

02402 Statistik (Polyteknisk grundlag)

Uge 1: Introduktion og R

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

Agenda

1 Praktiske informationer

2 Introduktion og motivation

3 Deskriptiv Statistik

- Middelværdi og median (centralitetsmål)
- Varians og standardafvigelse
- Fraktiler
- Kovarians og korrelation

4 Software: R & RStudio

Overview

1 Praktiske informationer

2 Introduktion og motivation

3 Deskriptiv Statistik

- Middelværdi og median (centralitetsmål)
- Varians og standardafvigelse
- Fraktiler
- Kovarians og korrelation

4 Software: R & RStudio

Praktiske informationer

• Undervisning

- Forelæsninger: Tirsdag 8-10
 - Bygning 303A, Aud. 41-42-43.
- Øvelser: Tirsdag 10-12
 - Bygning 303A, øvelsesområder HVEST/HOEST (HOEST til KID-studerende).
 - Bygning 324, stueetagen (Foyer øvelsesområder: 003, 004, 005 og 008. Lokaler: 020, 030, 040, 050, 060 og 070).

• Eksamens

- Lørdag den 16. december 2023.
- 4 timers multiple choice-prøve.

• Obligatoriske projekter

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

Praktiske informationer

• Generel ugeseddel

- Før undervisningen: Læs de relevante kapitler/afsnit i bogen/e-noten.
- Forelæsninger: Gennemgang af ugens pensum.
- Øvelser: Opgaveregning og online quizzers.
- Efter undervisningen: Area9 og “eksamensquizzers”.

• Undervisningsmateriale

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

Praktiske informationer

- Kursushjemmeside: 02402.compute.dtu.dk
 - Bog
 - Pensum
 - Undervisningsplan (agenda)
 - Øvelser og løsninger (engelsk)
 - Dias (dansk og engelsk)
 - Tidligere års forelæsninger (dansk og engelsk)
 - Quizzer
- DTU Learn
 - Beskeder
 - Projekter - formulering og aflevering
- Ed Discussion
 - Spørgsmål og diskussioner

Special for E2023

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

- 02323 lectures: Friday 8-10.

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

- 02402 lectures: Tirsdag 8-10.

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

Overview

1 Praktiske informationer

2 Introduktion og motivation

3 Deskriptiv Statistik

- Middelværdi og median (centralitetsmål)
- Varians og standardafvigelse
- Fraktiler
- Kovarians og korrelation

4 Software: R & RStudio

Indledning

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

Man vil uddrage viden og lære fra observerede data.

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

Man vil udlede viden og lære fra en teoretisk model.

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

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

Forskellige aspekter

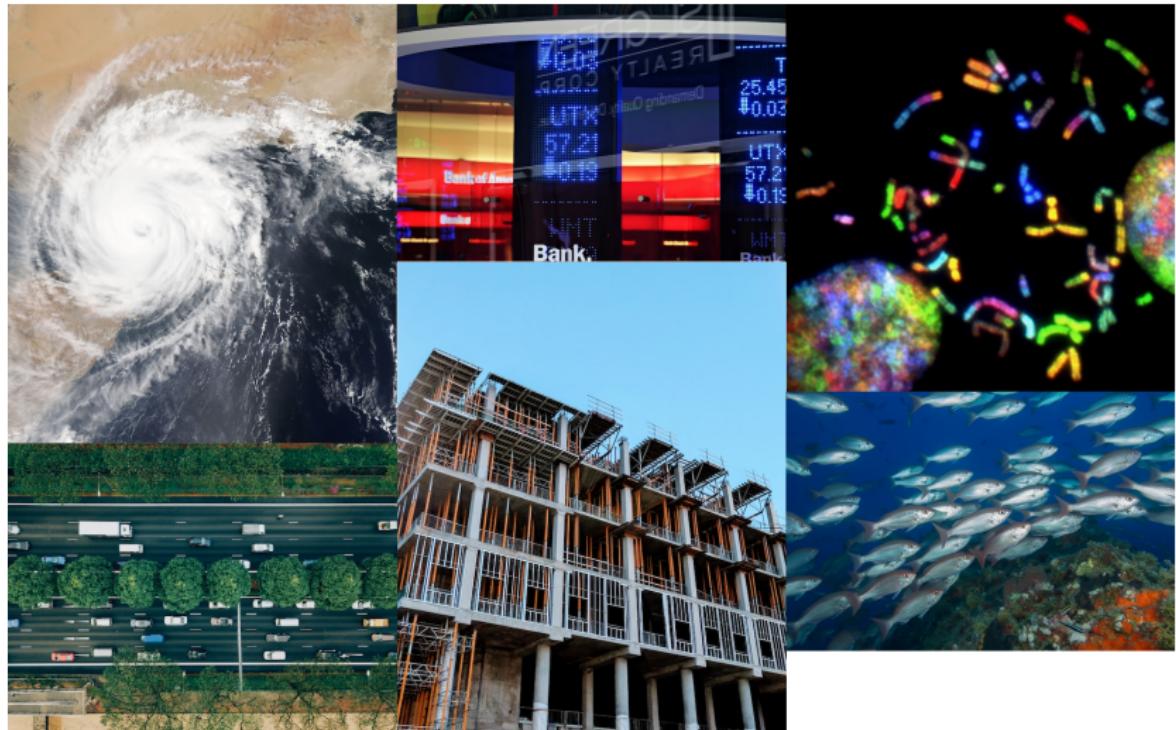
Der er mange spændende forskningsområder inden for både anvendt og teoretisk statistik.

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

- Matematisk analyse
- Numerisk optimering
- Operationsanalyse
- Kontrolteori

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

Anvendelse



Intro case-historier:

IBM big data, Novo Nordisk small data, Skive fjord

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

I hverdagen

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

- Nyheder
- Politik
- Reklamer
- Sport
- Arbejde

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

Almindelige fejslutninger og bias

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

Nogle typiske biases (systematiske skævvridninger) i statistik er:

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

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

Kursets overordnede mål og afgrænsning

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

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

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

Kursets indhold i store træk

En stor del af kurset omhandler:

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

i forskellige kontekster og setups.

Sandsynlighedsregningen bliver vores primære værktøj.

Grundlæggende om statistik

Statistik kan generelt opdeles i to dele:

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

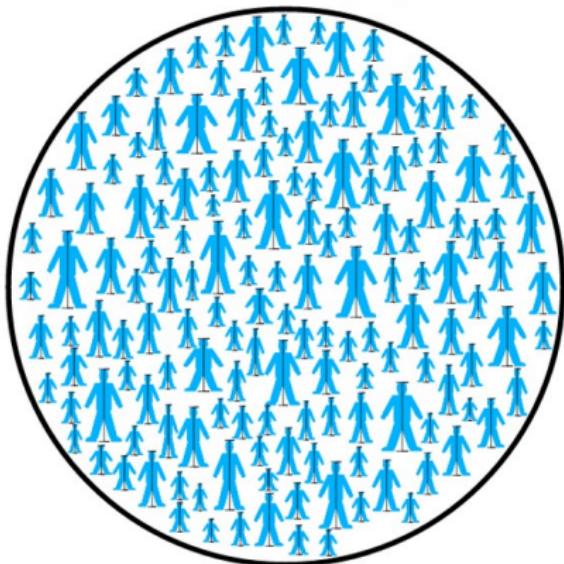
Statistik handler typisk om at analysere en *stikprøve*, taget ud af en *population*.

Ud fra stikprøven, udtaler vi os generelt om populationen.

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

Populationen og stikprøven

(Infinite) Statistical population

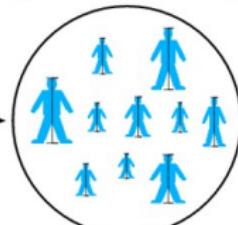


$$\text{Mean} \\ \mu$$

Statistical
Inference

Randomly
selected

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



$$\text{Sample mean} \\ \bar{x}$$

Overview

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

Det generelle setup

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

Stikprøven bliver almindeligvis repræsenteret med en vektor

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

Den sorterade stikprøve er så

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

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

Nøgletal (Summary statistics)

Nøgletal bruges til at opsummere og beskrive data.

