fair terning
- Udregn sandsynligheden for at få et udfald under 3:

$$\begin{aligned} P(X < 3) &= P(X \leq 2) \\ &= F(2) \text{ fordelingsfunktionen} \\ &= P(X = 1) + P(X = 2) \\ &= f(1) + f(2) \text{ tæthedsfunktionen} \\ &= \frac{1}{6} + \frac{1}{6} = \frac{1}{3} \end{aligned}$$

## Eksempel: Fair terning

- Udregn sandsynligheden for at få et udfald over eller lig 3:

$$\begin{aligned}P(X \geq 3) &= 1 - P(X \leq 2) \\&= 1 - F(2) \text{ fordelingsfunktionen} \\&= 1 - \frac{1}{3} = \frac{2}{3}\end{aligned}$$

# Spørgsmål: Sandsynlighed med fordelingsfunktionen

(socrative.com, room: PBAC)

Hvilket af følgende giver sandsynligheden  $P(X < 4)$  for terningeslag?

- A:  $F(2)$
- B:  $F(3)$
- C:  $F(4)$
- D:  $1 - F(2)$
- E:  $1 - F(3)$

# Spørgsmål: Sandsynlighed med fordelingsfunktionen

(socrative.com, room: PBAC)

Hvilket af følgende giver sandsynligheden  $P(X \geq 4)$  for terningeslag?

- A:  $F(2)$
- B:  $F(3)$
- C:  $F(4)$
- D:  $1 - F(2)$
- E:  $1 - F(3)$

# Konkrete statistiske fordelinger

- Der findes en række statistiske fordelinger, som kan bruges til at beskrive og analysere forskellige problemstillinger med
- I dag er det diskrete fordelinger:
  - Binomialfordelingen
  - Den hypergeometriske fordeling
  - Poissonfordelingen

## Binomialfordelingen

- Lad  $X$  repræsentere antal succeser efter  $n$  gentagelser af handling (eksperiment) med to udfald (succes eller ikke-succes)
- $X$  følger **binomialfordelingen**

$$X \sim B(n, p)$$

med parametre:

- $n$  antal gentagelser
- $p$  sandsynligheden for succes i hver gentagelse

- 
- Tæthedsfunktion: Sandsynlighed for  $x$  antal succeser

$$f(x; n, p) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

---

# Eksempel: Binomialfordelingen

Eksempel: Sandsynlighed for 2 plat ved 5 plat-eller-krone kast med mønt

$$f(2; 5, 0.5) = P(X = 2) = \binom{5}{2} 0.5^2 (1 - 0.5)^{5-2} = 0.3125$$

## Sandsynlighed for 2 plat (success) i 5 kast med mønt

## Slå op med binomial tæthedsfunktion

```
dbinom(x=2, size=5, prob=0.5)
```

## Binomialfordeling simuleringseksempel i R med terning:

```
## Fair terning eksempel

## Antal simulerede realiseringer
n <- 30
## Træk uafhængigt fra mængden (1,2,3,4,5,6) med ens sandsynlighed
xFair <- sample(1:6, size=n, replace=TRUE)
## Tæl sammen hvor mange seksere
sum(xFair == 6)

## Lav tilsvarende med rbinom()
rbinom(n=1, size=30, prob=1/6)
```

## Hypergeometrisk fordeling

- $X$  er igen antal succeser, men nu er det *uden tilbagelægning ved gentagelsen*
- $X$  følger en **hypergeometrisk fordeling**

$$X \sim H(n, a, N)$$

med parametrene

- $n$  er antallet af trækninger
- $a$  er antallet af succeser i populationen
- $N$  elementer store population

- Tæthedsfunktion: Sandsynlighed for at få  $x$  succeser

$$f(x; n, a, N) = P(X = x) = \frac{\binom{a}{x} \binom{N-a}{n-x}}{\binom{N}{n}}$$

# Binomial vs. hypergeometrisk

- Binomialfordelingen anvendes også for at analysere stikprøver med tilbagelægning (tænk på en terningekast)
- Når man vil analysere stikprøver uden tilbagelægning anvendes den hypergeometriske fordeling (tænk på træk fra en hat)

PAUSE

R navn	Betegnelse
binom	binomial
hyper	hypergeometrisk

- d** Tæthedsfunktion  $f(x)$  (probability density function).
- p** Fordelingsfunktion  $F(x)$  (cumulative distribution function).
- r** Tilfældige tal fra den anførte fordeling. (Forelæsning 10)
- q** Fraktil (quantile) i fordeling.

## Eksempel binomialfordelt:

Find

$$P(X \leq 5) = F(5; 10, 1/6)$$

```
## Binomial fordelingsfunktion (cdf)

## Sandsynlighed for at få 5 eller færre succeser i 10 kast med terning
pbinom(q=5, size=10, prob=1/6)
## Få hjælpen med
?pbinom
```

Husk at hjælp til funktion mm. fåes ved at sætte '?' foran navnet.

# Eksempel 1

Du skal afholde et selskab med 12 personer ialt. Du vil servere en frugt til hver person. Antag at der er 70% sandsynlighed for at en frugt du køber er god. Du skal købe ind og vælger at købe 20 frugter.

*Hvad er sandsynligheden for at der er mindst en god frugt til hver person?*

- **Step 1)** Hvad skal repræsenteres:  $X$  er
- **Step 2)** Fordeling:  $X$  følger A: binomial, B: hypergeometrisk?
- **Step 3)** Hvilken sandsynlighed:  $P(X \geq 12)$
- **Step 4)**
  - Hvad er antal trækninger?
  - Hvad er succes-sandsynligheden?

## Eksempel 2

Du spiller kortspillet casino med din ven. I skal til at dele ud til anden sidste runde, dvs. der er 16 kort tilbage. Der er blevet spillet 8 spar allerede, dvs. der er 5 spar tilbage i bunken. En hånd er på 4 kort.

*Hvad er sandsynligheden for at du får en hånd med udelukkende spar?*

- **Step 1)** Hvad skal repræsenteres:  $X$  er
- **Step 2)** Hvilken fordeling:  $X$  følger
- **Step 3)** Hvilken sandsynlighed:  $P(X = ?)$

- A:  $P(X = 0)$       B:  $1 - P(X \leq 4)$       C:  $P(X = 4)$       D:  $P(X \leq 4)$

- **Step 4)**

- Hvad er antal trækninger?
- Hvor mange succeser er der?
- Hvor mange er der i alt?

# Poissonfordelingen

- Poissonfordelingen anvendes ofte som en fordeling (model) for tælletal, hvor der ikke er nogen naturlig øvre grænse
- Poissonfordelingen karakteriseres ved en intensitet, dvs. på formen antal/enhed
- Parameteren  $\lambda$  angiver intensiteten
- $\lambda$  er typisk hændelser per tidsinterval
- Intervallerne mellem hændelserne er uafhængige, dvs. processen er hukommelsesløs

## Poissonfordelingen

- $X$  følger **Poissonfordelingen**

$$X \sim P(\lambda)$$

- Parameteren  $\lambda$  angiver intensiteten
- Tæthedsfunktion: Sandsynligheden for  $x$  antal i intervallet

$$f(x) = P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

## Eksempel 3.1: Poissonfordelingen

Det antages, at der i gennemsnit bliver indlagt 0.3 patienter pr. dag på københavnske hospitaler som følge af luftforurening.

*Hvad er sandsynligheden for at der på en vilkårlig dag bliver indlagt højst 2 patienter som følge af luftforurening?*

- **Step 1)** Hvad skal repræsenteres:  $X$  er
- **Step 2)** Hvilken fordeling:  $X$  følger
- **Step 3)** Hvilken sandsynlighed:  $P(X = 2)$
- **Step 4)** Hvad er raten:

## Eksempel 3.4: Skalering af intensiteten i Poissonfordeling

Hvad er sandsynligheden for at der i en periode på 3 dage bliver indlagt præcis 1 patient?

- **Step 1)** Hvad skal repræsenteres:
- **Step 2)** Hvilken fordeling følger  $X^{3\text{dage}}$ :
- **Step 3)** Hvilken sandsynlighed:  $P(X^{3\text{dage}} = 1)$
- **Step 4)** Skaler raten

# Middelværdi (mean) og forventningsværdi (expectation)

Definition: Middelværdi af stokastisk variabel

$$\mu = E(X) = \sum_{\text{alle } x} x f(x)$$

- Populationsgennemsnittet (det “rigtige gennemsnit”)
- Fortæller hvor “midten” af tæthedsfunktion for  $X$  er

# Eksempel: Middelværdi

Middelværdi af et terningekast

$$\begin{aligned}\mu = E(X) &= \sum_{x=1}^6 x f(x) \\ &= \sum_{x=1}^6 x \frac{1}{6} \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= 3.5\end{aligned}$$

# Eksempel: Simuler terningekast og beregn gennemsnit

```
## Simuler stikprøve af en fair terning og beregn gennemsnit
## Antal simulerede realiseringer (stikprøve på n elementer)
n <- 30
## Træk uafhængigt fra mængden (1,2,3,4,5,6) med ens sandsynlighed
xFair <- sample(1:6, size=n, replace=TRUE)
## Udregn stikprøvegennemsnit (sample mean)
mean(xFair)
```

# Spørgsmål om stikprøvevarians (socrative.com, room: PBAC)

Hvad sker der generelt med gennemsnittet af en stikprøve *når man får flere observationer?*

- A: Det er uafhængigt af antal observationer
- B: Det kommer generelt længere væk fra middelværdien
- C: Det kommer generelt tættere på middelværdien

# Varians (variance)

Definition: Varians af stokastisk variabel

$$\sigma^2 = \text{Var}(X) = \sum_{\text{alle } x} (x - \mu)^2 f(x)$$

- Et mål for spredningen
- Populationsvariansen
- Den “rigtige spredning“ af  $X$  tæthedsfunktion

# Eksempel: Varians

## Varians af terningekast

$$\begin{aligned}\sigma^2 &= E[(X - \mu)^2] = \\ &= (1 - 3.5)^2 \cdot \frac{1}{6} + (2 - 3.5)^2 \cdot \frac{1}{6} + (3 - 3.5)^2 \cdot \frac{1}{6} \\ &\quad + (4 - 3.5)^2 \cdot \frac{1}{6} + (5 - 3.5)^2 \cdot \frac{1}{6} + (6 - 3.5)^2 \cdot \frac{1}{6} \\ &\approx 2.92\end{aligned}$$

# Eksempel: Varians

```
## Simuler stikprøve med udfald af en fair terning og beregn stikprøvevarians

## Antal simulerede realiseringer
n <- 30
## Træk uafhængigt fra mængden (1,2,3,4,5,6) med ens sandsynlighed
xFair <- sample(1:6, size=n, replace=TRUE)

## Udregn empirisk varians (sample variance, læg mærke til
## at i R hedder funktionen "var")
var(xFair)
```

# Middelværdi og varians for de diskrete fordelinger

Fordeling	Middelværdi	Varians
Binomialfordelingen	$\mu = n \cdot p$	$\sigma^2 = n \cdot p \cdot (1 - p)$
Hypergeometrisk	$\mu = n \cdot \frac{a}{N}$	$\sigma^2 = \frac{na \cdot (N-a) \cdot (N-n)}{N^2 \cdot (N-1)}$
Poissonfordelingen	$\mu = \lambda$	$\sigma^2 = \lambda$

# Eksempel: Forskel på stikprøvegennemsnit (sample mean) og middelværdi (mean, dvs. populationsgennemsnittet)

Se stikprøvegennemsnittet i forhold til middelværdien:

```
## Simuler en binomialfordeling, terninge eksempel

## Gentag 10 gange: Tæl sammen for mange seksere på 30 slag
antalSeksere <- rbinom(n=10, size=30, prob=1/6)

## Endelig kan vi se på stikprøvegennemsnittet (sample mean)
mean(rbinom(n=10, size=30, prob=1/6))
## versus Middelværdien (mean)
n * 1/6
```

# Introduktion til Statistik

## Forelæsning 3: Kontinuerte fordelinger

Peder Bacher

DTU Compute, Dynamiske Systemer  
Bygning 303B, Rum 010  
Danmarks Tekniske Universitet  
2800 Lyngby – Danmark  
e-mail: pbac@dtu.dk

Forår 2021

# Kapitel 2: Kontinuerte fordelinger

Grundlæggende koncepter:

- Tæthedsfunktion:  $f(x)$  (*pdf*)
- Fordelingsfunktion:  $F(x) = P(X \leq x)$  (*cdf*)
- Middelværdi ( $\mu$ ) og varians ( $\sigma^2$ )
- Regneregler for stokastiske variabler (lineære funktioner)

Specifikke fordelinger:

- Uniform
- Normal
- Log-Normal
- Eksponential

Funktioner af normalfordeling (afsn. 2.10) (introduceres først i de næste uger):

- $t$ -fordelingen,  $\chi^2$ -fordelingen (*Chi-i-anden*) og  $F$ -fordelingen

# Chapter 2: Continuous Distributions

## General concepts:

- Density function:  $f(x)$  (*pdf*)
- Distribution:  $F(x) = P(X \leq x)$  (*cdf*)
- Mean ( $\mu$ ) and variance ( $\sigma^2$ )
- Calculation rules for random variables (linear functions)

## Specific distributions:

- Uniform
- Normal
- Log-Normal
- Exponential

## Funktions of normaldist. (Sec. 2.10) (introduced in the coming weeks):

- $t$ -distribution,  $\chi^2$ -distribution (*Chi-square*) og  $F$ -distribution

# Oversigt

## 1 Kontinuerte Stokastiske variable og fordelinger

- Tæthedsfunktion
- Fordelingsfunktion
- Middelværdi af en kontinuert stokastisk variabel
- Varians af en kontinuert stokastisk variabel

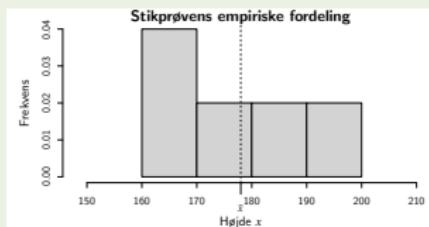
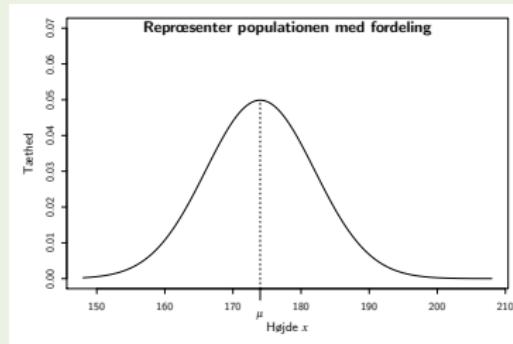
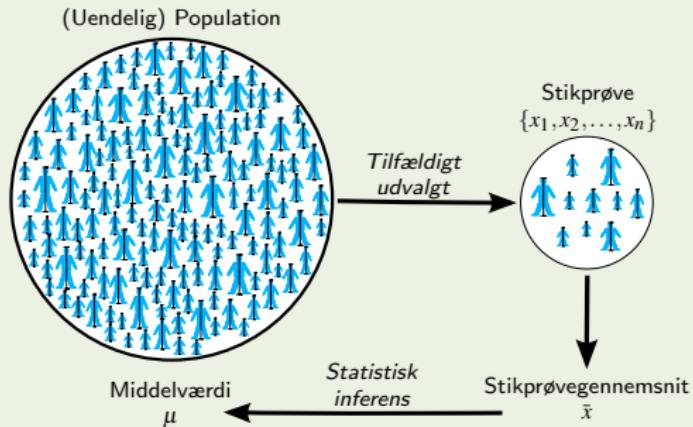
## 2 Konkrete Statistiske fordelinger

- Kontinuerte fordelinger i R
- Uniform fordeling
- Normalfordelingen
- Log-Normalfordelingen

## 3 Eksponentialfordelingen

## 4 Regneregler for middelværdi og varians

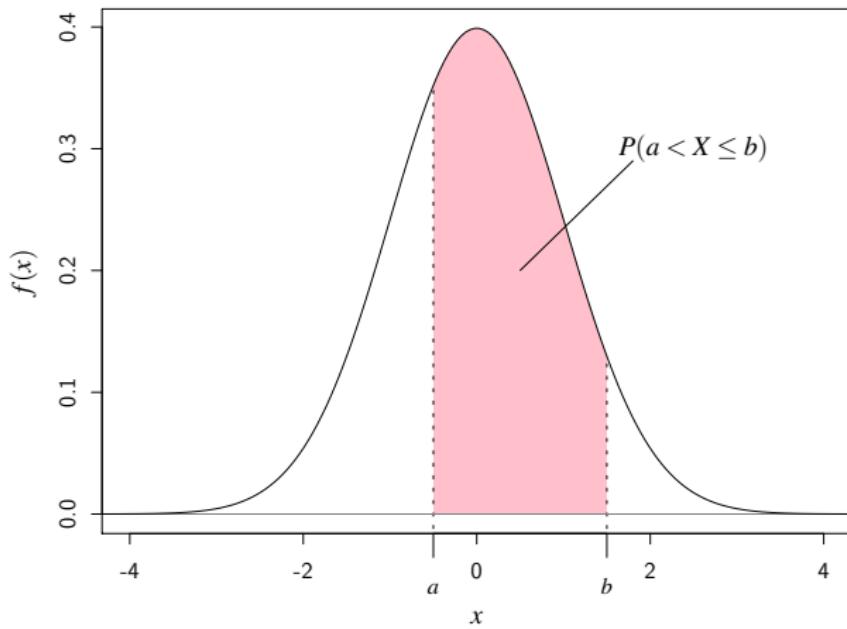
# Eksempel: Population og fordeling



# Tæthedsfunktion (probability density function (pdf))

- **Tæthedsfunktionen** for en stokastisk variabel betegnes ved  $f(x)$
- For kontinuerte variable svarer tætheden ikke til sandsynligheden, dvs.  $f(x) \neq P(X = x)$
- Et godt plot af  $f(x)$  er et histogram (kontinuert)

## Tæthedsfunktion for en kontinuert variabel



# Tæthedsfunktion for en kontinuert variabel

- Der gælder:

- Ingen negative værdier

$$f(x) \geq 0 \quad \text{for alle mulige } x$$

- Areal under kurven er een

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

# Fordelingsfunktion (distribution function eller cumulative density function (cdf))

- **Fordelingsfunktion** for en kontinuert stokastisk variabel betegnes ved

$$F(x)$$

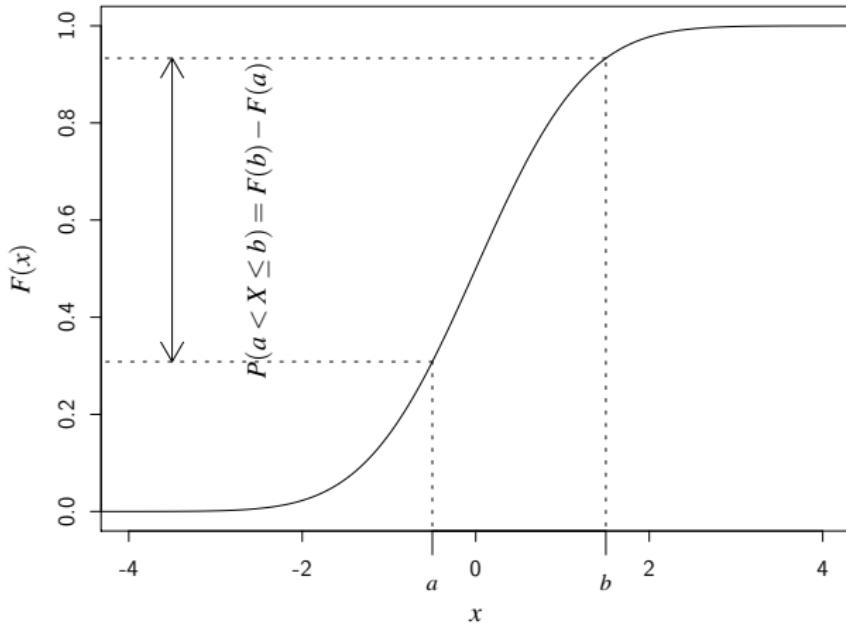
- **Fordelingsfunktionen** svarer til den kumulerede tæthedsfunktion ved

$$F(x) = P(X \leq x)$$

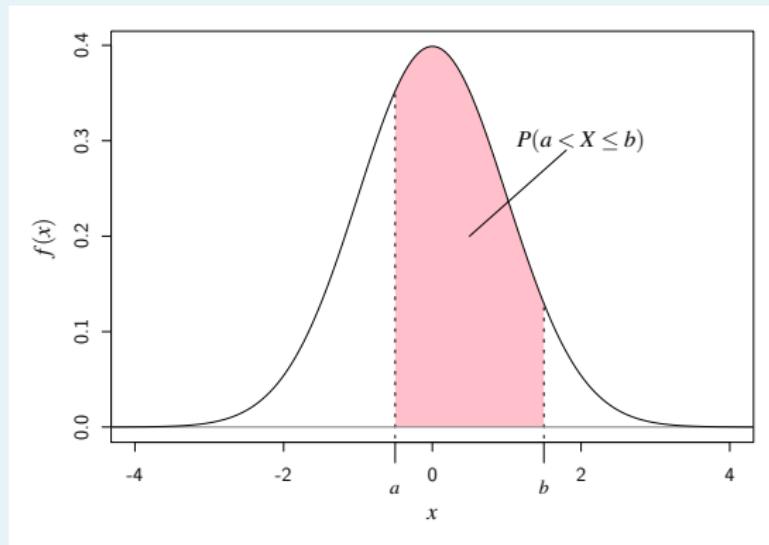
$$F(x) = \int_{-\infty}^x f(u)du$$

$$f(x) = F'(x)$$

# Fordelingsfunktion (distribution function eller cumulative density function (cdf))



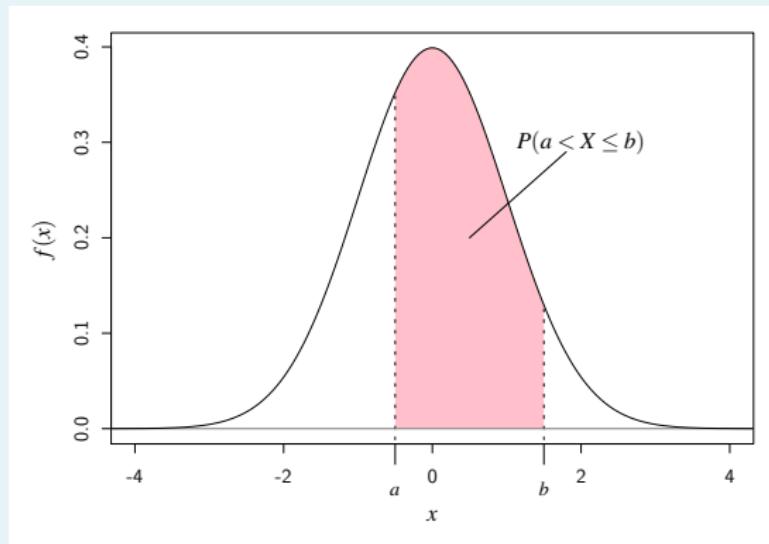
# Spørgsmål om sandsynligheder (socrative.com, room: PBAC)



Hvilket udtryk giver den markerede sandsynlighed? (arealet)

- A:  $\int_{-\infty}^b f(x)dx$       B:  $1 - \int_a^b f(x)dx$       C:  $\int_a^b f(x)dx$       D:  $1 - \int_a^{\infty} f(x)dx$

# Spørgsmål om sandsynligheder (socrative.com, room: PBAC)



Hvordan kan vi nemmest udregne den markerede sandsynlighed?

- A:  $\int_a^b f(x)dx$       B:  $\int_a^b F(x)dx$       C:  $f(b) - f(a)$       D:  $F(b) - F(a)$

# Middelværdi (mean) af en kontinuert stokastisk variabel

**Middelværdien** af en kontinuert stokastisk variabel

$$\mu = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Sammenlign med den diskrete definition:  $\mu = \sum_{\text{alle } x} x \cdot f(x)$

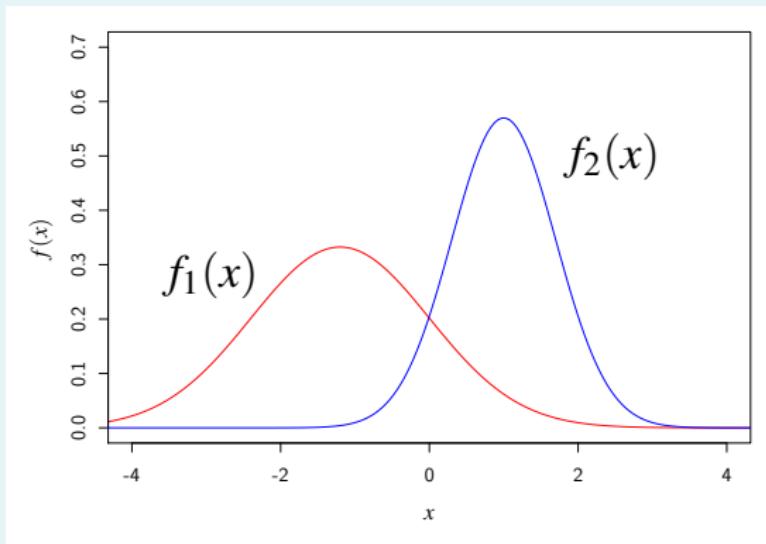
# Varians af en kontinuert stokastisk variabel

**Variansen** af en kontinuert stokastisk variabel:

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx$$

Sammenlign med den diskrete definition:  $\sigma^2 = \sum_{\text{alle } x} (x - \mu)^2 \cdot f(x)$

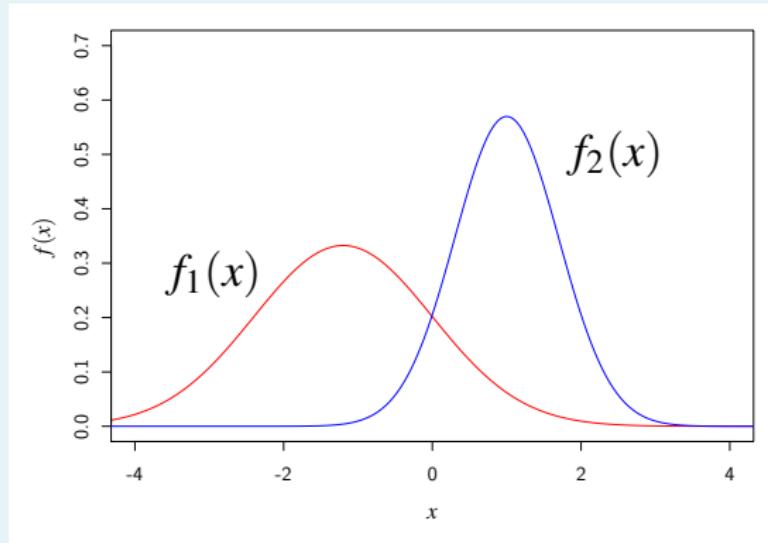
# Spørgsmål om middelværdi (socrative.com, room: PBAC)



Hvilken pdf har størst middelværdi (begge er symmetriske)?

- A:  $\mu_1 < \mu_2$
- B:  $\mu_1 > \mu_2$
- C:  $\mu_1 = \mu_2$
- D: Kan ikke afgøres

## Spørgsmål om spredning (socrative.com, room: PBAC)



Hvilken pdf har størst standard afvigelse (begge er symmetriske)?

- A:  $\sigma_1 < \sigma_2$
- B:  $\sigma_1 > \sigma_2$
- C:  $\sigma_1 = \sigma_2$
- D: Kan ikke afgøres

# Konkrete statistiske fordelinger

Der findes en række statistiske fordelinger, som kan bruges til at beskrive og analysere forskellige problemstillinger med

- Følgende kontinuerte fordelinger:
  - Uniform fordeling
  - Normalfordelingen
  - Log-normalfordelingen
  - Eksponentialfordelingen

# Kontinuerte fordelinger i R

R	Betegnelse
norm	Normalfordelingen
unif	Uniform fordeling
lnorm	Log-normalfordelingen
exp	Eksponentialfordelingen

- d** Tæthedsfunktion  $f(x)$  (probability density function).
- p** Fordelingsfunktion  $F(x)$  (cumulative distribution function).
- q** Fraktil (quantile) i fordeling.
- r** Tilfældige tal fra fordelingen.

# Uniform fordeling

Skrivemåde:

$X \sim U(\alpha, \beta)$  (Læses:  $X$  følger en uniform fordeling med parametre  $\alpha$  og  $\beta$ )

Tæthedsfunktion:

$$f(x) = \frac{1}{\beta - \alpha}$$

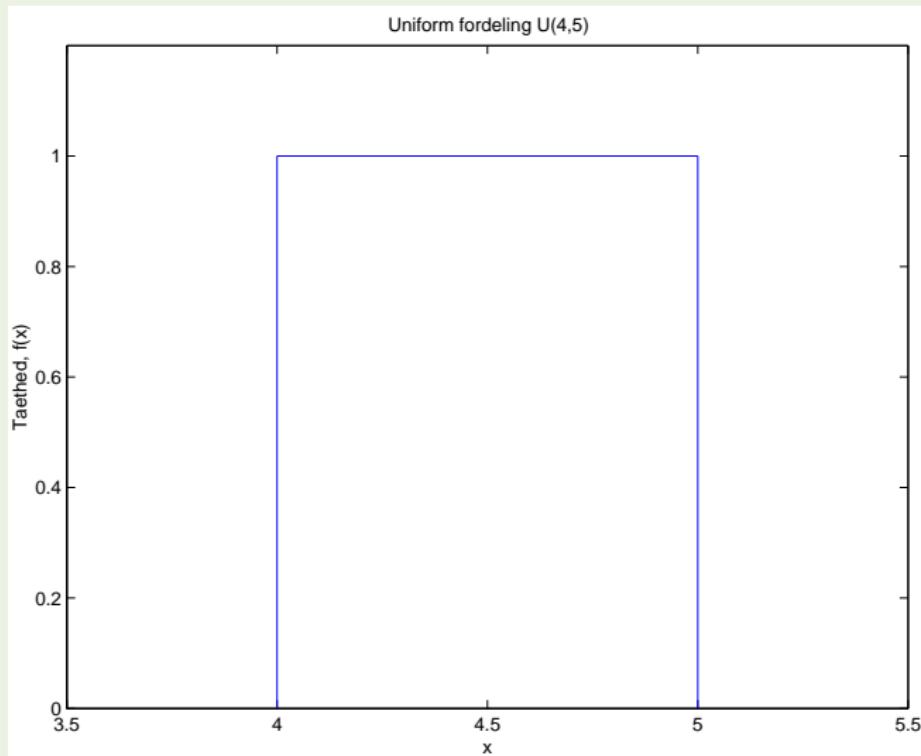
Middelværdi:

$$\mu = \frac{\alpha + \beta}{2}$$

Varians:

$$\sigma^2 = \frac{1}{12}(\beta - \alpha)^2$$

# Eksempel: Uniform fordeling



## Spørgsmål: Uniform fordeling (socrative.com, room: PBAC)

Medarbejdere på en arbejdsplads ankommer mellem klokken 8:00 og 8:30. Det antages, at ankomsttiden kan beskrives ved en uniform fordeling.

*Hvad er sandsynligheden for at en tilfældig udvalgt medarbejder ankommer mellem 8:20 og 8:30?*

- A: 1/2
- B: 1/6
- C: 1/3
- D: 0

## Spørgsmål: Uniform fordeling (socrative.com, room: PBAC)

Medarbejdere på en arbejdsplads ankommer mellem klokken 8:00 og 8:30. Det antages, at ankomsttiden kan beskrives ved en uniform fordeling.

*Hvad er sandsynligheden for at en tilfældig udvalgt medarbejder ankommer efter 8:30?*

- A: 1/2
- B: 1/6
- C: 1/3
- D: 0

# Normalfordelingen

Skrivemåde:

$$X \sim N(\mu, \sigma^2)$$

Tæthedsfunktion:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

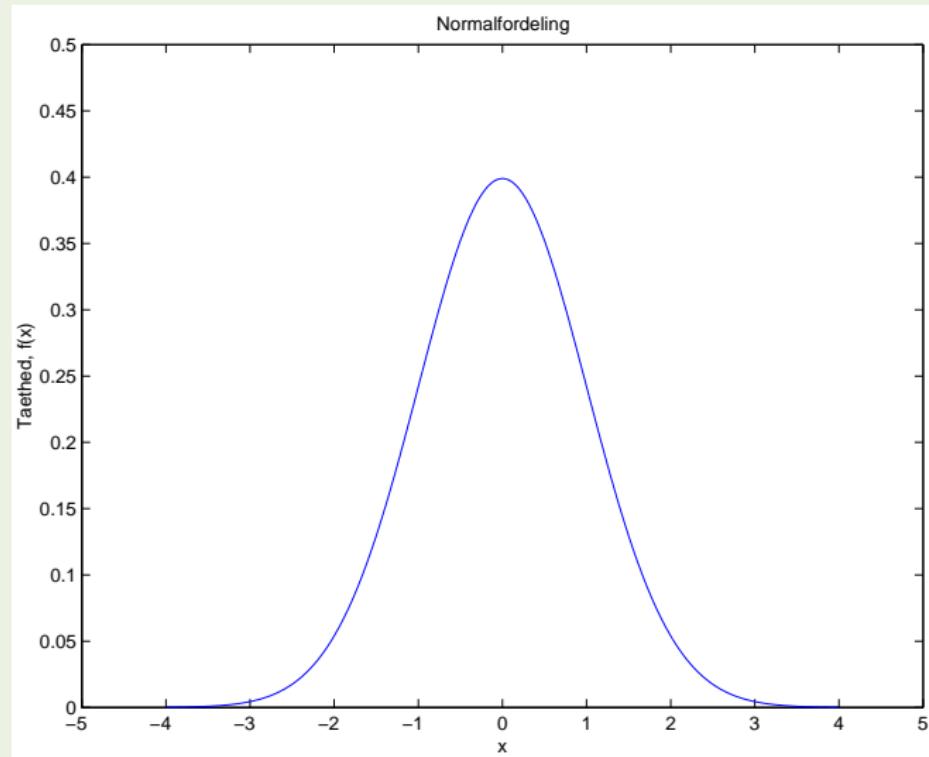
Middelværdi:

$$\mu = \mu$$

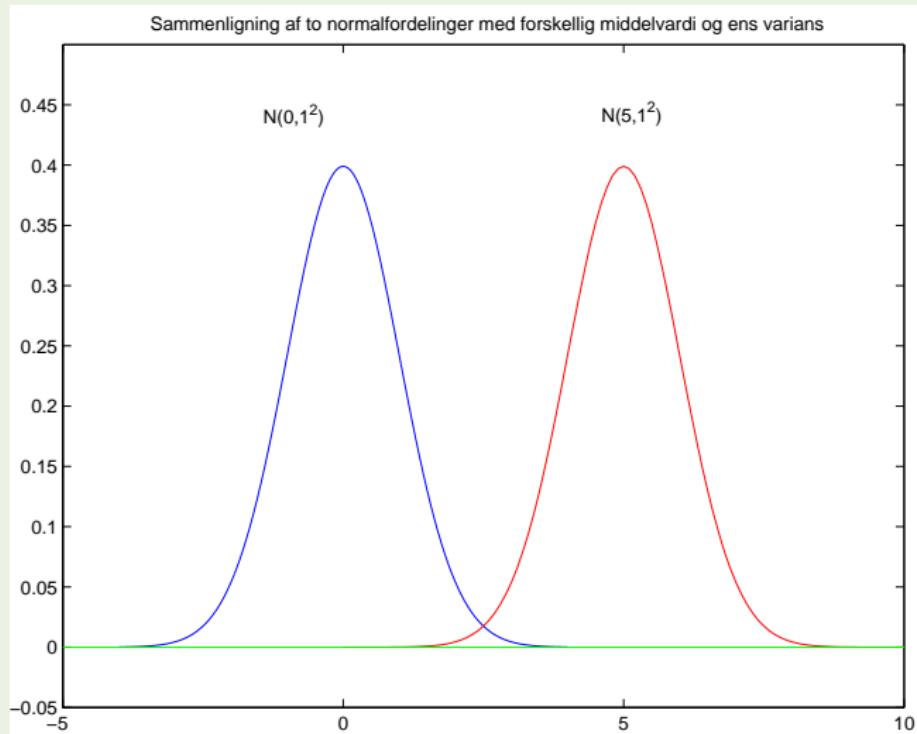
Varians:

$$\sigma^2 = \sigma^2$$

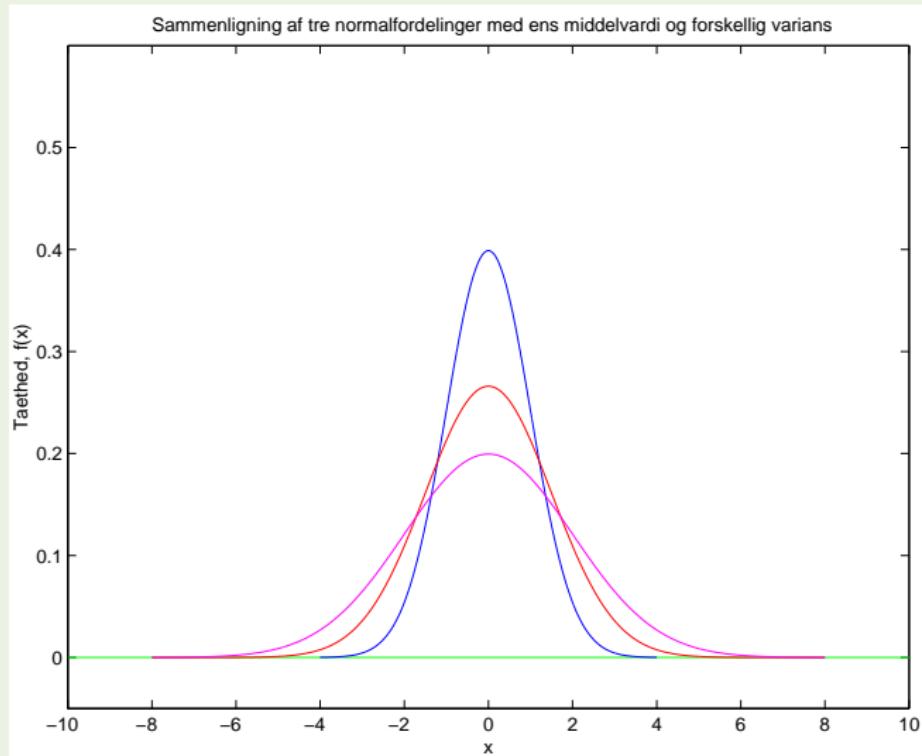
# Eksempel: Normalfordelingen



# Eksempel: Normalfordelingen



# Eksempel: Normalfordelingen



# Eksempel: Normalfordeling, sandsynligheder

Fordeling af vægt af rugbrød:

Antag at vægten af et rugbrød fra en produktionslinie kan beskrives med en normalfordeling

$$X \sim N(500, 10^2)$$

dvs. middelværdi  $\mu = 500$  gram og standardafvigelse  $\sigma = 10$  gram.

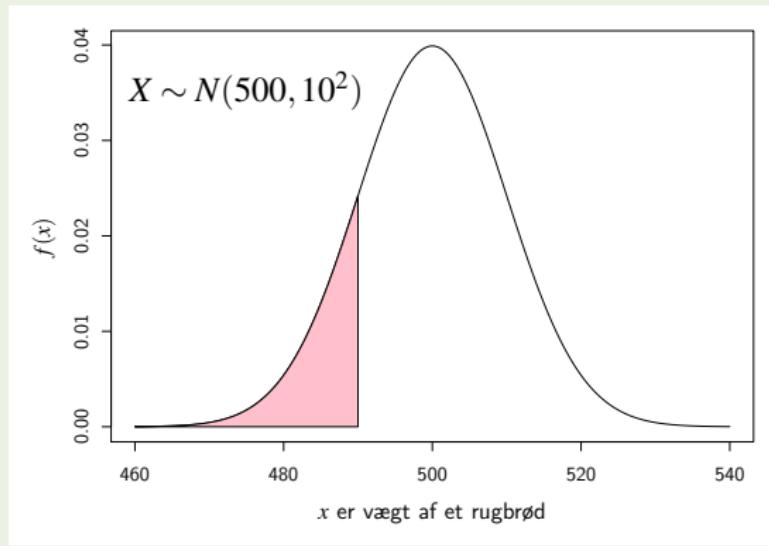
Vi vil måle vægten af ét tilfældigt udvalgt brød.

Spørgsmål:

1: Hvad er sandsynligheden for at brødet vejer under 490 g?

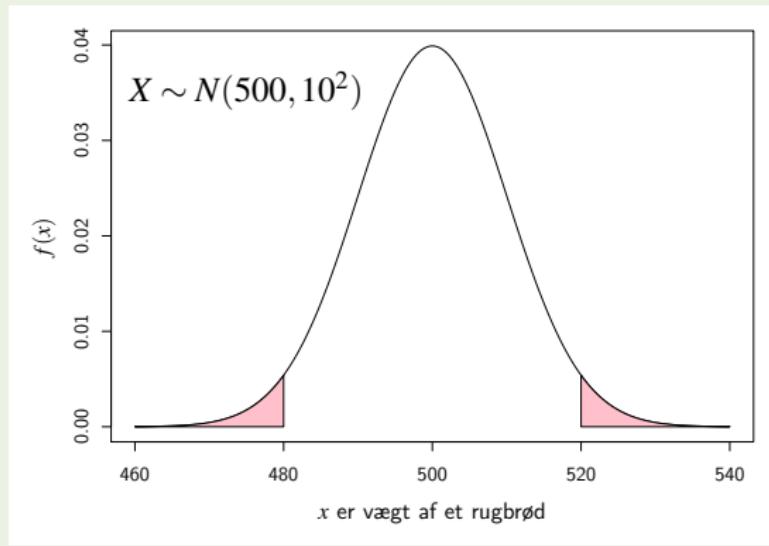
2: Hvad er sandsynligheden for at brødet vejer mere en 20 g forskelligt fra 500 g?

# Eksempel: Normalfordeling, spørgsmål 1



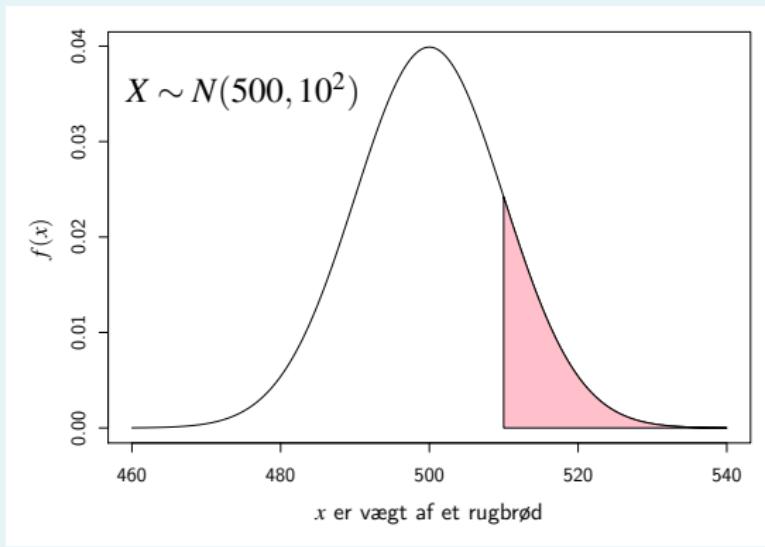
1: Hvad er sandsynligheden for at brødet vejer under 490 g?

## Eksempel: Normalfordeling, spørgsmål 2



1: Hvad er sandsynligheden for at brødet vejer mere end 20 g forskelligt fra 500 g?