- *Positionsmål*

- f.eks.: gennemsnit, median og fraktiler

- *Spredningsmål*

- f.eks.: varians og standardafvigelse

- *Sammenhængsmål*

- f.eks.: kovarians og korrelation

Husk at skelne mellem nøgletal for populationen og stikprøven!

Gennemsnit, definition 1.4

Gennemsnittet er et nøgletal, der angiver tyngdepunktet for data.

Middelværdien af en stikprøve (Stikprøvegennemsnittet):

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

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

Median, Definition 1.5

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

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

Medianen af en stikprøve (stikprøvemedianen):

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

Eksempel: Højde på studerende

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

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

- **Gennemsnit:**

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

- **Median:**

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

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

Stikprøvevariansen indikerer, hvor meget observationerne er spredt:

- Stikprøvevarians

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

- Stikprøvestandardafvigelse

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

Eksempel med spredning: Højde på studerende

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

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

- **Stikprøvevariанс:**

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

- **Stikprøvestandardafvigelse:**

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

Variationskoefficienten, Definition 1.12

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

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

Variationskoefficient:

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

Fraktiler (percentiles eller quantiles)

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

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

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

Bemærk:

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

Fraktiler, Definition 1.7

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

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

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

- ④ Hvis pn ikke er et heltal:

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

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

Eksempel: Højde på studerende

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

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

- **Nedre kvartil, Q1:**

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

- **Øvre kvartil, Q3:**

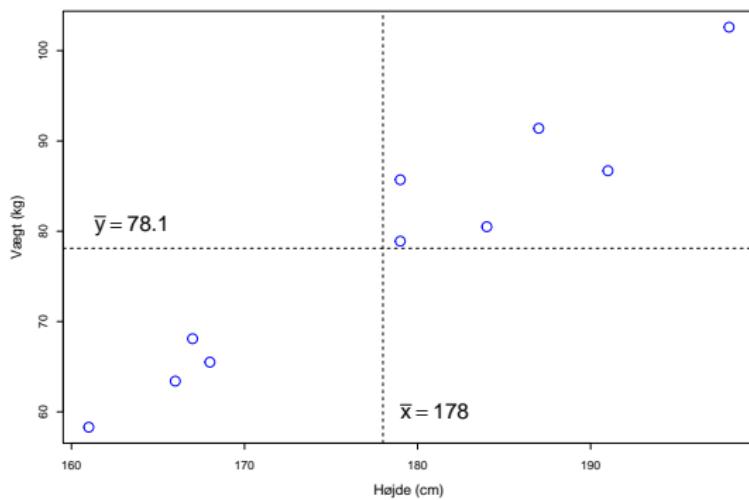
- Her er $p = 0.75$ og $n = 5$, hvorfor $np = 3.75$.
- Det mindste heltal større end np er 4.
- $Q3 = q_{0.75} = x_{(\lceil 3.75 \rceil)} = x_{(4)} = 185$.

- **IQR:**

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

Kovarians og korrelation - Sammenhængsmål

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



Stikprøvekovarians og -korrelation - Def 1.18 og 1.19

Stikprøvekovariansen er defineret ved

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

Stikprøvekorrelationskoefficienten er defineret ved

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

hvor s_x og s_y standardfvigelserne for hhv. x og y .

Stikprøvekovarians og -korrelation

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

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

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

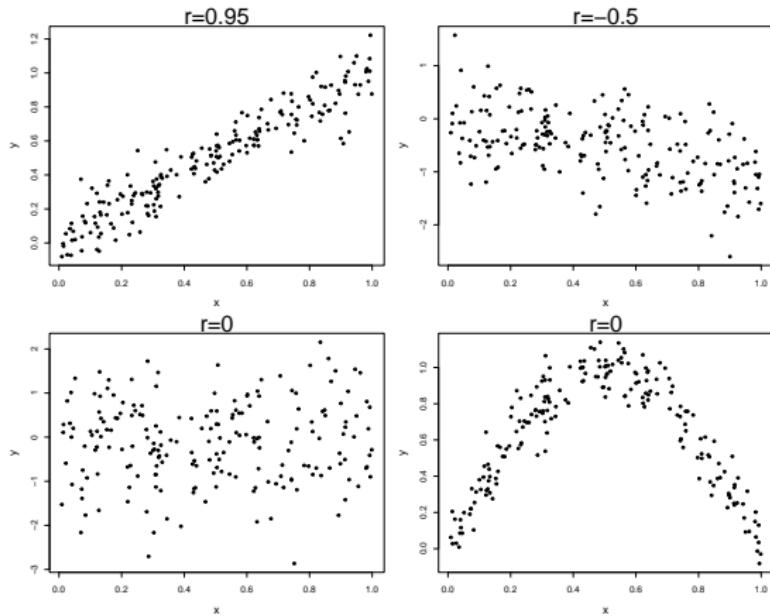
$$r = \frac{165.9}{12.21 \cdot 14.07} = 0.97$$

Egenskaber for korrelationskoefficienten

De vigtigste egenskaber for korrelationskoefficienten er:

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

Korrelation



Figurer/Tabeller

- Numeriske data

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

- Tælledata

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

Overview

1 Praktiske informationer

2 Introduktion og motivation

3 Deskriptiv Statistik

- Middelværdi og median (centralitetsmål)
- Varians og standardafvigelse
- Fraktiler
- Kovarians og korrelation

4 Software: R & RStudio

Software: R & RStudio

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

Software: R

```
> # Adding numbers in the console  
> 2 + 3  
  
## [1] 5
```

```
> # Assigning a number to a variable  
> x <- 3  
> x  
  
## [1] 3
```

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

```
> # A vector of integers from 1 to 10  
> ( x <- 1:10 )  
  
## [1] 1 2 3 4 5 6 7 8 9 10
```

Software: R

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

```
# Sample mean  
mean(x)  
  
## [1] 178
```

```
# Sample median  
median(x)  
  
## [1] 179
```

```
# Sample variance  
var(x)  
  
## [1] 149.1
```

Software: R

```
# Sample standard deviation
sd(x)

## [1] 12.21
```

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

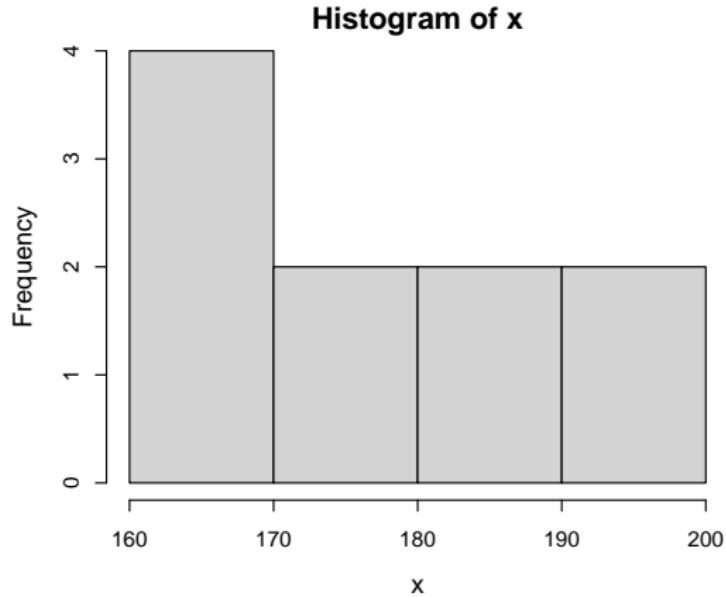
##    0%   25%   50%   75% 100%
## 161  167  179  187  198
```

```
# 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%
## 161.0 163.5 166.5 168.0 173.5 179.0 184.0 187.0 189.0 194.5
## 100%
## 198.0
```

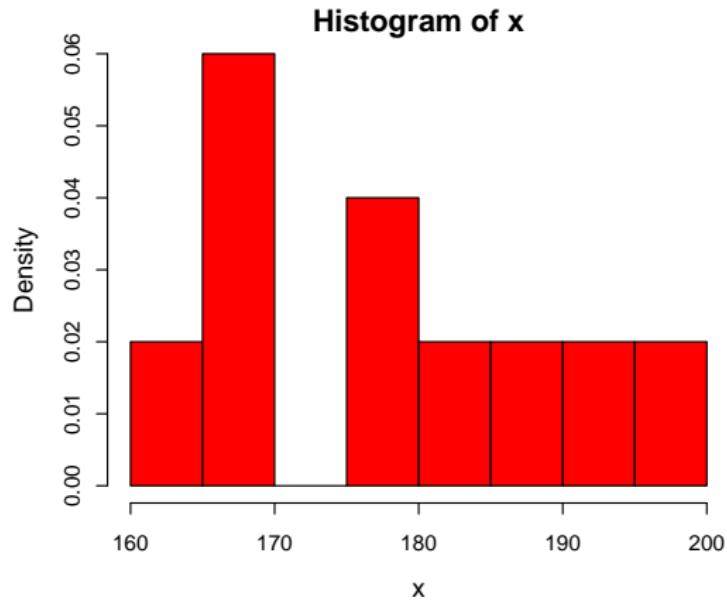
R: Histogram

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



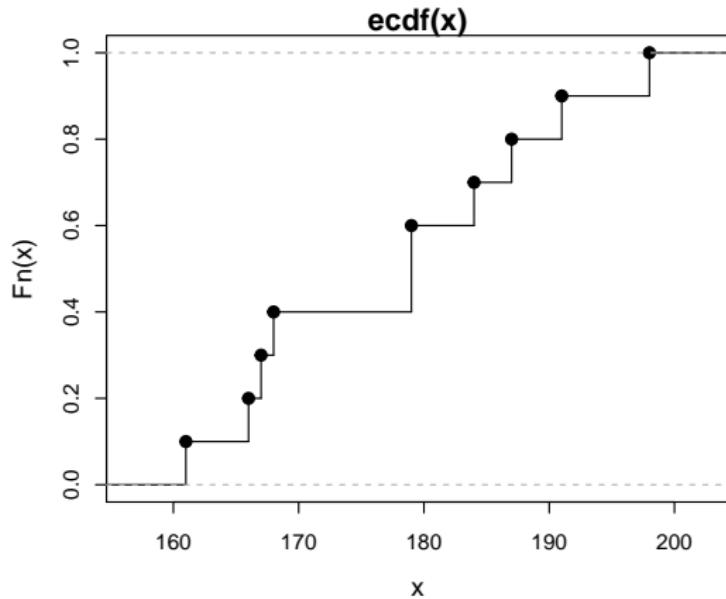
R: Empirisk tæthed

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



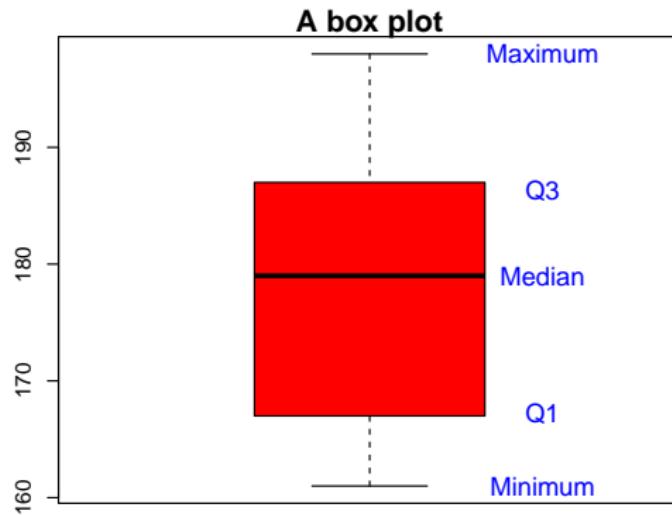
R: Empirisk kumuleret fordeling

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



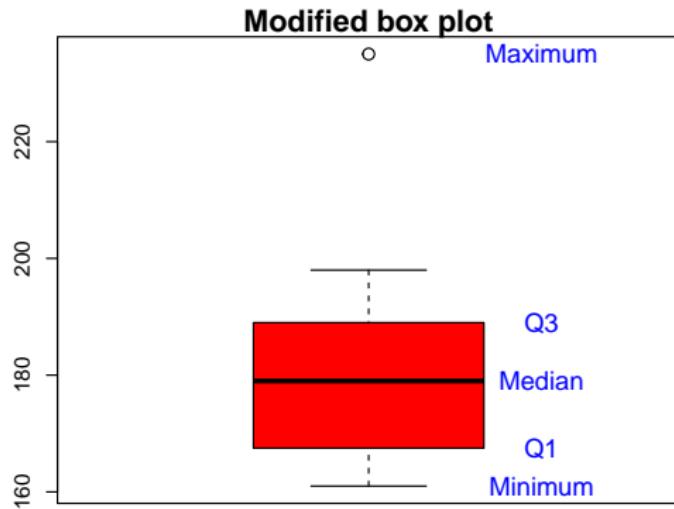
R: boxplot

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



Software: R

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



Næste uge:

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

Agenda

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

02402 Statistik (Polyteknisk grundlag)

Uge 2: Stokastiske variable og diskrete fordelinger

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

Dagsorden

- ① Opsummering: Uge 1
- ② Stokastiske variable og tæthedsfunktioner
- ③ Fordelingsfunktioner
- ④ Konkrete (diskrete) fordelinger I: Binomialfordelingen
 - Eksempel 1
- ⑤ Konkrete fordelinger II: Hypergeometrisk fordeling
 - Eksempel 2
- ⑥ Konkrete fordelinger III: Poissonfordelingen
 - Eksempel 3
- ⑦ Fordelinger i R
- ⑧ Middelværdi og varians (diskrete fordelinger)

Dagsorden

- 1 Opsummering: Uge 1
- 2 Stokastiske variable og tæthedsfunktioner
- 3 Fordelingsfunktioner
- 4 Konkrete (diskrete) fordelinger I: Binomialfordelingen
 - Eksempel 1
- 5 Konkrete fordelinger II: Hypergeometrisk fordeling
 - Eksempel 2
- 6 Konkrete fordelinger III: Poissonfordelingen
 - Eksempel 3
- 7 Fordelinger i R
- 8 Middelværdi og varians (diskrete fordelinger)

Opsummering: Uge 1

Vi ønsker at undersøge en population.

Populationen kan beskrives ved bl.a. positionsmål og spredningsmål. Hvis populationen består af N individer, kan populationsgennemsnittet og -variansen beregnes ved

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i,$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2.$$

Opsummering: Uge 1

Hvis vi har en repræsentativ stikprøve med n observationer, og vi ønsker at estimere populationsparametrene (lave statistisk inferens), kan man udregne stikprøvegennemsnittet og -variansen ved

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

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

Vi bemærker, at der divideres med $n - 1$ i beregningen af stikprøvevariansen, da vi benytter det estimerede gennemsnit $\hat{\mu}$ i stedet for μ . Hvis μ kendes, kan denne anvendes i formlen, og man dividerer med n .

Stikprøvevariansen er blot estimatet for populationsvariansen. Det er ikke variansen i stikprøven!

Dagsorden

- 1 Opsummering: Uge 1
- 2 Stokastiske variable og tæthedsfunktioner
- 3 Fordelingsfunktioner
- 4 Konkrete (diskrete) fordelinger I: Binomialfordelingen
 - Eksempel 1
- 5 Konkrete fordelinger II: Hypergeometrisk fordeling
 - Eksempel 2
- 6 Konkrete fordelinger III: Poissonfordelingen
 - Eksempel 3
- 7 Fordelinger i R
- 8 Middelværdi og varians (diskrete fordelinger)

Eksperimenter og stokastiske variable (random variables)

Setup: Et eksperiment.

Udfaldsrummet S er mængden af alle eksperimentets mulige udfald.

En stokastisk variabel X er en afbildung/funktion

$$X : S \rightarrow \mathbb{R}.$$

En stokastisk variabel repræsenterer værdien af udfaldet *før* det tilhørende *eksperiment* finder sted.

Eksempler

Nogle eksempler på stokastiske variable:

- Forelæsningens varighed
- Antallet af seksere i ti terningkast
- Andelen af stemmer til Det Republikanske Parti ved næste præsidentvalg
- En patients blodsukkerniveau
- Årsresultatet i Novo Nordisk
- Antal placeringer DTU er steget på QS University Ranking siden sidste år
- Ventetiden til Danmark vinder VM i fodbold

Hvilke egenskaber karakteriserer stokastiske variable?

Diskret eller kontinuert stokastisk variabel

Vi skelner mellem *diskrete* og *kontinuerte* stokastiske variable.

- Diskret: Værdimængden er tællelig
 - Antal personer, der bruger briller i lokalet
 - Antal passagerer, der letter fra Københavns Lufthavn inden for en time
- Kontinuert: Værdimængden er utællelig
 - Vindmåling
 - Transporttid til DTU
- I dag behandler vi diskrete variable, medens næste uges pensum omhandler kontinuerte variable.

Stokastisk variabel

Før eksperimentet udføres har vi en stokastisk variabel

X (eller X_1, \dots, X_n)

noteret med store bogstaver.

Stokastisk variabel

Før eksperimentet udføres har vi en stokastisk variabel

$$X \text{ (eller } X_1, \dots, X_n)$$

noteret med store bogstaver.

Så udføres eksperimentet, og vi har et udfald. Udfaldet giver anledning til en observation (observeret værdi)

$$x \text{ (eller } x_1, \dots, x_n)$$

noteret med små bogstaver.

Simulation: Kast en terning i R