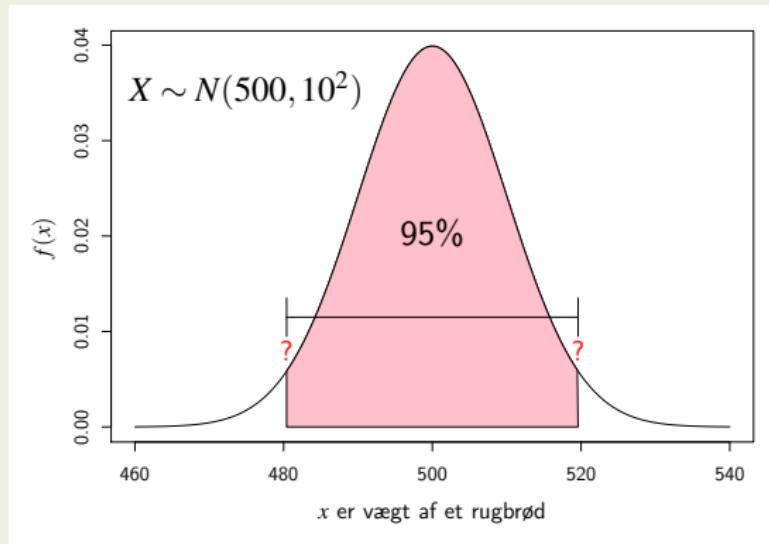
# Spørgsmål: Sandsynlighed i normalfordeling



Hvad er sandsynligheden for at rugbrødet vejer over 510 g?

- A:  $F(510)$
- B:  $1 - F(490)$
- C:  $1 - F(520)$
- D:  $1 - F(510)$

## Eksempel: Normalfordeling fraktiler



“Omvendt spørgsmål”: *Hvilket interval, symmetrisk om midten, dækker 95% af rugbrødene?*

# Standard normalfordelingen

## En standard normalfordeling

$$Z \sim N(0, 1^2)$$

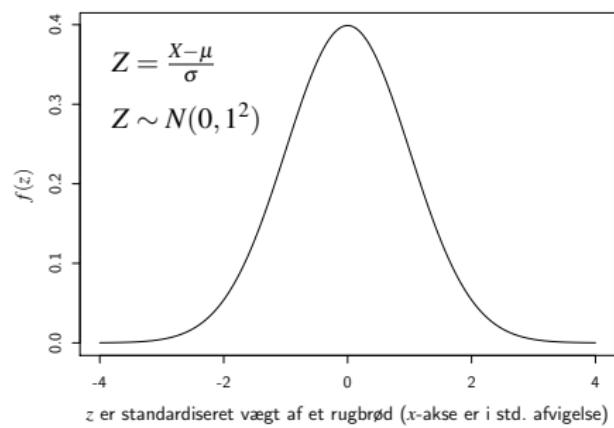
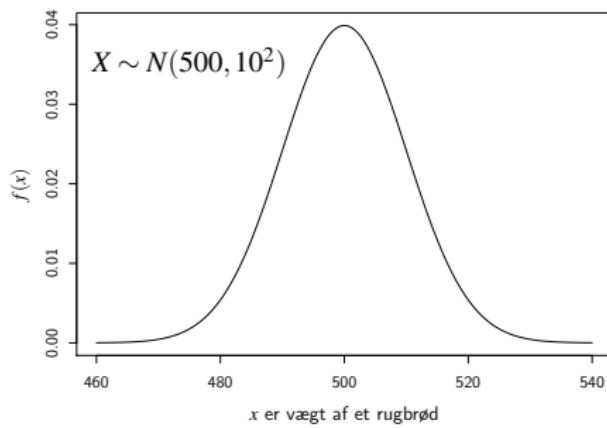
En normalfordeling med middelværdi 0 og varians 1.

## Standardisering

En vilkårlig normalfordelt variabel  $X \sim N(\mu, \sigma^2)$  kan standardiseres ved at beregne

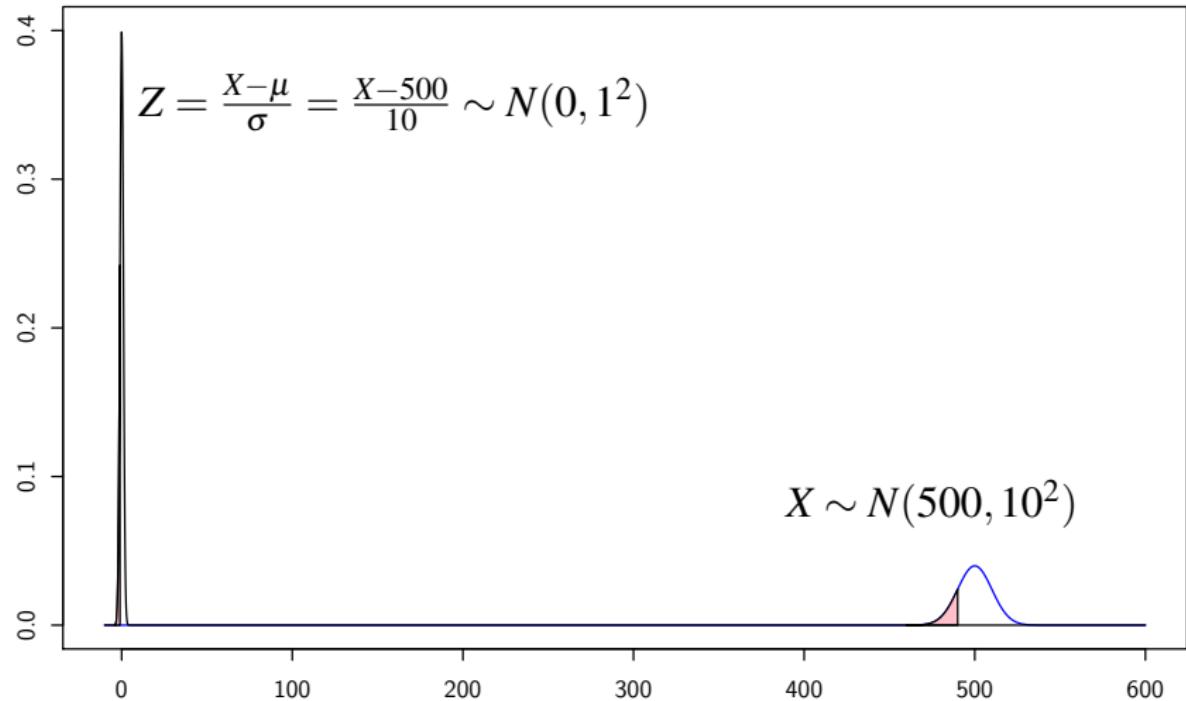
$$Z = \frac{X - \mu}{\sigma}$$

# Eksempel: Standard Normalfordeling

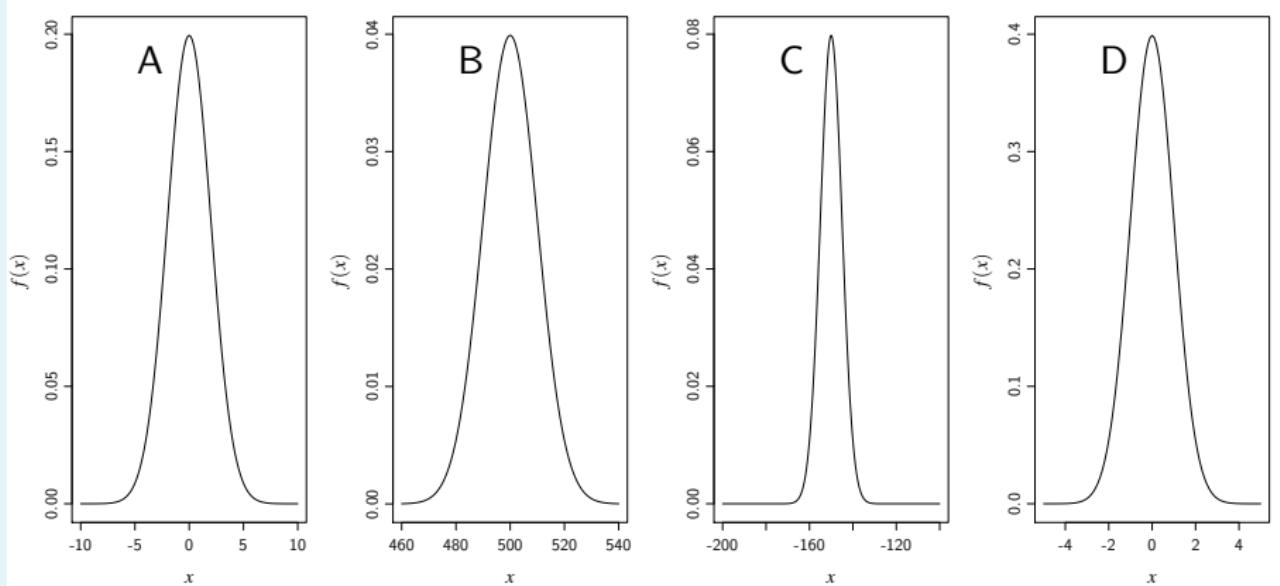


1: Hvad er sandsynligheden for at brødet vejer under 490 gram?

# Eksempel: Transformation til standard normalfordeling



# Eksempel: Transformation til standard normalfordeling



1: Hvilken af disse er standard normalfordelingens pdf?

# Log-Normalfordelingen

Skrivemåde:

$X \sim LN(\alpha, \beta^2)$  (Hvis  $X$  følger log-normal så følger  $\ln(X)$  normal)

Tæthedsfunktion:

$$f(x) = \begin{cases} \frac{1}{x\sqrt{2\pi}\beta} e^{-(\ln(x)-\alpha)^2/2\beta^2} & x > 0, \beta > 0 \\ 0 & \text{ellers} \end{cases}$$

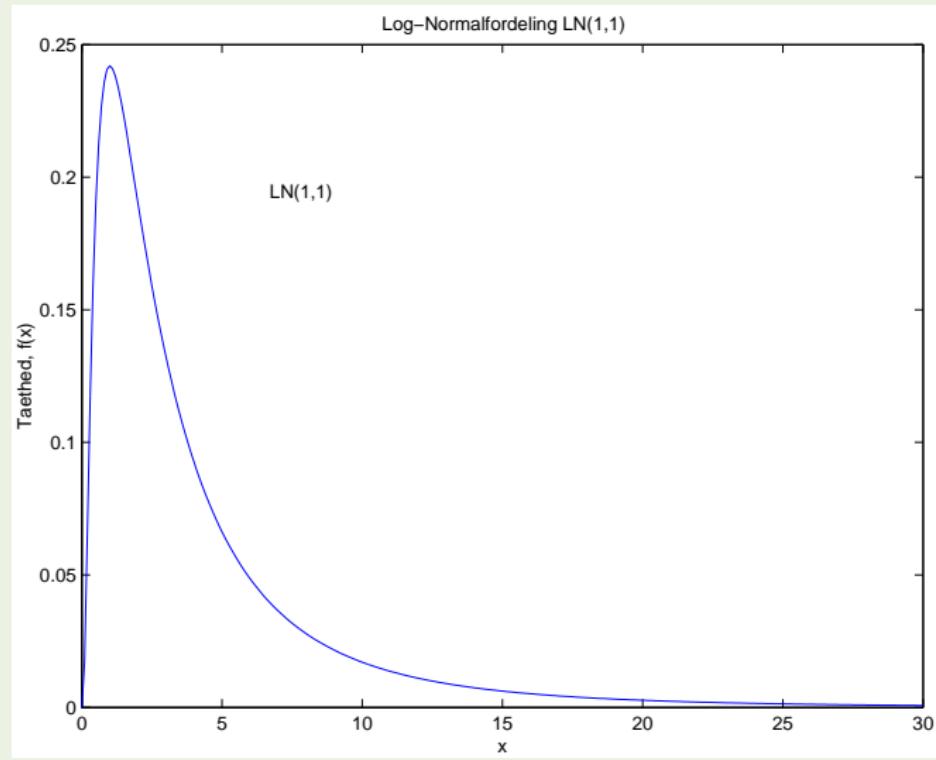
Middelværdi:

$$\mu = e^{\alpha + \beta^2/2}$$

Varians:

$$\sigma^2 = e^{2\alpha + \beta^2} (e^{\beta^2} - 1)$$

# Eksempel: Log-normalfordelingen



# Log-normalfordelingen

## Lognormal og Normalfordelingen:

En log-normalfordelt variabel  $Y \sim LN(\alpha, \beta^2)$ , kan transformeres til en standard normalfordelt variabel  $Z$  ved

$$Z = \frac{\ln(Y) - \alpha}{\beta}$$

dvs.

$$Z \sim N(0, 1^2)$$

# Eksponentialfordelingen

Skrivemåde:

$$X \sim Exp(\lambda)$$

Tæthedsfunktionen

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{ellers} \end{cases}$$

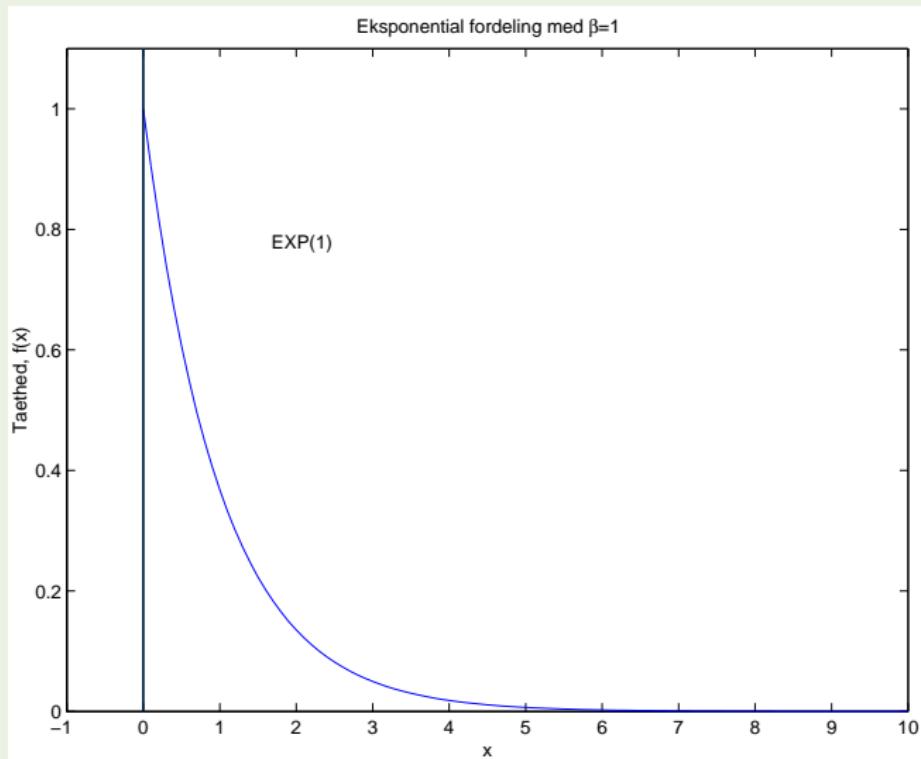
Middelværdi

$$\mu = \frac{1}{\lambda}$$

Varians

$$\sigma^2 = \frac{1}{\lambda^2}$$

# Eksempel: Eksponentialfordelingen



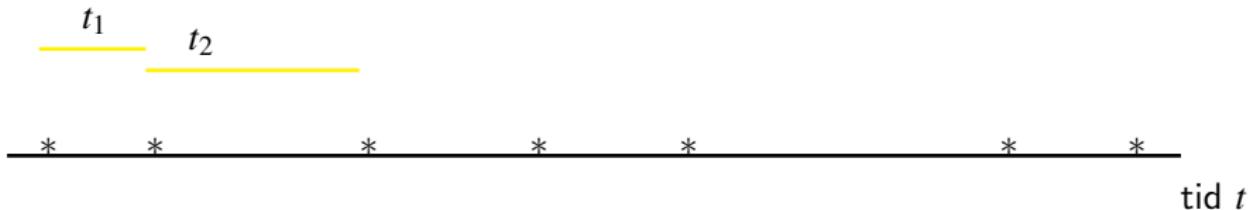
# Eksponentialfordelingen

- Eksponentialfordelingen er et special tilfælde af Gammafordelingen
- Eksponentialfordelingen anvendes f.eks. til at beskrive levetider og ventetider
- Eksponentialfordelingen kan bruges til at beskrive (vente)tiden mellem hændelser i poissonproces

# Sammenhæng mellem eksponential- og poissonfordelingen

Poisson: Diskrete hændelser pr. enhed

Eksponential: Kontinuert afstand mellem hændelser



# Eksempel: Eksponentielfordeling

## Kø-model - poissonproces

Tiden mellem kundeankomster på et posthus er eksponentielfordelt med middelværdi  $\mu = 2$  minutter, dvs.  $\lambda = \frac{1}{\mu} = \frac{1}{2} \frac{1}{\text{min}}$  (skaleret  $\lambda_{2\text{min}} = 1 \frac{1}{2\text{min}}$ ).

## Spørgsmål:

*En kunde er netop ankommet. Beregn sandsynligheden for at der ikke kommer flere kunder indefor en periode på 2 minutter vha. poissonfordelingen*

## Svar:

# Regneregler for lineær funktion af et $X$

Hvis:

- $X$  er en stokastisk variabel
- Vi antager at  $a$  og  $b$  er konstanter

Da gælder (gælder BÅDE kontinuert og diskret):

Middelværdi-regel:

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$

Varians-regel:

$$\mathbb{V}(aX + b) = a^2 \mathbb{V}(X)$$

# Eksempel: Regneregler for lineær funktion af et $X$

$X$  er en stokastisk variabel

En stokastisk variabel  $X$  har middelværdi 4 og varians 6.

Spørgsmål:

Beregn middelværdi og varians for  $Y = -3X + 2$

Svar:

# Regneregler for lineær funktion af flere $X$ er

Hvis:

- $X_1, \dots, X_n$  er stokastiske variable

Da gælder (når de er uafhængige) (gælder BÅDE kontinuert og diskret):

Middelværdi-regel:

$$E(a_1X_1 + a_2X_2 + \dots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \dots + a_nE(X_n)$$

Varians-regel:

$$V(a_1X_1 + a_2X_2 + \dots + a_nX_n) = a_1^2V(X_1) + a_2^2V(X_2) + \dots + a_n^2V(X_n)$$

# Eksempel: Regneregler for lineær funktion af flere X'er

## Flypassager-planlægning

Vægten af en passagerer på fly på en strækning antages normalfordelt  $X \sim N(70, 10^2)$ .

Et fly, der kan tage 55 passagerer, må max. lastes med 4000 kg (kun passageres vægt betragtes som last).

## Spørgsmål:

Beregn sandsynligheden for at flyet bliver overlastet.

## Hvad er den samlede passagervægt $Y$ på en afgang?

- A:  $Y = 55 \cdot X$
- B:  $Y = \sum_{i=1}^{55} X_i$
- C:  $Y = 55 + X$
- D: Ej A,B eller C

## Eksempel: Regneregler 3

Hvad er den samlede passagervægt  $Y$  på en afgang?

$$Y = \sum_{i=1}^{55} X_i, \text{ hvor } X_i \sim N(70, 10^2)$$

Middelværdi og varians for  $Y$ :

Bruger normalfordeling for  $Y$ :

```
1-pnorm(4000, mean = 3850, sd = sqrt(5500))
```

[1] 0.022

# Eksempel: Regneregler 3 - FORKERT ANALYSE

Hvad er  $Y$ ?

I hvert fald IKKE:  $Y = 55 \cdot X$  !!!!!!

Middelværdi og varians for  $Y$ :

$$E(Y) = 55 \cdot 70 = 3850$$

$$V(Y) = 55^2 V(X) = 55^2 \cdot 100 = 550^2 = 302500$$

Bruger normalfordeling for  $Y$ :

```
1-pnorm(4000, mean = 3850, sd = 550)
```

```
[1] 0.39
```

Konsekvens af forkert beregning:

MANGE spilde penge for flyselskabet!!!

# Lineær kombination af normalfordelte stokastiske variabler er også normalfordelt

- Lineær kombination af normalfordelte stokastiske variabler er også normalfordelt
- Theorem 2.40: Let  $X_1, \dots, X_n$  be independent normal random variables, then any linear combination of  $X_1, \dots, X_n$  will follow a normal distribution, with mean and variance given in Theorem 2.56.

# Introduktion til Statistik

## Forelæsning 4: Konfidensinterval for middelværdi (og varians)

Peder Bacher

DTU Compute, Dynamiske Systemer  
Bygning 303B, Rum 010  
Danmarks Tekniske Universitet  
2800 Lyngby – Danmark  
e-mail: pbac@dtu.dk

Forår 2021

# Kapitel 3: Konfidensintervaller for én gruppe/stikprøve

## Grundlæggende koncepter

- Population og tilfældig stikprøve
- Statistisk model
- Estimation (*f.eks.  $\hat{\mu}$  er estimat af  $\mu$* )
- Signifikansniveau  $\alpha$
- Konfidensintervaller (*fanger rigtige prm.  $1 - \alpha$  af gangene*)
- Stikprøvefordelinger (*stikprøvegennemsnit ( $t$ ) og empirisk varians ( $\chi^2$ )*)
- Centrale grænseværdidisætning

## Specifikke metoder, én gruppe/stikprøve

- Konfidensinterval for middelværdi ( $t$ -fordeling)
- Konfidensinterval for varians ( $\chi^2$ -fordeling)

# Chapter 3: One sample confidence intervals

## General concepts

- Population and a random sample
- Statistical model
- Estimation (*e.g.  $\hat{\mu}$  is estimate of  $\mu$* )
- Significance level  $\alpha$
- Confidence intervals (*Catches true value  $1 - \alpha$  times*)
- Sampling distributions (*sample mean ( $t$ ) and sample variance ( $\chi^2$ )*)
- Central Limit Theorem

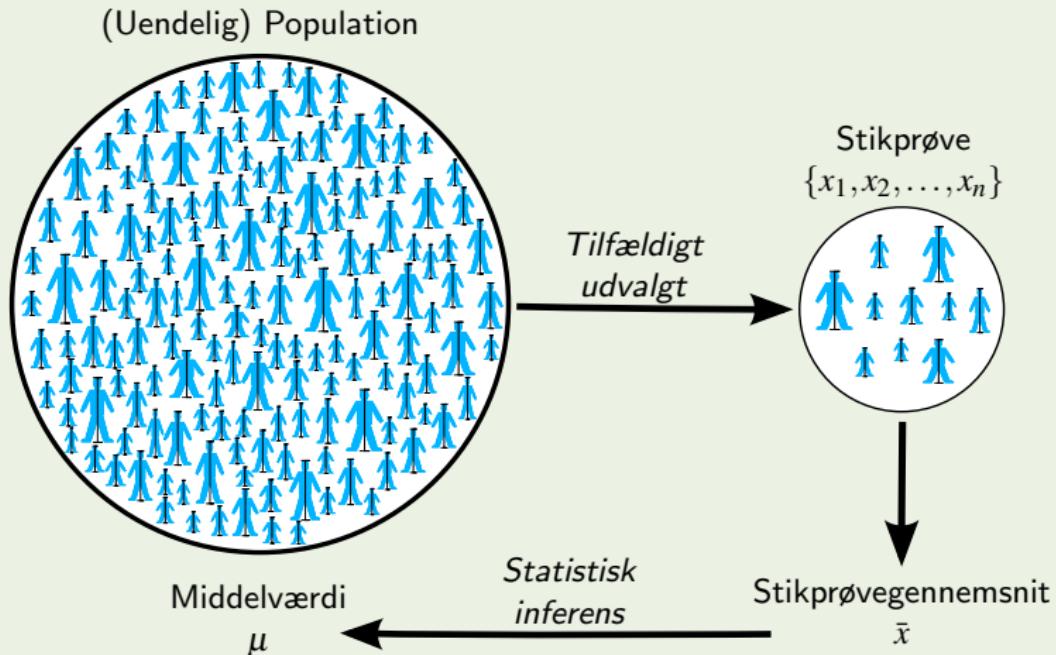
## Specific methods, one sample

- Confidence interval for the mean ( $t$ -distribution)
- Confidence interval for the variance ( $\chi^2$ -distribution)

# Oversigt

- 1 Fordelingen for gennemsnittet
  - $t$ -fordelingen
- 2 Konfidensintervallet for  $\mu$ 
  - Eksempel
- 3 Den statistiske sprogbrug og formelle ramme
- 4 Ikke-normale data, Central Grænseværdidisætning (CLT)
- 5 Konfidensinterval for varians og standardafvigelse

# Eksempel: Population og fordeling



Vi går nu på jagt efter  $\mu$  og  $\sigma$ !

# Stikprøveeksperiment 1

## Eksperiment

Tag en stikprøve på  $n = 10$  observationer fra populationen.

Kan det passe, at der er 95% sandsynlighed for at intervallet beregnet på stikprøven ved

$$\bar{X} \pm 2.26 \cdot \frac{S}{\sqrt{10}}$$

indeholder populationens gennemsnit  $\mu$  (dvs. middelværdien)?

Altså at følgende er sandt

$$P\left(\bar{X} - 2.26 \cdot \frac{S}{\sqrt{10}} < \mu < \bar{X} + 2.26 \cdot \frac{S}{\sqrt{10}}\right) = 0.95$$

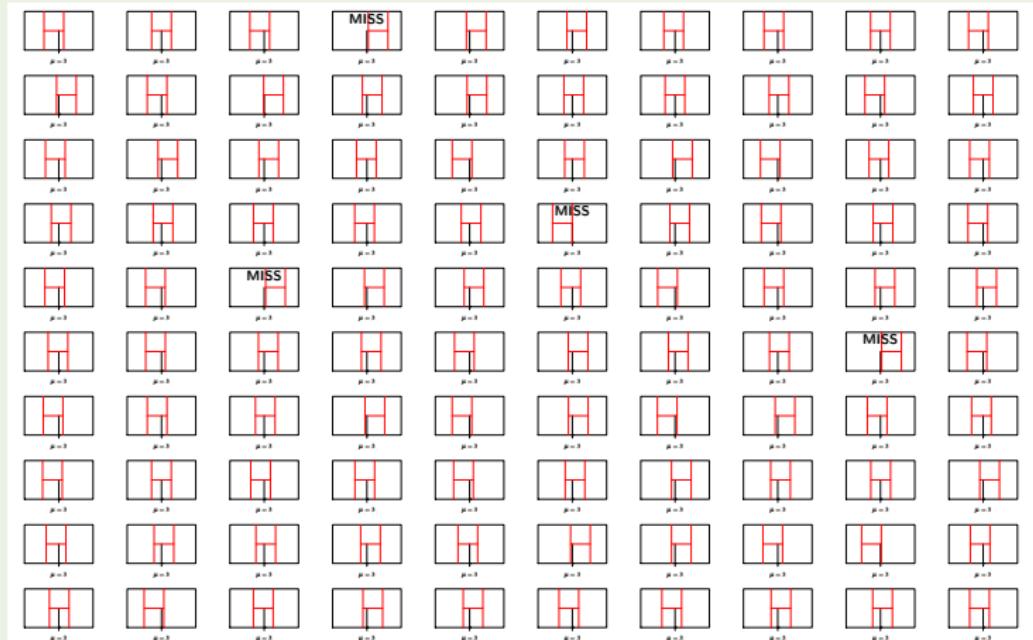
# Stikprøveeksperiment: Simulering af stikprøve og beregning af 95% konfidensinterval

```
## Middelværdien
mu <- 3
## Standardafvigelsen
sigma <- 1.8
## Stikprøvestørrelsen
n <- 10

## Simuler normalfordelte  $X_i$ 
x <- rnorm(n=n, mean=mu, sd=sigma)
## Se værdierne i den simulerede stikprøve
x
## Empirisk tæthed
hist(x, prob=TRUE, col='blue')

## Beregn stikprøvegennemsnittet  $\bar{x}$  (sample mean)
mean(x)
## Beregn stikprøvestandardafvigelsen  $s$  (sample standard deviation)
sd(x)
## Beregn 95% konfidensintervallet
mean(x) - 2.26 * sd(x)/sqrt(n)
mean(x) + 2.26 * sd(x)/sqrt(n)
```

# Stikprøveeksperiment 1: 100 simuleringer



## Theorem 3.2: Fordeling for gennemsnit af normalfordelinger

### (Stikprøve-) fordelingen for $\bar{X}$

Assume that  $X_1, \dots, X_n$  are independent and identically normally distributed random variables,  $X_i \sim N(\mu, \sigma^2)$  and  $i = 1, \dots, n$ , then:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \Rightarrow \quad \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

# Middelværdi og varians følger af regneregler

Theorem 2.40: Lineær funktion af normal distribuerede variable er også normalfordelt

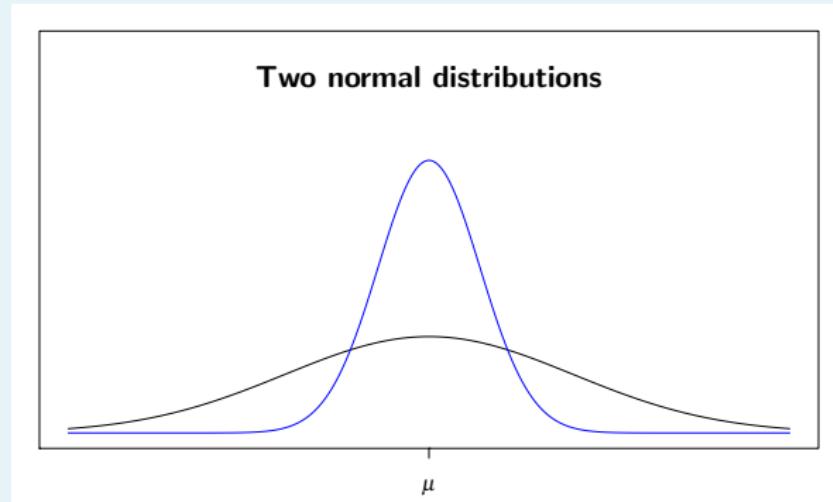
Theorem 2.53: Middelværdien af  $\bar{X}$

$$E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} n\mu = \mu$$

Theorem 2.53: Variansen for  $\bar{X}$

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

# Spørgsmål om stikprøvegennemsnittet (socrative.com, room: PBAC)



Den ene pdf hører til  $X_i$  og den anden til  $\bar{X}$ . Hvad kan konkluderes (for  $n > 1$ )?

- A: Den sorte hører til  $X_i$  og den blå til  $\bar{X}$
- B: Den sorte hører til  $\bar{X}$  og den blå til  $X_i$
- C: Det kan ikke afgøres
- D: Ved ikke

# Eksempel: Simuler middelværdi og standardafvigelse af stikprøvegennemsnit

```
## Middelværdien
mu <- 3
## Standardafvigelsen
sigma <- 1.8
## Stikprøvestørrelsen
n <- 10

## Simuler normalfordelte X_i
x <- rnorm(n=n, mean=mu, sd=sigma)
## Se den simulerede stikprøve
x
## Empirisk tæthed
hist(x, prob=TRUE, col='blue')

## Beregn stikprøvegennemsnittet (bar{x}: sample mean)
mean(x)
## Beregn stikprøvestandardafvigelsen (s: sample standard deviation)
sd(x)

## Gentag den simulerede stikprøvetagning mange gange
mat <- replicate(100, rnorm(n=n, mean=mu, sd=sigma))
## Beregn gennemsnittet for hver af dem
xbar <- apply(mat, 2, mean)
## Nu har vi mange realiseringer af stikprøvegennemsnittet
xbar
## Se deres fordeling
hist(xbar, prob=TRUE, col='blue')
## Deres gennemsnit
mean(xbar)
## og deres standardafvigelser
sd(xbar)
```

# Standardiseret fejl vi begår, Corollary 3.3:

Når vi bruger  $\bar{X}$  som estimat for  $\mu$ :

Så begår vi fejlen  $\bar{X} - \mu$

Fordelingen for den standardiserede fejl vi begår:

Assume that  $X_1, \dots, X_n$  are independent and identically normally distributed random variables,  $X_i \sim N(\mu, \sigma^2)$  where  $i = 1, \dots, n$ , then:

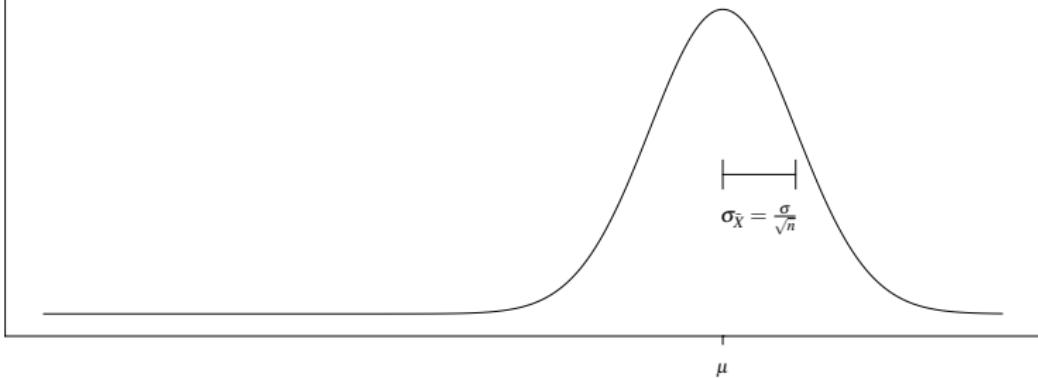
$$Z = \frac{\bar{X} - \mu}{\sigma_{(\bar{X} - \mu)}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$$

That is, the standardized sample mean  $Z$  follows a *standard normal distribution*.

# Transformation til standard normalfordeling:

Pdf for gennemsnittet  $\bar{X}$  når  $X_i \sim N(\mu, \sigma^2)$

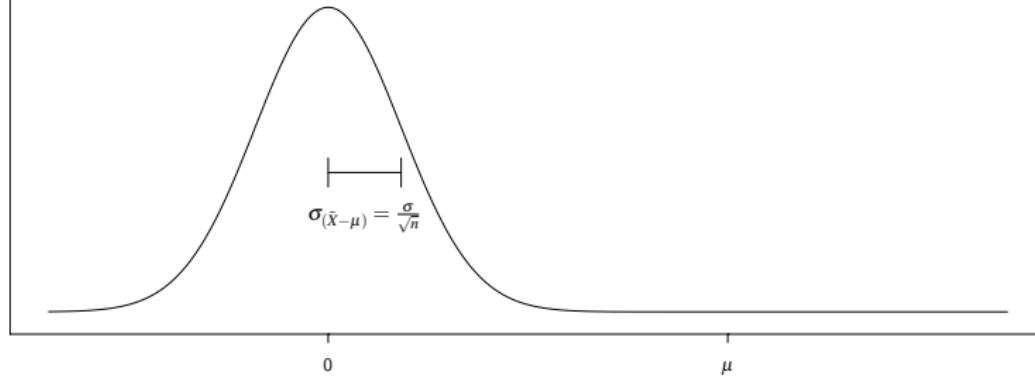
$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$



# Transformation til standard normalfordeling:

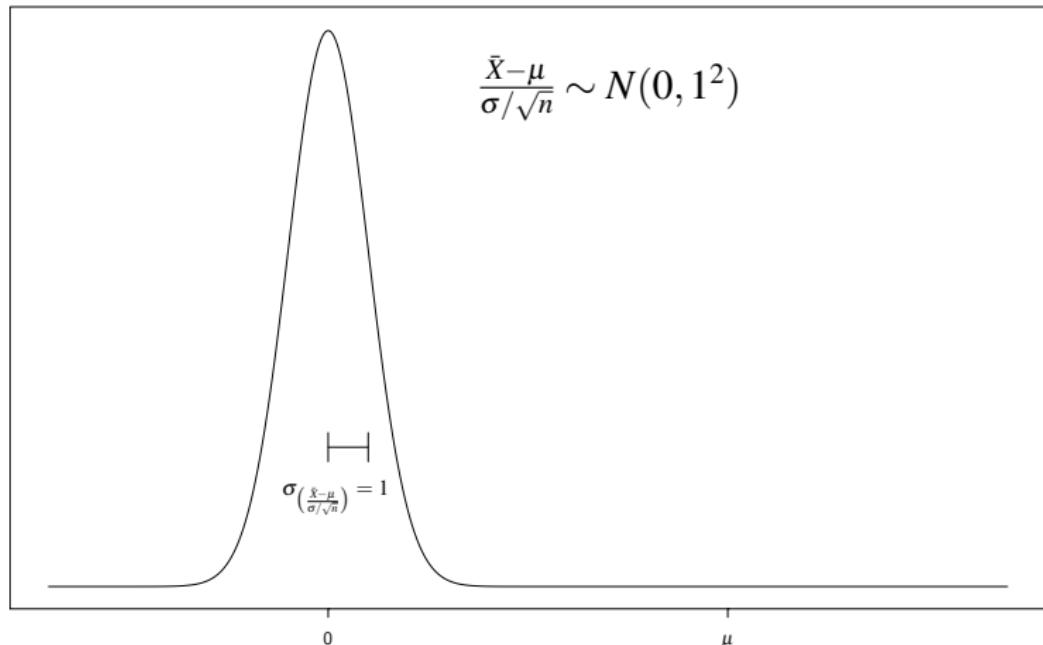
Pdf for *fejlen vi begår*  $\bar{X} - \mu$  når  $X_i \sim N(\mu, \sigma^2)$

$$\bar{X} - \mu \sim N(0, \frac{\sigma^2}{n})$$



# Transformation til standard normalfordeling:

Pdf for *den standardiserede fejl*  $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$  når  $X_i \sim N(\mu, \sigma^2)$



Standardiseret til *standard normalfordeling* (noteres  $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$ )

# Nu kan et 95% konfidensinterval udledes

95% konfidensinterval for  $\mu$ :

$$P(z_{0.025} < Z < z_{0.975}) = 0.95 \Leftrightarrow$$

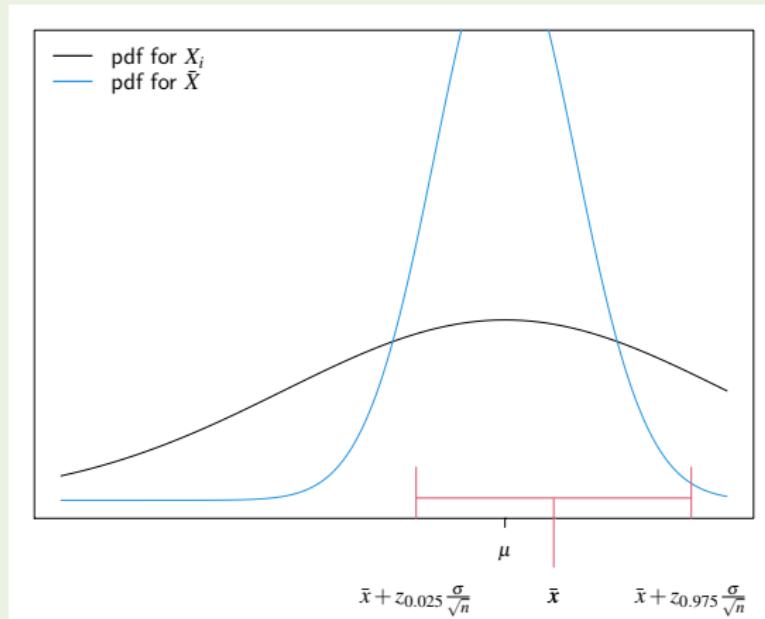
$$P\left(z_{0.025} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{0.975}\right) = 0.95 \Leftrightarrow$$

$$P\left(z_{0.025} \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{0.975} \frac{\sigma}{\sqrt{n}}\right) = 0.95 \Leftrightarrow$$

$$P\left(\bar{X} + z_{0.025} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{0.975} \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

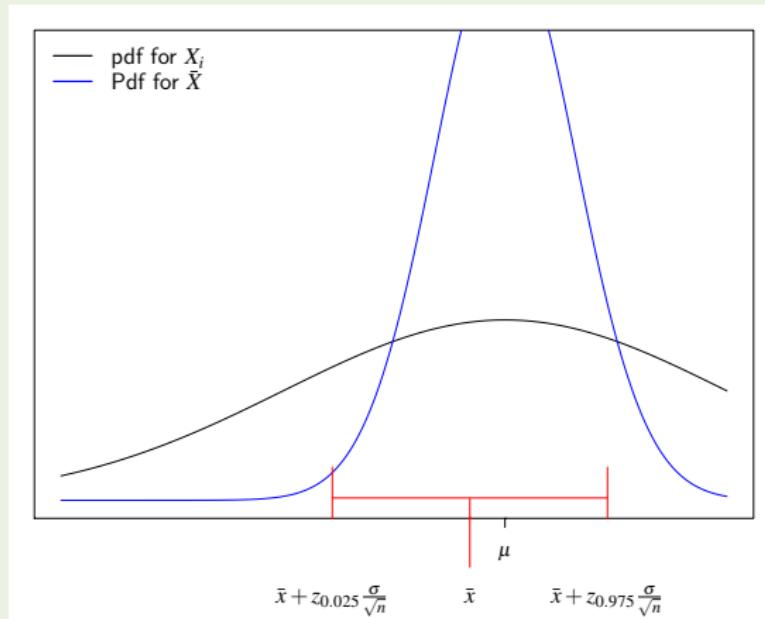
# 1. simulering: Beregning af 95% konfidensinterval

Konfidensintervallet er omkring  $\bar{x}$  og fanger her  $\mu$



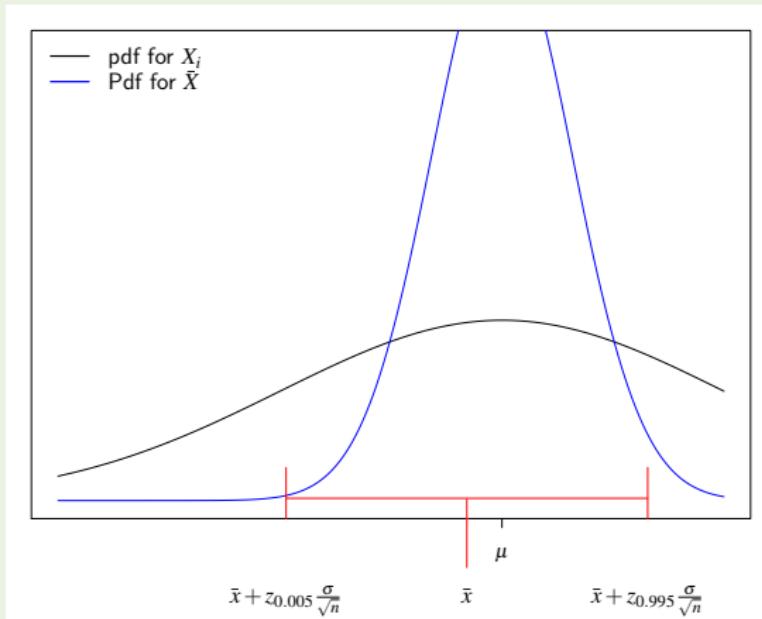
## 2. simulering: Beregning af 95% konfidensinterval

Konfidensintervallet er omkring  $\bar{x}$  og fanger her  $\mu$

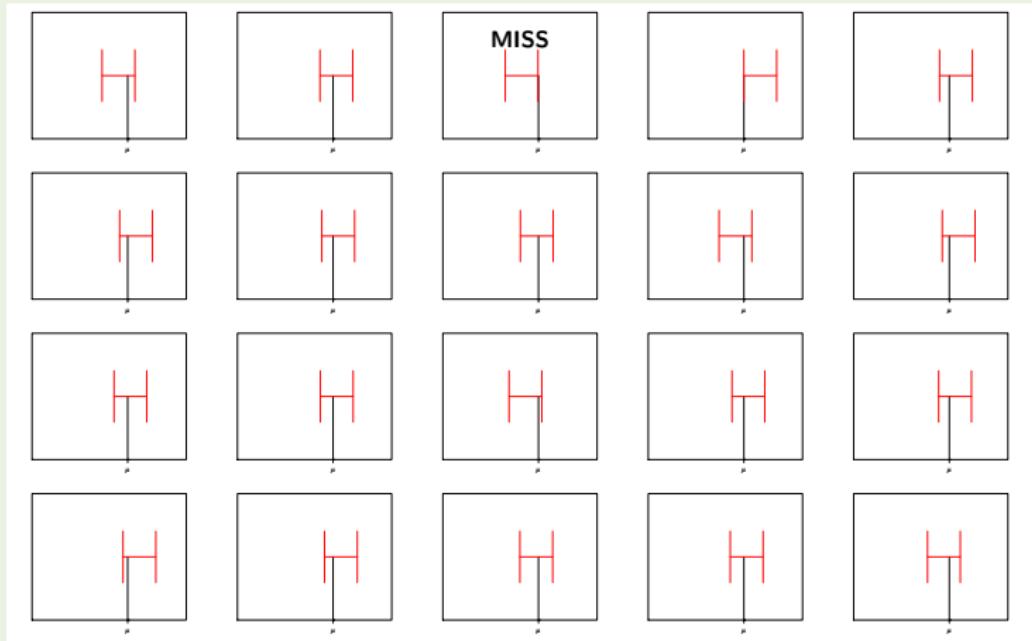


## 2. simulering: Beregning af 99% konfidensinterval

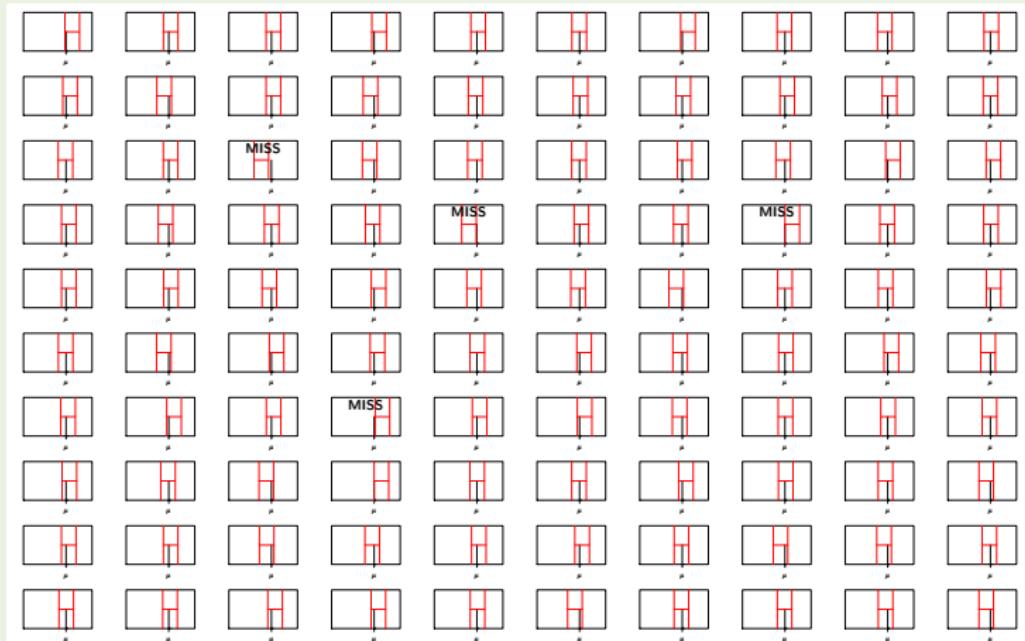
99% konfidensintervallet er breddere end 95% konfidensintervallet (det skal fange  $\mu$  oftere)



## 20 simuleringer: Beregning af 95% konfidensinterval



## 100 simuleringer: Beregning af 95% konfidensinterval



# Spørgsmål om konfidensinterval (socrative.com, room: PBAC)

Hvis vi planlægger at beregne et 98% konfidensinterval for middelværdien, hvad er da sandsynligheden for at middelværdien *ikke* ligger inde i intervallet?

- A: 1%
- B: 2%
- C: 4%
- D: Den kender vi ikke
- E: Ved ikke

# Spørgsmål om konfidensinterval (socrative.com, room: PBAC)

Når vi så har udført eksperimentet og har stikprøven, ved vi da om middelværdien er indeholdt i det konfidensinterval vi har beregnet?

- A: Ja
- B: Nej
- C: Ved ikke

# Praktisk problem!!

*Populationens standardafvigelse  $\sigma$  indgår i formlen og den kender vi ikke!!*

Oplagt løsning:

Anvend stikprøvens standardafvigelse  $S$  som estimatet af  $\sigma$  i stedet for!

MEN MEN:

Så bryder den givne teori faktisk sammen!!

HELDIGVIS:

Der findes en heldigvis udvidet teori, der kan klare det!!

## Theorem 3.4: More applicable extension of the same stuff: (kopi af Theorem 2.49)

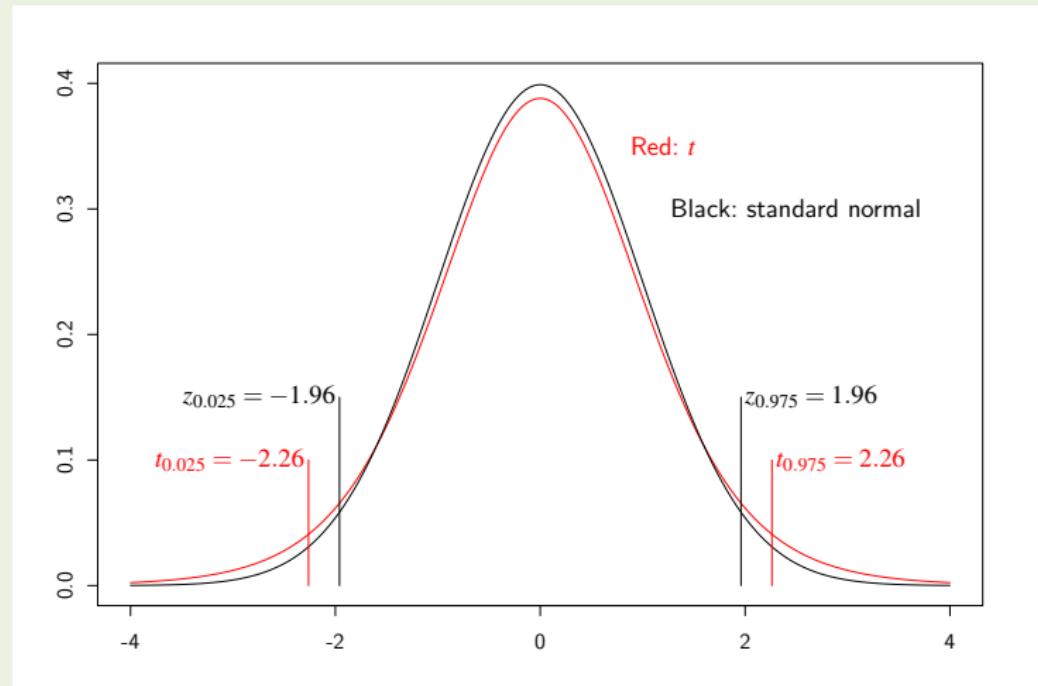
*t*-fordelingen tager højde for usikkerheden i at bruge  $s$ :

Assume that  $X_1, \dots, X_n$  are independent and identically normally distributed random variables, where  $X_i \sim N(\mu, \sigma^2)$  and  $i = 1, \dots, n$ , then

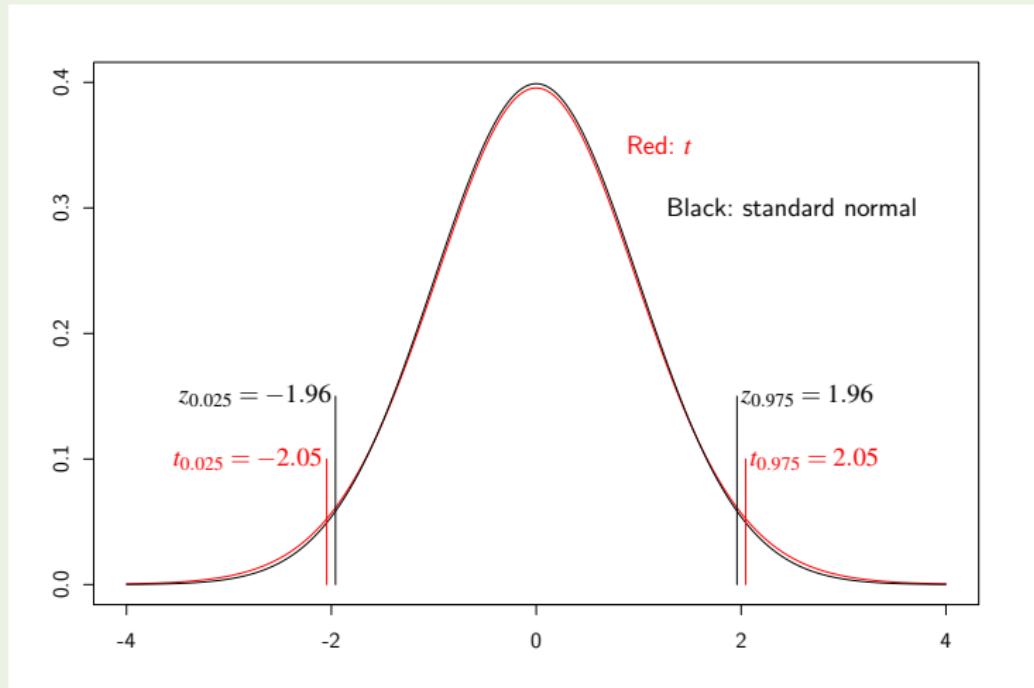
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t$$

where  $t$  is the  $t$ -distribution with  $n - 1$  degrees of freedom.

# *t*-fordelingen med 9 frihedsgrader ( $n = 10$ ) og standardnormalfordelingen



# *t*-fordelingen med 29 frihedsgrader ( $n = 30$ ) og standardnormalfordelingen



## Metodeboks 3.8: One-sample konfidensinterval for $\mu$

Brug den rigtige  $t$ -fordeling til at lave konfidensintervallet:

For a sample  $x_1, \dots, x_n$  the  $100(1 - \alpha)\%$  confidence interval is given by:

$$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

where  $t_{1-\alpha/2}$  is the  $100(1 - \alpha)\%$  quantile from the  $t$ -distribution with  $n - 1$  degrees of freedom.

Mest almindeligt med  $\alpha = 0.05$ :

The most commonly used is the 95%-confidence interval:

$$\bar{x} \pm t_{0.975} \cdot \frac{s}{\sqrt{n}}$$

## Eksempel - Højde af 10 studerende

Stikprøve,  $n = 10$ :

168 161 167 179 184 166 198 187 191 179

Sample mean og standard deviation:

$$\bar{x} = 178$$

$$s = 12.21$$

Estimer population mean og standard deviation:

$$\hat{\mu} = 178$$

$$\hat{\sigma} = 12.21$$

# Højde-eksempel, 95% konfidensinterval (CI)

```
## 97.5% fraktilen af t-fordelingen for n=10:  
qt(p=0.975, df=9)
```

```
## [1] 2.26
```

Indsat i formlen

$$178 \pm 2.26 \cdot \frac{12.21}{\sqrt{10}}$$

giver det

$$178 \pm 8.74 = [169.3; 186.7]$$

# Højde-eksempel, 99% Konfidensinterval (CI)

```
## 99.5% fraktilen af t-fordelingen for n=10:  
qt(p=0.995, df=9)  
  
## [1] 3.25
```

Indsat i formlen

$$178 \pm 3.25 \cdot \frac{12.21}{\sqrt{10}}$$

giver det

$$178 \pm 12.55 = [165.4; 190.6]$$

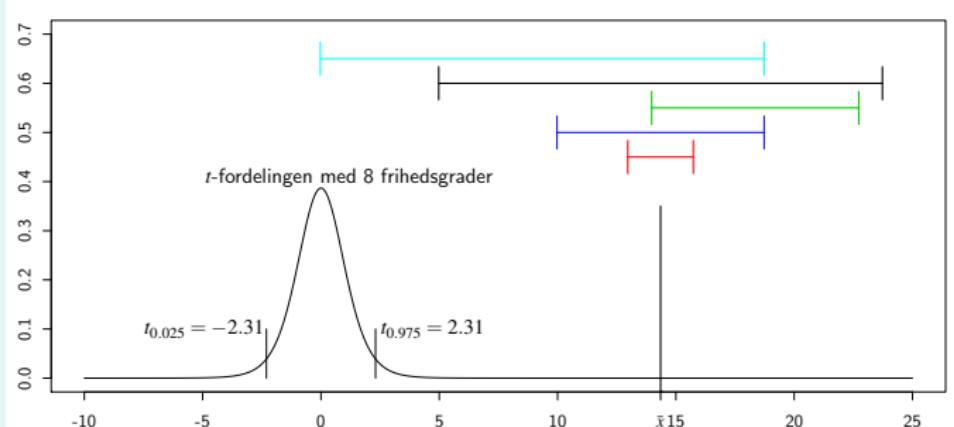
Der findes en R-funktion, der kan gøre det hele (med mere):

```
## Angiv data
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
## Beregn 99% konfidensinterval
t.test(x, conf.level=0.99)

##
##  One Sample t-test
##
## data: x
## t = 46, df = 9, p-value = 5e-12
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
## 165 191
## sample estimates:
## mean of x
## 178
```

## Svar via socrative.com eller Socrative app. Room: PBAC

- Gennemsnit  $\bar{x} = 14.4$ , stikprøvens standardafvigelse  $s = 6$ , antal obs. er  $n = 9$
- Formlen for konfidensintervallet er  $\bar{x} \pm t_{0.975} \frac{s}{\sqrt{n}}$



Hvilket af intervallerne er det rigtige 95% konfidensinterval?

- A: Turkise    B: Sorte    C: Grønne    D: Blå    E: Røde

# Den formelle ramme for *statistisk inferens*

Fra bogen, kapitel 1:

- An *observational unit* is the single entity/level about which information is sought (e.g. a person) (**Observationsenhed**)
- The *statistical population* consists of all possible “measurements” on each *observational unit* (**Population**)
- The *sample* from a statistical population is the actual set of data collected. (**Stikprøve**)

Sprogbrug og koncepter:

- $\mu$  og  $\sigma$  er parametre, som beskriver populationen
- $\bar{x}$  er *estimatet* for  $\mu$  (konkret udfald)
- $\bar{X}$  er *estimatoren* for  $\mu$  (nu set som stokastisk variabel)
- Begrebet 'statistic(s)' er en fællesbetegnelse for begge

# Den formelle ramme for *statistisk inferens* - Eksempel

Fra bogen, kapitel 1, højdeeksempel

Vi mäter højden for 10 tilfældige personer i Danmark

Stikprøven/The sample:

De 10 konkrete talværdier:  $x_1, \dots, x_{10}$

Populationen:

Højderne for alle mennesker i Danmark.

Observationsenheden:

En person

# Statistisk inferens = Learning from data

*Learning from data is learning about parameters of distributions that describe populations*

Vigtigt i den forbindelse:

Stikprøven skal på meningsfuld vis være repræsentativ for en eller anden veldefineret population

Hvordan sikrer man det

Ved at sikre at stikprøven er fuldstændig tilfældig udtaget

# Tilfældig stikprøveudtagning

## Definition 3.11:

- A random sample from an (infinite) population: A set of observations  $X_1, X_2, \dots, X_n$  constitutes a random sample of size  $n$  from the infinite population  $f(x)$  if:
  - 1 Each  $X_i$  is a random variable whose distribution is given by  $f(x)$
  - 2 These  $n$  random variables are independent

## Hvad betyder det????

- 1 Alle observationer skal komme fra den samme population
- 2 De må IKKE dele information med hinanden (f.eks. hvis man havde udtaget hele familier i stedet for enkeltindivider)

## Theorem 3.13: The Central Limit Theorem

Gennemsnittet af en tilfældig stikprøve følger altid en normalfordeling hvis  $n$  er stor nok:

Let  $\bar{X}$  be the mean of a random sample of size  $n$  taken from a population with mean  $\mu$  and variance  $\sigma^2$ , then

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

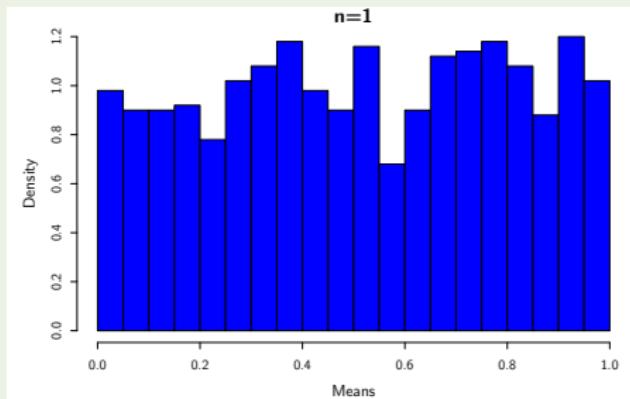
is a random variable whose distribution function approaches that of the standard normal distribution,  $N(0, 1^2)$ , as  $n \rightarrow \infty$

Dvs., hvis  $n$  er stor nok, kan vi (tilnærmelsesvist) antage:

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1^2) \text{ og } \frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t \text{ ved } t\text{-fordelingen med } n - 1 \text{ frihedsgrader}$$

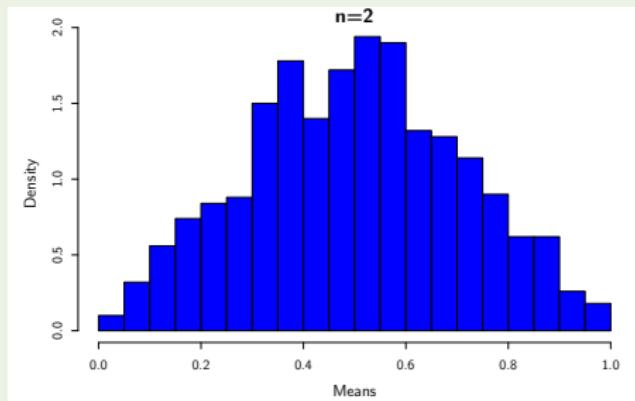
# CLT in action - gennemsnit af Uniform fordelte observationer

```
## Stikprøvestørrelse
n <- 1
## Antal gentagelser
k <- 1000
## Simuler værdier og sæt i k x n
u <- matrix(runif(n=k*n, min=0, max=1), ncol=n)
## Se empirisk tæthed
hist(apply(u, 1, mean), col='blue', main='n=1', xlab='Means', nclass=15, prob=TRUE, x
```



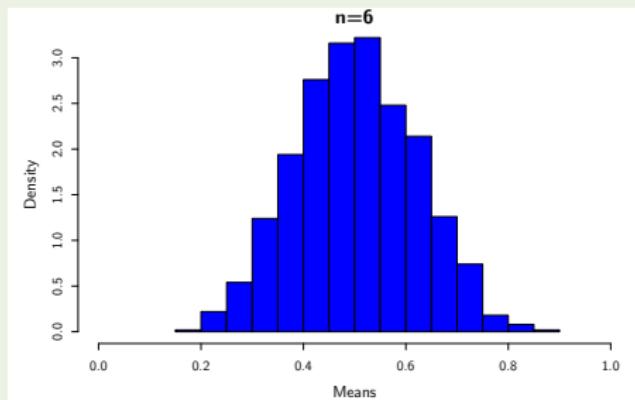
# CLT in action - gennemsnit af Uniform fordelte observationer

```
## Stikprøvestørrelse
n <- 2
## Antal gentagelser
k <- 1000
## Simuler
u <- matrix(runif(n=k*n, min=0, max=1), ncol=n)
## Se empirisk tæthed
hist(apply(u, 1, mean), col='blue', main='n=2', xlab='Means', nclass=15, prob=TRUE, x
```



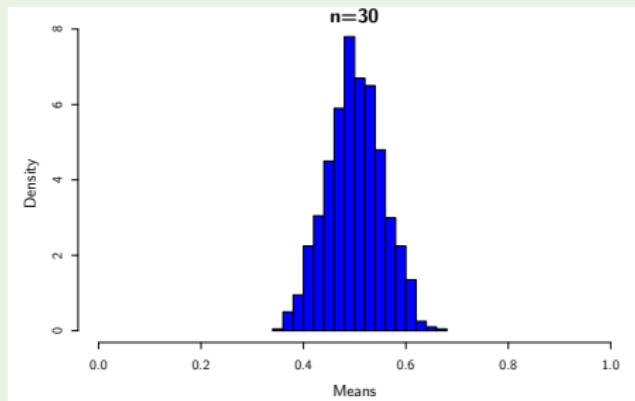
# CLT in action - gennemsnit af Uniform fordelte observationer

```
## Stikprøvestørrelse
n <- 6
## Antal gentagelser
k <- 1000
## Simuler
u <- matrix(runif(n=k*n, min=0, max=1), ncol=n)
## Se empirisk tæthed
hist(apply(u, 1, mean), col='blue', main='n=6', xlab='Means', nclass=15, prob=TRUE, x
```



# CLT in action - gennemsnit af Uniform fordelte observationer

```
## Stikprøvestørrelse
n <- 30
## Antal gentagelser
k <- 1000
## Simuler
u <- matrix(runif(n=k*n, min=0, max=1), ncol=n)
## Se empirisk tæthed
hist(apply(u, 1, mean), col='blue', main='n=30', xlab='Means', nclass=15, prob=TRUE, )
```



# Konsekvens af CLT:

Vores CI-metode virker OGSÅ for ikke-normale data:

Vi kan bruge konfidens-interval baseret på  $t$ -fordelingen i stort set alle situationer, blot  $n$  er "stør nok"

Hvad er "stør nok"?

Faktisk svært at svare præcist på, MEN:

- Tommelfingerregel:  $n \geq 30$
- Selv for mindre  $n$  kan formlen være (næsten) gyldig for ikke-normale data.

## Svar via socrative.com eller Socrative app. Room: PBAC

Er lydniveauet behageligt?

- A: Fino
- B: Nope, tal højere
- C: Nope, tal lavere
- D: Nope, der er bare dårlig lyd herinde

# Svar via socrative.com eller Socrative app. Room: PBAC

Bør Peder klæde sig mere nydeligt?

- A: Ja, for den da! Det er grimt det tøj
- B: Nej, han ser faktisk rigtig checket ud
- C: Nej, det kan være lige meget med tøjet, han skal barbere sig og rede sit hår først
- D: Ved ikke, jeg har simpelthen været for optaget af statistikken til at lægge mærke til hans påklædning

# Statistisk model

Statistisk model, se Remark 3.2

Der tages en stikprøve, som består af de stokastiske variable  $X_i$  hvor  $i = 1, \dots, n$ .

Der opstilles følgende model

$$X_i \sim N(\mu, \sigma^2) \text{ and i.i.d., where } i = 1, \dots, n$$

Dvs.

- $n$  observationer fra en normalfordelt population med parametre  $\mu$  og  $\sigma$
- observationerne er i.i.d.:
  - *independent*: de gøres uafhængigt af hinanden
  - *identically distributed*: de har samme fordeling

# Stikprøvefordelingen for varians-estimatet (Theorem 2.56)

Variansestimatorer opfører sig som en  $\chi^2$ -fordeling:

Let

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

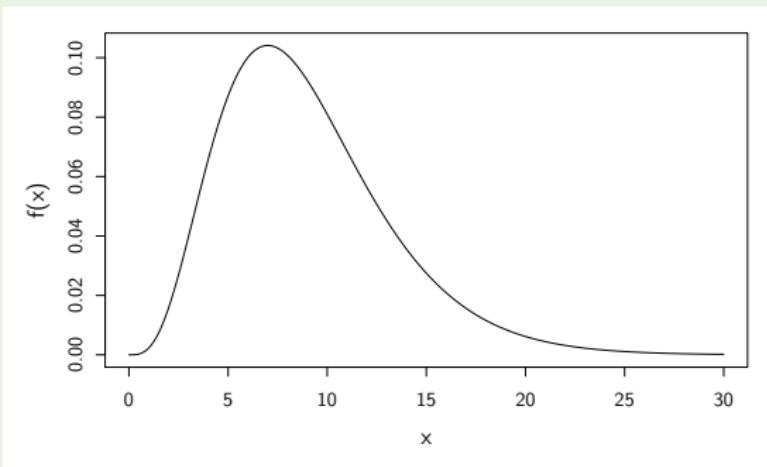
then:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

is a random variable following the  $\chi^2$ -distribution with  $v = n - 1$  degrees of freedom.

# $\chi^2$ -fordelingen med $v = 9$ frihedsgrader

```
## Plot chi^2 tæthedsfunktion med 9 frihedsgrader  
  
## En sekvens af x værdier  
x <- seq(0, 30, by = 0.1)  
## Plot chi^2 tæthedsfunktion  
plot(x, dchisq(x, df = 9), type = 'l', ylab="f(x)")
```



## Metode 3.18: Konfidensinterval for stikprøvevariens og stikprøvestandardafvigelse

### Variansen:

A  $100(1 - \alpha)\%$  confidence interval for the variance  $\sigma^2$  is:

$$\left[ \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}; \frac{(n-1)s^2}{\chi^2_{\alpha/2}} \right]$$

where the quantiles come from a  $\chi^2$ -distribution with  $v = n - 1$  degrees of freedom.

### Standardafvigelsen:

A  $100(1 - \alpha)\%$  confidence interval for the sample standard deviation  $\hat{\sigma}$  is:

$$\left[ \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}}; \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}} \right]$$

# Eksempel

## Produktion af tabletter

Vi producerer pulverblanding og tabletter deraf, så koncentrationen af det aktive stof i tabletterne skal være 1 mg/g med den mindst mulige spredning. En tilfældig stikprøve udtages, hvor vi mäter mængden af aktivt stof.

### Data:

En tilfældig stikprøve med  $n = 20$  tabletter er udtaget og fra denne får man:

$$\hat{\mu} = \bar{x} = 1.01, \hat{\sigma}^2 = s^2 = 0.07^2$$

95%-konfidensinterval for variansen - vi skal bruge  $\chi^2$ -fraktilerne:

$$\chi^2_{0.025} = 8.9065, \chi^2_{0.975} = 32.8523$$

```
## 2.5% og 97.5% fraktilerne i chi^2 fordelingen for n=20
qchisq(c(0.025, 0.975), df = 19)
```

# Eksempel

Så konfidensintervallet for variansen  $\sigma^2$  bliver:

$$\left[ \frac{19 \cdot 0.7^2}{32.85}; \frac{19 \cdot 0.7^2}{8.907} \right] = [0.002834; 0.01045]$$

Og konfidensintervallet for standardafvigelsen  $\sigma$  bliver:

$$\left[ \sqrt{0.002834}; \sqrt{0.01045} \right] = [0.053; 0.102]$$

# Højdeeksempel

Vi skal bruge  $\chi^2$ -fraktilerne med  $v = 9$  frihedsgrader:

$$\chi^2_{0.025} = 2.700389, \chi^2_{0.975} = 19.022768$$

```
## 2.5% og 97.5% fraktilerne i chi^2 fordelingen for n=10
qchisq(c(0.025, 0.975), df = 9)

## [1] 2.7 19.0
```

Så konfidensintervallet for højdens standardafvigelse  $\sigma$  bliver:

$$\left[ \sqrt{\frac{9 \cdot 12.21^2}{19.022768}}; \sqrt{\frac{9 \cdot 12.21^2}{2.700389}} \right] = [8.4; 22.3]$$

# Eksempel - Højde af 10 studerende - recap:

Stikprøve,  $n = 10$ :

168 161 167 179 184 166 198 187 191 179

Sample mean og standard deviation:

$$\bar{x} = 178$$

$$s = 12.21$$

Estimer population mean og standard deviation:

$$\hat{\mu} = 178$$

$$\hat{\sigma} = 12.21$$

NYT: Konfidensinterval,  $\mu$ :

$$178 \pm 2.26 \cdot \frac{12.21}{\sqrt{10}} \Leftrightarrow [169.3; 186.7]$$

NYT: Konfidensinterval,  $\sigma$ :

$$[8.4; 22.3]$$

# Svar via socrative.com eller Socrative app. Room: PBAC

Hvilket af følgende udsagn er korrekt?

- A: Statistik er virkelig skod, jeg tror ikke det kan bruges til noget
- B: Statistik er altså øv, man skal bare sidde og sætte en masse tal ind i nogle dumme formler
- C: Jeg burde ligge under min dyne og blive frisk til at feste igennem i aften
- D: Statistik er virkelig fedt, det er fascinerende, at man ikke bare kan regne et estimat ud, men man kan også regne ud hvor præcist det er

# Introduktion til Statistik

## Forelæsning 5: Hypotesetest og modelkontrol - one sample

Peder Bacher

DTU Compute, Dynamiske Systemer  
Bygning 303B, Rum 010  
Danmarks Tekniske Universitet  
2800 Lyngby – Danmark  
e-mail: pbac@dtu.dk

Forår 2021

# Kapitel 3: Hypotesetests for én gruppe/stikprøve

Grundlæggende koncepter:

- Hypoteser ( $H_0$  vs.  $H_1$ )
- $p$ -værdi (*Sandsynlighed for observeret eller mere ekstrem værdi af teststørrelsen, hvis  $H_0$  er sand, e.g.  $P(T > t_{\text{obs}})$* )
- Type I fejl (*I virkeligheden ingen effekt, men  $H_0$  afvises*)
  - $P(\text{Type I}) = \alpha$  (*Sandsynligheden for at begå type I fejl*)
- Type II fejl (*I virkeligheden effekt, men  $H_0$  afvises ikke*)
  - $P(\text{Type II}) = \beta$  (*Sandsynligheden for type II fejl*)
- Modelkontrol

Specifikke metoder, én gruppe:

- $t$ -test for middelværdiniveau
- Modelkontrol med normal qq-plot

# Chapter 3: One sample hypothesis testing

## General concepts:

- Hypotheses ( $H_0$  vs.  $H_1$ )
- $p$ -value (*Probability for observing the test value or more extreme, if  $H_0$  is true, e.g.  $P(T > t_{\text{obs}})$* )
- Type I error (*No effect in reality, but  $H_0$  is rejected*)
  - $P(\text{Type I}) = \alpha$  (*The probability for a Type I error*)
- Type II error: (*In reality an effect, but  $H_0$  is not rejected*)
  - $P(\text{Type II}) = \beta$  (*The probability for a Type II error*)
- Model validation

## Specific methods, one sample:

- $t$ -test for the mean
- Model validation with normal q-q plot

# Oversigt

- 1 One-sample  $t$ -test og  $p$ -værdi
- 2  $p$ -værdier og hypotesetest
- 3 Kritisk værdi og konfidensinterval
- 4 Hypotesetests (helt generelt)
  - Hypotesetest med alternativer
  - Den generelle metode
  - Type I og type II fejl
- 5 Check af normalfordelingsantagelse
  - The normal q-q plot
  - Transformation towards normality

# Spørgsmål om fordelingen af stikprøvegennemsnittet og standardisering (socrative.com - ROOM:PBAC)

Hvilken pdf representerer fordelingen af stikprøvegennemsnittet

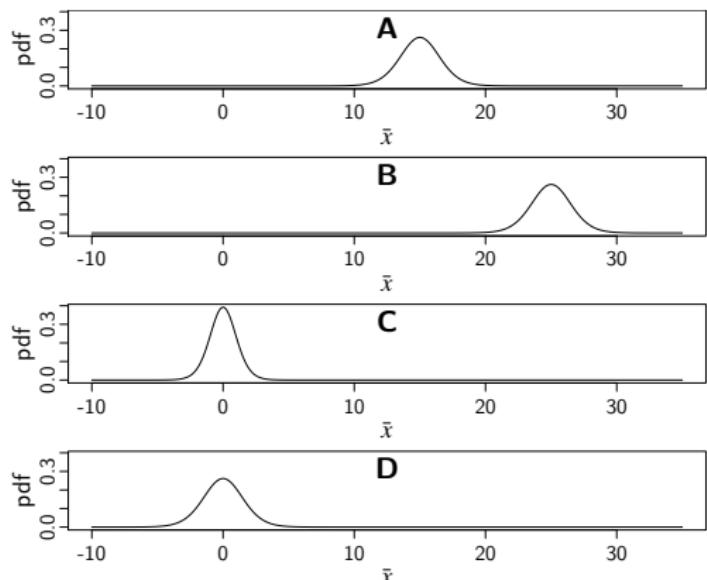
$$\bar{X}$$

for

$$\mu = 15$$

(stikprøvestørrelse  $n = 16$ )

(stikprøvestandardafvigelse  $s = 8$ )



A B C eller D?

# Spørgsmål om fordelingen af stikprøvegennemsnittet og standardisering (socrative.com - ROOM:PBAC)

Hvilken pdf representerer fordelingen af

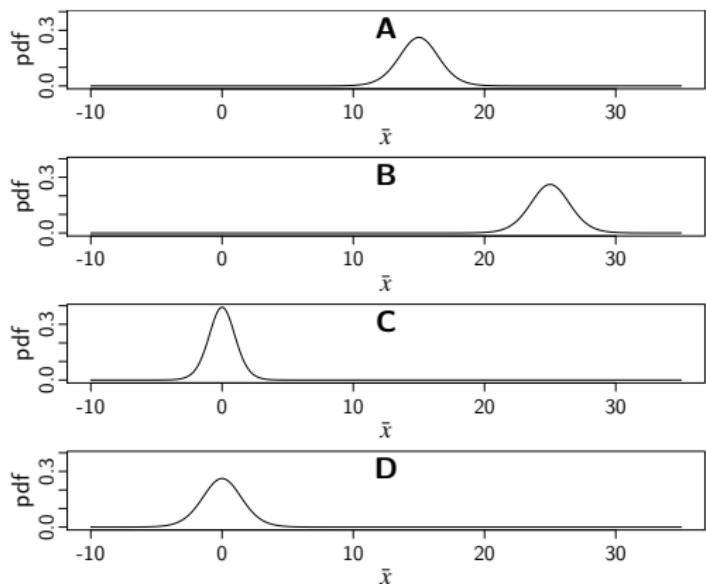
$$\bar{X} - \mu$$

for

$$\mu = 15$$

(stikprøvestørrelse  $n = 16$ )

(stikprøvestandardafvigelse  $s = 8$ )



A B C eller D?

# Spørgsmål om fordelingen af stikprøvegennemsnittet og standardisering (socrative.com - ROOM:PBAC)

Hvilken pdf representerer fordelingen af

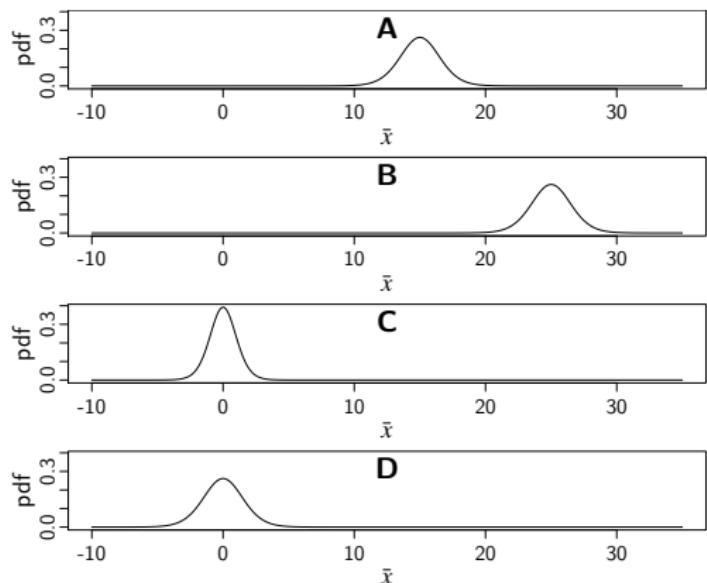
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

for

$$\mu = 15$$

(stikprøvestørrelse  $n = 16$ )

(stikprøvestandardafvigelse  $s = 8$ )



A B C eller D?

## Metode 3.23: One-sample $t$ -test og $p$ -værdi

Hvad er  $p$ -værdien og hvordan beregnes den?

Man fremsætter *nulhypotesen*

$$H_0: \mu = \mu_0$$

under hvilken man beregner *teststørrelsen*

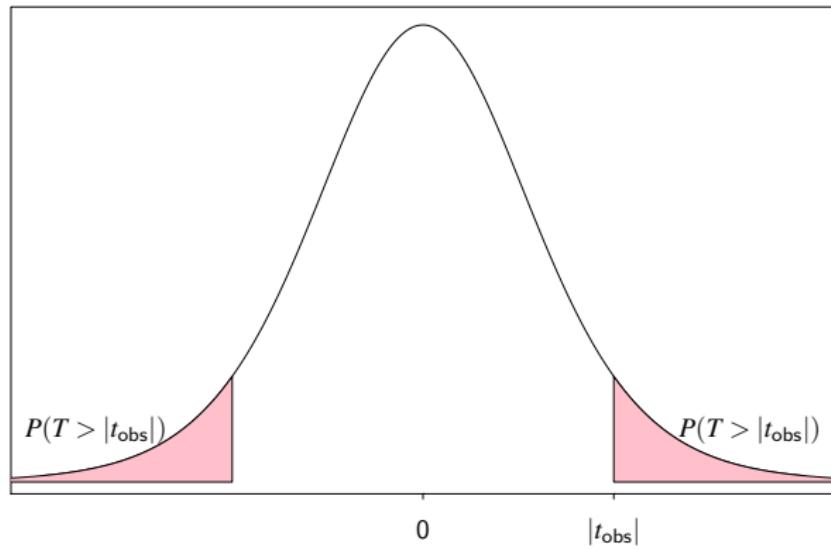
$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

som man derefter bruger til at beregne  $p$ -værdien

$$p\text{-værdi} = 2 \cdot P(T > |t_{\text{obs}}|)$$

- $p$ -værdien er altså: *Hvis nulhypotesen er sand, hvor sandsynligt er det da at få den observerede værdi af teststørrelsen ( $t_{\text{obs}}$ ) eller mere ekstremt?*

$$p\text{-værdi} = 2 \cdot P(T > |t_{\text{obs}}|)$$



Fortæller noget om: "hvor sandsynligt er det at få det observerede data under  $H_0$ " (dvs. hvis  $H_0$  er sand)

# Definition og fortolkning af $p$ -værdien (HELT generelt)

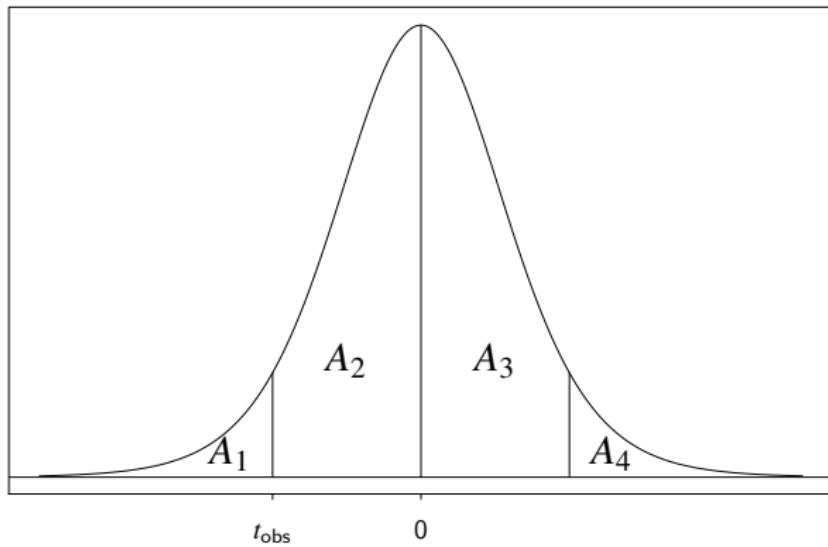
Definition 3.22 af  $p$ -værdien:

**The  $p$ -value** is the probability of obtaining a test statistic that is at least as extreme as the test statistic that was actually observed. This probability is calculated under the assumption that the null hypothesis is true.

$p$ -værdien udtrykker *evidence* imod nulhypotesen – Tabel 3.1:

$p < 0.001$	Very strong evidence against $H_0$
$0.001 \leq p < 0.01$	Strong evidence against $H_0$
$0.01 \leq p < 0.05$	Some evidence against $H_0$
$0.05 \leq p < 0.1$	Weak evidence against $H_0$
$p \geq 0.1$	Little or no evidence against $H_0$

## Spørgsmål om $p$ -værdi (socrative.com - ROOM:PBAC)



Hvad er  $2 \cdot P(T > |t_{\text{obs}}|)$ ?

- A:  $A_1 + A_2$
- B:  $A_3 + A_4$
- C:  $A_1 + A_4$
- D:  $A_2 + A_3$

## Motiverende eksempel - sovemedicin

### Forskel på sovemedicin?

I et studie er man interesseret i at sammenligne 2 sovemediciner *A* og *B*. For 10 testpersoner har man fået følgende resultater, der er givet i forlænget søvntid i timer (forskellen på effekten af de to midler er angivet):

Stikprøve,  $n = 10$ :

Person	$x = \text{Beffekt} - \text{Aeffekt}$	
1	1.2	
2	2.4	
3	1.3	
4	1.3	
5	0.9	Stikprøvens:
6	1.0	$\bar{x} = 1.67$ (gennemsnit)
7	1.8	$\bar{s} = 1.13$ (standardafvigelse)
8	0.8	
9	4.6	
10	1.4	

## Eksempel - sovemedicin

Hypotesen om ingen forskel på sovemedicin A og B ønskes undersøgt:

$$H_0 : \mu = 0$$

Er data i overenstemmelse med nulhypotesen  $H_0$ ?

Hvor "sandsynligt" er  $\bar{x} = 1.67$  hvis  $H_0 : \mu = 0$  er sand?

Beregne  $p$ -værdien:

Sandsynlighed for mere ekstremt data hvis  $H_0$  er sand:

$$\begin{aligned}2 \cdot P(T > |t_{\text{obs}}|) &= 2 \cdot P(T > 4.67) \\&= 0.00117\end{aligned}$$

Beregne teststørrelsen:

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1.67 - 0}{1.13/\sqrt{10}} = 4.67$$

**NYT: Konklusion:**

Idet data er usandsynligt under  $H_0$ , så **forkaster** vi  $H_0$  - vi har påvist en **signifikant effekt** af middel B ift. middel A.

## Eksempel - sovemedicin manuelt i R

```
## Angiv data
x <- c(1.2, 2.4, 1.3, 1.3, 0.9, 1.0, 1.8, 0.8, 4.6, 1.4)
n <- length(x)
## Beregn den observerede t værdi - den observerede test statistik
tobs <- (mean(x) - 0) / (sd(x) / sqrt(n))
## Beregn p-værdien, som sandsynligheden for at få tobs eller mere ekstremt
pvalue <- 2 * (1-pt(abs(tobs), df=n-1))
pvalue
## [1] 0.00117
```

## Eksempel - sovemedicin med indbygget funktion i R

```
## Kald funktionen med data x
t.test(x)

##
##  One Sample t-test
##
## data: x
## t = 4.67, df = 9, p-value = 0.00117
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.861 2.479
## sample estimates:
## mean of x
##      1.67
```

# Definition af hypotesetest og signifikans

## Definition 3.24 Hypotesetest:

We say that we carry out a hypothesis test when we decide against a null hypothesis or not, using the data.

A null hypothesis is *rejected* if the  $p$ -value, calculated after the data has been observed, is less than some  $\alpha$  (i.e.  $p$ -value  $< \alpha$ ), where  $\alpha$  is some pre-specified (so-called) *significance level*. And if not, then the null hypothesis is said to be *accepted*.