```
# One random draw from (1,2,3,4,5,6)
# with equal probability for each outcome
sample(1:6, size = 1)
```

```
[1] 1
```

Fordelinger (Distributions)

- Stokastiske variable beskriver udfaldet af et eksperiment før det udføres.
- Hvordan kan vi regne på eksperimentet før det er udført?

Fordelinger (Distributions)

- Stokastiske variable beskriver udfaldet af et eksperiment før det udføres.
- Hvordan kan vi regne på eksperimentet før det er udført?
- Løsning: *Fordelinger* (Distributions).

En univariat fordeling beskriver, hvordan sandsynlighedsmassen fordeles over de reelle tal.

Klassificering af fordelinger

Man kan klassificere en fordeling på flere måder:

- Fordelingsfunktionen
- Tæthedsfunktionen
- Laplacetransformationen
- Den momentgenererende funktion
- Den karakteristiske funktion

I dette kursus benytter vi kun de to første!

Tæthedsfunktion, diskret stokastisk variabel, Definition 2.6

Tæthedsfunktionen (density function / probability density function, forkortelse: pdf) for en diskret stokastisk variabel:

Definition

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

Sandsynligheden for at X antager værdien x , når eksperimentet udføres.

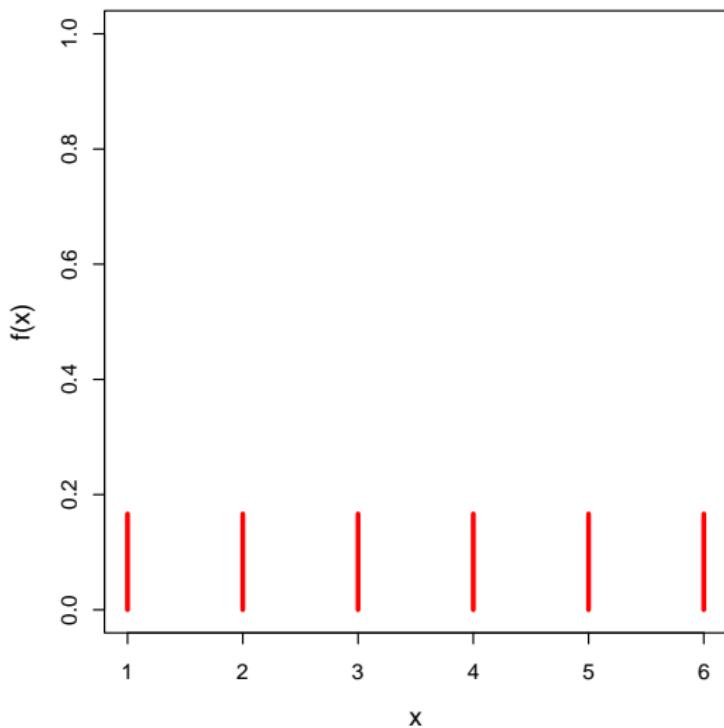
Tæthedsfunktion, diskret stokastisk variabel, Definition 2.6

Tæthedsfunktionen for en diskret stokastisk variabel opfylder følgende to betingelser:

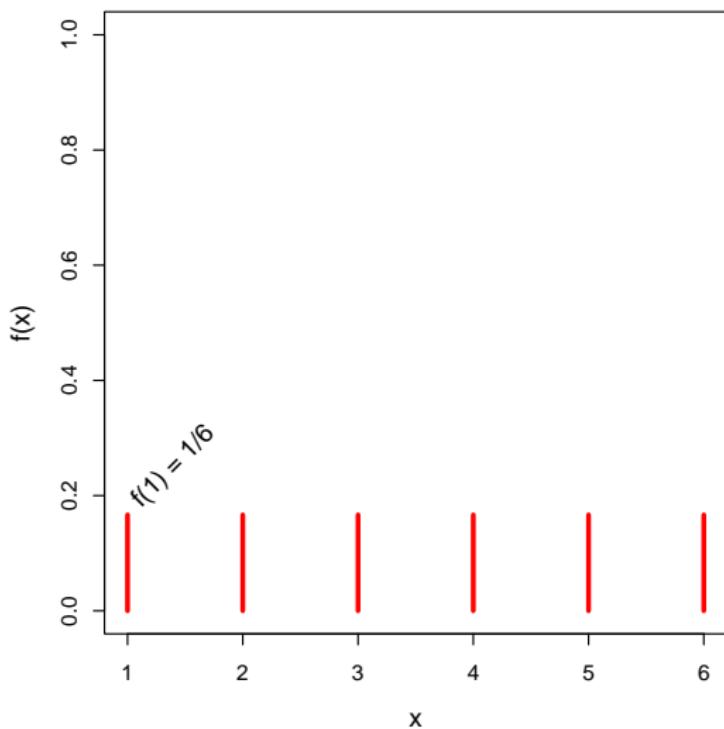
Definition

$$f(x) \geq 0 \text{ for alle } x \quad \text{og} \quad \sum_{\text{alle } x} f(x) = 1$$

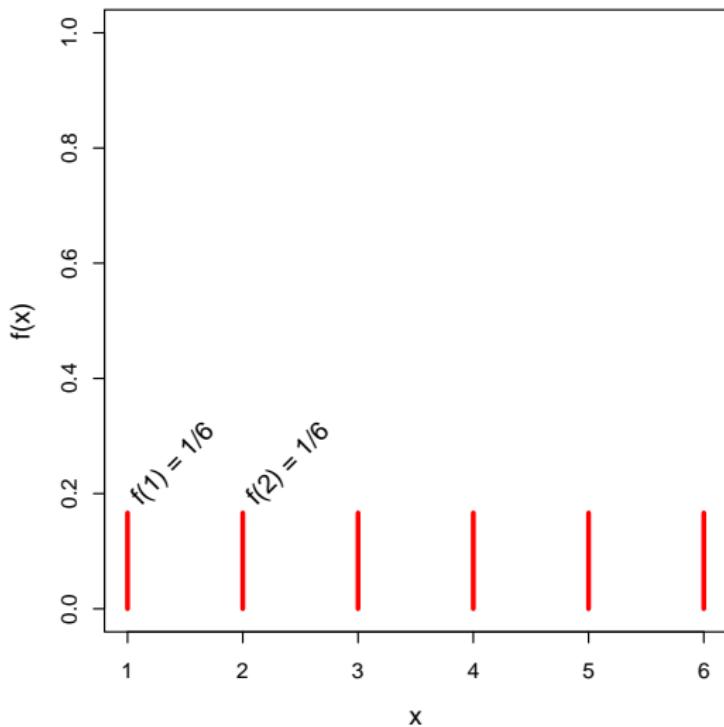
Eksempel: Tæthedsfunktion - Kast med en fair terning



Eksempel: Tæthedsfunktion - Kast med en fair terning



Eksempel: Tæthedsfunktion - Kast med en fair terning



Stikprøve

Hvad nu hvis vi kun har én observation. Kan vi da se fordelingen?

Stikprøve

Hvad nu hvis vi kun har én observation. Kan vi da se fordelingen? **Nej!**

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: Simulér n kast med en fair terning

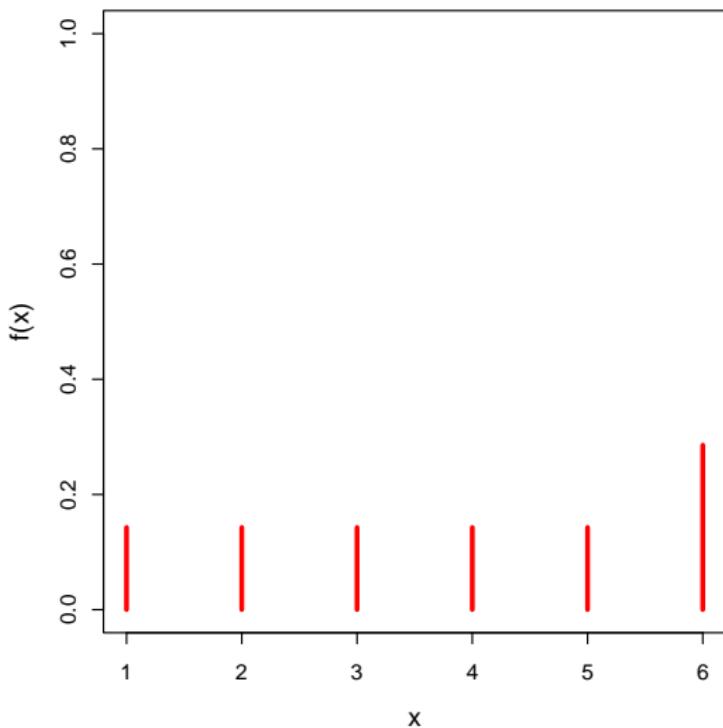
```
# Number of simulated realizations (sample size)
n <- 30