## Definition 3.29 Statistisk signifikans:

An *effect* is said to be (*statistically*) *significant* if the  $p$ -value is less than the significance level  $\alpha$ .

(OFTEN bruges  $\alpha = 0.05$ )

## Eksempel - sovemedicin

Konklusion for test af sovemedicin

Fortolkning med *p*-værdien.

Med  $\alpha = 0.05$  kan vi konkludere:

Idet *p*-værdien er mindre end  $\alpha$  så **forkaster** vi nulhypotesen.

Og dermed:

Vi har påvist en **signifikant effekt** af middel B ift. middel A. (Og dermed at B virker bedre end A)

# Spørgsmål om $p$ -værdi (socrative.com - ROOM:PBAC)

```
## Kald funktionen med nul-hypotesen  $H_0$ :  $\mu=1$ 
t.test(x, mu=1)

##
##  One Sample t-test
##
## data: x
## t = 1.87, df = 9, p-value = 0.0937
## alternative hypothesis: true mean is not equal to 1
## 95 percent confidence interval:
##  0.861 2.479
## sample estimates:
## mean of x
## 1.67
```

Signifikansniveau  $\alpha = 0.05$ . Bliver  $H_0$  afvist?

- A)  $H_0 : \mu = 1$  afvises ikke og må accepteres
- B)  $H_0 : \mu = 1$  afvises
- C) Ved ikke

# Kritisk værdi

Definition 3.31 - de kritiske værdier for  $t$ -testet:

The  $(1 - \alpha)100\%$  critical values for the (non-directional) one-sample  $t$ -test are the  $(\alpha/2)100\%$  and  $(1 - \alpha/2)100\%$  quantiles of the  $t$ -distribution with  $n - 1$  degrees of freedom:

$$t_{\alpha/2} \text{ and } t_{1-\alpha/2}$$

Metode 3.32: One-sample  $t$ -test vha. kritisk værdi:

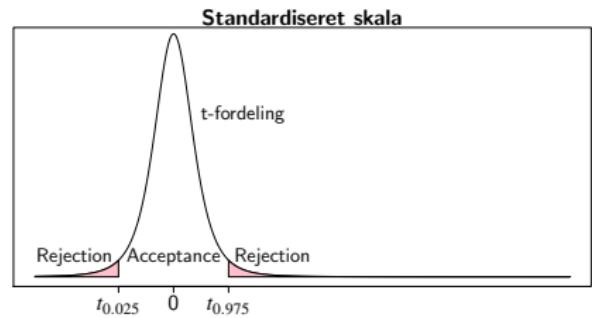
A null hypothesis is *rejected* if the observed test-statistic is more extreme than the critical values:

If  $|t_{\text{obs}}| > t_{1-\alpha/2}$  then *reject*  $H_0$

otherwise *accept*.

## Hypotesetests

Hvis  $t_{\text{obs}}$  er i acceptområdet, så accepteres  $H_0 : \mu = \mu_0$



# Kritisk værdi, konfidensinterval og hypotesetest

Theorem 3.33:

Kritisk-værdi-metode ækvivalent med konfidensinterval-metode

We consider a  $(1 - \alpha) \cdot 100\%$  confidence interval for  $\mu$

$$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

The confidence interval corresponds to the acceptance region for  $H_0$  when testing the (non-directional) hypothesis

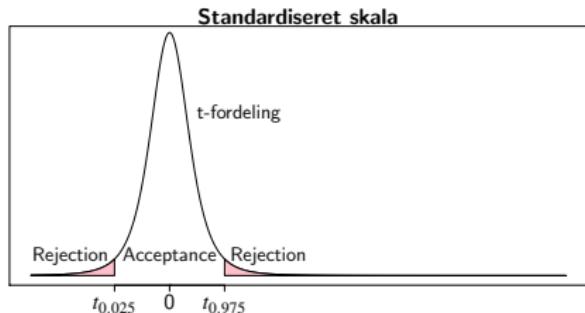
$$H_0 : \mu = \mu_0$$

(Ny) fortolkning af konfidensintervallet:

Nulhypoteser hvor  $\mu_0$  er udenfor konfidensintervallet ville være blevet afvist

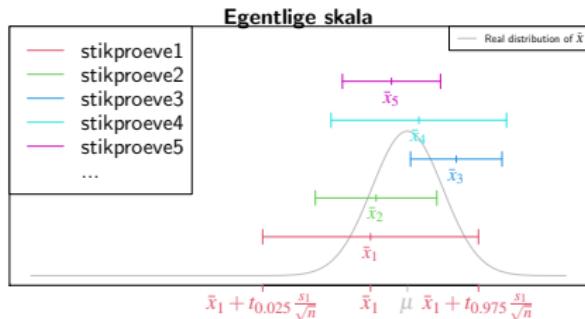
## Hypotesetests

Hvis  $t_{\text{obs}}$  er ude af acceptområdet, så afvises  $H_0 : \mu = \mu_0$



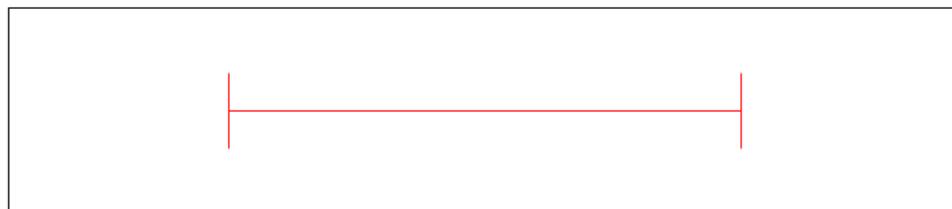
## Konfidensintervallet

Nulhypoteser med  $\mu_0$  udenfor konfidensintervallet ville være blevet afvist



# Spørgsmål om konfidensinterval (socrative.com - ROOM:PBAC)

Afgør på *signifikansniveau*  $\alpha = 5\%$  om en type PC skærm lever op til specifikationen af et effektforbrug på  $\mu = 83$  W. Der er taget en stikprøve af denne type skærm og et 95% konfidensinterval for middelværdien af effektforbruget  $\mu$  er beregnet til:



$$\bar{x} + t_{0.975} \frac{s}{\sqrt{n}} = 85.5$$

$$\bar{x} = 88.3$$

$$\bar{x} - t_{0.025} \frac{s}{\sqrt{n}} = 91.1$$

Hvilken af følgende hypoteser skal testes og hvilken konklusion er korrekt?

- A)  $H_0 : \mu = 0$  accepteres og signifikant højere effektforbrug er påvist
- B)  $H_0 : \mu = 0$  afvises og signifikant højere effektforbrug er påvist
- C)  $H_0 : \mu = 83$  accepteres og signifikant højere effektforbrug er ikke påvist
- D)  $H_0 : \mu = 83$  afvises og signifikant højere effektforbrug er påvist
- E) Ved ikke

# Den alternative hypotese

Den alternative hypotese  $H_1$  er negationen af nulhypotesen  $H_0$

Indtil nu - underforstået: (= non-directional)

Alternativet til  $H_0$ :  $\mu = \mu_0$  er :  $H_1$ :  $\mu \neq \mu_0$

MEN der kan være andre settings, f.eks. one-sided (=directional), less:

Alternativet til  $H_0$ :  $\mu \geq \mu_0$  er  $H_1$ :  $\mu < \mu_0$

I kurset er kun inkluderet opgaver med “non-directional”

# Steps ved hypotesetests - et overblik

Helt generelt består et hypotesetest af følgende trin:

- ① Formuler hypoteserne ( $H_0$  og  $H_1$ ) og vælg signifikansniveau  $\alpha$  (choose the "risk-level")
- ② Beregn med data værdien af teststatistikken
- ③ Beregn  $p$ -værdien med teststatistikken og den relevante fordeling, og sammenlign  $p$ -værdien med signifikansniveauet og drag en konklusion  
*eller*  
Lav konklusionen ved de relevante kritiske værdier

## Metode 3.36: The level $\alpha$ one-sample $t$ -test

- ① Compute  $t_{\text{obs}}$  using Equation (3-21):  $t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
- ② Compute the evidence against the *null hypothesis*

$$H_0: \mu = \mu_0,$$

vs. the *alternative hypothesis*

$$H_1: \mu \neq \mu_0,$$

by the

$$p\text{-value} = 2 \cdot P(T > |t_{\text{obs}}|),$$

where the  $t$ -distribution with  $n - 1$  degrees of freedom is used

- ③ If  $p\text{-value} < \alpha$ : We reject  $H_0$ , otherwise we accept  $H_0$ ,

or

The rejection/acceptance conclusion could alternatively, but equivalently, be made based on the critical value(s)  $\pm t_{1-\alpha/2}$ : If  $|t_{\text{obs}}| > t_{1-\alpha/2}$  we reject  $H_0$ , otherwise we accept  $H_0$

# Mulige fejl ved hypotesetests

To mulige sandheder vs. to mulige konklusioner:

	Reject $H_0$	Fail to reject $H_0$
$H_0$ is true	Type I error ( $\alpha$ )	Correct acceptance of $H_0$
$H_0$ is false	Correct rejection of $H_0$	Type II error ( $\beta$ )

# Eksempel - sovemedicin

To mulige sandheder vs. to mulige konklusioner:

	Reject $H_0$	Fail to reject $H_0$
<i>Sand <math>H_0</math>:</i> Ingen forskel på A og B	Type I fejl ( $\alpha$ )	Korrekt accept af $H_0$
<i>Falsk <math>H_0</math>:</i> Forskel på A og B	Korrekt afvisning af $H_0$	Type II fejl ( $\beta$ )

# Mulige fejl ved hypotesetests

Der findes to slags fejl (dog kun een af gangen!)

Type I: Rejection of  $H_0$  when  $H_0$  is true

Type II: Non-rejection of  $H_0$  when  $H_1$  is true

Risikoen for de to typer fejl kaldes sædvanligvis:

$$P(\text{Type I error}) = \alpha$$

$$P(\text{Type II error}) = \beta$$

Theorem 3.39: "Signifikansniveauet" = "Risikoen for Type I fejl":

The significance level  $\alpha$  in hypothesis testing is the overall Type I risk

$$P(\text{"Type I error"}) = P(\text{"Rejection of } H_0 \text{ when } H_0 \text{ is true"}) = \alpha$$

## Eksempel: Retsalsanalogi

En person står stillet for en domstol:

A man is standing in a court of law accused of criminal activity.

The null- and the alternative hypotheses are:

$H_0$  : The man is not guilty

$H_1$  : The man is guilty

At man ikke kan bevise skyldig er ikke det samme som at man er bevist uskyldig:

Absence of evidence is NOT evidence of absence! Or differently put:

*Accepting a null hypothesis is NOT a statistical proof of the null hypothesis being true!*

# Transport tid

Hypotesetests om studerendes transport tid til DTU fredag morgen

Tag link i meddelelse og indtast din transporttid i dag.

Kan det påvises at transporttiden for studerende på cykel er forskellig fra 20 minutter?

Kan det påvises at transporttiden for studerende på cykel er mere end 20 minutter?

Kan det påvises at tage mere end 20 minutter for studerende i bil?

Find de kritiske værdier for studerende i tog og bus.

# Normalfordelt data?

Teknikker til at undersøge om data kommer fra en normalfordeling:

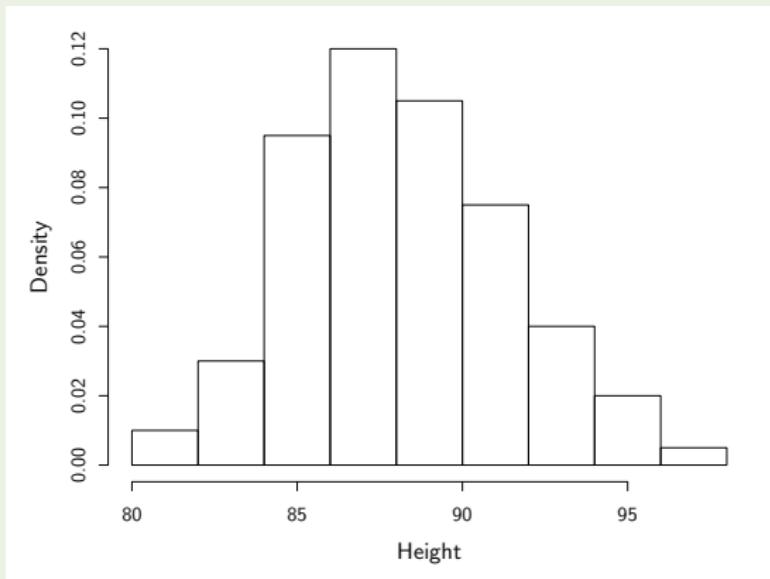
- Empirisk fordelings funktion (ecdf)
- Normal q-q plot

Transformer for at få mere normalfordelt data:

- Brug log-transformation til at få mere normalfordelte observationer

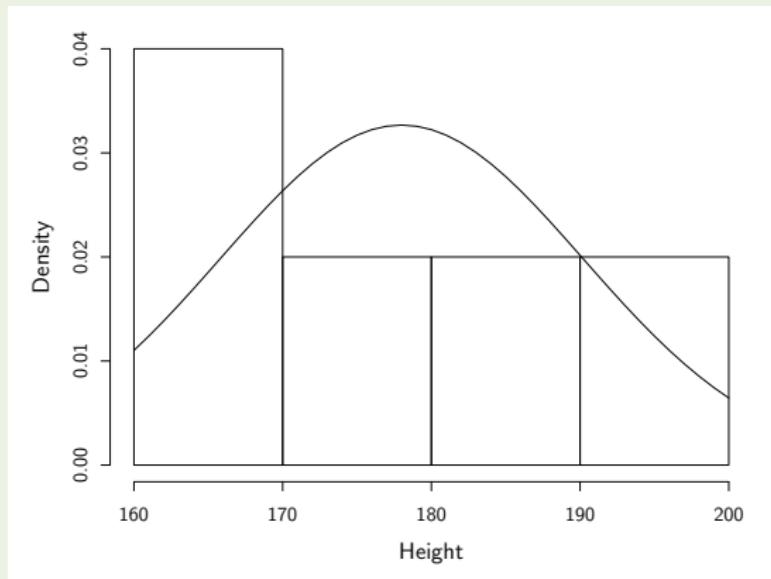
# Se på 100 sim. observationer fra en normal fordeling

```
## 100 simulerede observationer fra normalfordeling
xr <- rnorm(100, mean(x), sd(x))
hist(xr, xlab="Height", main="", prob=TRUE)
lines(seq(130, 230, 1), dnorm(seq(130, 230, 1), mean(x), sd(x)))
```



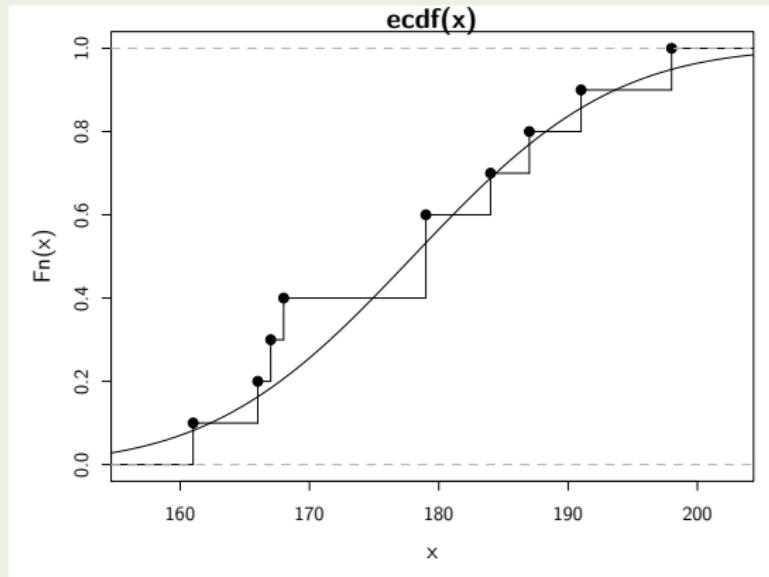
# Højde af studerende - er de normalfordelt?

```
## Empirisk og teoretisk pdf af højdeeksempel
x <- c(168,161,167,179,184,166,198,187,191,179)
hist(x, xlab="Height", main="", prob=TRUE)
lines(seq(160, 200, 1), dnorm(seq(160, 200, 1), mean(x), sd(x)))
```



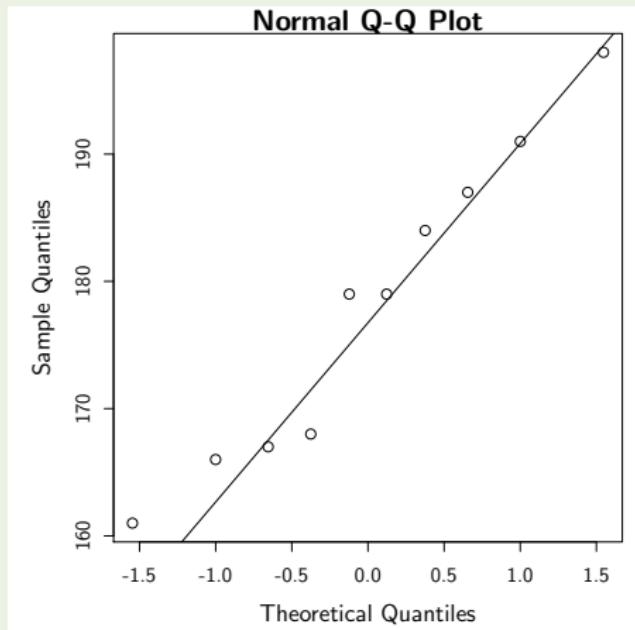
# Højde af studerende - ecdf

```
## Empirisk og teoretisk fordelingsfunktion (ecdf og cdf)
plot(ecdf(x), verticals = TRUE)
xp <- seq(0.9*min(x), 1.1*max(x), length.out = 100)
lines(xp, pnorm(xp, mean(x), sd(x)))
```



# Højde af studerende - Normal q-q plot

```
## q-q plot
qqnorm(x)
qqline(x)
```



# Normal q-q plot

## Metode 3.42 – Den formelle definition

The ordered observations  $x_{(1)}, \dots, x_{(n)}$ , called the sample quantiles, are plotted versus a set of expected normal quantiles  $z_{p_1}, \dots, z_{p_n}$ .

The usual definition of  $p_1, \dots, p_n$  to be used for finding the expected normal quantiles is

$$p_i = \frac{i - 0.5}{n}, \quad i = 1, \dots, n.$$

This is the default method in the `qqnorm` function in R, when  $n > 10$ , if  $n \leq 10$  instead

$$p_i = \frac{i - 3/8}{n + 1/4}, \quad i = 1, \dots, n,$$

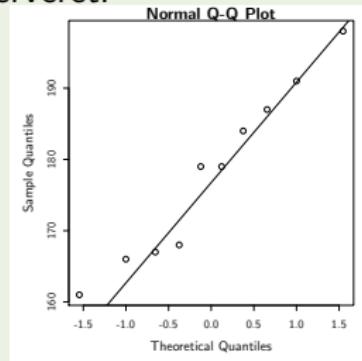
is used.

# Normal q-q plot

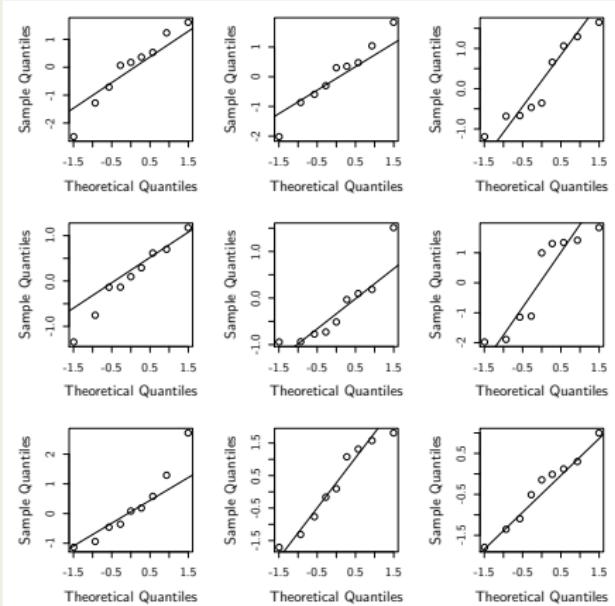
Vurder om de ligger på en ret linje:

Er observeret tydeligt forskelligt simulerede normalfordelte stikprøver?

Observeret:

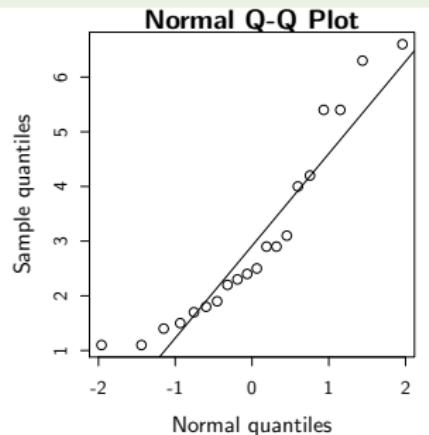
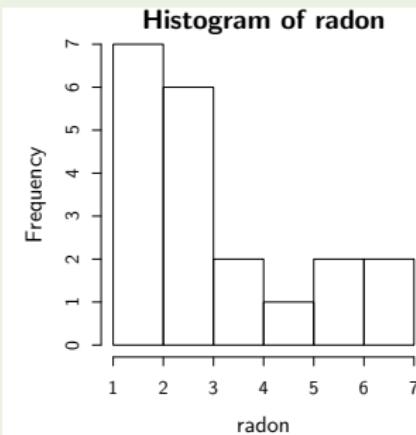


Simulerede:



# Eksempel - log-transformation af Radon data

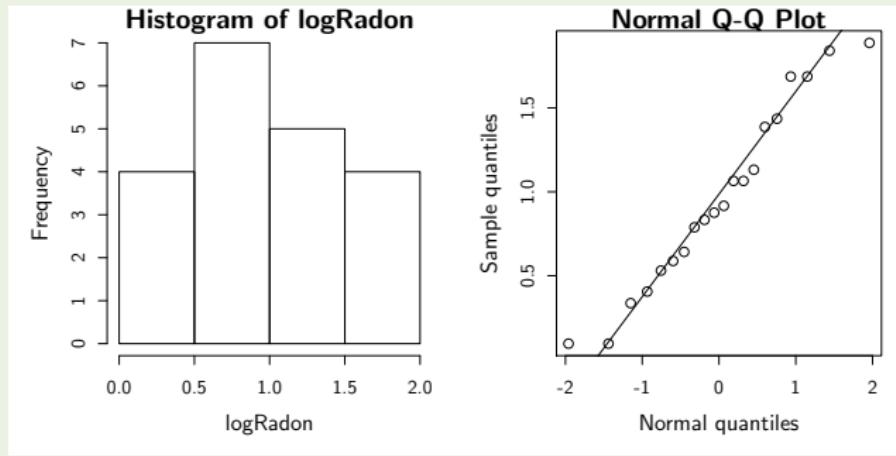
```
## Reading in the data
radon<-c(2.4, 4.2, 1.8, 2.5, 5.4, 2.2, 4.0, 1.1, 1.5, 5.4, 6.3,
       1.9, 1.7, 1.1, 6.6, 3.1, 2.3, 1.4, 2.9, 2.9)
## A histogram and q-q plot
par(mfrow=c(1,2))
hist(radon)
qqnorm(radon, ylab = "Sample quantiles", xlab = "Normal quantiles")
qqline(radon)
```



# Eksempel - Radon data - log-transformed are closer to a normal distribution

```
## Transformer med naturlig logaritme
logRadon <- log(radon)

hist(logRadon)
qqnorm(logRadon, ylab="Sample quantiles", xlab="Normal quantiles")
qqline(logRadon)
```



- Midtvejsevaluering launches lige om lidt...skyd løs!
- Projekterne, går det fremad?

# Kursus 02323: Introducerende Statistik

## Forelæsning 6: Sammenligning af to populationer

Peder Bacher

DTU Compute, Dynamiske Systemer  
Bygning 303B, Rum 010  
Danmarks Tekniske Universitet  
2800 Lyngby – Danmark  
e-mail: pbac@dtu.dk

Forår 2021

## Kapitel 3: Statistik for to populationer (2 stikprøver)

Specifikke metoder, to populationer:

- Konfidensinterval for forskel i middelværdi
- Test for forskel i middelværdi ( $t$ -test)
- To PARREDE grupper: "Tag differencen"  $\Rightarrow$  "Én gruppe"

# Chapter 3: Two Samples

Specific methods, two samples:

- Confidence interval for the mean difference
- Test for the mean difference ( $t$ -test)
- Two PAIRED samples: "Take difference"  $\Rightarrow$  "One sample"

# Oversigt

- 1 Motiverende eksempel - energiforbrug
- 2 Hypotesetest (Repetition)
- 3 Two-sample  $t$ -test og  $p$ -værdi
- 4 Konfidensinterval for forskellen
- 5 Overlappende konfidensintervaller?
- 6 Det parrede setup
- 7 Checking the normality assumptions
- 8 The pooled t-test - a possible alternative

# Motiverende eksempel - energiforbrug

## Forskel på energiforbrug?

I et ernæringsstudie ønsker man at undersøge om der er en forskel i energiforbrug for forskellige typer (moderat fysisk krævende) arbejde. In the study, the energy usage of 9 nurses from Hospital A and 9 (other) nurses from Hospital B have been measured. The measurements are given in the following table in mega Joule (MJ).

### Stikprøve fra hver hospital

$n_1 = n_2 = 9$ :

	Hospital A	Hospital B
	7.53	9.21
	7.48	11.51
	8.08	12.79
	8.09	11.85
	10.15	9.97
	8.40	8.79
	10.88	9.69
	6.13	9.68
	7.90	9.19

# Eksempel - energiforbrug

Hypotesen om ingen forskel ønskes undersøgt:

$$H_0 : \mu_1 = \mu_2$$

Sample means og standard deviations:

$$\hat{\mu}_1 = \bar{x}_1 = 8.293, (s_1 = 1.428)$$

$$\hat{\mu}_2 = \bar{x}_2 = 10.298, (s_2 = 1.398)$$

NYT: *p*-værdi for forskel:

$$p\text{-værdi} = 0.0083$$

(Beregnet under det scenarie, at  $H_0$  er sand)

Er data i overenstemmelse med nulhypotesen  $H_0$ ?

$$\text{Data: } \bar{x}_2 - \bar{x}_1 = 2.005$$

$$\text{Nulhypotese: } H_0 : \mu_2 - \mu_1 = 0$$

NYT: Konfidensinterval for forskel:

$$2.005 \pm 1.412 = [0.59; 3.42]$$

# Steps ved hypotesetests - et overblik (repetition)

Helt generelt består et hypotesetest af følgende trin:

- ① Formuler hypoteserne ( $H_0$  og  $H_1$ ) og vælg signifikansniveau  $\alpha$  (choose the "risk-level")
- ② Beregn med data værdien af teststatistikken
- ③ Beregn  $p$ -værdien med teststatistikken og den relevante fordeling, og sammenlign  $p$ -værdien med signifikansniveauet og drag en konklusion  
*eller*

Lav konklusionen ved de relevante kritiske værdier

# Definition og fortolkning af $p$ -værdien (repetition)

Definition 3.22 af  $p$ -værdien:

**The  $p$ -value** is the probability of obtaining a test statistic that is at least as extreme as the test statistic that was actually observed. This probability is calculated under the assumption that the null hypothesis is true.

$p$ -værdien udtrykker *evidence* imod nulhypotesen – Tabel 3.1:

$p < 0.001$	Very strong evidence against $H_0$
$0.001 \leq p < 0.01$	Strong evidence against $H_0$
$0.01 \leq p < 0.05$	Some evidence against $H_0$
$0.05 \leq p < 0.1$	Weak evidence against $H_0$
$p \geq 0.1$	Little or no evidence against $H_0$

## Metode 3.49: Two-sample $t$ -test

### Beregning af teststørrelsen

When considering the null hypothesis about the difference between the means of two *independent* samples

$$\delta = \mu_2 - \mu_1 \quad (\text{delta er forskellen i middelværdi})$$
$$H_0 : \delta = \delta_0 \quad (\text{typisk er } \delta_0 = 0)$$

the (Welch) two-sample  $t$ -test statistic is

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

## Theorem 3.50: Fordelingen af (Welch) $t$ -teststørrelsen

Welch  $t$ -teststørrelsen er  $t$ -fordelt

The (Welch) two-sample statistic seen as a random variable

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

approximately, under the null hypothesis, follows a  $t$ -distribution with  $v$  degrees of freedom, where

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

if the two population distributions are normal or if the two sample sizes are large enough.

Metode 3.51: The level  $\alpha$  two-sample  $t$ -test

- 1 Compute the test statistic using Equation (3-48) and  $v$  from Equation (3-50)

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \text{ and } v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

- 2 Compute the evidence against the *null hypothesis*

$$H_0: \mu_1 - \mu_2 = \delta_0,$$

vs. the *alternative hypothesis*

$$H_1: \mu_1 - \mu_2 \neq \delta_0,$$

by the

$$p\text{-value} = 2 \cdot P(T > |t_{\text{obs}}|),$$

where the  $t$ -distribution with  $v$  degrees of freedom is used

- 3 If  $p$ -value  $< \alpha$ : we reject  $H_0$ , otherwise we accept  $H_0$ ,

or

The rejection/acceptance conclusion can equivalently be based on the critical value(s)  $\pm t_{1-\alpha/2}$ :

if  $|t_{\text{obs}}| > t_{1-\alpha/2}$  we reject  $H_0$ , otherwise we accept  $H_0$

# Spørgsmål til fordelingen af forskellen i stikprøvegennemsnit (socrative.com - ROOM:PBAC)

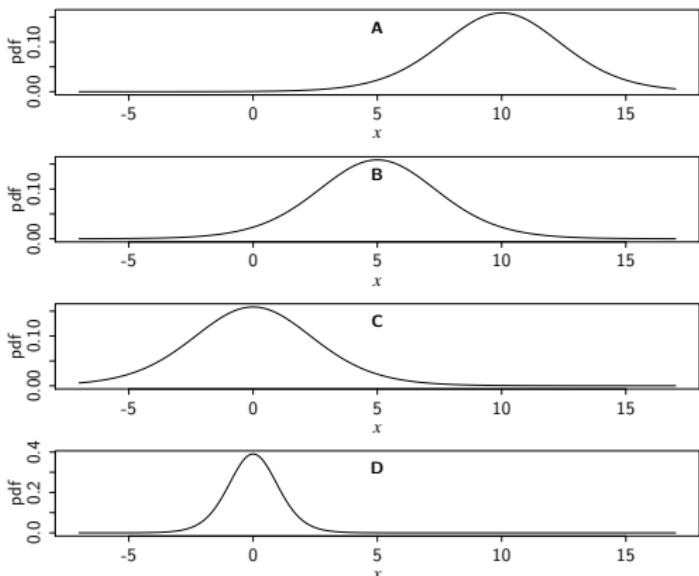
Hvilken af pdf'erne repræsenterer fordelingen af forskellen i stikprøvegennemsnit?

$$\bar{X}_2 - \bar{X}_1$$

*UNDER (dvs. antag er sand):*

$$H_0 : \delta = 10$$

(sample sizes  $n_1 = 7$  and  $n_2 = 8$ )  
 (sample std. dev.  $s_1 = 18$  and  $s_2 = 24$ )



A B C eller D?

# Spørgsmål til fordelingen af forskellen i stikprøvegennemsnit (socrative.com - ROOM:PBAC)

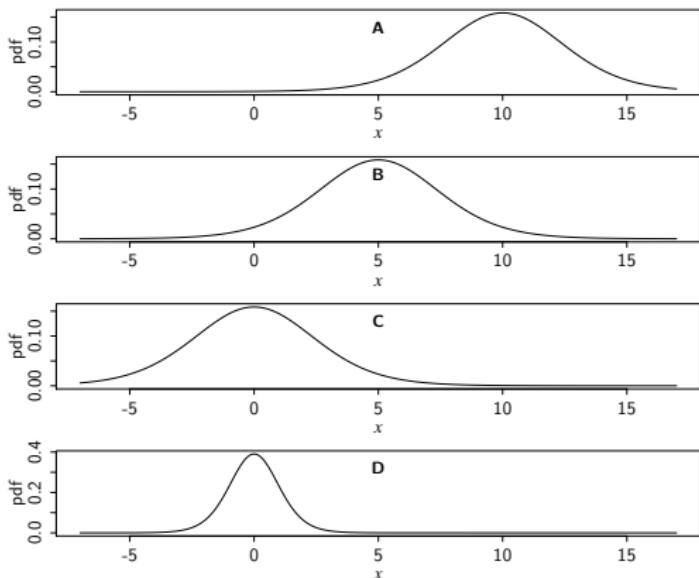
Hvilken af pdf'erne repræsenterer fordelingen af

$$\bar{X}_2 - \bar{X}_1 - \delta_0$$

under:

$$H_0 : \delta = 10$$

(sample sizes  $n_1 = 7$  and  $n_2 = 8$ )  
 (sample std. dev.  $s_1 = 18$  and  $s_2 = 24$ )



A B C eller D?

# Spørgsmål til fordelingen af forskellen i stikprøvegennemsnit (socrative.com - ROOM:PBAC)

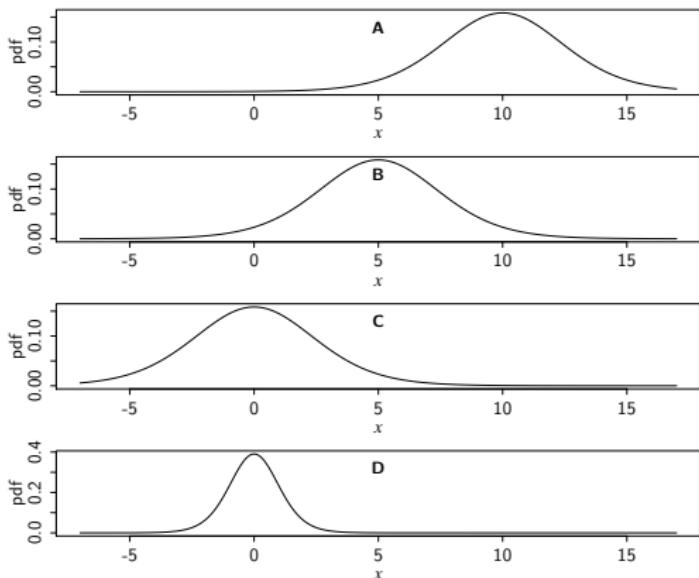
Hvilken af pdf'erne repræsenterer fordelingen af

$$T = \frac{\bar{X}_2 - \bar{X}_1 - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

under:

$$H_0 : \delta = 10$$

(sample sizes  $n_1 = 7$  and  $n_2 = 8$ )  
 (sample std. dev.  $s_1 = 18$  and  $s_2 = 24$ )



A B C eller D?

# Eksempel - energiforbrug

Hypotesen om ingen forskel ønskes undersøgt

$$H_0 : \delta = \mu_2 - \mu_1 = 0$$

versus the alternative

$$H_1 : \delta = \mu_2 - \mu_1 \neq 0$$

Først beregninger af  $t_{\text{obs}}$  og  $v$ :

$$t_{\text{obs}} = \frac{10.298 - 8.293}{\sqrt{2.0394/9 + 1.954/9}} = 3.01$$

and

$$v = \frac{\left(\frac{2.0394}{9} + \frac{1.954}{9}\right)^2}{\frac{(2.0394/9)^2}{8} + \frac{(1.954/9)^2}{8}} = 15.99$$

# Eksempel - energiforbrug

Dernæst findes *p*-værdien:

$$p\text{-value} = 2 \cdot P(T > |t_{\text{obs}}|) = 2 \cdot P(T > 3.01) = 2 \cdot 0.00415 = 0.0083$$

```
## p-værdi for nulhypotese om ingen forskel mellem sygeplejeskers energiforbrug
2 * (1 - pt(3.01, df = 15.99))

## [1] 0.0083
```

# Eksempel - energiforbrug - brug funktion i R:

```
#####
## t-test for forskel i middelværdi på sygeplejeskers energiforbrug
xA <- c(7.53, 7.48, 8.08, 8.09, 10.15, 8.4, 10.88, 6.13, 7.9)
xB <- c(9.21, 11.51, 12.79, 11.85, 9.97, 8.79, 9.69, 9.68, 9.19)
## Default i t.test() er  $H_0$ :  $\mu_1 = \mu_2$  (ingen forskel i middelværdi)
t.test(xB, xA)

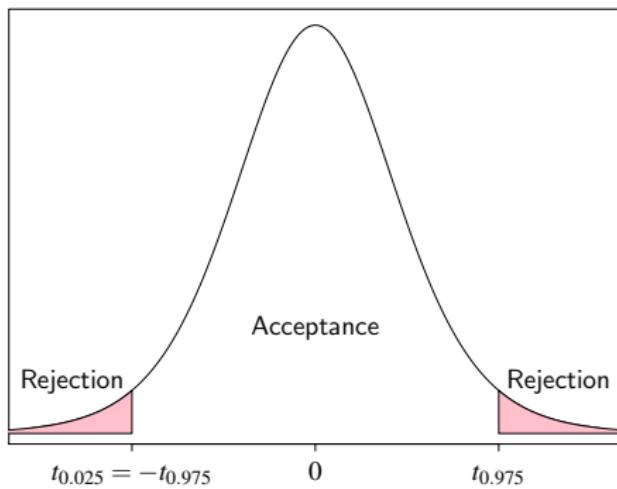
##
##  Welch Two Sample t-test
##
## data: xB and xA
## t = 3.009, df = 15.99, p-value = 0.00832
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.59228 3.41661
## sample estimates:
## mean of x mean of y
## 10.29778 8.29333
```

I pausen: Installer *Space Frontier* (ikke 2'eren) på jeres device (Android eller iphone), men vent med at spille.

(Spring igennem menuer, så I ikke giver dem data)!

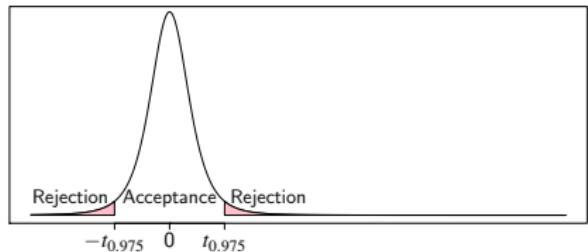
# Kritiske værdier og hypotesetest

Acceptområdet er værdier for teststatistikken  $t_{\text{obs}}$  som ligger indenfor de kritiske værdier:



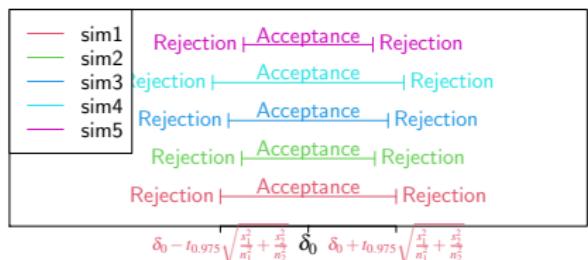
## Den standardiserede skala

Hvis  $t_{\text{obs}}$  er i acceptområdet, så accepteres  $H_0$



## Den egentlige skala

Hvis  $\bar{x} - \bar{y}$  er i acceptområdet, så accepteres  $H_0$



## Metode 3.47: Konfidensinterval for $\mu_1 - \mu_2$

Konfidensintervallet for middelforskellen bliver:

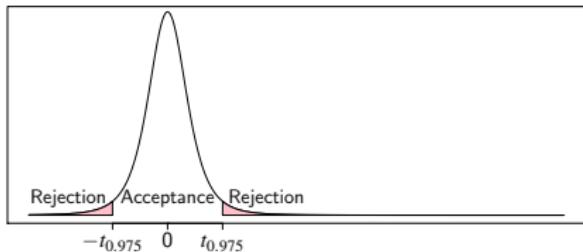
For two samples  $x_1, \dots, x_{n_1}$  and  $y_1, \dots, y_{n_2}$  the  $100(1 - \alpha)\%$  confidence interval for  $\mu_1 - \mu_2$  is given by

$$\bar{x} - \bar{y} \pm t_{1-\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where  $t_{1-\alpha/2}$  is the  $100(1 - \alpha/2)\%$ -quantile from the  $t$ -distribution with  $v$  degrees of freedom given from Equation (3-50).

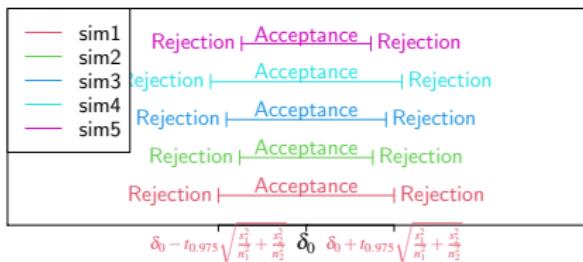
## Den standardiserede skala

Hvis  $t_{\text{obs}}$  er i acceptområdet, så accepteres  $H_0$



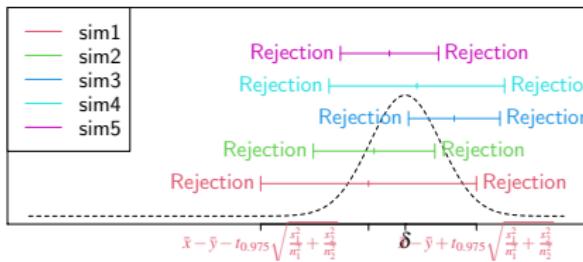
## Den egentlige skala

Hvis  $\bar{x} - \bar{y}$  er i acceptområdet, så accepteres  $H_0$



## Konfidensintervallet

Nulhypoteser med  $\delta_0$  udenfor konfidensintervallet ville være blevet afvist



## Eksempel - energiforbrug - det hele i R:

Let us find the 95% confidence interval for  $\mu_2 - \mu_1$ :

Since the relevant  $t$ -quantile is, using  $v = 15.99$ ,

$$t_{0.975} = 2.120$$

the confidence interval becomes:

$$10.298 - 8.293 \pm 2.120 \cdot \sqrt{\frac{2.0394}{9} + \frac{1.954}{9}}$$

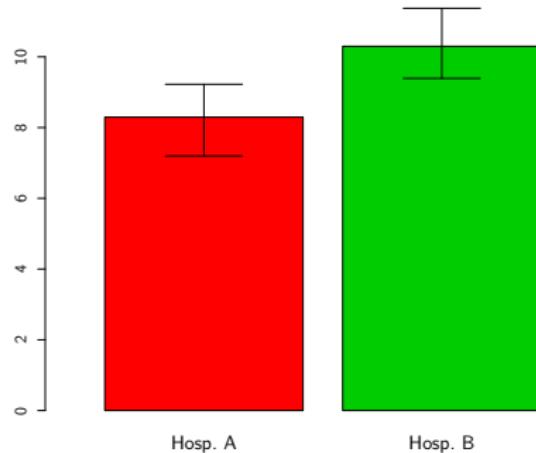
which then gives the result as also seen above:

$$[0.59; 3.42]$$

# Eksempel - energiforbrug - Præsentation af resultat

Barplot med *error bars* ses ofte

Et grupperet barplot med nogle "error bars" - herunder er 95%-konfidensintervallerne for hver gruppe vist:



# Vær varsom med at bruge "overlappende konfidensintervaller"

Remark 3.73. Regel for brug af "overlappende konfidensintervaller":

When two CIs DO NOT overlap: The two groups are significantly different

When two CIs DO overlap: We do not know what the conclusion is

# Motiverende eksempel - sovemedicin

## Forskel på sovemedicin?

I et studie er man interesseret i at sammenligne 2 sovemedler  $A$  og  $B$ . For 10 testpersoner har man fået følgende resultater, der er givet i forlænget søvntid (i timer) (Forskellen på effekten af de to midler er angivet):

Stikprøve,  $n = 10$ :

Person	$A$	$B$	$D = B - A$
1	+0.7	+1.9	+1.2
2	-1.6	+0.8	+2.4
3	-0.2	+1.1	+1.3
4	-1.2	+0.1	+1.3
5	-1.0	-0.1	+0.9
6	+3.4	+4.4	+1.0
7	+3.7	+5.5	+1.8
8	+0.8	+1.6	+0.8
9	0.0	+4.6	+4.6
10	+2.0	+3.4	+1.4

# Parret setup og analyse: Brug one-sample analyse

```
## Det parrede setup: Tag forskellen og brug one-sample test
x1 <- c(.7,-1.6,-.2,-1.2,-1,3.4,3.7,.8,0,2)
x2 <- c(1.9,.8,1.1,.1,-.1,4.4,5.5,1.6,4.6,3.4)
dif <- x2-x1
t.test(dif)

##
## One Sample t-test
##
## data: dif
## t = 5, df = 9, p-value = 0.001
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 0.86 2.48
## sample estimates:
## mean of x
## 1.7
```

# Parret setup og analyse: Brug one-sample analyse

```
## Eller angiv at testen er parret med "paired=TRUE"
t.test(x2, x1, paired=TRUE)

##
##  Paired t-test
##
## data: x2 and x1
## t = 5, df = 9, p-value = 0.001
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.86 2.48
## sample estimates:
## mean of the differences
##                           1.7
```

# Parret versus independent eksperiment

## Completely Randomized (independent samples)

20 patients are used and completely at random allocated to one of the two treatments (but usually making sure to have 10 patients in each group). Hence: *different persons in the different groups.*

## Paired (dependent samples)

10 patients are used, and each of them tests both of the treatments. Usually this will involve some time in between treatments to make sure that it becomes meaningful, and also one would typically make sure that some patients do A before B and others B before A. (and doing this allocation at random). Hence: *the same persons in the different groups.*

# Eksempel - Sovemedicin - FORKERT analyse

```
##  
## Welch Two Sample t-test  
##  
## data: x1 and x2  
## t = -2, df = 18, p-value = 0.07  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -3.49 0.15  
## sample estimates:  
## mean of x mean of y  
## 0.66 2.33
```

# Undersøgelse af computerspil

Undersøgelse om et computerspil er designet så man forbedrer sig når man spiller:

- Forsøg: Personer spiller samme bane i spillet tre gange i træk
- Nogle har spillet det før og er derfor erfarne. Alle angiver deres erfaring ved: 'nybegynder', 'mellem' og 'øvet'
- Scoren måles for hver person de tre gange de spiller banen

Der testes for forskellen mellem *nybegyndere* og *øvede personer*:

Hvilket setup skal benyttes? A: Parret B: Ikke parret C: Ved ikke

Der testes for forskellen i score *fra første til tredje gang de spiller banen*:

Hvilket setup skal benyttes? A: Parret B: Ikke parret C: Ved ikke

# Undersøgelse af computerspil

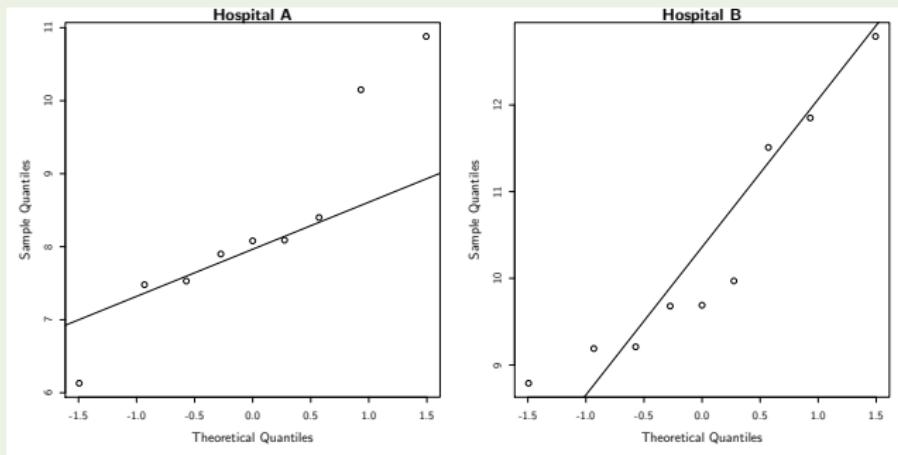
Gå ind i første level og spil i så indtil der bliver sagt stop. Noter bedste score. Dette gentager vi 3 gange ialt.

Download "analyserGame.R" (følg link under uge6 på "Course material"):

- Kan der påvises en signifikant forskel fra *nybegyndere* til *meget øvede* på  $\alpha = 5\%$  niveau?
- Kan der påvises en signifikant forbedring mellem første og tredje gang banen spilles på  $\alpha = 5\%$  niveau?

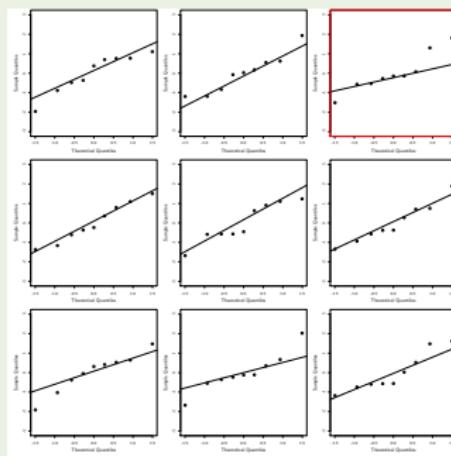
# Eksempel: q-q plot inden for hver stikprøve

```
## Check af normalitetsantagelsen med q-q plots
par(mfrow=c(1,2))
qqnorm(xA, main="Hospital A")
qqline(xA)
qqnorm(xB, main="Hospital B")
qqline(xB)
```



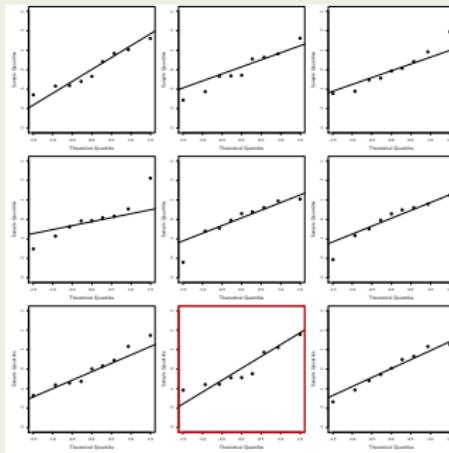
# Eksempel - Sammenligning med simulerede, A

```
## Define the plotting function
qqwrap <- function(x, y, ...){
  stdy <- (y-mean(y))/sd(y)
  qqnorm(stdy, main="", ...)
  qqline(stdy)}
## Do the Wally plot
wallyplot(xA, FUN=qqwrap, ylim=c(-3,3))
```



# Eksempel - Sammenligning med simulerede, B

```
## Check af normalitetsantagelsen med q-q plots og Wally-plot
## Do the Wally plot
wallyplot(xB, FUN=qqwrap, ylim=c(-3,3))
```



## Metode 3.52: The pooled two-sample estimate of variance

### Det poolede variansestimat

Under the assumption that  $\sigma_1^2 = \sigma_2^2$  the *pooled* estimate of variance is the weighted average of the two sample variances

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

## Metode 3.53: The pooled two-sample $t$ -test statistic

### Beregning af den poolede teststørrelse

When considering the null hypothesis about the difference between the means of two *independent* samples

$$\delta = \mu_2 - \mu_1$$

$$H_0: \delta = \delta_0$$

the pooled two-sample  $t$ -test statistic is

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}$$

## Theorem 3.54: Fordelingen af den poolede teststørrelse

Fordelingen af den poolede teststørrelse er en *t*-fordeling

The pooled two-sample statistic seen as a random variable

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{S_p^2/n_1 + S_p^2/n_2}}$$

follows, under the null hypothesis and under the assumption that  $\sigma_1^2 = \sigma_2^2$ , a *t*-distribution with  $n_1 + n_2 - 2$  degrees of freedom if the two population distributions are normal.

# Vi bruger altid "Welch" versionen (den "ikke-poolede")

Nogenlunde (idiot)sikkert at bruge Welch-versionen altid

- if  $s_1^2 = s_2^2$  the Welch and the Pooled test statistics are the same
- Only when the two variances become really different the two test-statistics may differ in any important way, and if this is the case, we would not tend to favour the pooled version, since the assumption of equal variances appears questionable then
- Only for cases with a small sample sizes in at least one of the two groups the pooled approach may provide slightly higher power if you believe in the equal variance assumption. And for these cases the Welch approach is then a somewhat cautious approach

# Kursus 02323: Introducerende Statistik

## Forelæsning 7: Simuleringsbaseret statistik

Peder Bacher

DTU Compute, Dynamiske Systemer  
Bygning 303B, Rum 010  
Danmarks Tekniske Universitet  
2800 Lyngby – Danmark  
e-mail: pbac@dtu.dk

Forår 2021

# Kapitel 4: Statistik ved simulering

## Simulering:

- Træk tilfældige værdier og beregn statistik mange gange
- Fejlforplantning (error propagation rules)  
*(F.eks. igennem ikke-lineær funktion)*
- Bootstrapping af konfidensintervaller:
  - Parametrisk (*Simuler mange udfald af stokastisk var.*)
  - Ikke-parametrisk (*Træk direkte fra data*)

## Specifikke bootstrap setups: (4 versioner af konfidensintervaller)

- Én gruppe/stikprøve og to grupper/stikprøver data
- Parametrisk vs. ikke-parametrisk

# Chapter 4: Statistics by simulation

## Simulation:

- Draw random values and calculate the statistic many times
- Error propagation rules  
(e.g. *through a non-linear function*)
- Bootstrapping of confidence intervals:
  - Parametric (*Simulate many outcomes of random var.*)
  - Non-parametric (*Draw values directly from data*)

## Specific bootstrap set ups: (4 versions of confidence intervals)

- One-sample and Two-sample data
- Parametric vs. non-parametric

# Oversigt

- 1 Introduktion til simulation
  - Hvad er simulering egentlig?
- 2 Fejlphobningslove
- 3 Parametrisk bootstrap
  - Introduction to bootstrap
  - One-sample konfidensinterval for  $\mu$
  - One-sample konfidensinterval for en vilkårlig størrelse
  - Two-sample konfidensintervaller for en vilkårlig fordeling
- 4 Ikke-parametrisk bootstrap
  - One-sample konfidensinterval for  $\mu$
  - One-sample konfidensinterval for en vilkårlig størrelse
  - Two-sample konfidensintervaller

# Motivation

- Mange relevant statistikker ("computed features") har komplikerede samplingfordelinger:
  - Medianen
  - Fraktiler generelt, dvs. f.eks. også  $IQR = Q_3 - Q_1$
  - Enhver ikke-lineær funktion af en eller flere input variable
  - ...
- Populations (og stikprøve) fordelingen kan være ikke-normal (komplicerer den statistiske teori)
- MEN: Nogle gange kan vi ikke være helt sikre på om det er godt nok - simulering kan hjælpe til at verificere!
- Kræver: Brug af computer - R er et super værktøj til dette!

# Anvendelser

*Stokastisk* simulering anvendes mange steder:

- Trafiks simulering
- Kø simulering, f.eks. call-center
- Agent baseret simulering, f.eks. evakuering og markeder
- ...

Generelt, kan bruges til at modellere komplekse stokastiske processer ved generere tilfældige udfald

# Hvad er simulering egentlig?

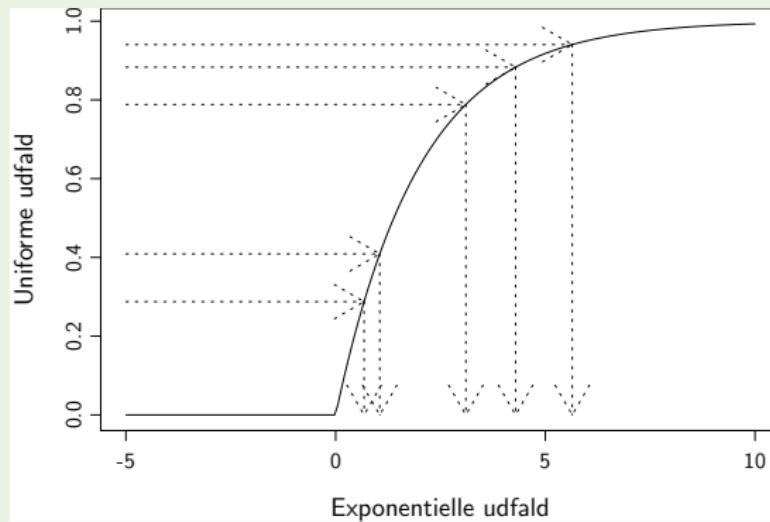
- (Pseudo)tilfældige tal genereret af en computer
- En tilfældighedsgenerator er en algoritme der kan generere  $x_{i+1}$  ud fra  $x_i$
- En sekvens af tal "ser tilfældige ud"
- Kræver en "start" - kaldet "seed" (Bruger typisk uret i computeren)
- Kapitel 2.6: Grundlæggende simuleres den uniforme fordeling, og så bruges:

Remark 2.51:

Hvis  $U \sim \text{Uniform}(0, 1)$  og  $F$  er en fordelingsfunktion for en eller anden sandsynlighedsfordeling, så vil  $F^{-1}(U)$  følge fordelingen givet ved  $F$

# Eksempel: Exponentialfordelingen med $\lambda = 0.5$

$$F(x) = \int_0^x f(t)dt = 1 - e^{-0.5x}$$



# Øvelse: Simuler på papir

Tegn på papir (el. pc)

- Tegn eksponentiel fordelingsfunktion (altså cdf:  $F(x)$ )
- Træk uniform fordelte værdier og “put dem igennem”  $F^{-1}(x)$

Gentag med normal fordelingsfunktion

# I praksis i R

De forskellige fordelinger er gjort klar til simulering:

---

rbinom	Binomialfordelingen
rpois	Poissonfordelingen
rhyper	Den hypergeometriske fordeling
rnorm	Normalfordelingen
rlnorm	Lognormalfordelingen
rexp	Eksponentialfordelingen
runif	Den uniforme(lige) fordeling
rt	t-fordelingen
rchisq	$\chi^2$ -fordelingen
rf	F-fordelingen

---

# Eksempel: Simuler 100 binomialfordelte værdier

```
## Simuler fra nogle fordelinger

## Sæt et seed for at få samme udfald hver gang
set.seed(123)

## 100 realisationer fra binomialfordelingen:
## Antal successer fra 25 trækninger med 0.2 sandsynlighed for succes
rbinom(n=100, size=25, prob=0.2)

## Exponential fordelte
hist(rexp(n=100, rate=2), prob=TRUE)
## Plot the theoretical pdf
xseq <- seq(0,10,len=1000)
lines(xseq, dexp(xseq, rate=2))
```

# Eksempel: Areal af plader

En virksomhed producerer rektangulære plader

- Bredden  $X$  af pladerne (i meter) antages at kunne beskrives med en normalfordeling  $N(2, 0.01^2)$  og længden  $Y$  af pladerne (i meter) antages at kunne beskrives med en normalfordeling  $N(3, 0.02^2)$
- Man er interesseret i arealet, som jo så givet ved

$$A = XY$$

Spørgsmål som kan stilles er f.eks.:

- Hvor ofte sådanne plader har et areal, der afviger mere end  $0.1\text{m}^2$  fra de  $6\text{m}^2$ ?
- Sandsynligheden for andre mulige hændelser?
- Generelt: Hvad er *sandsynlighedsfordelingen* for  $A$ ?

# Eksempel: Areal af plader

- Hvor ofte sådanne plader har et areal, der afviger mere end  $0.1\text{m}^2$  fra de  $6\text{m}^2$ ?

Løsning ved simulering:

```
## Sæt et seed for at få samme udfald hver gang
set.seed(345)
## Antal simuleringer
k = 10000
## Simuler længderne af siderne
x = rnorm(k, 2, 0.01)
y = rnorm(k, 3, 0.02)
## Beregn arealet for hver simulering
area = x*y

## Beregn statistikker
## (stikprøve)Gennemsnit
mean(area)
## (stikprøve)spredning
sd(area)
## Fraktion af arealer afviger mere end 0.1
sum(abs(area-6) > 0.1) / k
```

# Fejlophobning

## Fejlophobning (error propagation)

- Vi har  $n$  stokastiske variabler  $X_1, X_2, \dots, X_n$  og kender deres varianser  $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$
- Og vil finde variansen af en ikke-lineær funktion af dem  $\sigma_{f(X_1, \dots, X_n)}^2 = \text{Var}(f(X_1, \dots, X_n))$

3 måder at løse på for ikke-lineær funktion  $f()$ :

- Simulation (Method 4.4)
- Lineær approksimation (Method 4.3)
- Teoretisk udledning

# Fejlophobning - ved simulering

## Method 4.4: Error propagation by simulation

Assume we have actual measurements  $x_1, \dots, x_n$  with known/assumed error variances  $\sigma_1^2, \dots, \sigma_n^2$ .

- ① Simulate  $k$  outcomes of all  $n$  measurements from assumed error distributions, e.g.  $N(x_i, \sigma_i^2)$ :  $X_i^{(j)}, j = 1, \dots, k$
- ② Calculate the standard deviation directly as the observed standard deviation of the  $k$  simulated values of  $f$

$$s_{f(X_1, \dots, X_n)}^{\text{sim}} = \sqrt{\frac{1}{k-1} \sum_{i=1}^k (f_j - \bar{f})^2}$$

where

$$f_j = f(X_1^{(j)}, \dots, X_n^{(j)})$$

# Eksempel: Varians af areal (Simulering)

Beregn variansen af arealet med simulering:

```
## Beregn variansen af arealet med simulering
## Antal simuleringer
k = 10000
## Simuler længderne af siderne
x = rnorm(k, 2, 0.01)
y = rnorm(k, 3, 0.02)
## Beregn arealet for hver simulering
area = x*y
## Beregn variansen af de simulerede arealer
var(area)
```

# Fejlophobning - ved lineær approksimation

Vi kender allerede regneregel for lineære funktioner (Theorem 2.56)

$$\sigma_{f(X_1, \dots, X_n)}^2 = \sum_{i=1}^n a_i^2 \sigma_i^2 \quad \text{hvis } f(X_1, \dots, X_n) = \sum_{i=1}^n a_i X_i$$

Method 4.3: for ikke-lineære funktioner

$$\sigma_{f(X_1, \dots, X_n)}^2 \approx \sum_{i=1}^n \left( \frac{\partial f}{\partial x_i} \right)^2 \sigma_i^2$$

# Spørgsmål om hvilken metode til beregning af varians

(socrative.com - ROOM:PBAC)

Hvordan beregnes:  $\text{Var}(2^2X - Y)$ ?

- A: Som lineær funktion
- B: Som ikke-lineær funktion
- C: Ved ikke

Hvordan beregnes:  $\text{Var}(X + Y^{-1})$ ?

- A: Som lineær funktion
- B: Som *ikke-lineær* funktion
- C: Ved ikke

## Eksempel: Varians af areal (ikke-lineære fejlophobningslov)

- Vi har allerede brugt eksemplet med areal af en plade
- Nu spørger vi: Hvad er "fejlen" på  $A = 2.00 \times 3.00 = 6.00 \text{ m}^2$  fundet ved den ikke-lineære fejlophobningslov?

# Eksempel: Varians af areal (ikke-lineære fejlophobningslov)

Varianserne er

$$\sigma_X^2 = \text{Var}(X) = 0.01^2 \quad \text{og} \quad \sigma_Y^2 = \text{Var}(Y) = 0.02^2$$

Funktionen og de partielt afledte er

$$f(x, y) = xy, \quad \frac{\partial f}{\partial x} = y, \quad \frac{\partial f}{\partial y} = x$$

Så resultatet bliver

$$\begin{aligned} \text{Var}(A) &\approx \left( \frac{\partial f}{\partial x} \right)^2 \sigma_X^2 + \left( \frac{\partial f}{\partial y} \right)^2 \sigma_Y^2 & \sigma_{f(X_1, \dots, X_n)}^2 &\approx \sum_{i=1}^n \left( \frac{\partial f}{\partial x_i} \right)^2 \sigma_i^2 \\ &= y^2 \sigma_X^2 + x^2 \sigma_Y^2 \\ &= 3.00^2 \cdot 0.01^2 + 2.00^2 \cdot 0.02^2 \\ &= 0.0025 \end{aligned}$$

Method 4.3:

# Eksempel: Varians af areal (Teoretisk udledning)

Faktisk kan man finde variansen for  $A = XY$  teoretisk

$$\begin{aligned}\text{Var}(XY) &= \mathbb{E} \left[ (XY)^2 \right] - [\mathbb{E}(XY)]^2 \\ &= \mathbb{E}(X^2)\mathbb{E}(Y^2) - \mathbb{E}(X)^2\mathbb{E}(Y)^2 \\ &= [\text{Var}(X) + \mathbb{E}(X)^2] [\text{Var}(Y) + \mathbb{E}(Y)^2] - \mathbb{E}(X)^2\mathbb{E}(Y)^2 \\ &= \text{Var}(X)\text{Var}(Y) + \text{Var}(X)\mathbb{E}(Y)^2 + \text{Var}(Y)\mathbb{E}(X)^2 \\ &= 0.01^2 \times 0.02^2 + 0.01^2 \times 3^2 + 0.02^2 \times 2^2 \\ &= 0.00000004 + 0.0009 + 0.0016 \\ &= 0.00250004\end{aligned}$$

# Et summary

Igen: 3 måder til *beregning af varians af ikke-lineær funktion (husk, det er rent teoretisk, dvs. ingen data):*

- ① Simuleringsbaseret
- ② Den analytiske, men approksimative, fejlophobningslov
- ③ Teoretisk udledning

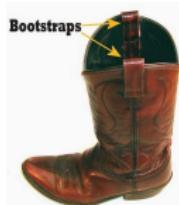
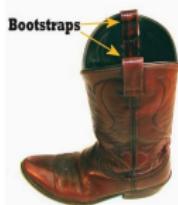
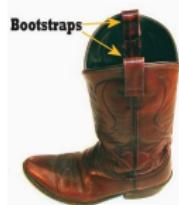
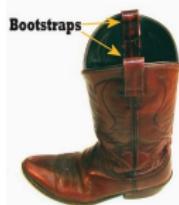
Simulering har en række fordele:

- ① Simpel måde at beregne andre størrelser (kan være mere komplikerede, f.eks. udtryk sværere at differentiere)
- ② Simpel måde at bruge andre fordelinger end normalfordelingen
- ③ Afhænger ikke af en lineær approksimation (som error propagation) til den underliggende ikke-lineære funktion

# Bootstrapping

Bootstrapping findes i to versioner:

- ① Parametrisk bootstrap: Simuler gentagne samples fra en antagede (og estimerede) fordeling
- ② Ikke-parametrisk bootstrap: Simuler gentagne samples direkte fra data



## Eksempel: Konfidensinterval for middelværdien i en eksponentialfordeling

Antag at vi har observeret følgende 10 opkaldsventetider (i sekunder) i et call center

32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0

Vores fordelingsantagelse:

Ventetiderne kommer fra en eksponentialfordeling

Vi estimerer fra data (eksponential fordelingen har en parameter *raten*)

$$\hat{\mu} = \bar{x} = 26.08 \text{ og dermed er raten: } \hat{\lambda} = 1/26.08 = 0.03834356$$

Hvad er konfidensintervallet for  $\mu$ ?

Baseret på tidligere indhold i dette kursus: Det ved vi ikke!

# Eksempel: Konfidensinterval for middelværdien i en eksponentialfordeling

```
## Beregn konfidensinterval for middelværdien med simulering

## Sæt seed hvis sammen resultat ønskes
set.seed(758)

## Første gang: Simuler 10 observationer og beregn gennemsnit
simSample <- rexp(10, 1/26.08)
mean1 <- mean(simSample)

## Anden gang: Simuler 10 observationer og beregn gennemsnit
simSample <- rexp(10, 1/26.08)
mean2 <- mean(simSample)

## Tredje gang: Simuler 10 observationer og beregn gennemsnit
simSample <- rexp(10, 1/26.08)
mean3 <- mean(simSample)

## Gør det 100000 gange!
```

Alright, det må kunne gøres smartere!!

# Eksempel: Konfidensinterval for middelværdien i en eksponentialfordeling

```
## Beregn konfidensinterval for middelværdien med simulering
## Set the number of simulations:
k <- 100000
set.seed(321)

## 1. Simulate 10 exponentials k times and keep in a matrix:
simSamples <- replicate(k, rexp(10, 1/26.08))

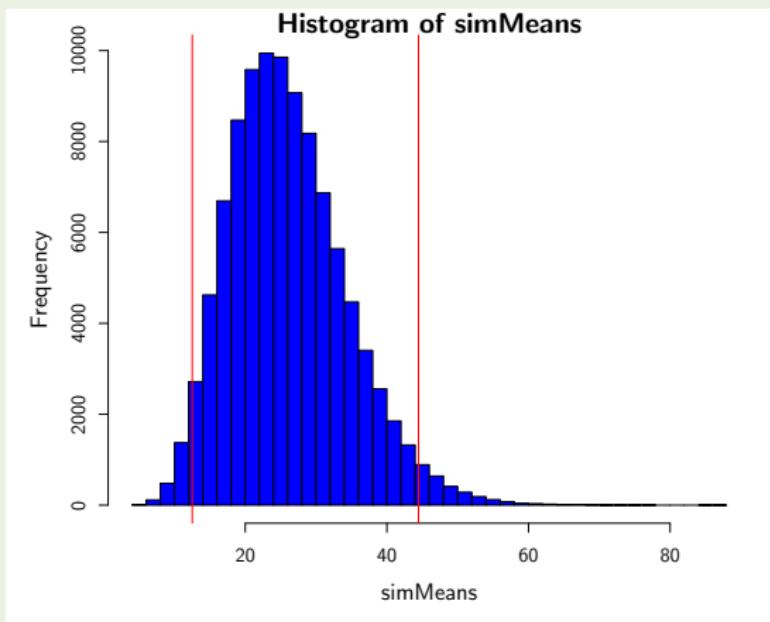
## 2. Compute the mean of the 10 simulated observations k times:
simMeans <- apply(simSamples, 2, mean)

## 3. Find the two relevant quantiles of the k simulated means:
quantile(simMeans, c(0.025, 0.975))

## 2.5% 97.5%
## 12.52 44.47
```

## Example: Konfidensinterval for middelværdien i en eksponentialfordeling

```
hist(simMeans, col="blue", nclass=30)
abline(v=quantile(simMeans, c(0.025, 0.975)), col="red")
```



# Eksempel: Konfidensinterval for medianen i en eksponentialfordeling

Antag at vi har observeret følgende 10 opkaldsventetider (i sekunder) i et call center

32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0

Vores fordelingsantagelse

Ventetiderne kommer fra en eksponentialfordeling

Vi estimerer fra data (som før)

$$\text{Median} = 21.4 \text{ og } \hat{\mu} = \bar{x} = 26.08$$

Hvad er konfidensintervallet for medianen?

Baseret på tidligere indhold i dette kursus: Det ved vi ikke!

# Eksempel: Konfidensinterval for medianen i en eksponentialfordeling

```
## Beregn konfidensinterval for medianen med parametrisk bootstrapping

## Set the number of simulations:
k <- 100000
set.seed(543)

## 1. Simulate 10 exponentials with the right mean k times:
simSamples <- replicate(k, rexp(10, 1/26.08))

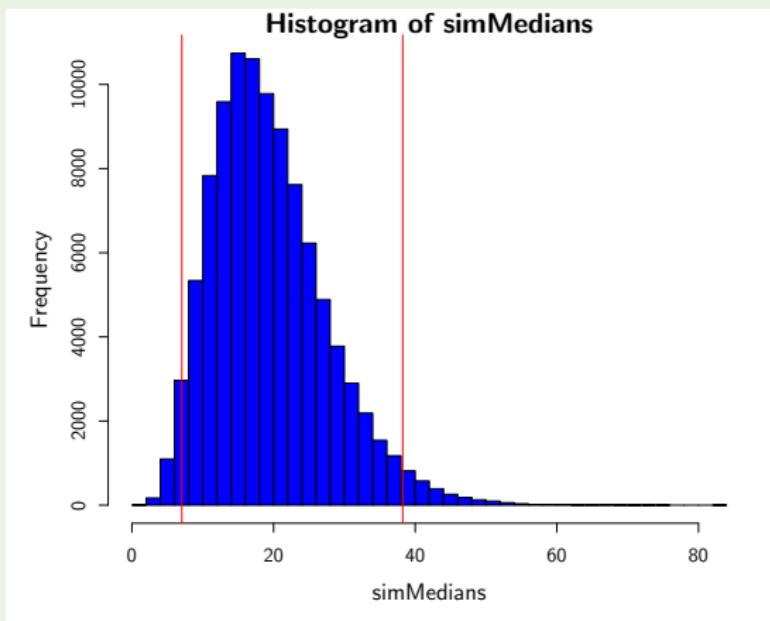
## 2. Compute the median of the n=10 simulated observations k times:
simMedians <- apply(simSamples, 2, median)

## 3. Find the two relevant quantiles of the k simulated medians:
quantile(simMedians, c(0.025, 0.975))

##    2.5% 97.5%
##    7.045 38.235
```

# Eksempel: Konfidensinterval for medianen i en eksponentialfordeling

```
hist(simMedians, col="blue", nclass=30)
abline(v=quantile(simMedians, c(0.025, 0.975)), col="red")
```



# Konfidensinterval for en vilkårlig beregningsstørrelse

Method 4.7: Confidence interval for any feature  $\theta$  by parametric bootstrap

Assume we have actual observations  $x_1, \dots, x_n$  and assume that they stem from some probability distribution with density  $f$ .

- ① Simulate  $k$  samples of  $n$  observations from the assumed distribution  $f$  where the mean<sup>a</sup> is set to  $\bar{x}$
- ② Calculate the statistic  $\hat{\theta}$  in each of the  $k$  samples  $\hat{\theta}_1^*, \dots, \hat{\theta}_k^*$
- ③ Find the  $100(\alpha/2)\%$  and  $100(1 - \alpha/2)\%$  quantiles for these,  $q_{100(\alpha/2)\%}^*$  and  $q_{100(1 - \alpha/2)\%}^*$  as the  $100(1 - \alpha)\%$  confidence interval:  
$$[q_{100(\alpha/2)\%}^*, q_{100(1 - \alpha/2)\%}^*]$$

---

<sup>a</sup>And otherwise chosen to match the data as good as possible: Some distributions have more than just a single mean related parameter, e.g. the normal or the log-normal. For these one should use a distribution with a variance that matches the sample variance of the data. Even more generally the approach would be to match the chosen distribution to the data by the so-called *maximum likelihood* approach

# Et andet eksempel: 99% konfidensinterval for $Q_3$ for en normalfordeling

```
## Konfidensinterval for den øvre kvartil (Q_3) i en normalfordeling
## Read in the heights data:
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
n <- length(x)
## Define a (upper-quartile) Q3-function:
Q3 <- function(x){ quantile(x, 0.75)}
## Set the number of simulations:
k <- 100000
set.seed(543)
## 1. Simulate k samples of n=10 normals with the right mean and variance:
simSamples <- replicate(k, rnorm(n, mean(x), sd(x)))
## 2. Compute the Q3 of the n=10 simulated observations k times:
simQ3s <- apply(simSamples, 2, Q3)
## 3. Find the two relevant quantiles of the k simulated medians:
quantile(simQ3s, c(0.005, 0.995))

## 0.5% 99.5%
## 172.9 198.0
```

# Two-sample konfidensinterval for en vilkårlig feature sammenligning $\theta_X - \theta_Y$ (inkl. $\mu_X - \mu_Y$ )

Method 4.10: Two-sample confidence interval for any feature comparison  $\theta_X - \theta_Y$  by parametric bootstrap

Assume we have actual observations  $x_1, \dots, x_n$  and  $y_1, \dots, y_n$  and assume that they stem from some probability distributions with density  $f_X$  and  $f_Y$ .

- ① Simulate  $k$  sets of 2 samples of  $n_X$  and  $n_Y$  observations from the assumed distributions setting the means <sup>a</sup> to  $\hat{\mu}_X = \bar{x}$  and  $\hat{\mu}_Y = \bar{y}$ , respectively
- ② Calculate the difference between the features in each of the  $k$  samples  $\hat{\theta}_{x1}^* - \hat{\theta}_{y1}^*, \dots, \hat{\theta}_{xk}^* - \hat{\theta}_{yk}^*$
- ③ Find the  $100(\alpha/2)\%$  and  $100(1 - \alpha/2)\%$  quantiles for these,  $q_{100(\alpha/2)\%}^*$  and  $q_{100(1 - \alpha/2)\%}^*$  as the  $100(1 - \alpha)\%$  confidence interval:  
$$[q_{100(\alpha/2)\%}^*, q_{100(1 - \alpha/2)\%}^*]$$

---

<sup>a</sup>As before

# Eksempel: Konfidensinterval for the forskellen mellem to exponentielle middelværdier

```
## Konfidensinterval for the forskellen mellem to exponentielle middelværdier
## Day 1 data:
x <- c(32.6, 1.6, 42.1, 29.2, 53.4, 79.3,
      2.3, 4.7, 13.6, 2.0)
## Day 2 data:
y <- c(9.6, 22.2, 52.5, 12.6, 33.0, 15.2,
      76.6, 36.3, 110.2, 18.0, 62.4, 10.3)
## Keep sample sizes
n1 <- length(x)
n2 <- length(y)
```

# Eksempel: Konfidensinterval for the forskellen mellem to exponentielle middelværdier

```
## Konfidensinterval for the forskellen mellem to exponentielle middelværdier
## Set the number of simulations:
k <- 100000
set.seed(987)
## 1. Simulate k samples of each n1=10 and n2=12
## exponentials with the right means:
simxSamples <- replicate(k, rexp(n1, 1/mean(x)))
simySamples <- replicate(k, rexp(n2, 1/mean(y)))
## 2. Compute the difference between the simulated
## means k times:
simDifMeans <- apply(simxSamples, 2, mean) -
                  apply(simySamples, 2, mean)
## 3. Find the two relevant quantiles of the
## k simulated differences of means:
quantile(simDifMeans, c(0.025, 0.975))

##    2.5% 97.5%
## -40.53 13.94
```

# Parametrisk bootstrap - et overblik

Vi antager en fordeling!

Der er parametre i en fordeling, derfor *parametrisk bootstrap*

To konfidensinterval-metodeboxe blev givet:

	One-sample	Two-sample
For any feature	Method 4.7	Method 4.10

# Ikke-parametrisk bootstrap - et overblik

Vi antager IKKE noget om nogen fordelinger!

Altså der er ingen parametre, derfor *ikke-parametrisk bootstrap*

To konfidensinterval-metodeboxe bliver givet:

	One-sample	Two-sample
For any feature	Method <a href="#">4.15</a>	Method <a href="#">4.17</a>

## Eksempel: Kvinders cigaretforbrug

I et studie undersøgte man kvinders cigaretforbrug før og efter fødsel

Man fik følgende observationer af antal cigaretter pr. dag:

før	efter	før	efter
8	5	13	15
24	11	15	19
7	0	11	12
20	15	22	0
6	0	15	6
20	20		

Sammenlign før og efter! Er der sket nogen ændring i gennemsnitsforbruget!

# Eksempel: Kvinders cigaretforbrug

Et parret  $t$ -test setup, MEN med tydeligvis ikke-normale data!

```
## Parret test af middelværdiforskel med ikke-parametrisk bootstrapping
## Input the two cigaret use samples
x1 <- c(8, 24, 7, 20, 6, 20, 13, 15, 11, 22, 15)
x2 <- c(5, 11, 0, 15, 0, 20, 15, 19, 12, 0, 6)
## Calculate the difference
dif <- x1 - x2
dif

## [1]  3 13  7  5  6  0 -2 -4 -1 22  9

## And the sample mean
mean(dif)

## [1] 5.273
```

# Eksempel: Kvinders cigaretforbrug - bootstrapping

```
#####
## Resample several times
sample(dif, replace = TRUE)

## [1] 7 5 -2 -4 22 13 5 9 -2 -2 9

sample(dif, replace = TRUE)

## [1] 5 9 6 -1 -2 -2 -2 0 13 9 6

sample(dif, replace = TRUE)

## [1] 7 -1 0 22 5 -1 -1 -4 5 -4 -2

sample(dif, replace = TRUE)

## [1] -1 -4 6 22 9 5 -2 9 -2 9 9
```

# Eksempel: Kvinders cigaretforbrug - de ikke-parametriske bootstrap resultater:

```
## Resample calculate mean statistic many time, and find 95% confidence interval
k = 100000
simSamples = replicate(k, sample(dif, replace = TRUE))
## Take the mean for every resample
simMeans = apply(simSamples, 2, mean)
## Take the two quantiles to get the confidence interval
quantile(simMeans, c(0.025,0.975))

## 2.5% 97.5%
## 1.273 9.818
```

# One-sample konfidensinterval for en vilkårlig feature $\theta$ (inkl. $\mu$ )

Method 4.15: Confidence interval for any feature  $\theta$  by non-parametric bootstrap

Assume we have actual observations  $x_1, \dots, x_n$ .

- ① Simulate  $k$  samples of size  $n$  by randomly sampling among the available data (with replacement)
- ② Calculate the statistic  $\hat{\theta}$  in each of the  $k$  samples  $\hat{\theta}_1^*, \dots, \hat{\theta}_k^*$ .
- ③ Find the  $100(\alpha/2)\%$  and  $100(1 - \alpha/2)\%$  quantiles for these,  $q_{100(\alpha/2)\%}^*$  and  $q_{100(1 - \alpha/2)\%}^*$  as the  $100(1 - \alpha)\%$  confidence interval:  
$$[q_{100(\alpha/2)\%}^*, q_{100(1 - \alpha/2)\%}^*]$$

# Eksempel: Kvinders cigaretforbrug

Lad os finde 95% konfidensintervallet for ændringen af median cigaretforbruget

```
## Konfidensintervallet for ændringen af median cigaretforbruget
## Simulate many times
k = 100000
simsamples = replicate(k, sample(dif, replace = TRUE))
## Take the median for each resample
simMedians = apply(simsamples, 2, median)
## Take the two quantiles to get the confidence interval
quantile(simMedians, c(0.025,0.975))

## 2.5% 97.5%
## -1 9
```

## Eksempel: Tandsundhed og flaskebrug

I et studie ville man undersøge, om børn der havde fået mælk fra flaske som barn havde dårligere eller bedre tænder end dem, der ikke havde fået mælk fra flaske. Fra 19 tilfældigt udvalgte børn registrerede man hvornår de havde haft deres første tilfælde af karies.

flaske	alder	flaske	alder	flaske	alder
nej	9	nej	10	ja	16
ja	14	nej	8	ja	14
ja	15	nej	6	ja	9
nej	10	ja	12	nej	12
nej	12	ja	13	ja	12
nej	6	nej	20		
ja	19	ja	13		

Find konfidensintervallet for forskellen!

# Eksempel: Tandsundhed og flaskebrug - et 95% konfidensinterval for $\mu_X - \mu_Y$

```
## Tandsundhed og flaskebrug. Konfidensinterval for forskel i middelværdi
## Reading in no group:
x <- c(9,10,12,6,10,8,6,20,12)
## Reading in yes group:
y <- c(14,15,19,12,13,13,16,14,9,12)

## Resample many times
k <- 100000
simxSamples <- replicate(k, sample(x, replace = TRUE))
simySamples <- replicate(k, sample(y, replace = TRUE))
## Take the mean for each time and subtract
simmeandiffs <- apply(simxSamples, 2, mean) -
apply(simySamples, 2, mean)
## Take the two quantiles to get the confidence interval
quantile(simmeandiffs, c(0.025,0.975))

##      2.5%    97.5%
## -6.2000 -0.1444
```

# Two-sample konfidensinterval for $\theta_X - \theta_Y$ (inkl. $\mu_X - \mu_Y$ ) med ikke-parametrisk bootstrap

Method Method 4.17: Two-sample confidence interval for  $\theta_X - \theta_Y$  by non-parametric bootstrap

Assume we have actual observations  $x_1, \dots, x_n$  and  $y_1, \dots, y_n$ .

- 1 Simulate  $k$  sets of 2 samples of  $n_X$  and  $n_Y$  observations from the respective groups (with replacement)  
 $\hat{\theta}_{x1}^*, \hat{\theta}_{y1}^*, \dots, \hat{\theta}_{xk}^*, \hat{\theta}_{yk}^*$ .
- 2 Calculate the difference between the features in each of the  $k$  samples  
 $\hat{\theta}_{x1}^* - \hat{\theta}_{y1}^*, \dots, \hat{\theta}_{xk}^* - \hat{\theta}_{yk}^*$ .
- 3 Find the  $100(\alpha/2)\%$  and  $100(1 - \alpha/2)\%$  quantiles for these,  $q_{100(\alpha/2)\%}^*$  and  $q_{100(1 - \alpha/2)\%}^*$  as the  $100(1 - \alpha)\%$  confidence interval:  
$$[q_{100(\alpha/2)\%}^*, q_{100(1 - \alpha/2)\%}^*]$$

# Eksempel: Tandsundhed og flaskebrug - et 99% confidence interval for median-forskellen

```
## Tandsundhed og flaskebrug. Konfidensinterval for forskel i median
## Resample many times
k <- 100000
simxSamples <- replicate(k, sample(x, replace = TRUE))
simySamples <- replicate(k, sample(y, replace = TRUE))
## Take the difference in medians
simmedianDiffs <- apply(simxSamples, 2, median)-
                    apply(simySamples, 2, median)
## Take the two quantiles to get the confidence interval
quantile(simmedianDiffs, c(0.005,0.995))

## 0.5% 99.5%
##      -8      0
```

# Spørgsmål: bootstrapping methods (socrative.com - ROOM:PBAC)

```
## Resample many times
k <- 100000
simxSamples <- replicate(k, sample(x, replace = TRUE))
simySamples <- replicate(k, sample(y, replace = TRUE))
## Define a (upper-quartile) Q3-function:
Q3 <- function(x){ quantile(x, 0.75)}
## Calculate the simulated statistic
simQ3Diffs <- apply(simxSamples, 2, Q3) - apply(simySamples, 2, Q3)
## Find the two relevant quantiles of the k simulated differences
quantile(simQ3Diffs, c(0.005, 0.995))
```

Hvilken analyse udføres?

- A: One-sample parametric
- B: Two-sample parametric
- C: One-sample non-parametric
- D: Two-sample non-parametric

# Spørgsmål: bootstrapping methods (socrative.com - ROOM:PBAC)

```
## Resample many times
k <- 100000
simxSamples <- replicate(k, sample(x, replace = TRUE))
## Define a (upper-quartile) Q3-function:
Q3 <- function(x){ quantile(x, 0.75)}
## Calculate the simulated statistic
simQ3 <- apply(simxSamples, 2, Q3)
## Find the two relevant quantiles of the k simulated differences
quantile(simQ3, c(0.005, 0.995))
```

Hvilken analyse udføres?

- A: One-sample parametric
- B: Two-sample parametric
- C: One-sample non-parametric
- D: Two-sample non-parametric

# Spørgsmål: bootstrapping methods (socrative.com - ROOM:PBAC)

```
## Resample many times
k <- 100000
simxSamples <- replicate(k, rnorm(length(x), mean(x), sd(x)))
## Define a (upper-quartile) Q3-function:
Q3 <- function(x){ quantile(x, 0.75)}
## Calculate the simulated statistic
simQ3 <- apply(simxSamples, 2, Q3)
## Find the two relevant quantiles of the k simulated differences
quantile(simQ3, c(0.005, 0.995))
```

Hvilken analyse udføres?