# n independent random draws from the set (1,2,3,4,5,6)
# with equal probability of each outcome
xFair <- sample(1:6, size = n, replace = TRUE)
xFair

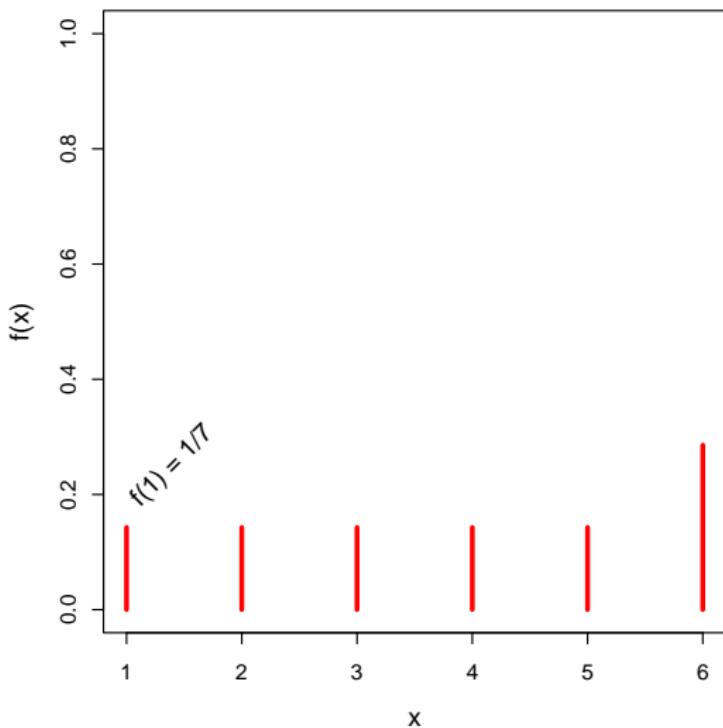
# Count number of each outcome using the 'table' function
table(xFair)

# Plot the empirical pdf
plot(table(xFair)/n, lwd = 10, ylim = c(0,1), xlab = "x",
      ylab = "Density f(x)")
# Add the true pdf to the plot
lines(rep(1/6,6), lwd = 4, type = "h", col = 2)
# Add a legend to the plot
legend("topright", c("Empirical pdf","True pdf"), lty = 1, col = c(1,2),
       lwd = c(5, 2), cex = 0.8)
```

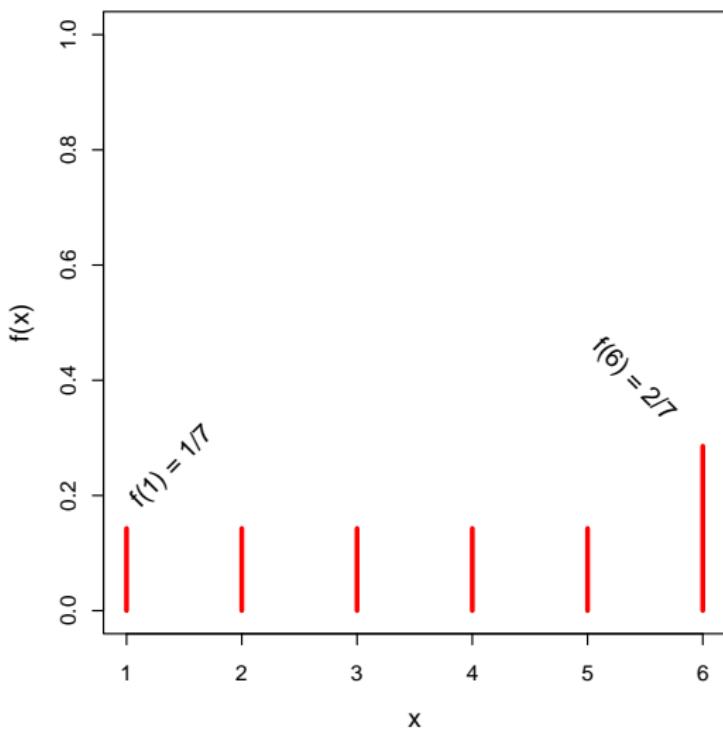
Eksempel: Tæthedsfunktion - Kast med en unfair terning



Eksempel: Tæthedsfunktion - Kast med en unfair terning



Eksempel: Tæthedsfunktion - Kast med en unfair terning



Eksempel: Simulér n kast med en unfair terning

```
# Number of simulated realizations (sample size)
n <- 30

# n independent random draws from the set (1,2,3,4,5,6)
# with higher probability of getting a six
xUnfair <- sample(1:6, size = n, replace = TRUE, prob = c(rep(1/7,5),2/7))
xUnfair

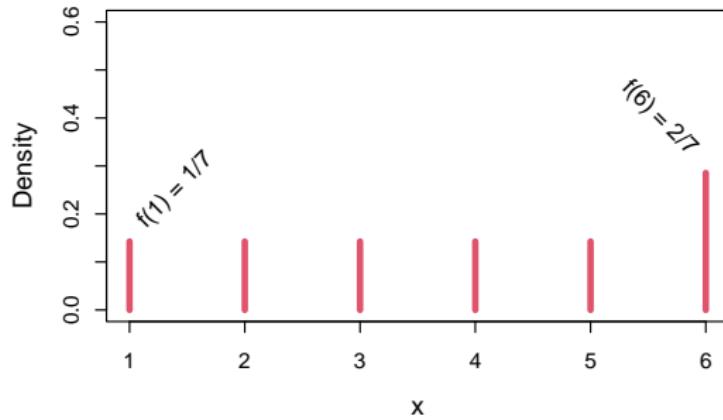
# Plot the empirical pdf
plot(table(xUnfair)/n, lwd = 10, ylim = c(0,1), xlab = "x",
      ylab = "Density f(x)")
# Add the true pdf to the plot
lines(c(rep(1/7,5),2/7), lwd = 4, type = "h", col = 2)
# Add a legend to the plot
legend("topright", c("Empirical pdf","True pdf"), lty = 1, col = c(1,2),
       lwd = c(5, 2), cex = 0.8)
```

Spørgsmål

Lad X beskrive det antal øjne, der fås ved et kast med den *unfair* terning.

Hvad er:

- Sandsynligheden for at få 4?
- Sandsynligheden for at få 5 eller 6?
- Sandsynligheden for at få mindre end 3?



Dagsorden

- 1 Opsummering: Uge 1
- 2 Stokastiske variable og tæthedsfunktioner
- 3 Fordelingsfunktioner
- 4 Konkrete (diskrete) fordelinger I: Binomialfordelingen
 - Eksempel 1
- 5 Konkrete fordelinger II: Hypergeometrisk fordeling
 - Eksempel 2
- 6 Konkrete fordelinger III: Poissonfordelingen
 - Eksempel 3
- 7 Fordelinger i R
- 8 Middelværdi og varians (diskrete fordelinger)

Fordelingsfunktion for en diskret stokastisk variabel: Definition 2.9

Fordelingsfunktionen (cumulative distribution function, cdf) for en diskret stokastisk variabel:

Definition

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

Der gælder for en fordelingsfunktion (cdf):

- Det er en 'ikke-aftagende' funktion
- Den nærmer sig (konvergerer mod) 1, når $x \rightarrow \infty$

Eksempel: Kast med en fair terning

Lad X repræsentere værdien af et kast med en fair terning.

Find sandsynligheden for at observere en værdi mindre end 3:

Eksempel: Kast med en fair terning

Lad X repræsentere værdien af et kast med en fair terning.

Find sandsynligheden for at observere en værdi mindre end 3:

$$P(X < 3)$$

Eksempel: Kast med en fair terning

Lad X repræsentere værdien af et kast med en fair terning.

Find sandsynligheden for at observere en værdi mindre end 3:

$$P(X < 3) = P(X \leq 2)$$

Eksempel: Kast med en fair terning

Lad X repræsentere værdien af et kast med en fair terning.

Find sandsynligheden for at observere en værdi mindre end 3:

$$\begin{aligned}P(X < 3) &= P(X \leq 2) \\&= F(2) \text{ fordelingsfunktionen}\end{aligned}$$

Eksempel: Kast med en fair terning

Lad X repræsentere værdien af et kast med en fair terning.

Find sandsynligheden for at observere en værdi mindre end 3:

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

Eksempel: Kast med en fair terning

Lad X repræsentere værdien af et kast med en fair terning.

Find sandsynligheden for at observere en værdi mindre end 3:

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

Eksempel: Kast med en fair terning

Lad X repræsentere værdien af et kast med en fair terning.

Find sandsynligheden for at observere en værdi mindre end 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: Kast med en fair terning

Find sandsynligheden for at observere en værdi større end eller lig 3:

Eksempel: Kast med en fair terning

Find sandsynligheden for at observere en værdi større end eller lig 3:

$$P(X \geq 3)$$

Eksempel: Kast med en fair terning

Find sandsynligheden for at observere en værdi større end eller lig 3:

$$P(X \geq 3) = 1 - P(X \leq 2)$$

Eksempel: Kast med en fair terning

Find sandsynligheden for at observere en værdi større end eller lig 3:

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

Eksempel: Kast med en fair terning

Find sandsynligheden for at observere en værdi større end eller lig 3:

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

Dagsorden

- 1 Opsummering: Uge 1
- 2 Stokastiske variable og tæthedsfunktioner
- 3 Fordelingsfunktioner
- 4 Konkrete (diskrete) fordelinger I: Binomialfordelingen
 - Eksempel 1
- 5 Konkrete fordelinger II: Hypergeometrisk fordeling
 - Eksempel 2
- 6 Konkrete fordelinger III: Poissonfordelingen
 - Eksempel 3
- 7 Fordelinger i R
- 8 Middelværdi og varians (diskrete fordelinger)

Konkrete (diskrete) statistiske fordelinger

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

Binomialfordelingen

- Vi betragter et eksperiment med to udfald: "succes" og "fiasko", som gentages et vist antal gange (uafhængige gentagelser).
- Lad X være antallet af succeser efter n gentagelser.

Binomialfordelingen

- Vi betragter et eksperiment med to udfald: "succes" og "fiasko", som gentages et vist antal gange (uafhængige gentagelser).
- Lad X være antallet af succeser efter n gentagelser.
- Så følger X en binomialfordeling m. antalsparameter n og succesparameter p :

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

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

Binomialfordelingens tæthedsfunktion

Sandsynligheden for at observere x antal succeser gives ved

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

hvor binomialkoefficienten kan beregnes som

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}.$$

Eksempel - Binomialfordelingen

Antag $X \sim B(4, p)$, dvs. $n = 4$. Find sandsynligheden for at observere 3 succeser.

Eksempel - Binomialfordelingen

Antag $X \sim B(4, p)$, dvs. $n = 4$. Find sandsynligheden for at observere 3 succeser.

- Sandsynligheden for 3 succeser er $P(X = 3)$.

Eksempel - Binomialfordelingen

Antag $X \sim B(4, p)$, dvs. $n = 4$. Find sandsynligheden for at observere 3 succeser.

- Sandsynligheden for 3 succeser er $P(X = 3)$.
- De tre succeser kan fremkomme på fire måder:
SSSF, SSFS, SFSS, FSSS.

Eksempel - Binomialfordelingen

Antag $X \sim B(4, p)$, dvs. $n = 4$. Find sandsynligheden for at observere 3 succeser.

- Sandsynligheden for 3 succeser er $P(X = 3)$.
- De tre succeser kan fremkomme på fire måder:
SSSF, SSFS, SFSS, FSSS.
- Altså:

$$\binom{n}{x} = \binom{4}{3} = \frac{4!}{3!(4-3)!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \cdot 1} = 4,$$

og

$$P(X = 3) = 4p^3(1-p).$$

Simulation med binomialfordeling

```
## Probability of success
p <- 0.1

## Number of repetitions
nRepeat <- 30

## Simulate Bernoulli experiment 'nRepeat' times
tmp <- sample(c(0,1), size = nRepeat, prob = c(1-p,p), replace = TRUE)

# Compute 'x'
sum(tmp)

## Or: Use the binomial distribution simulation function
rbinom(1, size = 30, prob = p)
```

Eksempel: Kast med en fair terning

```
# Number of simulated realizations (sample size)
n <- 30

# n independent random draws from the set (1,2,3,4,5,6)
# with equal probability for each outcome
xFair <- sample(1:6, size = n, replace = TRUE)