- A: One-sample parametric
- B: Two-sample parametric
- C: One-sample non-parametric
- D: Two-sample non-parametric

# Spørgsmål: bootstrapping methods (socrative.com - ROOM:PBAC)

```
## Resample many times
k <- 100000
simxSamples <- replicate(k, rnorm(length(x), mean(x), sd(x)))
simySamples <- replicate(k, rnorm(length(y), mean(y), sd(y)))
## Define a (upper-quartile) Q3-function:
Q3 <- function(x){ quantile(x, 0.75)}
## Calculate the simulated statistic
simQ3Diffs <- apply(simxSamples, 2, Q3) - apply(simySamples, 2, Q3)
## Find the two relevant quantiles of the k simulated differences
quantile(simQ3Diffs, c(0.005, 0.995))
```

Hvilken analyse udføres?

- A: One-sample parametric
- B: Two-sample parametric
- C: One-sample non-parametric
- D: Two-sample non-parametric

# Bootstrapping - et overblik

Vi har fået 4 ret ens forskellige metode-bokse:

- ① Med eller uden fordeling (parametrisk eller ikke-parametrisk)
- ② For one- eller two-sample analyse (en eller to grupper)

Bemærk:

*Middelværdier* (means) er inkluderet i *vilkårlige beregningsstørrelser* (other features). Eller: Disse metoder kan også anvendes for andre analyser end for means!

Hypotesetest også muligt

Vi kan udføre hypotese test ved at kigge på konfidensintervallerne!

# Kursus 02323: Introducerende Statistik

## Forelæsning 8: Simpel lineær regression

Peder Bacher

DTU Compute, Dynamiske Systemer  
Bygning 303B, Rum 010  
Danmarks Tekniske Universitet  
2800 Lyngby – Danmark  
e-mail: pbac@dtu.dk

Forår 2021

# Kapitel 5: Simpel lineær regressions analyse

To variable:  $x$  og  $y$

- Beregn mindstekvadraters estimat af ret linje

Inferens med simpel lineær regressionsmodel

- Statistisk model:  $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Estimation, konfidensintervaller og tests for  $\beta_0$  og  $\beta_1$
- $1 - \alpha$  konfidensinterval for linjen (*stør sikkerhed for den rigtige linje ligger indenfor*)
- $1 - \alpha$  prædiktionsinterval for punkter (*stør sikkerhed for at nye punkter er indenfor*)

$\rho$ ,  $R$  og  $R^2$

- $\rho$  er korrelationen ( $= \text{sign}_{\beta_1} R$ ) er graden af lineær sammenhæng mellem  $x$  og  $y$
- $R^2$  er andelen af den totale variation som er forklaret af modellen
- Afvises  $H_0 : \beta_1 = 0$  så afvises også  $H_0 : \rho = 0$

# Chapter 5: Simple linear Regression Analysis

Two quantitative variables:  $x$  and  $y$

- Calculate the least squares line

Inferences for a simple linear regression model

- Statistical model:  $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Estimation, confidence intervals and tests for  $\beta_0$  and  $\beta_1$ .
- $1 - \alpha$  confidence interval for the line (*high certainty that the real line will be inside*)
- $1 - \alpha$  prediction interval for punkter (*high certainty that new points will be inside*)

$\rho$ ,  $R$  and  $R^2$

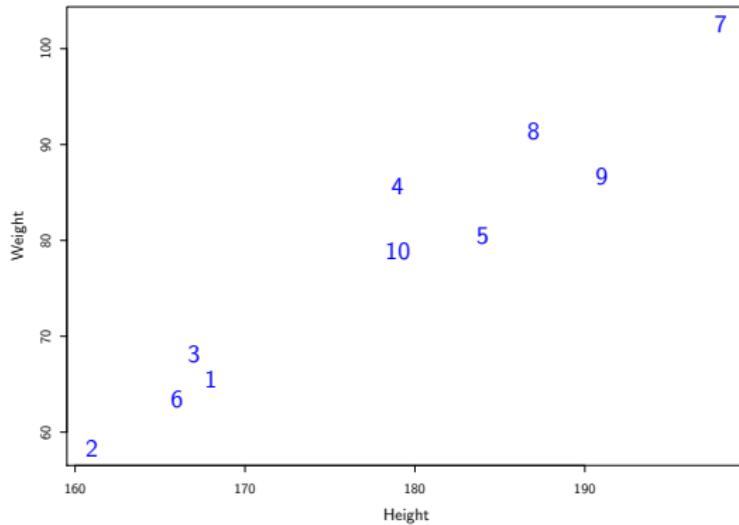
- $\rho$  is the correlation ( $= \text{sign}_{\beta_1} R$ ) is the strength of linear relation between  $x$  and  $y$
- $R^2$  is the fraction of the total variation explained by the model
- If  $H_0 : \beta_1 = 0$  is rejected, then  $H_0 : \rho = 0$  is also rejected

# Oversigt

- 1 Lineær regressionsmodel
- 2 Mindste kvadraters metode (least squares)
- 3 Statistik og lineær regression
- 4 Hypotesetests og konfidensintervaller for  $\hat{\beta}_0$  og  $\hat{\beta}_1$
- 5 Konfidensinterval og prædiktionsinterval
  - Konfidensinterval for linien
  - Prædiktionsinterval
- 6 `summary(lm())` wrap up
- 7 Korrelation
- 8 Model validering: Residual analyse

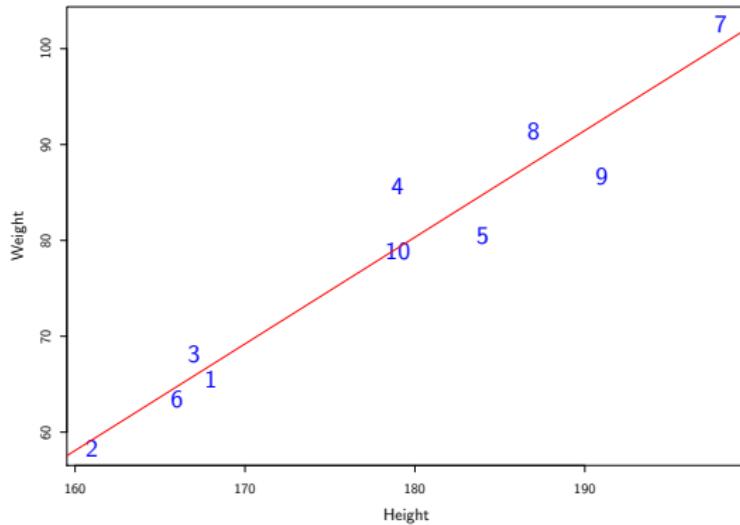
## Motiverende eksempel: Højde-vægt

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



# Motiverende eksempel: Højde-vægt

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



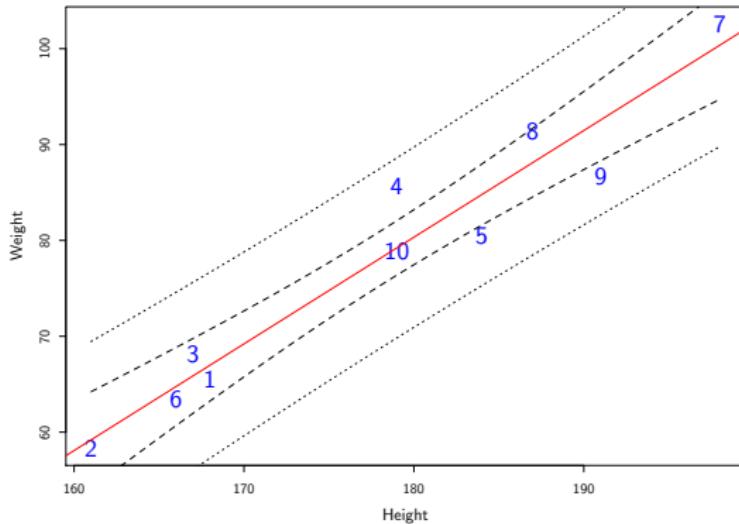
# Motiverende eksempel: Højde-vægt

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

```
##  
## Call:  
## lm(formula = y ~ x)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -5.876 -1.451 -0.608  2.234  6.477  
##  
## Coefficients:  
##                 Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -119.958     18.897   -6.35  0.00022 ***  
## x             1.113      0.106   10.50  5.9e-06 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 3.9 on 8 degrees of freedom  
## Multiple R-squared:  0.932, Adjusted R-squared:  0.924  
## F-statistic: 110 on 1 and 8 DF,  p-value: 5.87e-06
```

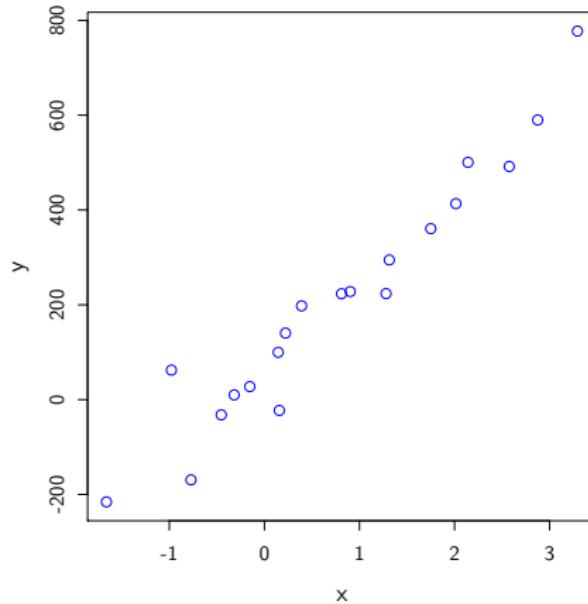
# Motiverende eksempel: Højde-vægt

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



# Et scatter plot af nogle punkter. Hvilken model?

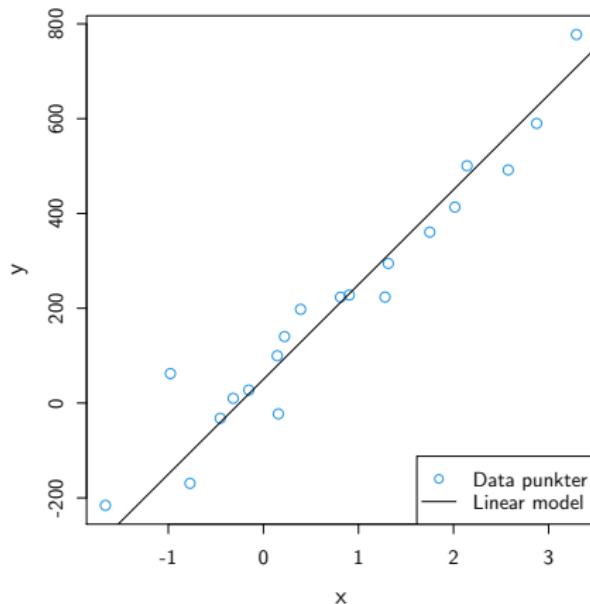
- Datapunkter  $(x_i, y_i)$



# Kommer de fra en almindelig lineær model?

- Opstil en lineær model:

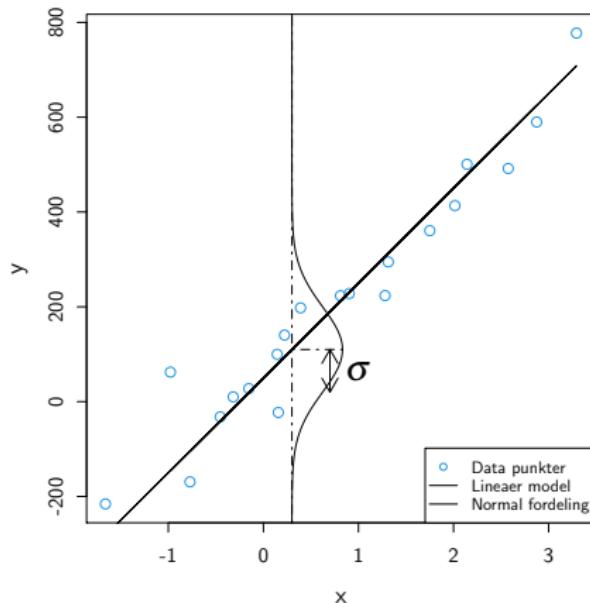
$$y_i = \beta_0 + \beta_1 x_i$$



men den der mangler noget til at beskrive den *tilfældige variation!*

# De kommer fra en lineær regressionsmodel

- Opstil en lineær regressionsmodel:  $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$  hvor  $\varepsilon_i \sim N(0, \sigma^2)$



Den tilfældige variation er beskrevet med en normalfordeling om linien

# Opstil en lineær regressionsmodel

- Opstil den *lineære regressionsmodel*

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- $Y_i$  er den *afhængige variabel* (dependent variable). En stokastisk variabel
- $x_i$  er en *forklarende variabel* (explanatory variable)
- $\varepsilon_i$  (epsilon) er afvigelsen (deviation). En stokastisk variabel

og vi antager

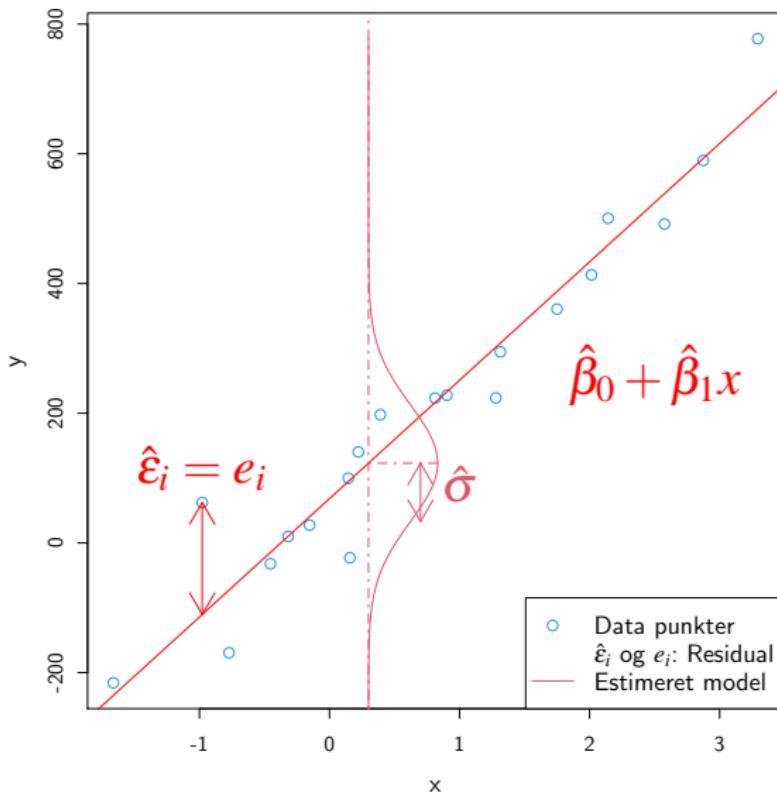
$\varepsilon_i$  er *independent and identically distributed* (i.i.d.) og  $N(0, \sigma^2)$

# Mindste kvadraters metode

- Hvis vi kun har datapunkterne, hvordan kan vi estimere parametrene  $\beta_0$  og  $\beta_1$ ?
- But how!?

Dvs. estimaterne  $\hat{\beta}_0$  og  $\hat{\beta}_1$  er dem som minimerer RSS

# Simuleret eksempel af model, data og fit

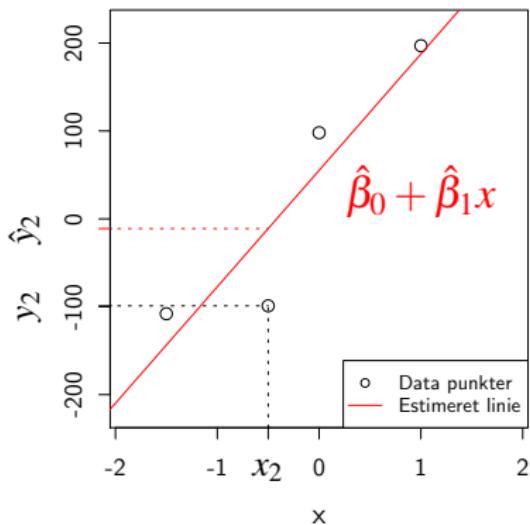


# Spørgsmål om beregning af residual (socrative.com-ROOM:PBAC)

Udregning af residual for punkt  $i$ :

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i = \hat{y}_i + e_i \Leftrightarrow$$

$$e_i = y_i - \hat{y}_i$$



Hvad er  $e_2$ ?

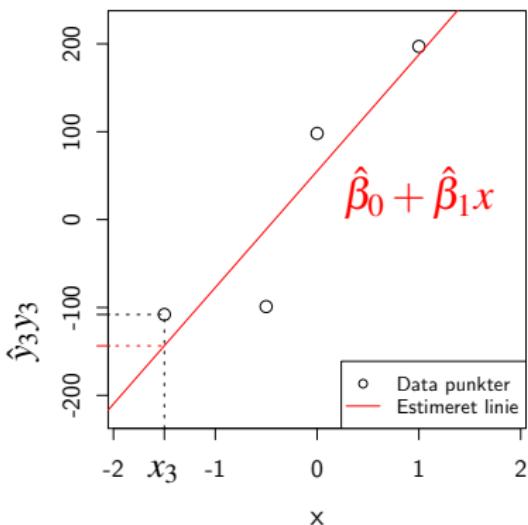
- A: ca. 131      B: ca. 36      C: ca. -88      D: Ved ikke

# Spørgsmål om beregning af residual (socrative.com-ROOM:PBAC)

Udregning af residual for punkt  $i$ :

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i = \hat{y}_i + e_i \Leftrightarrow$$

$$e_i = y_i - \hat{y}_i$$



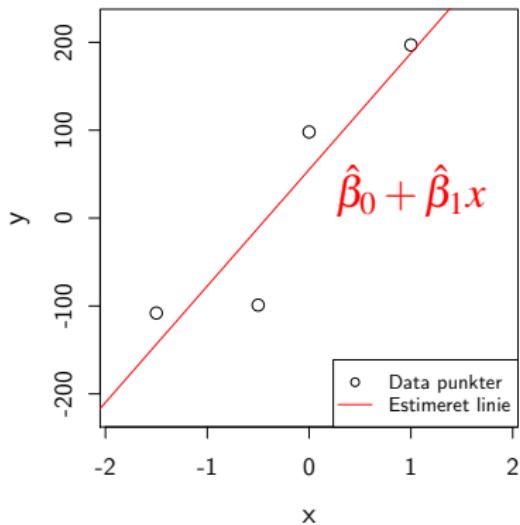
Hvad er  $e_3$ ?

- A: ca. 131      B: ca. 36      C: ca. -88      D: Ved ikke

# Spørgsmål om beregning af RSS (socrative.com-ROOM:PBAC)

Beregn:  
Residual Sum of Squares (RSS)

Fire punkter, så  $n=4$



Hvad er  $RSS = \sum_{i=1}^n e_i^2$  her?

- A: ca. 10917      B: ca. 165      C: ca. -3467      D: Ved ikke

# Least squares estimator minimerer RSS

Theorem 5.4 (her for estimatorer som i bogen)

The least squares estimators of  $\beta_0$  and  $\beta_1$  are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

where  $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ .

# Least squares estimator minimerer RSS

## Theorem 5.4 (her for estimatorer)

The least squares estimatates of  $\beta_0$  and  $\beta_1$  are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{S_{xx}}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where  $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ .

*Vi går ikke dybere ind forskellen mellem estimatorer og estimatorer her i kurset*

# R eksempel

```
## Simuler en lineær model med normalfordelt afvigelse og estimer parametrene

## FØRST LAV DATA:
## Generer n værdier af input x som uniform fordelte
x <- runif(n=20, min=-2, max=4)

## Simuler lineær regressionsmodel
beta0=50; beta1=200; sigma=90
y <- beta0 + beta1 * x + rnorm(n=length(x), mean=0, sd=sigma)

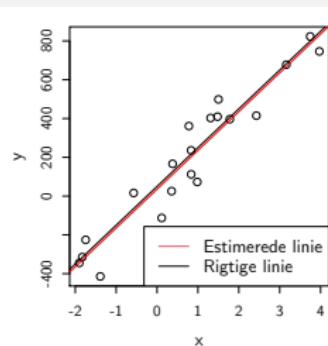
## HERFRA ligesom virkeligheden, vi har dataen i x og y:
## Et scatter plot af x og y
plot(x, y)

## Udregn least squares estimeraterne, brug Theorem 5.4
(beta1hat <- sum( (y-mean(y))*(x-mean(x)) ) / sum( (x-mean(x))^2 ) )
(beta0hat <- mean(y) - beta1hat*mean(x))

## Brug lm() til at udregne estimeraterne
lm(y ~ x)

## Plot den estimerede linie
abline(lm(y ~ x), col="red")

## Tilføj den "rigtige" linie
abline(a=beta0, b=beta1)
legend("bottomright", c("Estimerede linie","Rigtige linie"), lty=1, col=c(2,1))
```



# Parameter estimerterne er stokastiske variabler

Hvis vi gentager forsøget vil estimerterne  $\hat{\beta}_0$  og  $\hat{\beta}_1$  have samme udfald hver gang?

Hvordan er parameter estimerterne fordelt (givet normalfordelte afvigelser)?

- Hvordan er parameter estimaterne i en lineær regressionsmodel fordelt (givet normalfordelte afvigelser)?

### Theorem 5.8 (første del)

$$V[\hat{\beta}_0] = \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{S_{xx}}$$

$$V[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}$$

$$\text{Cov}[\hat{\beta}_0, \hat{\beta}_1] = -\frac{\bar{x} \sigma^2}{S_{xx}}$$

- Kovariansen  $\text{Cov}[\hat{\beta}_0, \hat{\beta}_1]$  (covariance) gør vi ikke mere ud af her.

# Estimater af standardafvigelserne på $\hat{\beta}_0$ og $\hat{\beta}_1$

## Theorem 5.8 (anden del)

Where  $\sigma^2$  is usually replaced by its estimate ( $\hat{\sigma}^2$ ). The central estimator for  $\sigma^2$  is

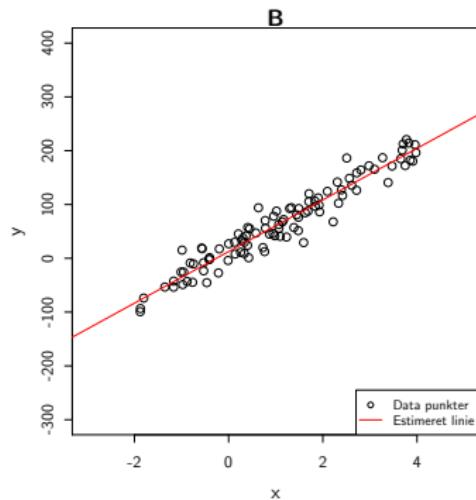
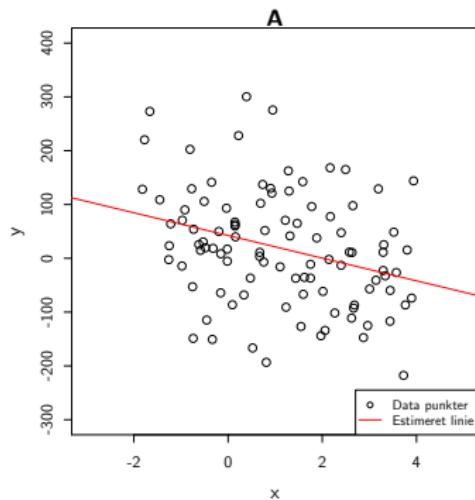
$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}.$$

When the estimate of  $\sigma^2$  is used the variances also become estimates and we'll refer to them as  $\hat{\sigma}_{\beta_0}^2$  and  $\hat{\sigma}_{\beta_1}^2$ .

- Estimat af standardafvigelserne for  $\hat{\beta}_0$  og  $\hat{\beta}_1$  (ligningerne (5-43) og (5-44))

$$\hat{\sigma}_{\beta_0} = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}; \quad \hat{\sigma}_{\beta_1} = \hat{\sigma} \sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

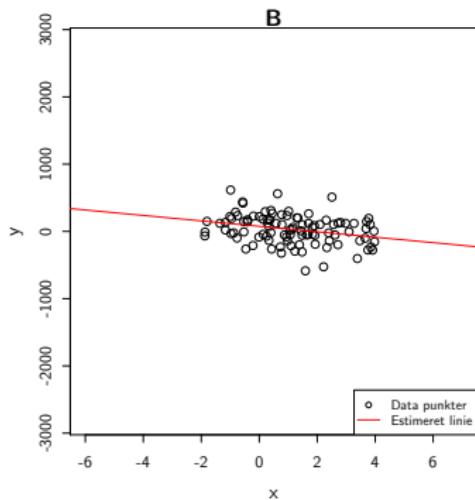
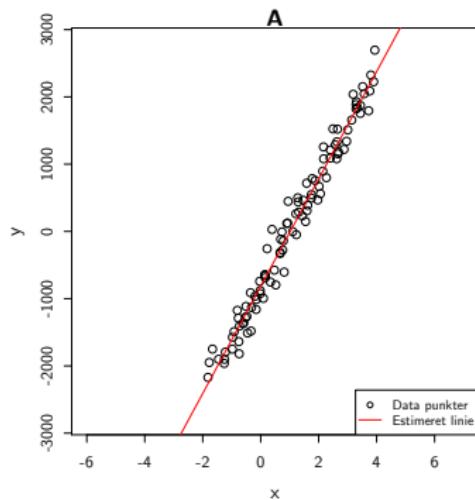
# Spørgsmål: Om fejlenes spredning $\sigma$ (socrative.com-ROOM:PBAC)



For hvilken er residual variansen  $\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}$  størst?

- A: For fit i plot A      B: For fit i plot B      C: Lige stor for begge      D: Ved ikke

# Spørgsmål: Om fejlenes spredning $\sigma$ (socrative.com-ROOM:PBAC)



For hvilken er residual variansen  $\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}$  størst?

- A: For fit i plot A
- B: For fit i plot B
- C: Lige stor for begge
- D: Ved ikke

# Hypotesetests for parameter parametrene

- Vi kan altså udføre hypotesetests for parameter estimatorer i en lineær regressionsmodel:

$$H_{0,i} : \beta_i = \beta_{0,i}$$

$$H_{1,i} : \beta_i \neq \beta_{1,i}$$

- Vi bruger de  $t$ -fordelte statistikker:

## Theorem 5.12

Under the null-hypothesis ( $\beta_0 = \beta_{0,0}$  and  $\beta_1 = \beta_{0,1}$ ) the statistics

$$T_{\beta_0} = \frac{\hat{\beta}_0 - \beta_{0,0}}{\hat{\sigma}_{\beta_0}}; \quad T_{\beta_1} = \frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}},$$

are  $t$ -distributed with  $n - 2$  degrees of freedom, and inference should be based on this distribution.

## Eksempel: Hypotesetest for parametrene

- Se Eksempel 5.13 for eksempel på hypotesetest, samt Metode 5.14
- Test om parametrene er signifikant forskellige fra 0

$$H_{0,i} : \beta_i = 0$$

$$H_{1,i} : \beta_i \neq 0$$

- Se resultatet med simulering i R

```
## Hypotesetests for signifikante parametre

## Generer x
x <- runif(n=20, min=-2, max=4)
## Simuler Y
beta0=50; beta1=200; sigma=90
y <- beta0 + beta1 * x + rnorm(n=length(x), mean=0, sd=sigma)

## Brug lm() til at udregne estimatorne
fit <- lm(y ~ x)

## Se summary, deri står hvad vi har brug for
summary(fit)
```

# Konfidensintervaller for parametrene

## Method 5.15

$(1 - \alpha)$  confidence intervals for  $\beta_0$  and  $\beta_1$  are given by

$$\hat{\beta}_0 \pm t_{1-\alpha/2} \hat{\sigma}_{\beta_0}$$

$$\hat{\beta}_1 \pm t_{1-\alpha/2} \hat{\sigma}_{\beta_1}$$

where  $t_{1-\alpha/2}$  is the  $(1 - \alpha/2)$ -quantile of a  $t$ -distribution with  $n - 2$  degrees of freedom.

- husk at  $\hat{\sigma}_{\beta_0}$  og  $\hat{\sigma}_{\beta_1}$  findes ved ligningerne (5-43) og (5-44)
- i R kan  $\hat{\sigma}_{\beta_0}$  og  $\hat{\sigma}_{\beta_1}$  aflæses ved "Std. Error" ved "summary(fit)"

# Simuleringseksempel: Konfidensintervaller for parametrene

```
## Lav konfidensintervaller for parametrene

## Antal gentagelser
nRepeat <- 100

## Fangede vi den rigtige parameter
TrueValInCI <- logical(nRepeat)

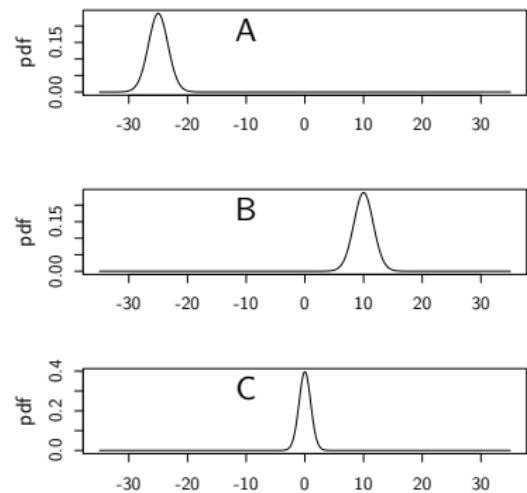
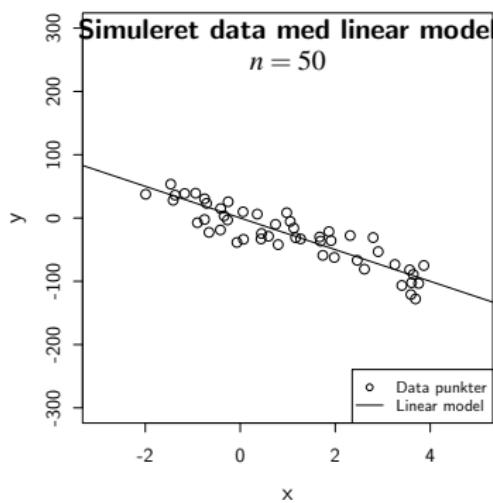
## Gentag simuleringen og estimeringen nRepeat gange
for(i in 1:nRepeat){
  ## Generer x
  x <- runif(n=20, min=-2, max=4)
  ## Simuler y
  beta0=50; beta1=200; sigma=90
  y <- beta0 + beta1 * x + rnorm(n=length(x), mean=0, sd=sigma)

  ## Brug lm() til at udregne estimaterne
  fit <- lm(y ~ x)

  ## Heldigvis kan R beregne konfidensintervallet (level=1-alpha)
  (ci <- confint(fit, "(Intercept)", level=0.95))

  ## Var den rigtige parameterværdi "fanget" af intervallet?
  (TrueValInCI[i] <- ci[1] < beta0 & beta0 < ci[2])
}

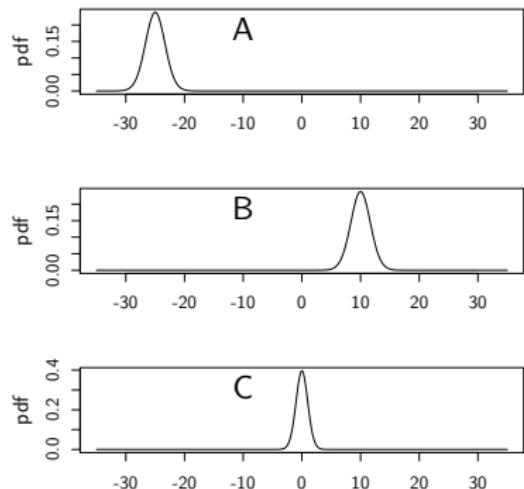
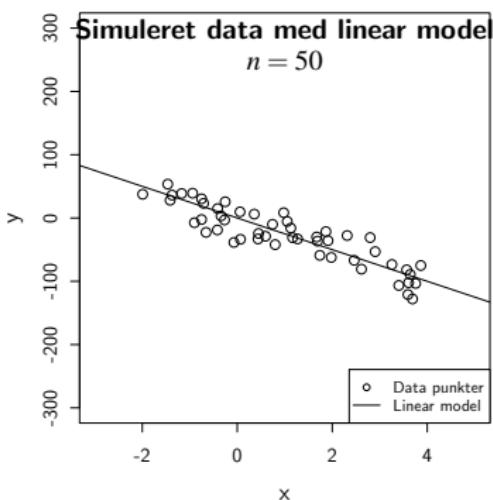
## Hvor ofte blev den rigtige værdi "fanget"?
sum(TrueValInCI) / nRepeat
```

Spørgsmål: Om fordelingen af  $\hat{\beta}_1$  (socrative.com-ROOM:PBAC)

Hvilket plot repræsenterer fordelingen af  $\hat{\beta}_1$ ?

- A: Plot A      B: Plot B      C: Plot C      D: Ved ikke

# Spørgsmål: Om fordelingen af $\frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\hat{\beta}_1}}$ (socrative.com-ROOM:PBAC)



Hvilket plot repræsenterer fordelingen af  $\frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\hat{\beta}_1}}$  under  $H_0: \beta_{0,1} = -25$ ?

- A: Plot A      B: Plot B      C: Plot C      D: Ved ikke

## Method 5.18: Konfidensinterval for $\beta_0 + \beta_1 x_0$

- Konfidensinterval for  $\beta_0 + \beta_1 x_0$  svarer til et konfidensinterval for linien i punktet  $x_0$
- Beregnes med

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2} \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

- Der er  $100(1 - \alpha)\%$  sandsynlighed for at den rigtige linie, altså  $\beta_0 + \beta_1 x_0$ , er inde i konfidensintervallet

## Method 5.18: Prædiktionsinterval for $\beta_0 + \beta_1 x_0 + \varepsilon_0$

- Prædiktionsintervallet (prediction interval) for  $Y_0$  beregnes for en "ny" værdi af  $x_i$ , her kaldt  $x_0$
- Dette gøres *før*  $Y_0$  observeres ved

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

- Der er  $100(1 - \alpha)\%$  sandsynlighed for at den observerede  $y_0$  vil falde inde i prædiktionsintervallet
- Et prædiktionsinterval bliver altid større end et konfidensinterval for fastholdt  $\alpha$

# Eksempel med konfidensinterval for linien

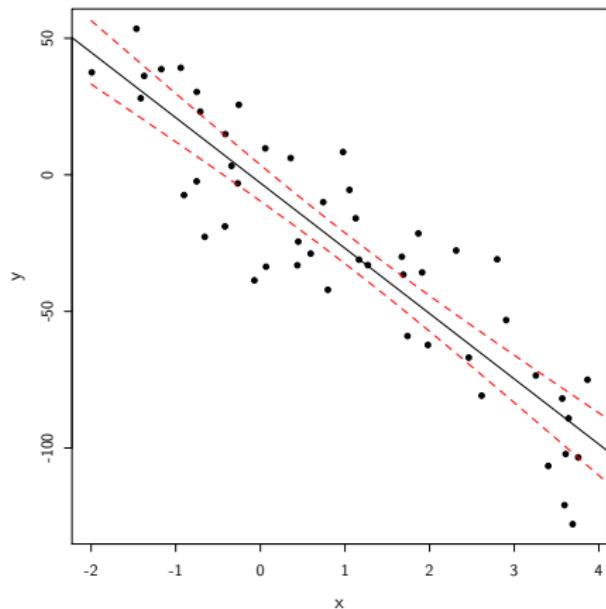
```
## Eksempel med konfidensinterval for linien

## Lav en sekvens af x værdier
xval <- seq(from=-2, to=6, length.out=100)

## Brug predict funktionen
CI <- predict(fit, newdata=data.frame(x=xval),
interval="confidence",
level=.95)

## Se lige hvad der kom
head(CI)

## Plot data, model og intervaller
plot(x, y, pch=20)
abline(fit)
lines(xval, CI[, "lwr"], lty=2, col="red", lwd=2)
lines(xval, CI[, "upr"], lty=2, col="red", lwd=2)
```



# Eksempel med prædiktionsinterval

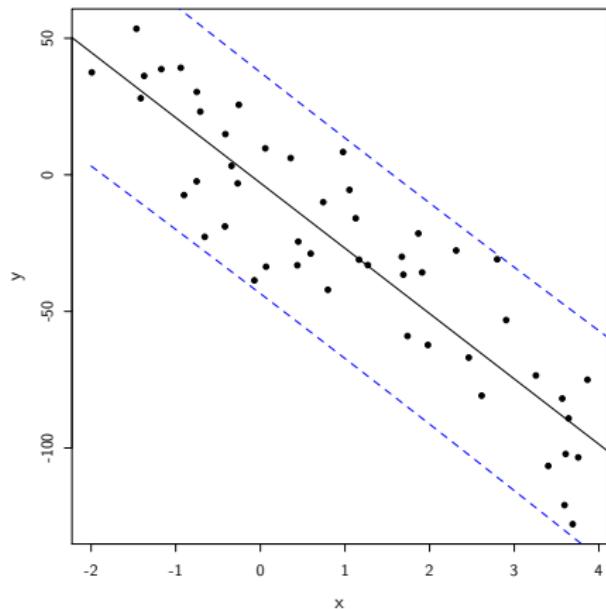
```
## Eksempel med prædiktionsinterval

## Lav en sekvens af x værdier
xval <- seq(from=-2, to=6, length.out=100)

## Beregn interval for hvert x
PI <- predict(fit, newdata=data.frame(x=xval),
interval="prediction",
level=.95)

## Se lige hvad der kom tilbage
head(PI)

## Plot data, model og intervaller
plot(x, y, pch=20)
abline(fit)
lines(xval, PI[, "lwr"], lty=2, col="blue", lwd=2)
lines(xval, PI[, "upr"], lty=2, col="blue", lwd=2)
```



# Hvad bliver mere skrevet ud af summary?

```
##  
## Call:  
## lm(formula = y ~ x)  
##  
## Residuals:  
##    Min     1Q Median     3Q    Max  
## -37.35 -14.08   0.61  14.05  38.96  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -2.99      3.29   -0.91   0.37  
## x           -23.91     1.67  -14.34  <2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 20 on 48 degrees of freedom  
## Multiple R-squared:  0.811, Adjusted R-squared:  0.807  
## F-statistic: 206 on 1 and 48 DF,  p-value: <2e-16
```

## summary(lm(y~x)) wrap up

- Residuals: Min 1Q Median 3Q Max
- Coefficients:

Estimate	Std. Error	t value	Pr(> t )	"stjerner"
----------	------------	---------	----------	------------
- Residual standard error: XXX on XXX degrees of freedom
- Multiple R-squared: XXX

Resten bruger vi ikke i det her kursus

# Forklaret varians og korrelation

- Forklaret varians af en model er  $r^2$ , i summary "Multiple R-squared"
- Beregnes med

$$r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

hvor  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

- *Andel af den totale varians i data ( $y_i$ ) der er forklaret med modellen*

# Forklaret varians og korrelation

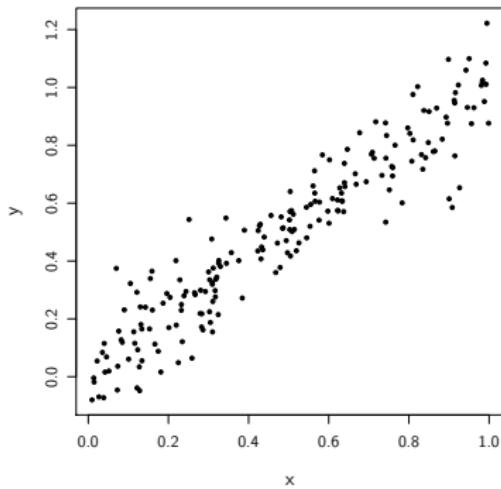
- Korrelationen  $\rho$  er et mål for *lineær sammenhæng* mellem to stokastiske variable
- Estimeret (i.e. empirisk) korrelation

$$\hat{\rho} = r = \sqrt{r^2} \operatorname{sgn}(\hat{\beta}_1)$$

hvor  $\operatorname{sgn}(\hat{\beta}_1)$  er:  $-1$  for  $\hat{\beta}_1 \leq 0$  og  $1$  for  $\hat{\beta}_1 > 0$

- Altså:
  - Positiv korrelation ved positiv hældning
  - Negativ korrelation ved negativ hældning

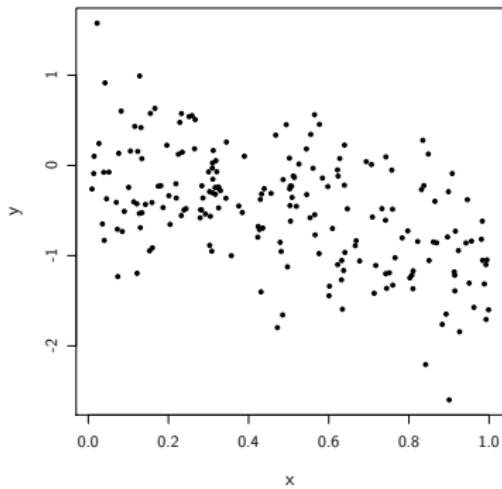
# Spørgsmål om korrelation (socrative.com-ROOM:PBAC)



Hvad er korrelationen mellem  $x$  og  $y$ ?

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

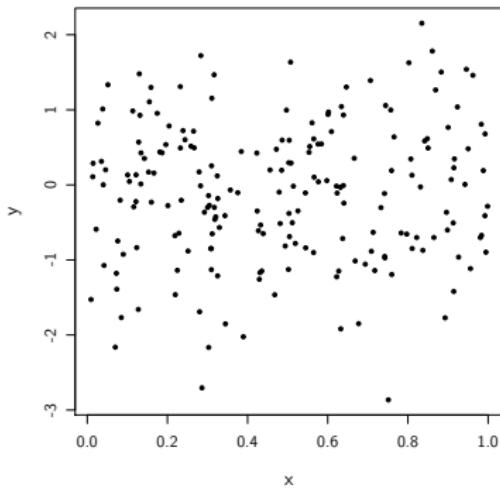
# Spørgsmål om korrelation (socrative.com-ROOM:PBAC)



Hvad er korrelationen mellem  $x$  og  $y$ ?

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

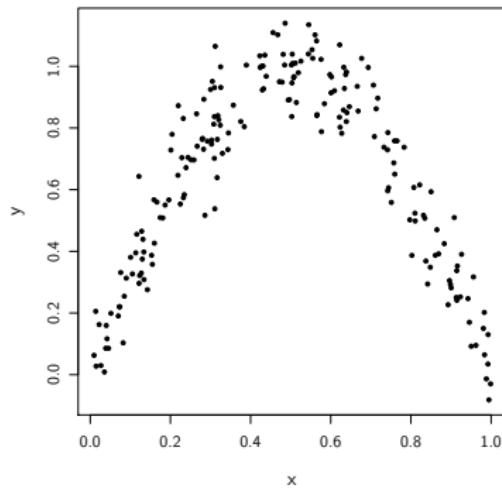
# Spørgsmål om korrelation (socrative.com-ROOM:PBAC)



Hvad er korrelationen mellem  $x$  og  $y$ ?

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

# Spørgsmål om korrelation (socrative.com-ROOM:PBAC)



Hvad er korrelationen mellem  $x$  og  $y$ ?