# Count the number of six'es
sum(xFair == 6)

## Do the same using 'rbinom()' instead
rbinom(n = 1, size = 30, prob = 1/6)
```

Eksempel 1

I et kundecenter i et telefonselskab prøver man at forbedre kundetilfredsheden. Det er især vigtigt, at når der indrapporteres en fejl, bliver fejlen udbedret i løbet af samme dag.

Antag at sandsynligheden for, at en fejl bliver udbedret i løbet af samme dag, er 70%.

I løbet af en dag indrapporteres 6 fejl. Hvad er sandsynligheden for at samtlige fejl udbedres?

Eksempel 1

I et kundecenter i et telefonselskab prøver man at forbedre kundetilfredsheden. Det er især vigtigt, at når der indrapporteres en fejl, bliver fejlen udbedret i løbet af samme dag.

Antag at sandsynligheden for, at en fejl bliver udbedret i løbet af samme dag, er 70%.

I løbet af en dag indrapporteres 6 fejl. Hvad er sandsynligheden for at samtlige fejl udbedres?

- **Trin 1)** Hvad skal repræsenteres af den stokastiske variabel X ?

Eksempel 1

I et kundecenter i et telefonselskab prøver man at forbedre kundetilfredsheden. Det er især vigtigt, at når der indrapporteres en fejl, bliver fejlen udbedret i løbet af samme dag.

Antag at sandsynligheden for, at en fejl bliver udbedret i løbet af samme dag, er 70%.

I løbet af en dag indrapporteres 6 fejl. Hvad er sandsynligheden for at samtlige fejl udbedres?

- **Trin 1)** Hvad skal repræsenteres af den stokastiske variabel X ?

Antallet af udbedrede fejl.

Eksempel 1

I et kundecenter i et telefonselskab prøver man at forbedre kundetilfredsheden. Det er især vigtigt, at når der indrapporteres en fejl, bliver fejlen udbedret i løbet af samme dag.

Antag at sandsynligheden for, at en fejl bliver udbedret i løbet af samme dag, er 70%.

I løbet af en dag indrapporteres 6 fejl. Hvad er sandsynligheden for at samtlige fejl udbedres?

- **Trin 1)** Hvad skal repræsenteres af den stokastiske variabel X ?
Antallet af udbedrede fejl.
- **Trin 2)** Hvad er fordelingen af X ?

Eksempel 1

I et kundecenter i et telefonselskab prøver man at forbedre kundetilfredsheden. Det er især vigtigt, at når der indrapporteres en fejl, bliver fejlen udbedret i løbet af samme dag.

Antag at sandsynligheden for, at en fejl bliver udbedret i løbet af samme dag, er 70%.

I løbet af en dag indrapporteres 6 fejl. Hvad er sandsynligheden for at samtlige fejl udbedres?

- **Trin 1)** Hvad skal repræsenteres af den stokastiske variabel X ?

Antallet af udbedrede fejl.

- **Trin 2)** Hvad er fordelingen af X ?

En binomialfordeling med $n = 6$ og $p = 0.7$.

Eksempel 1

I et kundecenter i et telefonselskab søger man at forbedre kundetilfredsheden. Især er det vigtigt at når der indrapporteres en fejl, bliver fejlen udbedret i løbet af samme dag.

Antag at sandsynligheden for at en fejl bliver udbedret i løbet af samme dag er 70%.

I løbet af en dag indrapporteres 6 fejl. Hvad er sandsynligheden for at samtlige fejl udbedres?

- **Trin 3)** Hvilken sandsynlighed skal udregnes

Eksempel 1

I et kundecenter i et telefonselskab søger man at forbedre kundetilfredsheden. Især er det vigtigt at når der indrapporteres en fejl, bliver fejlen udbedret i løbet af samme dag.

Antag at sandsynligheden for at en fejl bliver udbedret i løbet af samme dag er 70%.

I løbet af en dag indrapporteres 6 fejl. Hvad er sandsynligheden for at samtlige fejl udbedres?

- **Trin 3)** Hvilken sandsynlighed skal udregnes

$$\underline{P(X = 6) = f(6; 6, 0.7)}$$

Dagsorden

- 1 Opsummering: Uge 1
- 2 Stokastiske variable og tæthedsfunktioner
- 3 Fordelingsfunktioner
- 4 Konkrete (diskrete) fordelinger I: Binomialfordelingen
 - Eksempel 1
- 5 Konkrete fordelinger II: Hypergeometrisk fordeling
 - Eksempel 2
- 6 Konkrete fordelinger III: Poissonfordelingen
 - Eksempel 3
- 7 Fordelinger i R
- 8 Middelværdi og varians (diskrete fordelinger)

Den hypergeometriske fordeling

- X er igen antallet succeser, men nu *uden tilbagelægning* ved trækningen.

Den hypergeometriske fordeling

- X er igen antallet succeser, men nu *uden tilbagelægning* ved trækningen.
- X følger da den hypergeometriske fordeling

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

- n er antallet af trækninger (gentagelser)
- a er antallet af succeser i populationen
- N er antallet af elementer i (hele) populationen

Den hypergeometriske fordeling

- Sandsynligheden for at få x succeser er

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

- n er antallet af trækninger (gentagelser)
- a er antallet af succeser i populationen
- N er antallet af elementer i (hele) populationen

Eksempel 2

I en forsendelse med 10 harddiske har 2 af dem mindre skrammer.

Vi udtager en (tilfældig) stikprøve på 3 harddiske. **Hvad er sandsynligheden for at mindst en af dem har skrammer?**

Eksempel 2

I en forsendelse med 10 harddiske har 2 af dem mindre skrammer.

Vi udtager en (tilfældig) stikprøve på 3 harddiske. **Hvad er sandsynligheden for at mindst en af dem har skrammer?**

- **Trin 1)** Hvad skal repræsenteres af den stokastiske variabel X ?

Eksempel 2

I en forsendelse med 10 harddiske har 2 af dem mindre skrammer.

Vi udtager en (tilfældig) stikprøve på 3 harddiske. **Hvad er sandsynligheden for at mindst en af dem har skrammer?**

- **Trin 1)** Hvad skal repræsenteres af den stokastiske variabel X ?

Antallet af harddiske med skramme i stikprøven.

Eksempel 2

I en forsendelse med 10 harddiske har 2 af dem mindre skrammer.

Vi udtager en (tilfældig) stikprøve på 3 harddiske. **Hvad er sandsynligheden for at mindst en af dem har skrammer?**

- **Trin 1)** Hvad skal repræsenteres af den stokastiske variabel X ?
Antallet af harddiske med skramme i stikprøven.
- **Trin 2)** Hvad er fordelingen af X ?

Eksempel 2

I en forsendelse med 10 harddiske har 2 af dem mindre skrammer.

Vi udtager en (tilfældig) stikprøve på 3 harddiske. **Hvad er sandsynligheden for at mindst en af dem har skrammer?**

- **Trin 1)** Hvad skal repræsenteres af den stokastiske variabel X ?

Antallet af harddiske med skramme i stikprøven.

- **Trin 2)** Hvad er fordelingen af X ?

En hypergeometrisk fordeling med $n = 3$, $a = 2$ og $N = 10$.

Eksempel 2

I en forsendelse med 10 harddiske har 2 af dem mindre skrammer.

Vi udtager en (tilfældig) stikprøve på 3 harddiske. **Hvad er sandsynligheden for at mindst en af dem har skrammer?**

- **Trin 1)** Hvad skal repræsenteres af den stokastiske variabel X ?

Antallet af harddiske med skramme i stikprøven.

- **Trin 2)** Hvad er fordelingen af X ?

En hypergeometrisk fordeling med $n = 3$, $a = 2$ og $N = 10$.

- **Trin 3)** Hvilken sandsynlighed skal udregnes?

Eksempel 2

I en forsendelse med 10 harddiske har 2 af dem mindre skrammer.

Vi udtager en (tilfældig) stikprøve på 3 harddiske. **Hvad er sandsynligheden for at mindst en af dem har skrammer?**

- **Trin 1)** Hvad skal repræsenteres af den stokastiske variabel X ?

Antallet af harddiske med skramme i stikprøven.

- **Trin 2)** Hvad er fordelingen af X ?

En hypergeometrisk fordeling med $n = 3$, $a = 2$ og $N = 10$.

- **Trin 3)** Hvilken sandsynlighed skal udregnes?

$$\underline{P(X \geq 1) = 1 - P(X = 0) = 1 - f(0; 3, 2, 10)}$$

Binomial vs. hypergeometrisk

- Binomialfordelingen bruges til at analysere stikprøver med tilbagelægning.
- Den hypergeometriske fordeling bruges til at analysere stikprøver uden tilbagelægning.

Dagsorden

- 1 Opsummering: Uge 1
- 2 Stokastiske variable og tæthedsfunktioner
- 3 Fordelingsfunktioner
- 4 Konkrete (diskrete) fordelinger I: Binomialfordelingen
 - Eksempel 1
- 5 Konkrete fordelinger II: Hypergeometrisk fordeling
 - Eksempel 2
- 6 Konkrete fordelinger III: Poissonfordelingen
 - Eksempel 3
- 7 Fordelinger i R
- 8 Middelværdi og varians (diskrete fordelinger)

Poissonfordelingen

- Poissonfordelingen anvendes ofte som en fordeling (model) for tælletal, hvor der ikke er nogen naturlig øvre grænse.
- Poissonfordelingen karakteriseres/defineres normalt ved en *intensitet*, som har formen "antal/enhed", ofte benævnt λ .
- Typisk *hændelser per tidsinterval*.

Poissonfordelingen

$$X \sim Po(\lambda)$$

Tæthedsfunktion:

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

Fordelingsfunktion:

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

Eksempel 3

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?

Eksempel 3

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?

- Trin 1) Hvad skal repræsenteres af den stokastiske variabel X ?

Eksempel 3

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?

- **Trin 1)** Hvad skal repræsenteres af den stokastiske variabel X ?

Antal patienter, der indlægges som følge af luftforurening på en vilkårlig dag.

Eksempel 3

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?

- **Trin 1)** Hvad skal repræsenteres af den stokastiske variabel X ?

Antal patienter, der indlægges som følge af luftforurening på en vilkårlig dag.

- **Trin 2)** Hvad er fordelingen af X ?

Eksempel 3

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?

- **Trin 1)** Hvad skal repræsenteres af den stokastiske variabel X ?

Antal patienter, der indlægges som følge af luftforurening på en vilkårlig dag.

- **Trin 2)** Hvad er fordelingen af X ?

En poissonfordeling med $\lambda = 0.3$.

Eksempel 3

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?

- **Trin 1)** Hvad skal repræsenteres af den stokastiske variabel X ?

Antal patienter, der indlægges som følge af luftforurening på en vilkårlig dag.

- **Trin 2)** Hvad er fordelingen af X ?

En poissonfordeling med $\lambda = 0.3$.

- **Trin 3)** Hvilken sandsynlighed skal udregnes?

Eksempel 3

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?

- **Trin 1)** Hvad skal repræsenteres af den stokastiske variabel X ?

Antal patienter, der indlægges som følge af luftforurening på en vilkårlig dag.

- **Trin 2)** Hvad er fordelingen af X ?

En poissonfordeling med $\lambda = 0.3$.

- **Trin 3)** Hvilken sandsynlighed skal udregnes?

$$\underline{P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)}$$

Dagsorden

- 1 Opsummering: Uge 1
- 2 Stokastiske variable og tæthedsfunktioner
- 3 Fordelingsfunktioner
- 4 Konkrete (diskrete) fordelinger I: Binomialfordelingen
 - Eksempel 1
- 5 Konkrete fordelinger II: Hypergeometrisk fordeling
 - Eksempel 2
- 6 Konkrete fordelinger III: Poissonfordelingen
 - Eksempel 3
- 7 Fordelinger i R
- 8 Middelværdi og varians (diskrete fordelinger)

Fordelinger i R

R	Name
binom	Binomialfordeling
hyper	Hypergeometrisk fordeling
pois	Poissonfordeling

- d Tæthedsfunktion (density)
- p Fordelingsfunktion (probability)
- r Tilfældighedsgenerator: Simulerer tilfælige tal (random number)
- q Fraktilfunktion ("invers" af fordelingsfunktionen) (quantile)

Eksempel: Binomialfordeling, $P(X \leq 5) = F(5; 10, 0.6)$