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

# Test for signifikant korrelation

- Test for signifikant korrelation (lineær sammenhæng) mellem to variable

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

er ækvivalent med

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

hvor  $\hat{\beta}_1$  er estimatet af hældningen i simpel lineær regressionsmodel

# Simuleringseksempel om korrelation

```
## Korrelation

## Generer x
x <- runif(n=20, min=-2, max=4)
## Simuler y
beta0=50; beta1=200; sigma=90
y <- beta0 + beta1 * x + rnorm(n=length(x), mean=0, sd=sigma)

## Scatter plot
plot(x,y)

## Brug lm() til at udregne estimaterne
fit <- lm(y ~ x)

## Den rigtige linie
abline(beta0, beta1)
## Plot fittet
abline(fit, col="red")

## Se summary, deri står hvad vi har brug for
summary(fit)

## Korrelation mellem x og y
cor(x,y)

## Kvadreret er den "Multiple R-squared" fra summary(fit)
cor(x,y)^2
```

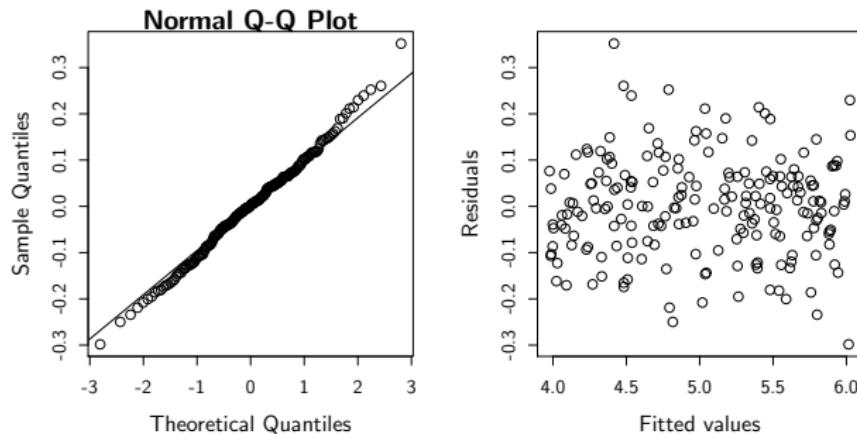
# Model validering: Residual analyse

Method 5.28 (can it be rejected that  $\hat{\varepsilon}_i$  is i.i.d.?)

- Check normality assumption with q-q plot (less important with many observations).
- Check (non)systematic behavior by plotting the residuals  $e_i$  as a function of fitted values  $\hat{y}_i$

# Residual Analysis in R (er $\hat{\varepsilon}_i$ i.i.d.?)

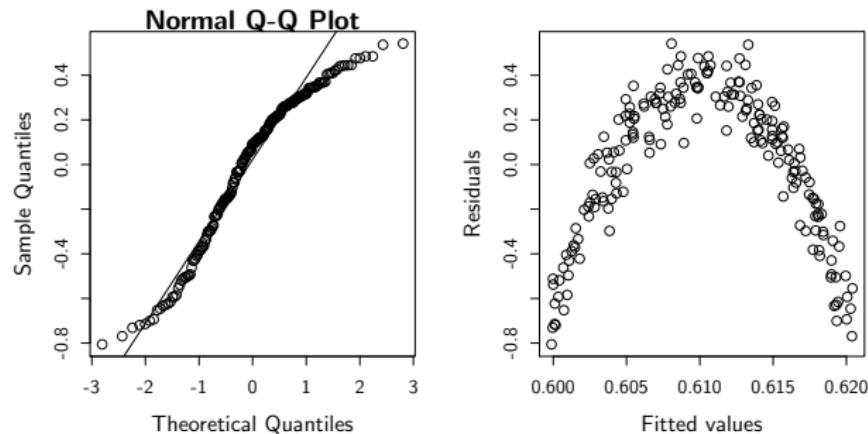
```
## Model validering: residual analysis
fit <- lm(y1 ~ x)
par(mfrow = c(1, 2))
qqnorm(fit$residuals)
qqline(fit$residuals)
plot(fit$fitted, fit$residuals, xlab="Fitted values", ylab="Residuals")
```



Hvor fitted values er  $\hat{y}_i$  og residuals er  $\hat{\varepsilon}_i$ . **Her ser det fint ud!**

# Residual Analysis in R (er $\hat{\varepsilon}_i$ i.i.d.?)

```
## Model validering: residual analysis
fit <- lm(y4 ~ x)
par(mfrow = c(1, 2))
qqnorm(fit$residuals)
qqline(fit$residuals)
plot(fit$fitted, fit$residuals, xlab="Fitted values", ylab="Residuals")
```



Hvor fitted values er  $\hat{y}_i$  og residuals er  $\hat{\varepsilon}_i$ . **Her ser det ikke fint ud:**  
 $\hat{\varepsilon}_i$  ikke normalfordelt, samt klar sammenhæng mellem  $\hat{y}_i$  og  $\varepsilon_i$ !

# Introduktion til Statistik

## Forelæsning 9: Multipel lineær regression

Peder Bacher

DTU Compute, Dynamiske Systemer  
Bygning 303B, Rum 010  
Danmarks Tekniske Universitet  
2800 Lyngby – Danmark  
e-mail: pbac@dtu.dk

Forår 2021

# Kapitel 6: Multipel lineær regressions analyse

## Multipel lineær regressionsmodel

- Flere variabler:  $Y, x_1, x_2, \dots$   
(*y afhængig/respons var. og x'er er forklarende/uafhængige var.*)
- Mindstekvadraters rette plan (*et plan da der er >2 dimensioner*)

## Inferens for en multipel lineær regressionmodel

- Statistisk model:  $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_p x_{p,i} + \varepsilon_i$
- Estimation af konfidensintervaller og tests for  $\beta$ 'er
- Konfidensintervaller for modellen (middelplanet)
- Prædiktionsintervaller for nye punkter
- $R^2$  er andelen af den totale variationen som er forklaret af modellen

## Model validering af antagelser ved residual analyse

- Normalfordeling? q-q plots af residualer
- Uafhængighed? Plot residualer mod prædikterede værdier  $\hat{y}_i$  og inputs  $x_{j,i}$

# Chapter 6: Multiple linear Regression Analysis

## Multipel lineær regressionsmodel

- Many quantitative variables:  $y, x_1, x_2, \dots$   
(*y is the dependent/response var. and x's are explanatory/independent var.*)
- Calculating least squares surface (*a plane surface since there are >2 dimensions*)

## Inferences for the multiple linear regression model

- Statistical model:  $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_p x_{p,i} + \varepsilon_i$
- Confidence interval estimation and test for the  $\beta$ 's
- Confidence interval for the model (the mean surface)
- Prediction interval for new points
- $R^2$  expresses the proportion of the total variation explained by the linear fit

## Model validation of assumptions with residual analysis

- Normal distribution? q-q plots of residuals
- Independence? Plot residuals against predicted values  $\hat{y}_i$  and inputs  $x_{j,i}$

# Oversigt

- 1 Warm up med lidt simpel lineær reg.
- 2 Multipel lineær regression
- 3 Modeludvælgelse
- 4 Residual analyse (model kontrol)
- 5 Kurvilinearitet
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet

## Eksempel: Ozon koncentration

Vi har givet et sæt af sammenhængende målinger af: ozon koncentration (ppb), temperatur, solindstråling og vindhastighed:

ozone	radiation	wind	temperature	month	day
41	190	7.4	67	5	1
36	118	8.0	72	5	2
:	:	:	:	:	:
18	131	8.0	76	9	29
20	223	11.5	68	9	30

# Eksempel: Ozon koncentration

```

## Se info om data
?airquality
## Indlæs data
Air <- airquality
## Fjern rækker hvor der er mindst en NA værdi
Air <- na.omit(Air)
## Fjern en outlier
Air <- Air[-which(Air$Ozone == 1), ]

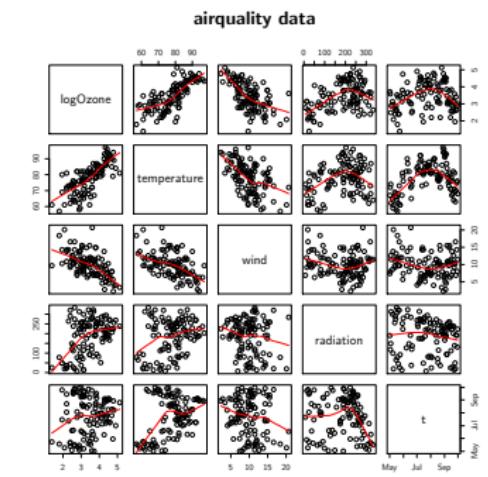
## Se lige på empirisk tæthedsfunktion
hist(Air$Ozone, probability=TRUE, xlab="Ozon", main="")
## Koncentrationer er positive og meget højre-skæv fordeling, derfor log transformere
Air$logOzone <- log(Air$Ozone)
## Bedre epdf?
hist(Air$logOzone, probability=TRUE, xlab="log Ozon", main="")

## Lav en tid (R tidsklasse, se ?POSIXct)
Air$t <- ISOdate(1973, Air$Month, Air$Day)
## Behold kun nogle af kolonnerne
Air <- Air[, c(7, 4, 3, 2, 8)]
## Nye navne på kolonnerne
names(Air) <- c("logOzone", "temperature", "wind", "radiation", "t")

## Hvad er der i Air?
str(Air)
Air
head(Air)
tail(Air)

## Typisk vil man starte med et pairs plot
pairs(Air, panel = panel.smooth, main = "airquality data")

```



# Eksempel: Ozonkoncentration

- Lad os se på sammenhængen mellem log ozon koncentrationen og temperaturen
- Brug en *simpel lineær regressionsmodel*

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2) \text{ og i.i.d.}$$

hvor

- $Y_i$  er log ozonkoncentrationen for måling  $i$
- $x_i$  er temperaturen ved måling  $i$

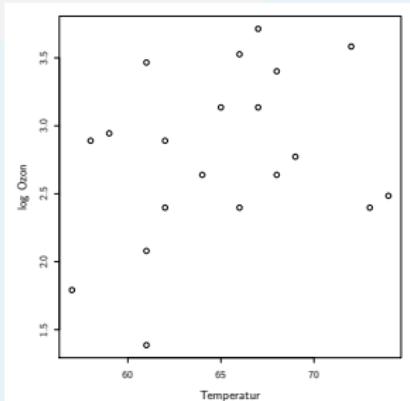
# Fit simpel lineær regressions model i R

Antag at vi kun havde de første 20 datapunkter  
Er der afhængighed?

```
## Start med at sige at vi har 20 datapunkter
Air20 <- Air[1:20, ]

## Se på sammenhængen mellem log(ozon) og temperatur
plot(Air20$temperature, Air20$logOzone, xlab="Temperatur", ylab="log Ozon")

## Korrelation
cor(Air20$logOzone, Air20$temperature)
```



# Spørgsmål om signifikant korrelation (socrative.com-ROOM:PBAC)

```
## Se om der er signifikant korrelation mellem 20 observationer
summary(lm(logOzone ~ temperature, data=Air20))

##
## Call:
## lm(formula = logOzone ~ temperature, data = Air20)
##
## Residuals:
##       Min     1Q     Median      3Q     Max
## -1.2476 -0.4560  0.0559  0.4154  0.8550
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 0.3494     1.8483    0.19    0.85    
## temperature 0.0375     0.0284    1.32    0.20    
## 
## Residual standard error: 0.6 on 18 degrees of freedom
## Multiple R-squared:  0.0883, Adjusted R-squared:  0.0377 
## F-statistic: 1.74 on 1 and 18 DF,  p-value: 0.203
```

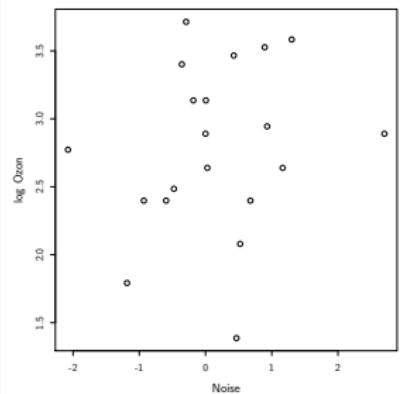
Er der signifikant korrelation mellem logOzone og temperature på 5% signifikansniveau?

- A: Ja
- B: Nej
- C: Ved ikke

# Tilføj en vektor med tilfældige værdier

Simuler 20 *uafhængige stokastiske variable* og tilføj til data.frame  
Er der sammenhæng?

```
## Er der signifikant lineær sammenhæng (korrelation)?  
Air20$noise <- rnorm(20)  
plot(Air20$noise, Air20$logOzone, xlab="Noise", ylab="log Ozon")  
cor(Air20$noise, Air20$logOzone)
```



# Spørgsmål om signifikant korrelation (socrative.com-ROOM:PBAC)

```
## Se om der er signifikant korrelation mellem 20 observationer
summary(lm(logOzone ~ noise, data=Air20))

##
## Call:
## lm(formula = logOzone ~ noise, data = Air20)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4344 -0.2721 -0.0303  0.4557  0.9819
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  2.766     0.140   19.71  1.2e-13 ***
## noise        0.117     0.138    0.85    0.41    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.62 on 18 degrees of freedom
## Multiple R-squared:  0.0382, Adjusted R-squared:  -0.0152 
## F-statistic: 0.715 on 1 and 18 DF,  p-value: 0.409
```

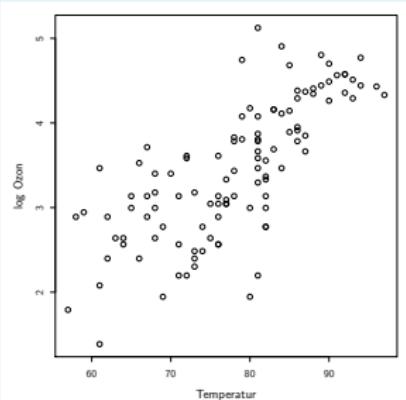
Er der signifikant korrelation mellem logOzone og noise på 5% signifikansniveau?

- A: Ja
- B: Nej
- C: Ved ikke

# Fit simpel lineær regressions model i R

Nu tager vi alle observationer med.  
Er der sammenhæng?

```
## Er der signifikant lineær sammenhæng (korrelation)?  
plot(Air$temperature, Air$logOzone, xlab="Temperatur", ylab="log Ozon")  
cor(Air$temperature, Air$logOzone)
```



# Spørgsmål om signifikant korrelation (socrative.com-ROOM:PBAC)

```
## Se om der er signifikant korrelation med ALLE 110 OBSERVATIONER
summary(lm(logOzone ~ temperature, data=Air))

##
## Call:
## lm(formula = logOzone ~ temperature, data = Air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6303 -0.3331  0.0115  0.3455  1.4843
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -1.49997   0.43485  -3.45   8e-04 ***
## temperature  0.06345   0.00554   11.46  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.54 on 108 degrees of freedom
## Multiple R-squared:  0.549, Adjusted R-squared:  0.544 
## F-statistic: 131 on 1 and 108 DF,  p-value: <2e-16
```

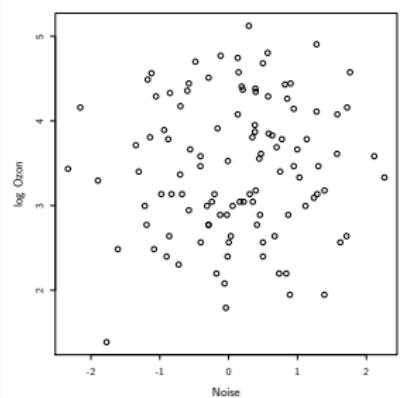
Er der signifikant korrelation mellem logOzone og temperature på 5% signifikansniveau?

- A: Ja
- B: Nej
- C: Ved ikke

# Tilføj en vektor med tilfældige værdier

Simuler 110 uafhængige stokastiske variabler og tilføj.  
Er der sammenhæng?

```
## Tilføj en vektor med normalfordelte tilfældige værdier
## Er der signifikant lineær sammenhæng (korrelation)?
Air$noise <- rnorm(nrow(Air))
plot(Air$noise, Air$logOzone, xlab="Noise", ylab="log Ozon")
cor(Air$noise, Air$logOzone)
```



# Spørgsmål om signifikant korrelation (socrative.com-ROOM:PBAC)

```
## Test om der er signifikant korrelation med ALLE 110 OBSERVATIONER
summary(lm(logOzone ~ noise, data=Air))

##
## Call:
## lm(formula = logOzone ~ noise, data = Air)
##
## Residuals:
##       Min     1Q   Median     3Q    Max
## -1.9409 -0.5705 -0.0068  0.6247  1.6657
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  3.4396    0.0775  44.38   <2e-16 ***
## noise        0.0634    0.0820    0.77    0.44    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.81 on 108 degrees of freedom
## Multiple R-squared:  0.0055, Adjusted R-squared:  -0.0037 
## F-statistic: 0.598 on 1 and 108 DF,  p-value: 0.441
```

Er der signifikant korrelation mellem logOzone og noise på 5% signifikansniveau?

- A: Ja
- B: Nej
- C: Ved ikke

# Simpel lineær regressionsmodel til de to andre

Vi kan også lave en simpel lineær regressionsmodel med de to andre

```
## Simpel lineær regressionsmodel med vindhastigheden
plot(Air$wind, Air$logOzone, xlab="Vindhastighed", ylab="log Ozon")
cor(Air$wind, Air$logOzone)
summary(lm(logOzone ~ wind, data=Air))

## Simpel lineær regressionsmodel med indstrålingen
plot(Air$radiation, Air$logOzone, ylab="log Ozon", xlab="Indstråaling")
cor(Air$radiation, Air$logOzone)
summary(lm(logOzone ~ radiation, data=Air))
```

# Multipel lineær regression

- $Y$  er den *afhængige variabel* (dependent variable)
- Vi er interesseret i at modellere  $Y$ 's afhængighed af de *forklarende eller uafhængige variabler* (explanatory eller independent variables)  $x_1, x_2, \dots, x_p$
- Vi undersøger en *lineær sammenhæng* mellem  $Y$  og  $x_1, x_2, \dots, x_p$ , ved en regressionsmodel på formen

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i} + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2) \text{ og i.i.d.}$$

- $Y_i$  og  $\varepsilon_i$  er stokastiske variabler og  $x_{j,i}$  er variabler

Vi har altså  $i = 1, 2, \dots, n$  observationer og tilsvarende  $Y_i$  og  $\varepsilon_i$  stokastiske variabler

# Mindste kvadraters metode (least squares)

- Residualerne findes ved at prædiktionen

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_p x_{i,p}$$

indsættes

$$y_i = \hat{y}_i + e_i$$

”observation = prædiktion + residual”

og trækkes fra

$$e_i = y_i - \hat{y}_i$$

”residual = observation – prædiktion”

# Mindste kvadraters metode (least squares)

- Ved det bedste estimat for  $\beta_0, \beta_1, \dots, \beta_p$  forstås de værdier  $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$  der minimerer residual sum of squares (RSS)

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- og estimatet for afvigelsernes ( $\varepsilon_i$ ) standard afvigelse er

$$\hat{\sigma}^2 = \frac{1}{n - (p + 1)} \sum_{i=1}^n e_i^2$$

# Mindste kvadraters metode

$\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$  findes ved at løse de såkaldte normalligninger, der for  $p = 2$  er givet ved

$$\sum_{i=1}^n y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_{i,1} + \hat{\beta}_2 \sum_{i=1}^n x_{i,2}$$

$$\sum_{i=1}^n x_{i,1}y_i = \hat{\beta}_0 \sum_{i=1}^n x_{i,1} + \hat{\beta}_1 \sum_{i=1}^n x_{i,1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i,1}x_{i,2}$$

$$\sum_{i=1}^n x_{i,2}y_i = \hat{\beta}_0 \sum_{i=1}^n x_{i,2} + \hat{\beta}_1 \sum_{i=1}^n x_{i,1}x_{i,2} + \hat{\beta}_2 \sum_{i=1}^n x_{i,2}^2$$

Man skal simpelthen gange nogle matricer sammen!

# Udvid modellen (forward selection)

Forward selection (*ikke beskrevet i bogen*):

- Start med *mindste model* med den mest signifikante (mest forklarende) variabel
- *Udvid modellen* med de andre forklarende variabler (inputs) en ad gangen
- *Stop* når der ikke er flere signifikante udvidelser

```
## Forward selection:  
## Tilføj vind, indstråling eller støj input til modellen  
summary(lm(logOzone ~ temperature + wind, data=Air))  
summary(lm(logOzone ~ temperature + radiation, data=Air))  
summary(lm(logOzone ~ temperature + noise, data=Air))  
## Tilføj indstråling eller støj input til modellen  
summary(lm(logOzone ~ temperature + wind + radiation, data=Air))  
summary(lm(logOzone ~ temperature + wind + noise, data=Air))  
## Udvid yderligere med støj input?  
summary(lm(logOzone ~ temperature + wind + radiation + noise, data=Air))
```

# Formindsk modellen (model reduction eller backward selection)

Backward selection (*beskrevet i bogen, Method 6.16*):

- Start med den fulde model
- Fjern den mest insignifikante forklarende variabler
- Stop hvis alle prm. estimatorer er signifikante

```
## Fit den fulde model
summary(lm(log0zone ~ temperature + wind + radiation + noise, data=Air))
## Fjern det mest ikke-signifikante input, er alle nu sigifikante?
summary(lm(log0zone ~ temperature + wind + radiation, data=Air))
```

# Modeludvælgelse

*Der er ikke noget sikker metode til automatisk at finde den bedste model!*

- Det vil kræve subjektive beslutninger at udvælge en model
- Bedste procedure (forward eller backward): det afhænger af forholdene
- Statistiske tests og mål til at sammenligne modeller
- Her i kurset kun backward procedure inkluderet

# Simulate an MLR model with 2 inputs

```
aspect3d(c(1,1,1))
axes3d(c('x--','y--','z--'))
mtext3d('x1', edge=c('x--'), line=2)
mtext3d('x2', edge=c('y--'), line=2)
mtext3d('y', edge=c('z--'), line=2)

## Estimate a plane
fit <- lm(y ~ x1 + x2)

## Make predictions for a grid to see the estimated plane
nplot <- 20
x1plot <- seq(min(x1),max(x1),len=nplot)
x2plot <- seq(min(x2),max(x2),len=nplot)
yprd <- outer(x1plot, x2plot, function(x1,x2){predict(fit, data.frame(x1=x1, x2=x2))})
## 'jet.colors' is as in Matlab, alternatives see ?rainbow
jet.colors <- colorRampPalette(c("#00007F", "blue", "#007FFF", "cyan",
 "#7FFF7F", "yellow", "#FF7F00", "red", "#7F0000"))
## Use 100 different colors
colors <- jet.colors(100)
## Set the colors for z values
color <- colors[((yprd-min(yprd))/(max(yprd)-min(yprd))*100]
rgl.viewpoint(fov=40, theta=0, phi=-90)
## Make a surface with jet colors and grid
surface3d(x1plot, x2plot, yprd, color=color, alpha=0.5)
surface3d(x1plot, x2plot, yprd, front="lines", back="lines", alpha=0.5)
```

# Spørgsmål om MLR estimat (socrative.com-ROOM:PBAC)

Hvordan ligger estimerne af  $\beta_0$ ,  $\beta_1$  og  $\beta_2$ ?

- A:  $\hat{\beta}_0 < 0$ ,  $\hat{\beta}_1 < 0$  og  $\hat{\beta}_2 < 0$
- B:  $\hat{\beta}_0 > 0$ ,  $\hat{\beta}_1 > 0$  og  $\hat{\beta}_2 > 0$
- C:  $\hat{\beta}_0 > 0$ ,  $\hat{\beta}_1 > 0$  og  $\hat{\beta}_2 < 0$
- D:  $\hat{\beta}_0 < 0$ ,  $\hat{\beta}_1 > 0$  og  $\hat{\beta}_2 > 0$
- E:  $\hat{\beta}_0 > 0$ ,  $\hat{\beta}_1 < 0$  og  $\hat{\beta}_2 > 0$

# Spørgsmål om modelreduktion (socrative.com-ROOM:PBAC)

```
##  
## Call:  
## lm(formula = y ~ x1 + x2 + x3)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -0.9195 -0.1555  0.0104  0.1465  0.6304  
##  
## Coefficients:  
##                 Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -0.0528     0.2285  -0.23   0.8201  
## x1          -0.7357     0.3034  -2.42   0.0275 *  
## x2           0.2618     0.2937   0.89   0.3859  
## x3           1.1817     0.3553   3.33   0.0043 **  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.37 on 16 degrees of freedom  
## Multiple R-squared:  0.507, Adjusted R-squared:  0.414  
## F-statistic: 5.48 on 3 and 16 DF,  p-value: 0.00878
```

Skal modellen reduceres i backward selection step?

- A: Nej      B: Ja,  $x_1$  skal væk      C: Ja,  $x_2$  skal væk      D: Ja,  $x_3$  skal væk

# Spørgsmål om $\sigma$ på afvigelserne (socrative.com-ROOM:PBAC)

```
##  
## Call:  
## lm(formula = y ~ x1 + x2 + x3)  
##  
## Residuals:  
##      Min    1Q  Median    3Q   Max  
## -0.9195 -0.1555  0.0104  0.1465  0.6304  
##  
## Coefficients:  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -0.0528    0.2285   -0.23   0.8201  
## x1          -0.7357    0.3034   -2.42   0.0275 *  
## x2          0.2618    0.2937    0.89   0.3859  
## x3          1.1817    0.3553    3.33   0.0043 **  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.37 on 16 degrees of freedom  
## Multiple R-squared:  0.507, Adjusted R-squared:  0.414  
## F-statistic: 5.48 on 3 and 16 DF,  p-value: 0.00878
```

Hvad er den estimerede standard deviation på afvigelserne  $\hat{\sigma}$ ?

- A:  $\hat{\sigma} = 0.2285$       B:  $\hat{\sigma} = 0.0104$       C:  $\hat{\sigma} = 0.37$       D: Ved ikke

# Residual analyse (model kontrol)

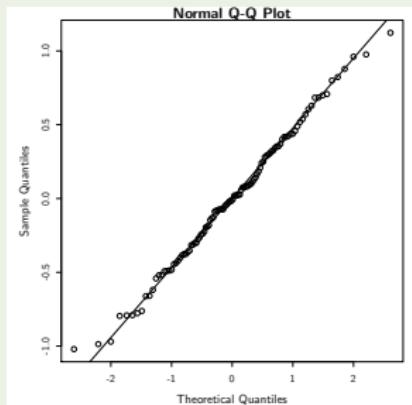
- Model kontrol: Analyser residualerne for at checke at forudsætningerne er opfyldt
- $\varepsilon_i \sim N(0, \sigma^2)$  og er independent and identically distributed (i.i.d.)
  - Husk:  $\varepsilon_i$  er afvigelsen (en stokastisk variabel)
  - Husk:  $e_i = \hat{\varepsilon}_i$  er residualen (realisationen eller observationen af afvigelsen)
- Samme som for simpel lineær model, dog også plot med residualer vs. inputs

# Antagelse om normalfordelte residualer

Lav et q-q plot for at se om de ikke afviger fra at være normalfordelt

```
## Gem det udvalgte fit
fitSel <- lm(logOzone ~ temperature + wind + radiation, data=Air)

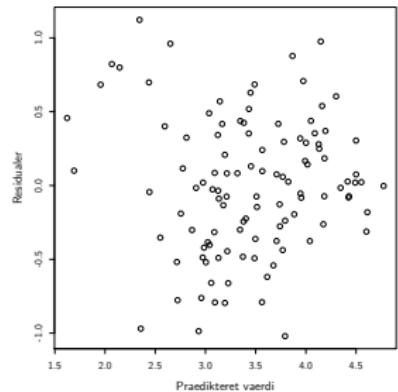
## qq-normalplot
qqnorm(fitSel$residuals)
qqline(fitSel$residuals)
```



# Antagelse om identisk distribution

Plot residualerne ( $e_i$ ) mod de prædikterede (fittede) værdier ( $\hat{y}_i$ )

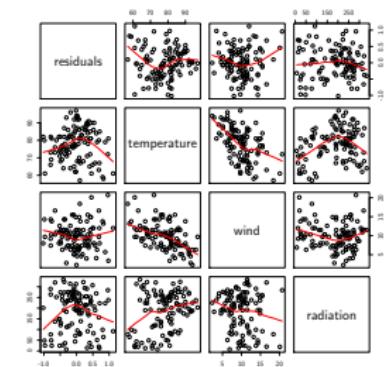
```
plot(fitSel$fitted.values, fitSel$residuals,
      xlab="Prædikteret værdi", ylab="Residualer")
```



Plot residualer mod de forklarende variabler

```
pairs(cbind(fitSel$residuals, Air[,c("temperature", "wind",
  "radiation")]), panel = panel.smooth)
```

Kan måske forbedres med ikke-lineær sammenhæng  
temperature eller vindhastighed.



# Kurvelineær (Curvilinear)

Hvis vi ønsker at estimere en model af typen

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

kan vi benytte multipel lineær regression i modellen

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$$

hvor

- $x_{i,1} = x_i$
- $x_{i,2} = x_i^2$

og benytte samme metoder som ved multipel lineær regression.

# Udvid ozon modellen med passende kurvelineær regression

```

## Lav den kvadrerede vind
Air$windSq <- Air$wind^2
## Tilføj den til modellen
fitWindSq <- lm(logOzone ~ temperature + wind + windSq + radiation, data=Air)
summary(fitWindSq)

## Gør tilsvarende for temperatur
Air$temperatureSq <- Air$temperature^2
## Tilføj
fitTemperatureSq <- lm(logOzone ~ temperature + temperatureSq + wind + radiation, data=Air)
summary(fitTemperatureSq)

## Gør tilsvarende for indstråling
Air$radiationSq <- Air$radiation^2
## Tilføj
fitRadiationSq <- lm(logOzone ~ temperature + wind + radiation + radiationSq, data=Air)
summary(fitRadiationSq)

## Hvilken en var bedst!?
summary(fitWindSq)
summary(fitTemperatureSq)

## Her kunne man prøve at udvide yderligere
fitWindSqTemperatureSq <- lm(logOzone ~ temperature + temperatureSq + wind + windSq + radiation, data=Air)
summary(fitWindSqTemperatureSq)

## Model kontrol
qqnorm(fitWindSq$residuals)
qqline(fitWindSq$residuals)
plot(fitWindSq$fitted.values, fitWindSq$residuals, pch=19)

#####
## Plot residualerne vs. de forklarende variabler
pairs(cbind(fitWindSq$residuals, Air[,c("temperature","wind","radiation")]), panel=panel.smooth)

```

# Konfidens- og prædiktionsintervaller

```
## Generer et nyt data.frame med konstant temperatur og instråling, men varierende vindhastighed
wind <- seq(1,20.3,by=0.1)
setTemperature <- 78
setRadiation <- 186
AirForPred <- data.frame(temperature=setTemperature, wind=wind, windSq=wind^2, radiation=setRadiation)

## Udregn konfidens- og prædiktionsintervaller (-bånd)
## Læg mærke til at der transformeres tilbage
CI <- exp(predict(fitWindSq, newdata=AirForPred, interval="confidence", level=0.95))
PI <- exp(predict(fitWindSq, newdata=AirForPred, interval="prediction", level=0.95))

## Plot them
Air$ozone <- exp(Air$logOzone)
plot(Air$wind, Air$ozone, ylim=range(CI,PI,Air$ozone), xlab="", ylab="")
title(xlab="Vindhastighed (MpH)", ylab="Ozon (ppb)", main=paste("Ved temperatur =",setTemperature, "F og indstr.
lines(wind, CI[, "fit"])
lines(wind, CI[, "lwr"], lty=2, col=2)
lines(wind, CI[, "upr"], lty=2, col=2)
lines(wind, PI[, "lwr"], lty=2, col=3)
lines(wind, PI[, "upr"], lty=2, col=3)
## legend
legend("topright", c("Prædiktionsbånd", "95% konfidensbånd"), lty=c(1,2,2), col=1:3)
```

# Kollinearitet (Colinearity)

Der opstår problemer hvis de forklarende variabler er stærkt korrelerede

```
## Lav en variabel, som er meget korreleret f.eks. endnu en vindmåling
set.seed(367)
Air$wind2 <- Air$wind + rnorm(nrow(Air), sd=1)
cor(Air$wind, Air$wind2)
plot(Air$wind, Air$wind2)
## Tilføj den til modellen
fitWind2 <- lm(logOzone ~ temperature + wind + wind2 + radiation, data=Air)
summary(fitWind2)

## Sammenlign med modellen med kun den ene
fitWind <- lm(logOzone ~ temperature + wind + radiation, data=Air)
summary(fitWind)
```

Få feedback fra TA om jeres projekt 1

# Introduktion til Statistik

## Forelæsning 10: Inferens for andele

Peder Bacher

DTU Compute, Dynamiske Systemer  
Bygning 303B, Rum 010  
Danmarks Tekniske Universitet  
2800 Lyngby – Danmark  
e-mail: pbac@dtu.dk

Forår 2021

# Kapitel 7: Inferens for andele

## Statistik for andele:

- Andel:  $p = \frac{x}{n}$  (*x successer ud af n observationer*)
- Specifikke metoder, én, to og  $k > 2$  grupper
  - Binær/kategorisk respons

## Specifikke metoder:

- Estimation og konfidensintervaller for andele
  - Metoder korrektion ved små stikprøver
- Hypoteser for én andel ( $p$ )
- Hypoteser for to andele
- Analyse af antalstabeller ( $\chi^2$ -test) (alle forventede antal  $> 5$ )

# Chapter 7: Inferences for Proportions

## Statistics for proportions:

- Proportion:  $p = \frac{x}{n}$  (*x successes out of n observations*)
- Specific methods: one, two and  $k > 2$  samples:
  - Binary/categorical response

## Specific methods:

- Estimation and confidence interval of proportions
  - Methods for correction for small samples
- Hypotheses for one proportion
- Hypotheses for two proportions
- Analysis of contingency tables ( $\chi^2$ -test) (all expected  $> 5$ )

# Oversigt

- 1 Intro
- 2 Konfidensinterval for én andel
  - Eksempel 1
- 3 Hypotesetest for én andel
  - Eksempel 1 - fortsat
- 4 Konfidensinterval og hypotesetest for forskel på to andele
  - Eksempel 2
- 5 Hypotesetest for flere andele
  - Eksempel 2 - fortsat
- 6 Analyse af antalstabeller

# Forskellige analyse/data-situationer

Hypotesetests og konfidensintervaller for:

- Én middelværdi (*one-sample, i.e. one group/population*)
- To middelværdier (*two-sample, i.e. two groups/populations*)
- Næste uge: For flere middelværdier (*k-sample, i.e. k groups/populations*)

I dag: Hypotesetests og konfidensintervaller for:

- Én andel
- To andele
- Flere andele (kun hypotesetest)
- Flere "multi-categorical" andele (kun hypotesetest)

# Estimation af andele

Estimation af andele fås ved at observere antal gange  $x$  en hændelse har indtruffet ud af  $n$  forsøg:

$$\hat{p} = \frac{x}{n}$$

$$\hat{p} \in [0; 1]$$

# Spørgsmål om andel (socrative.com, ROOM: pbac)

Hvilken kan ikke en være en andel?

- A: 103/900
- B: 12/80
- C: 0.957
- D: 202/154
- E: 0.224

# Konfidensinterval for én andel

## Method 7.3

Såfremt der haves en stor stikprøve, fås et  $(1 - \alpha)\%$  konfidensinterval for  $p$

$$[\hat{p} - z_{1-\alpha/2} \cdot \hat{\sigma}_{\hat{p}}, \quad \hat{p} + z_{1-\alpha/2} \cdot \hat{\sigma}_{\hat{p}}] \quad \left[ \hat{p} - z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \hat{p} + z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

(Vi siger: Med stor sikkerhed vælger vi at tro at  $p$  i dette interval)

## Hvordan?

Følger af at approximere binomialfordelingen med normalfordelingen

## As a rule of thumb

The normal distribution gives a good approximation of the binomial distribution if  $np$  and  $n(1-p)$  are both greater than 15

# Konfidensinterval for én andel

Middelværdi og varians i binomialfordelingen, kapitel 2:

$$\mathbb{E}(X) = np$$

$$\text{Var}(X) = np(1 - p)$$

Derfor får man

$$\mathbb{E}(\hat{p}) = \mathbb{E}\left(\frac{X}{n}\right) = \frac{np}{n} = p$$

$$\text{Var}(\hat{p}) = \sigma_{\hat{p}}^2 = \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2} \text{Var}(X) = \frac{p(1-p)}{n}$$

# Eksempel 1

Venstrehåndede:

$p$  = Andelen af venstrehåndede i Danmark

eller:

Kvindelige ingeniørstuderende:

$p$  = Andelen af kvindelige ingeniørstuderende

# Eksempel 1

Venstrehåndede ( $x = 10$  ud af  $n = 100$ ):

$$\hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{10/100(1-10/100)}{100}} = 0.03$$

$$0.10 \pm 1.96 \cdot 0.03 \Leftrightarrow 0.10 \pm 0.06 \Leftrightarrow [0.04, 0.16]$$

Bedre "small sample" metode - "plus 2-approach" (Remark 7.7):

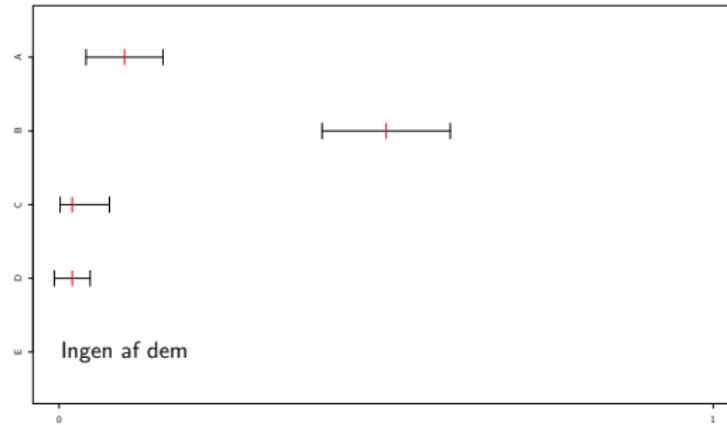
Anvend samme formel på  $\tilde{x} = 10 + 2 = 12$  og  $\tilde{n} = 104$ :

$$\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}} = \sqrt{\frac{12/104(1-12/104)}{104}} = 0.0313$$

$$0.115 \pm 1.96 \cdot 0.0313 \Leftrightarrow 0.115 \pm 0.061 \Leftrightarrow [0.054, 0.18]$$

# Spørgsmål om plus 2-approach (socrative.com, ROOM: pbac)

Hvilket af følgende intervaller er med plus 2-approach?



Ingen af dem

# Trin ved Hypotesetest

## Trin ved Hypotesetest:

1. Opstil hypoteser og vælg signifikansniveau  $\alpha$
2. Beregn teststørrelse
3. Beregn  $p$ -værdi (eller kritisk værdi)
4. Fortolk  $p$ -værdi og/eller sammenlign  $p$ -værdi og signifikansniveau, og derefter drag en konklusion

(Alternativ 4. Sammenlign teststørrelse og kritisk værdi og drag en konklusion)

# Hypotesetest for én andel

Vi betragter en nul- og alternativ hypotese for én andel  $p$ :

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

Man vælger som sædvanligt enten at acceptere  $H_0$  eller at forkaste  $H_0$

## Theorem 7.10 og Method 7.11

Såfremt stikprøven er tilstrækkelig stor ( $np_0 > 15$  og  $n(1 - p_0) > 15$ ) bruges teststørrelsen:

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}}$$

Under nulhypotesen gælder at den tilsvarende tilfældige variabel  $Z$  følger en standard normalfordeling, dvs.  $Z \sim N(0, 1^2)$

# Test ved brug af $p$ -værdi (Method 7.11)

Find  $p$ -værdien (bevis mod nulhypotesen):

- We only use two-sided:  $2P(Z > |z_{\text{obs}}|)$  in exercises and exams
- Remark 7.9 om one-sided "less" og "greater"

Kritiske værdier

Alternativ hypoteze	Afvis nulhypoteze hvis
$p \neq p_0$	$z_{\text{obs}} < -z_{1-\alpha/2}$ eller $z_{\text{obs}} > z_{1-\alpha/2}$

## Eksempel 1 - fortsat

Er halvdelen af alle danskere venstrehåndede?

$$H_0 : p = 0.5, \quad H_1 : p \neq 0.5$$

Teststørrelse:

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}} = \frac{10 - 100 \cdot 0.5}{\sqrt{100 \cdot 0.5(1 - 0.5)}} = -8$$

Er  $p$ -værdien under 0.05? (dvs. skal nulhypotesen forkastes ved  $\alpha = 0.05$ )

- A: Ja    B: Nej    C: Ved ikke

## R: prop.test - een andel

```
## Single proportion

## Testing the probability = 0.5 with a two-sided alternative
## We have observed 518 out of 1154
## Without continuity corrections

prop.test(x=518, n=1154, p = 0.5, correct = FALSE)
```

# Konfidensinterval for forskel på to andele

## Method 7.15

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2} \cdot \hat{\sigma}_{\hat{p}_1 - \hat{p}_2}$$

hvor

$$\hat{\sigma}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

## Rule of thumb:

Både  $n_i p_i \geq 10$  and  $n_i(1 - p_i) \geq 10$  for  $i = 1, 2$

# Hypotesetest for forskel på to andele, Method 7.18

## Two sample proportions hypothesis test

Såfremt man ønsker at sammenligne to andele (her vist for et tosidedt alternativ)

$$H_0: p_1 = p_2$$

$$H_1: p_1 \neq p_2$$

Fås teststørrelsen:

$$z_{\text{obs}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad \text{hvor} \quad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Og for passende store stikprøver:

Brug standardnormalfordelingen igen

## Eksempel 2

Sammenhæng mellem brug af p-piller og risikoen for blodprob i hjertet (hjerteinfarkt)

I et studie (USA, 1975) undersøgte man dette. Fra et hospital havde man indsamlet følgende to stikprøver

	p-piller	Ikke p-piller
Blodprob	23	35
Ikke blodprob	34	132

Er der sammenhæng mellem brug af p-piller og sygdomsrisiko

Udfør et test for om der er sammenhæng mellem brug af p-piller og risiko for blodprob i hjertet. Anvend signifikansniveau  $\alpha = 5\%$ .

## Eksempel 2

Sammenhæng mellem brug af p-piller og risikoen for blodprob i hjertet

	p-piller	Ikke p-piller
Blodprob	$x_1 = 23$	$x_2 = 35$
Ikke blodprob	34	132
Sum	$n_1 = 57$	$n_2 = 167$

Estimater i hver stikprøve

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{23}{57} = 0.40, \quad \hat{p}_2 = \frac{x_2}{n_1} = \frac{35}{167} = 0.21$$

## R: prop.test - to andele

```
## Pill study: two proportions

## Reading the table into R
pill.study <- matrix(c(23, 34, 35, 132), ncol = 2)
rownames(pill.study) <- c("Blood Clot", "No Clot")
colnames(pill.study) <- c("Pill", "No pill")

## Testing that the probabilities for the two groups are equal
prop.test(t(pill.study), correct = FALSE)
```

```
## Or simply directly by
prop.test(x=c(23,35), n=c(57,167), correct = FALSE)
```

# Hypotesetest for flere andele

## Sammenligning af $c$ andele

I nogle tilfælde kan man være interesseret i at vurdere om to eller flere binomialfordlinger har den samme parameter  $p$ , dvs. man er interesseret i at teste nulhypotesen

$$H_0 : p_1 = p_2 = \dots = p_c = p$$

mod en alternativ hypotese at disse andele ikke er ens

# Hypotesetest for flere andele

Tabel af observerede antal for  $c$  stikprøver:

	stikprøve 1	stikprøve 2	...	stikprøve $c$	Total
Succes	$x_1$	$x_2$	...	$x_c$	$x$
Fiasco	$n_1 - x_1$	$n_2 - x_2$	...	$n_c - x_c$	$n - x$
Total	$n_1$	$n_2$	...	$n_c$	$n$

Fælles (gennemsnitlig) estimat:

Under nulhypotesen fås et estimat for  $p$

$$\hat{p} = \frac{x}{n}$$

# Hypotesetest for flere andele

Fælles (gennemsnitlig) estimat:

Under nulhypotesen fås et estimat for  $p$

$$\hat{p} = \frac{x}{n}$$

"Brug" dette fælles estimat i hver gruppe:

såfremt nulhypotesen gælder, vil vi forvente at den  $j$ 'te gruppe har  $e_{1j}$  successer og  $e_{2j}$  fiaskoer, hvor

$$e_{1j} = n_j \cdot \hat{p} = n_j \cdot \frac{x}{n}$$

$$e_{2j} = n_j(1 - \hat{p}) = n_j \cdot \frac{n - x}{n}$$

# Hypotesetest for flere andele

Generel formel for beregning af forventede værdier i antalstabeller:

$$e_{ij} = (j\text{'th column total}) \cdot \frac{(i\text{'th row total})}{(\text{total})}$$

# Beregning af teststørrelse - Method 7.20

Teststørrelsen bliver

$$\chi^2_{\text{obs}} = \sum_{i=1}^2 \sum_{j=1}^c \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

hvor  $o_{ij}$  er observeret antal i celle  $(i,j)$  og  $e_{ij}$  er forventet antal i celle  $(i,j)$

# Find $p$ -værdi eller brug kritisk værdi - Method 7.20

Stikprøvefordeling for test-størrelse:

$\chi^2$ -fordeling med  $(c - 1)$  frihedsgrader

Kritisk værdi metode

Såfremt  $\chi^2_{\text{obs}} > \chi^2_{1-\alpha}(c - 1)$  forkastes nulhypotesen

Rule of thumb for validity of the test:

Alle forventede værdier  $e_{ij} \geq 5$

## Eksempel 2 - fortsat

De OBSERVEREDE værdier  $o_{ij}$

	p-piller	Ikke p-piller	Total
Blodprob	23	35	
Ikke blodprob	34	132	

## Eksempel 2 - fortsat

Beregn de **FORVENTEDE** værdier  $e_{ij}$  (altså forventede *under  $H_0$* )

	p-piller	Ikke p-piller	Total
Blodprob			$x = 58$
Ikke blodprob			
	$n_1 = 57$	$n_2 = 167$	$n = 224$

## Eksempel 2 - fortsat

Beregn de **FORVENTEDE** værdier  $e_{ij}$  (altså forventede *under  $H_0$* )

	p-piller	Ikke p-piller	Total
Blodprob	14.76	43.24	$x = 58$
Ikke blodprob	42.24	123.76	
	$n_1 = 57$	$n_2 = 167$	$n = 224$

Brug "reglen" for forventede værdier fire gange, f.eks. :

$$e_{12} = 167 \cdot \frac{58}{224} = 43.24$$

## Eksempel 2 - fortsat

Teststørrelsen:

$$\chi^2_{\text{obs}} = \frac{(o_{11} - e_{11})^2}{e_{11}} + \frac{(o_{12} - e_{12})^2}{e_{12}} + \frac{(o_{21} - e_{21})^2}{e_{21}} + \frac{(o_{22} - e_{22})^2}{e_{22}}$$

=

$$\chi^2_{\text{obs}} = \frac{(23 - 14.76)^2}{14.76} + \frac{(35 - 43.24)^2}{43.24} + \frac{(34 - 42.24)^2}{42.24} + \frac{(132 - 123.76)^2}{123.76}$$

= 8.33

Kritisk værdi og  $p$ -værdi:

```
## Kritisk værdi
qchisq(0.95, 1)
```

```
## [1] 3.8
```

```
## p-værdi
1 - pchisq(8.33, df=1)
```

```
## [1] 0.0039
```

## R: chisq.test - to andele

```
## Pill study: two proportions, chi-square test

## Chi2 test for testing the probabilities for the two groups are equal
chisq.test(pill.study, correct = FALSE)
## If we want the expected numbers save the test in an object
chi <- chisq.test(pill.study, correct = FALSE)
## The expected values
chi$expected
```

# Antalstabeller

## Antalstabbel

- Flere end 2 kategorier (f.eks. fire.: rød, grøn, blå, sort)
- Beregningerne er ens for begge følgende setups

## To mulige setups

- Setup 1:  $c$  stikprøver med  $r$  kategorier:
  - Test om der er forskel i fordelingen mellem kategorierne for hver stikprøve
- Setup 2: To kategoriske variabel ( $r$  kategorier) målt på samme individer (parret setup):
  - Test om der er forskel i fordelingen mellem de to grupper

# Setup 1: c stikprøver med r kategorier

En  $3 \times 3$  tabel - 3 stikprøver, 3-kategori udfald

	4 uger før	2 uger før	1 uge før
Kandidat I	79	91	93
Kandidat II	84	66	60
ved ikke	37	43	47
	$n_1 = 200$	$n_2 = 200$	$n_3 = 200$

Er stemmefordelingen ens?

$$H_0 : p_{i1} = p_{i2} = p_{i3}, i = 1, 2, 3$$

## Setup 2: To kategoriske variabel (r kategorier) målt på samme individer (parret setup)

En  $3 \times 3$  tabel - 1 stikprøve, to stk. 3-kategori variable:

	dårlig	middel	god
dårlig	23	60	29
middel	28	79	60
god	9	49	63

Er der uafhængighed mellem inddelingskriterier?

$$H_0 : p_{ij} = p_{i \cdot} \cdot p_{\cdot j}$$

f.eks. er der sammenhæng mellem den måde elever klarer sig i matematik som i dansk?

# Beregning af teststørrelse – uanset type af tabel

I en antalstable med  $r$  rækker og  $c$  søjler, fås teststørrelsen

$$\chi^2_{\text{obs}} = \sum_{i=1}^r \sum_{j=1}^c \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

hvor  $o_{ij}$  er observeret antal i celle  $(i,j)$  og  $e_{ij}$  er forventet antal i celle  $(i,j)$

Generel formel for beregning af forventede værdier i antalstabeller:

$$e_{ij} = (j\text{'th column total}) \cdot \frac{(i\text{'th row total})}{(\text{total})}$$

# Spørgsmål (socrative.com, ROOM: pbac)

En  $3 \times 4$  tabel - 4 stikprøver, 3-kategori udfald

	Gruppe A	Gruppe B	Gruppe C	Gruppe D	$n_j$
Han	3	3	2	2	10
Hun	3	3	5	2	13
Tvekøn	4	4	3	6	17
$n_i$	10	10	10	10	40

Hvad er  $e_{23}$ ? (forventning af hunner i gruppe C under  $H_0$ )

- A:  $10 \cdot 10/40$
- B: 3
- C:  $10 \cdot 13/40$
- D:  $17 \cdot 4/40$
- E: Ved ikke

# Find $p$ -værdi eller brug kritisk værdi – Method 7.22

Stikprøvefordeling for test-størrelse:

$\chi^2$ -fordeling med  $(r - 1)(c - 1)$  frihedsgrader

Kritisk værdi metode

Såfremt  $\chi^2_{\text{obs}} > \chi^2_{1-\alpha}$  med  $(r - 1)(c - 1)$  frihedsgrader forkastes nulhypotesen

Rule of thumb for validity of the test:

Alle forventede værdier  $e_{ij} \geq 5$

# R: chisq.test - antalstabeller

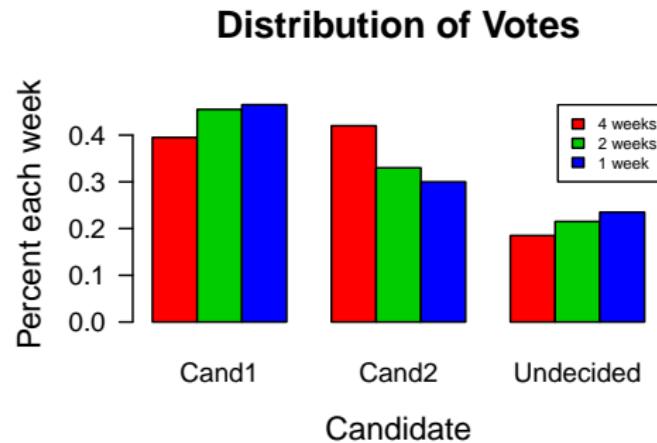
```
## Poll study: contingency table, chi-square test

## Reading the table into r
poll <- matrix(c(79, 91, 93, 84, 66, 60, 37, 43, 47), ncol = 3, byrow = TRUE)
colnames(poll) <- c("4 weeks", "2 weeks", "1 week")
rownames(poll) <- c("Cand1", "Cand2", "Undecided")

## Column percentages
colpercent <- prop.table(poll, 2)
colpercent
```

# R: chisq.test - antalstabeller

```
barplot(t(colpercent), beside = TRUE, col = 2:4, las = 1,  
       ylab = "Percent each week", xlab = "Candidate",  
       main = "Distribution of Votes")  
legend( legend = colnames(poll), fill = 2:4, "topright", cex = 0.5)  
par(mar=c(5,4,4,2)+0.1)
```



# R: chisq.test - antalstabeller

```
## Testing same distribution in the three populations
chi <- chisq.test(poll, correct = FALSE)
chi
## Expected values
chi$expected
```

# Kursus 02323: Introduktion til Statistik

## Forelæsning 11: Envejs variansanalyse, ANOVA

Peder Bacher

DTU Compute, Dynamiske Systemer  
Bygning 303B, Rum 010  
Danmarks Tekniske Universitet  
2800 Lyngby – Danmark  
e-mail: pbac@dtu.dk

Forår 2021

# Kapitel 8: Envejs variansanalyse (envejs ANOVA)

## $k$ UAFHÆNGIGE grupper

- Test om middelværdi for mindst en gruppe er forskellig fra de andre gruppers middelværdi
- Model  $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$

## Specifikke metoder, envejs variansanalyse:

- ANOVA-tabel:  $SST = SS(Tr) + SSE$
- $F$ -test
- Post hoc test(s): Parvise  $t$ -test med pooleret varians estimat
  - Hvis planlagt på forhånd, så uden Bonferroni korrektion
  - Hvis alle sammenligninger udføres, så med Bonferroni korrektion

# Chapter 8: One-way Analysis of Variance

$k$  INDEPENDENT samples (groups)

- Test if the mean of at least one of the groups is different from the mean of the other groups
- Model  $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

Specific methods, one-way analysis of variance:

- ANOVA-table:  $SST = SS(Tr) + SSE$
- $F$ -test
- Post hoc test(s): pairwise  $t$ -test with pooled variance estimate
  - If planned on beforehand, then without Bonferroni correction
  - If all samples are compared, then with Bonferroni correction

# Oversigt

- 1 Intro eksempel
- 2 Model og hypotese
- 3 Beregning - variationsopspaltning og ANOVA tabellen
- 4 Hypotesetest (F-test)
- 5 Post hoc sammenligninger
- 6 Model kontrol

## Envejs variansanalyse - eksempel

Gruppe A	Gruppe B	Gruppe C
2.8	5.5	5.8
3.6	6.3	8.3
3.4	6.1	6.9
2.3	5.7	6.1

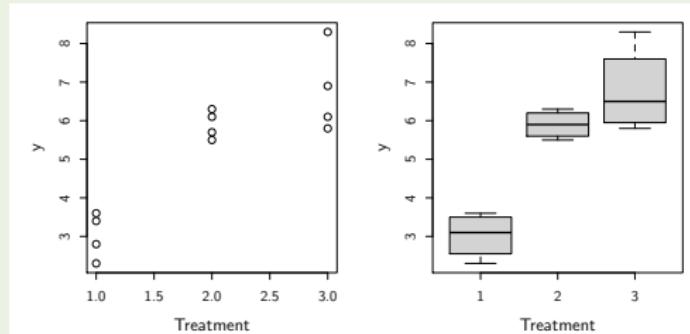
- Er der forskel på grupperne A, B og C?  
(dvs. forskel i middelværdi på population A, B og C?)
- Variansanalyse (ANOVA) kan anvendes til analysen såfremt observationerne i hver gruppe kan antages at være normalfordelte (vigtigt når man har få observationer, men jo flere man observationer man har des mindre vigtigt ifølge CLT)

# Envejs variansanalyse - eksempel

```
## Observationer
y <- c(2.8, 3.6, 3.4, 2.3,
      5.5, 6.3, 6.1, 5.7,
      5.8, 8.3, 6.9, 6.1)

## Grupper (behandlinger)
treatm <- factor(c(1, 1, 1, 1,
                   2, 2, 2, 2,
                   3, 3, 3, 3))

## Plot som punkter
plot(as.numeric(treatm), y, xlab="Treatment", ylab="y")
## Plot som box-plot
plot(treatm, y, xlab="Treatment", ylab="y")
```



# Envejs variansanalyse, model og hypotese

## Opstil en model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

hvor det antages, at

$$\varepsilon_{ij} \sim N(0, \sigma^2) \text{ and i.i.d.}$$

- $\mu$  er samlet middelværdi
- $\alpha_i$  angiver effekt af gruppe (behandling)  $i$
- $j$  tæller målinger i grupperne, fra 1 til  $n_i$  i hver gruppe

# Envejs variansanalyse, model og hypotese

## Hypotese

- Vi vil nu sammenligne (flere end to) middelværdier  $\mu + \alpha_i$  i modellen

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

- så vi opsætter hypotesen

$$H_0 : \alpha_i = 0 \quad \text{for alle } i$$

$$H_1 : \alpha_i \neq 0 \quad \text{for mindst et } i$$

# Envejs variansanalyse, opspaltnings og ANOVA tabellen

Med modellen

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

kan den totale variation i  $Y$  opspaltes

$$SST = SS(Tr) + SSE$$

hvor

- $SST$ : Kvadratafvigelsessum ("den totale varians")
  - $SSE$ : Kvadratafvigelsessum af residualer ("variens tilbage efter model")
  - $SS(Tr)$ : Kvadratafvigelsessum af gruppering ("variens forklaret af model")
- 
- "Envejs" hentyder til, at der kun er én faktor (én opdeling) i forsøget, på i alt  $k$  nivauer
  - Metoden kaldes variansanalyse, fordi testningen foregår ved at sammenligne varianser

# Formler for kvadratafvigelsessummer

- Kvadratafvigelsessum ( “*den totale varians*” )

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

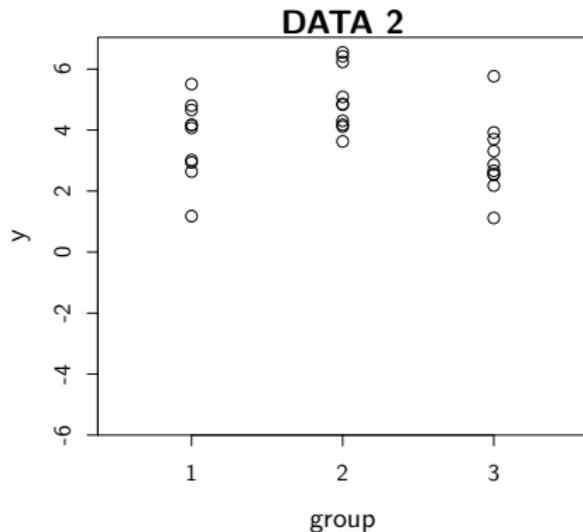
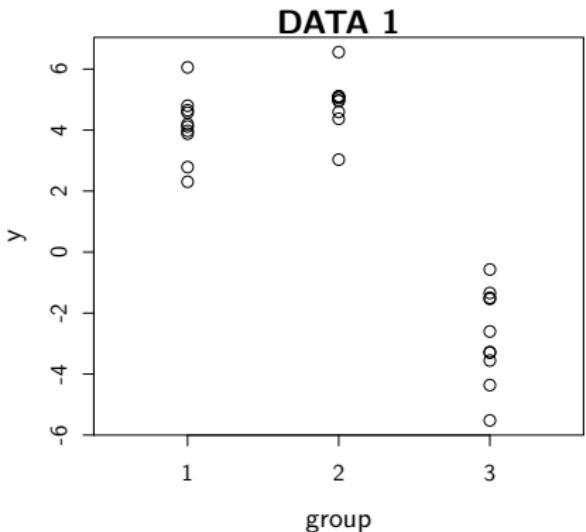
- Kvadratafvigelsessum af residualer ( “*varians tilbage efter model*” )

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

- Kvadratafvigelsessum af gruppering ( “*varians forklaret af model*” )

$$SS(Tr) = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 = SST - SSE$$

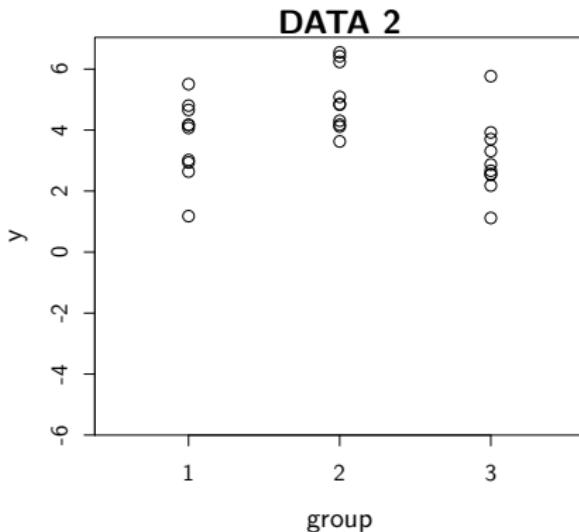
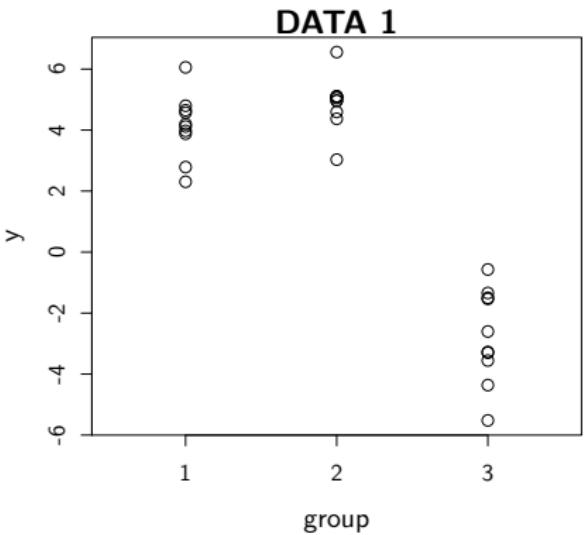
## Spørgsmål den totale varians (SST) Socrative.com, room: PBAC



For hvilken data er SST (totale variation) størst?

- A: DATA1
- B: DATA2
- C: Omtrent lige stor
- D: Ved ikke

## Spørgsmål: residual variansen (SSE) Socrative.com, room: PBAC



For hvilken data er SSE (residual variationen) størst?

- A: DATA1
- B: DATA2
- C: Omtrent lige stor
- D: Ved ikke

# Envejs variansanalyse, F-test

- Vi har altså

$$SST = SS(Tr) + SSE$$

- og  $H_0 : \alpha_i = 0$  for alle  $i$  (dvs. ingen forskel i middelværdi), da vil teststatistikken

$$F = \frac{SS(Tr)/(k-1)}{SSE/(n-k)}$$

følge en  $F$ -fordeling, hvor

- $k$  er antal nivauer af faktoren (antal grupper)
- $n$  er antal observationer
- Signifikansniveau  $\alpha$  vælges og teststatistikken  $F_{obs}$  beregnes
- Teststatistikken sammenlignes med en fraktil i  $F$  fordelingen

$$F \sim F_{\alpha}(k-1, n-k)$$

# F-fordeling

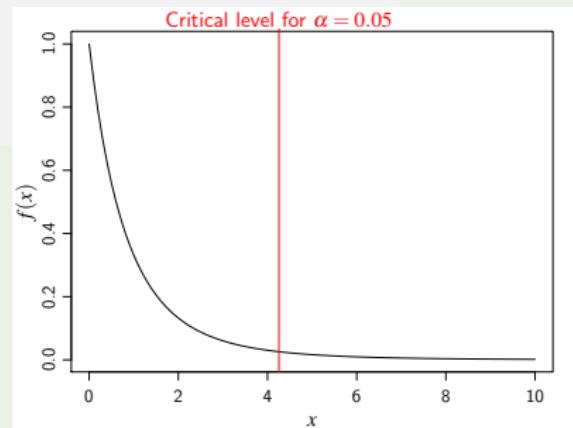
```

## Husk, dette er under  $H_0$  (altså vi regner som om  $H_0$  er sand):
## Antal grupper
k <- 3
## Antal punkter
n <- 12
## Sekvens til plot
xseq <- seq(0, 10, by=0.1)

## Plot F fordelingens tæthedsfunktion
plot(xseq, df(xseq, df1=k-1, df2=n-k), type="l", xlab="x", ylab="f(x)")
## Kritisk værdi for signifikans niveau 5 %
cr <- qf(0.95, df1=k-1, df2=n-k)
## Tegn den i plottet
abline(v=cr, col="red")

## Test statistikkens værdi
(Fobs <- (SSTr/(k-1)) / (SSE/(n-k)))
## p-værdien er da
(1 - pf(Fobs, df1=k-1, df2=n-k))

```



# Variansanalysetabel

Variations-kilde	Frihedsgrader	Kvadrat-afvig. sum	Gns. kvadratafv. sum	Test-størrelse $F$	p-værdi
<i>Source of variation</i>	Deg. of freedom	Sums of squares	Mean sum of squares	Test-statistic $F$	$p$ -value
<i>Gruppering</i>	$k - 1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F_{\text{obs}} = \frac{MS(Tr)}{MSE}$	$P(F > F_{\text{obs}})$
<i>Residual</i>	$n - k$	$SSE$	$MSE = \frac{SSE}{n-k}$		
<i>Total</i>	$n - 1$	$SST$			

```
## Alt dette beregnes med lm() og anova()

anova(lm(y ~ treatm))

## Analysis of Variance Table
##
## Response: y
##              Df Sum Sq Mean Sq F value    Pr(>F)
## treatm      2   30.8   15.40    26.7 0.00017 ***
## Residuals  9    5.2    0.58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Spørgsmål ANOVA table Socrative.com, room: PBAC

```
anova(lm(y ~ treatm))

## Analysis of Variance Table
##
## Response: y
##             Df  Sum Sq Mean Sq F value Pr(>F)
## treatm      3    37.6   12.54    4.51  0.024 *
## Residuals  12    33.3    2.78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hvad er den totale variation SST?

- A: 12.54
- B: 37.6
- C: 70.9
- D: Ved ikke

# Spørgsmål ANOVA table Socrative.com, room: PBAC

```
anova(lm(y ~ treatm))

## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq  Mean Sq F value Pr(>F)
## treatm     3    37.6    12.54    4.51   0.024 *
## Residuals 12    33.3     2.78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Husk antagelsen om normalfordelte afvigelser  $\varepsilon_{ij} \sim N(0, \sigma^2)$

Hvad er  $\hat{\sigma}^2$ ?

- A:  $\frac{33.3}{12}$
- B:  $\frac{37.6}{3}$
- C: 4.51
- D: Ved ikke

# Spørgsmål ANOVA table Socrative.com, room: PBAC

```
anova(lm(y ~ treatm))

## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value Pr(>F)
## treatm     3    37.6   12.54    4.51  0.024 *
## Residuals 12    33.3    2.78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Konklusionen på 5% signifikansniveau test af:  $H_0: \alpha_i = 0$  for alle  $i$ ?

- A:  $H_0$  accepteres      B:  $H_0$  afvises      C: Ved ikke

# Post hoc konfidensinterval

*Enkelt forudplanlagt* konfidensinterval for forskel på to grupper

- En enkelt forudplanlagt sammenligning af forskelle på gruppe  $i$  og  $j$  findes ved

$$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2} \sqrt{MSE \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}$$

hvor  $t_{1-\alpha/2}$  er fra  $t$ -fordelingen med  $n - k$  frihedsgrader

- Forskel fra Welch two-sample test: Alle observationer er anvendt i beregningen af  $MSE = SSE/(n - k) = s_p^2$  (i.e. pooled varians estimat med alle observationer)

*Mange* konfidensintervaller

- Hvis alle  $M = k(k - 1)/2$  kombinationer af parvise konfidensintervaller udføres, brug da formlen  $M$  gange, men hver gang med  $\alpha_{\text{Bonferroni}} = \alpha/M$

# Post hoc parvis hypotesetest

*Enkelt forudplanlagt t-test for forskel på grupper*

- En enkelt forudplanlagt hypotesetest på  $\alpha$  signifikansniveau om forskel af gruppe  $i$  og  $j$

$$H_0: \mu_i = \mu_j, \quad H_1: \mu_i \neq \mu_j$$

udføres ved

$$t_{\text{obs}} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}}$$

og

$$p\text{-value} = 2P(t > |t_{\text{obs}}|)$$

hvor  $t$ -fordelingen med  $n - k$  frihedsgrader anvendes

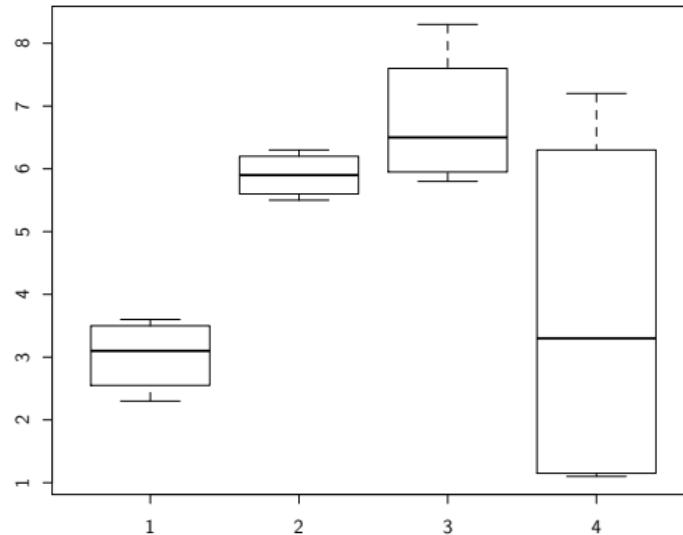
## *Mange t-tests*

- Hvis alle  $M = k(k - 1)/2$  kombinationer af hypotesetests udføres, da bruges det korrigerede signifikansniveau  $\alpha_{\text{Bonferroni}} = \alpha/M$

# Varians homogenitet

Se på box-plot om spredning ser meget forskellig ud for hver gruppe

```
## Box plot  
plot(treatm,y)
```

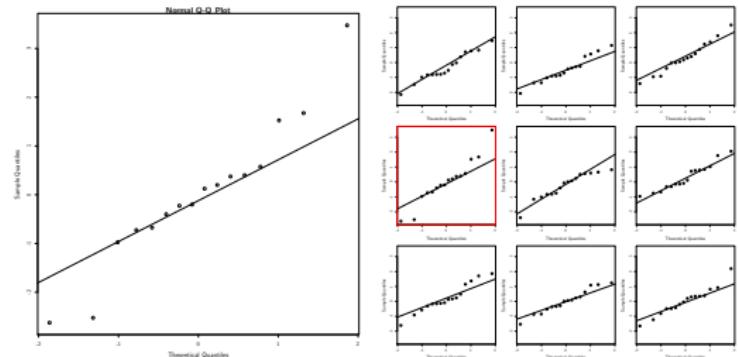


# Normalfordelingsantagelse

Se på qq-normal plot

```
## qq-normal plot af residualer
fit1 <- lm(y ~ treatm)
qqnorm(fit1$residuals)
qqline(fit1$residuals)

## Eller med et Wally plot
library(MESS)
qqwrap <- function(x, y, ...) {qqnorm(y, main="", ...); qqline(y)}
## Kan vi se et afvigende qq-norm plot?
wallyplot(fit1$residuals, FUN = qqwrap)
```



# Kursus 02323: Introducerende Statistik

## Forelæsning 12: Forsøgsplanlægning

Peder Bacher

DTU Compute, Dynamiske Systemer  
Bygning 303B, Rum 010  
Danmarks Tekniske Universitet  
2800 Lyngby – Danmark  
e-mail: pbac@dtu.dk

Forår 2021

## Afsnit 3.3 og 7.2.2: Forsøgsplanlægning

Grundlæggende koncepter for forsøgsplanlægning:

- Testens styrke er  $1 - \beta$  (*hvor  $\beta$  er sandsynligheden for at begå Type II fejl*)

Specifikke metoder, forsøgsplanlægning  
(middelværdi, både one og two sample setup):

- Stikprøvestørrelse  $n$  for ønsket præcision af konfidensintervaller
- Stikprøvestørrelse  $n$  for ønsket styrke af tests

Specifikke metoder, forsøgsplanlægning  
(andel, one sample setup):

- Stikprøvestørrelse  $n$  for ønsket præcision af konfidensintervaller

## Section 3.3 and 7.2.2: Design of experiments

General concepts for design of experiments:

- Power of a test is  $1 - \beta$  (*where  $\beta$  is the probability of making a Type II error*)

Specific methods, design of experiments  
(mean, both one and two sample setup):

- Sample size  $n$  for wanted precision of confidence intervals
- Sample size  $n$  for wanted power of tests

Specific methods, design of experiments  
(proportion, one-sample setup):

- Sample size  $n$  for wanted precision of confidence intervals

# Oversigt

- 1 Planlægning af studie med krav til præcision
- 2 Planlægning: Power og sample size
- 3 Andele: Bestemmelse af stikprøvestørrelse

# Planlægning af studie med krav til præcision

Man vil gerne tænke sig om inden eksperimentet udføres:

- Brug for at vide hvor præcise resultater (f.eks. konfidensinterval) forventes at blive med et fremtidigt eksperiment
- Hvor stor en effekt forventes at kunne opdages (e.g. hvis sovemedicinen virker 2 timer bedre, hvad er sandsynligheden for at det opdages?)
- Spørgsmål om at optimere økonomiske ressourcer og etik!

## Method 3.63: The one-sample CI sample size formula

Antal observationer og den forventede (halve) bredde af konfidensintervallet

When  $\sigma$  is known or guessed at some value, we can calculate the sample size  $n$  needed to achieve a given margin of error,  $ME$ , with probability  $1 - \alpha$  as

$$n = \left( \frac{z_{1-\alpha/2} \cdot \sigma}{ME} \right)^2$$

- Margin of error  $ME$  er den *halve bredde* af konfidensintervallets forventede bredde

# Eksempel på to mulige fejl ved hypotesetest – sovemedicin

To mulige sandheder vs. to mulige konklusioner:

<i>Test udfald</i>		
<i>Virkelighed</i>	<b>Afvis <math>H_0</math></b>	<b>Afvis ikke <math>H_0</math></b>
<b>Sand <math>H_0</math>:</b> Medicinen virker ikke	Type I fejl ( $\alpha$ )	Korrekt: Ingen effekt
<b>Falsk <math>H_0</math>:</b> Medicinen virker	Korrekt: Påvist effekt	Type II fejl ( $\beta$ )

# Eksempel på to mulige fejl ved hypotesetest (repetition)

To mulige sandheder vs. to mulige konklusioner:

<i>Test udfald</i>			
<i>Virkelighed</i>	<b>Afvis <math>H_0</math></b>	<b>Afvis ikke <math>H_0</math></b>	
<b>Sand <math>H_0</math></b>	Type I fejl ( $\alpha$ )	Korrekt accept af $H_0$	
<b>Falsk <math>H_0</math></b>	Korrekt afvisning af $H_0$	Type II fejl ( $\beta$ )	

# Mulige fejl ved hypotesetests (repetition)

Der findes to slags fejl (dog kun een af gangen!)

Type I: Rejection of  $H_0$  when  $H_0$  is true

(Medicinen virker ikke, men vi kommer til at tro den virker)

Type II: Non-rejection of  $H_0$  when  $H_1$  is true

(Medicinen virker, men vi opdager det ikke)

Risikoen for de to typer fejl kaldes sædvanligvis:

$$P(\text{Type I error}) = \alpha$$

$$P(\text{Type II error}) = \beta$$

# Testens styrke (power)

Hvad er styrken  $1 - \beta$  for et kommende studie/eksperiment:

- Sandsynligheden for at opdage en effekt (af en vis størrelse  $|\mu_0 - \mu_1|$ )
- Probability of correct rejection of  $H_0$
- $P(\text{"Accept af } H_0\text{"})$  når  $H_1$  er sand

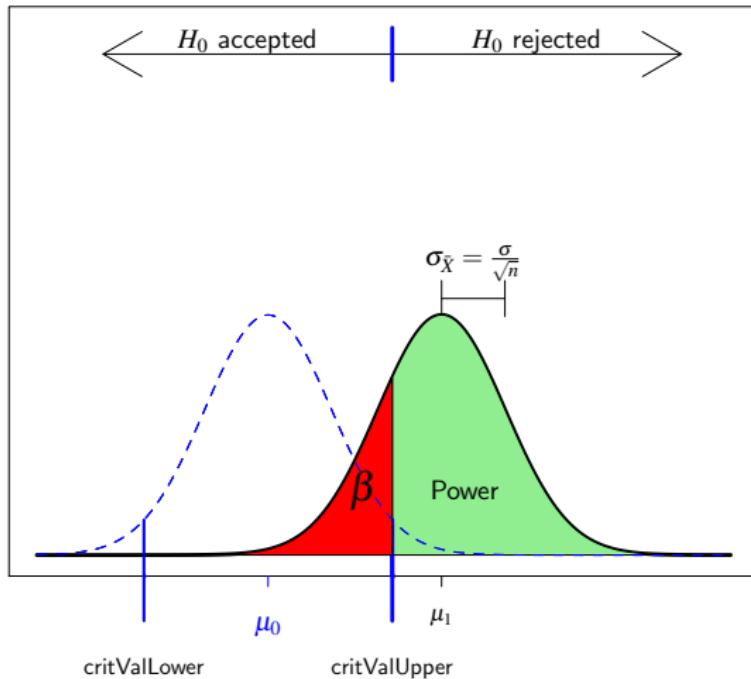
Udfordring: Nulhypotesen kan være forkert på mange måder!

I praksis, scenarie-baseret approach:

- F.eks. "Hvis nu sovemedicinen *rent faktisk virker 2 timer bedre*, hvor godt vil mit studie være til at opdage dette? "
- Eller, jeg vil gerne *hvis min sovemedicin virker 3 timer bedre*, opdage dette med en bestemt sandsynlighed (power)

$\bar{X} \sim N(\mu_1, \frac{\sigma^2}{n})$  er den antagede fordeling (dvs. om  $\mu_1$  og  $ME = |\mu_1 - \mu_0|$ )  
 $H_0: \mu = \mu_0$  er nulhypotesen

Vi kan se hvad  $\beta$  er:  $P(H_0 \text{ accepteres forkert}) = P(\text{Type II}) = \beta$

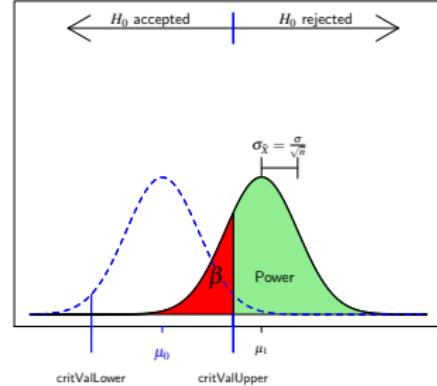


## Spørgsmål om power (socrative.com - ROOM:PBAC)

Vi antager at  $\bar{X} \sim N(\mu_1, \frac{\sigma^2}{n})$  (dvs. fordelt om  $\mu_1$ , som er effekten vi vil kunne påvise),  $H_0 : \mu = \mu_0$  er nulhypotesen

Hvordan kan vi opnå en større power *uden at kompromitere noget ved testen* (dvs. ikke ændre på hypotesen eller risikoen for type I fejl)?

- A: Mindske  $\mu_0$  så  $ME = |\mu_0 - \mu_1|$  øges
- B: Øge  $\alpha$  (på figur vil 'critvalUpper' mindskes)
- C: Øge  $n$  antallet af observationer
- D: Desværre kan det ikke lade sig gøre
- E: Ved ikke



# Planlægning, find sample size $n$

Det store spørgsmål i praksis: HVAD skal  $n$  være?

Forsøget skal være stort nok til at kunne opdage en relevant effekt med stor power (som regel mindst 80%):

Metode 3.65: Tilnærmet svar for en one-sample  $t$ -test:

For the one-sample  $t$ -test for given  $\alpha$ ,  $\beta$  and  $\sigma$

$$n = \left( \sigma \frac{z_{1-\beta} + z_{1-\alpha/2}}{(\mu_0 - \mu_1)} \right)^2$$

where  $\mu_0 - \mu_1$  is the difference in means that we would want to detect and  $z_{1-\beta}$ ,  $z_{1-\alpha/2}$  are quantiles of the standard normal distribution.

# Planlægning, sæt 4 prm. og beregn den sidste

Når man har fastlagt hvilket test, der skal bruges:

Kender man (eller fastlægger/gætter på) fire ud af følgende fem oplysninger, kan man regne sig frem til den femte:

- $n$  Sample size
- $\alpha$  Significance level of the test
- $\delta$  A difference in mean that you would want to detect (effect size)  
(dvs.  $\mu_2$  er her den middelværdi med afstand til  $\mu_1$  som vi "mindst" vil kunne påvise)
- $\sigma$  The population standard deviation
- $1 - \beta$  The power

# Eksempel - The sample size for power= 0.80

```
## Stikprøvestørrelse for t-test
power.t.test(power = .80, delta = 4, sd = 12.21, sig.level=0.05,
             type = "one.sample")

##
##      One-sample t test power calculation
##
##              n = 75.08
##              delta = 4
##              sd = 12.21
##              sig.level = 0.05
##              power = 0.8
##      alternative = two.sided
```

Svar:  $n = 76$  (*husk at runde op*)

### Metode 3.65: *Tilnærmet* svar for en one-sample *t*-test:

Formlen giver lidt lavere resultat, fordi normalfordelingen bruges i stedet for *t*-fordelingen:

$$\begin{aligned} n &= \left( \sigma \frac{z_{1-\beta} + z_{1-\alpha/2}}{(\mu_0 - \mu_1)} \right)^2 \\ &= \left( 12.21 \frac{0.84 + 1.96}{4} \right)^2 \\ &= 73.05 \end{aligned}$$

I opgaver bruges resultatet fra `power.t.test()`, hvis der ikke henvises konkret til formlen fra bogen.

# Eksempel - The power for $n = 40$

```
## Beregn power for t-test
power.t.test(n = 40, delta = 4, sd = 12.21, sig.level=0.05,
             type = "one.sample")

##
##      One-sample t test power calculation
##
##              n = 40
##              delta = 4
##              sd = 12.21
##              sig.level = 0.05
##              power = 0.5242
##              alternative = two.sided
```

# Styrke og stikprøvestørrelse - two-sample

Finding the sample size for detecting a group difference of 2 with  $\sigma = 1$  and power= 0.9:

```
## Beregn stikprøvestørrelsen
power.t.test(power = 0.90, delta = 2, sd = 1, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##              n = 6.387
##              delta = 2
##              sd = 1
##              sig.level = 0.05
##              power = 0.9
##      alternative = two.sided
##
## NOTE: n is number in *each* group
```

Svar:  $n = 7$  (*husk at runde op*)

# Styrke og stikprøvestørrelse - two-sample

Finding the power of detecting a group difference of 2 with  $\sigma = 1$  for  $n = 10$ :

```
## Power beregning
power.t.test(n = 10, delta = 2, sd = 1, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##              n = 10
##              delta = 2
##              sd = 1
##              sig.level = 0.05
##              power = 0.9882
##              alternative = two.sided
##
## NOTE: n is number in *each* group
```

# Styrke og stikprøvestørrelse - two-sample

Finding the detectable effect size (delta) with  $\sigma = 1$ ,  $n = 10$  and power= 0.9:

```
## Beregn margin of error
power.t.test(power = 0.90, n = 10, sd = 1, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##              n = 10
##              delta = 1.534
##              sd = 1
##              sig.level = 0.05
##              power = 0.9
##      alternative = two.sided
##
## NOTE: n is number in *each* group
```

# Andele: Stikprøvestørrelse: "Margin of Error" (ME):

Margin of Error på estimat kan siges at være:

- Forventningsværdi af "halvdelen af konfidensintervallets bredde"
- "Den forskel i middelværdi" man gerne vil være i stand til at påvise
- Under  $H_0$ : Forventningsværdi af afstanden mellem middelværdien og det kritiske niveau

Margin of Error

med  $(1 - \alpha)\%$  konfidens bliver

$$ME = z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

hvor et estimat af  $p$  fås ved  $p = \frac{x}{n}$

## Spørgsmål om Margin of Error (socrative.com, ROOM: pbac)

Hvad er "Margin of Error" (ME) hvis man vil have et konfidensinterval med bredde på 0.2?

- A: 0.1
- B: -0.15
- C: 0.2
- D: 0.4
- E: Ved ikke

## Spørgsmål om forskel i middel (socrative.com, ROOM: pbac)

Hvad er "Margin of Error" (ME) hvis man vil være i stand til at påvise forskel i middelværdi på 0.2?

- A: 0.1
- B: -0.15
- C: 0.2
- D: 0.4
- E: Ved ikke

# Bestemmelse af stikprøvestørrelse

## Method 7.13

Såfremt man højst vil tillade en Margin of Error ( $ME$ ) med  $(1 - \alpha)\%$  konfidens, bestemmes den nødvendige stikprøvestørrelse ved

$$n = p(1 - p) \left[ \frac{z_{1-\alpha/2}}{ME} \right]^2$$

## Method 7.13

Såfremt man højst vil tillade en Margin of Error  $ME$  med  $(1 - \alpha)\%$  konfidens, og  $p$  ikke kendes, bestemmes den nødvendige stikprøvestørrelse ved

$$n = \frac{1}{4} \left[ \frac{z_{1-\alpha/2}}{ME} \right]^2$$

idet man får den mest konservative stikprøvestørrelse ved at vælge  $p = \frac{1}{2}$

## Eksempel 1 - fortsat

Venstrehåndede:

Antag vi ønsker  $ME = 0.01$  (med  $\alpha = 0.05$ ) - hvad skal  $n$  være?

Antag  $p \approx 0.10$ :

$$n = 0.1 \cdot 0.9 \left( \frac{1.96}{0.01} \right)^2 = 3467.4 \approx 3468$$

UDEN antagelse om størrelsen af  $p$ :

$$n = \frac{1}{4} \left( \frac{1.96}{0.01} \right)^2 = 9604$$

# Spørgsmål om stikprøvestørrelse (socrative.com, ROOM: pbac)

Ved test af hvilken af følgende nulhypoteser skal bruges den største stikprøvestørrelse ( $n$ ) ved samme  $\alpha$  signifikansniveau?

- A:  $H_0 : p = 0.2$
- B:  $H_0 : p = 0.1$
- C:  $H_0 : p = 0.4$
- D:  $H_0 : p = 0.95$
- E: Ved ikke

# Spørgsmål om stikprøvestørrelse (socrative.com, ROOM: pbac)

Kan I nu beregne hvor mange gange man skal slå med en terning for at teste om den har sandsynlighed  $1/6$  indenfor  $0.01$  for at slå en sekser?

- A: Ja
- B: Nej
- C: Ved ikke

```
## Andel (sandsynlighed) vi vil teste for
p <- 1/6
## Signifikansniveau
## (hvor ofte vil vi lave denne fejl: Terningen er fair, men
## vi konkluderer den ikke er fair)
alpha <- 0.05
## Fejlmargen vi vil tillade
ME <- 0.01
## Beregn antal gange vi skal slå med terningen
p * (1-p) * (qnorm(1-alpha/2)/ME)^2
## [1] 5335
```

Husk også at sige at Exercise 3.10 spørgsmål c) er ret abstrakt og man kan springe den over.