```
pbinom(q = 5, size = 10, prob = 0.6)
```

```
[1] 0.3669
```

```
# Get help with:  
?pbinom
```

Dagsorden

- 1 Opsummering: Uge 1
- 2 Stokastiske variable og tæthedsfunktioner
- 3 Fordelingsfunktioner
- 4 Konkrete (diskrete) fordelinger I: Binomialfordelingen
 - Eksempel 1
- 5 Konkrete fordelinger II: Hypergeometrisk fordeling
 - Eksempel 2
- 6 Konkrete fordelinger III: Poissonfordelingen
 - Eksempel 3
- 7 Fordelinger i R
- 8 Middelværdi og varians (diskrete fordelinger)

Middelværdi (expectation, expected value)

Middelværdien af en diskret stokastisk variabel,
definition 2.13:

Definition

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

- Det *"sande gennemsnit"* af X (i modsætning til stikprøvegennemsnittet).

Eksempel: Kast med en fair terning

Lad X repræsentere antallet af øjne ved et kast med en fair terning. Så følger X en diskret uniform fordeling (diskret ligefordeling) på intervallet $[1, 6]$ og har middelværdi:

$$\mu = E(X)$$

$$\begin{aligned} &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= 3.5 \end{aligned}$$

Sammenligning med stikprøvegennemsnittet - lær fra simulationer

```
# Number of simulated realizations (sample size)
n <- 30

# Sample independently from the set (1,2,3,4,5,6)
# with equal probability of outcomes
xFair <- sample(1:6, size = n, replace = TRUE)

# Compute the sample mean
mean(xFair)
```

```
[1] 3.3
```

Asymptotisk resultat: Store tals lov

Des flere observationer, des tættere kommer vi på den sande middelværdi:

$$\lim_{n \rightarrow \infty} \hat{\mu} = \mu$$

- Kaldes *store tals lov* (law of large numbers).

Varians

Variansen af en diskret stokastisk variabel, Definition 2.16:

Definition

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

- Måler den gennemsnitlige spredning rundt om middelværdien.
- Den “rigtige varians“ af X (i modsætning til stikprøvevariansen).

Eksempel: Kast med en fair terning

Lad X repræsentere antallet af øjne ved et kast med en fair terning. Så følger X en diskret uniform fordeling (diskret ligefordeling) på intervallet $[1, 6]$ og har varians:

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

Sammenligning med stikprøvevariansen - lær fra simulationer

```
# Number of simulated realizations (sample size)
n <- 30

# Sample independently from the set (1,2,3,4,5,6)
# with equal probability of outcomes
xFair <- sample(1:6, size = n, replace = TRUE)

# Compute the sample variance
var(xFair)
```

```
[1] 2.437
```

Middelværdi og varians for konkrete fordelinger

Binomialfordelingen:

- Middelværdi:

$$\mu = n \cdot p$$

- Varians:

$$\sigma^2 = n \cdot p \cdot (1 - p)$$

Middelværdi og varians for konkrete fordelinger

Den hypergeometriske fordeling

- Middelværdi:

$$\mu = n \cdot \frac{a}{N}$$

- Varians:

$$\sigma^2 = \frac{n \cdot a \cdot (N-a) \cdot (N-n)}{N^2 \cdot (N-1)}$$

Middelværdi og varians for konkrete fordelinger

Poissonfordelingen

- Middelværdi:

$$\mu = \lambda$$

- Varians:

$$\sigma^2 = \lambda$$

Dagsorden

- 1 Opsummering: Uge 1
- 2 Stokastiske variable og tæthedsfunktioner
- 3 Fordelingsfunktioner
- 4 Konkrete (diskrete) fordelinger I: Binomialfordelingen
 - Eksempel 1
- 5 Konkrete fordelinger II: Hypergeometrisk fordeling
 - Eksempel 2
- 6 Konkrete fordelinger III: Poissonfordelingen
 - Eksempel 3
- 7 Fordelinger i R
- 8 Middelværdi og varians (diskrete fordelinger)

02402 Statistik (Polyteknisk grundlag)

Uge 3: Stokastiske variable og kontinuerte fordelinger

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

Dagsorden

- 1 Opsummering
- 2 Kontinuerte fordelinger
 - Tæthed- og fordelingsfunktioner
 - Middelværdi, varians og kovarians
- 3 Vigtige kontinuerte fordelinger
 - Den uniforme fordeling
 - Normalfordelingen
 - Log-normalfordelingen
 - Eksponentialfordelingen
- 4 Regneregler for stokastiske variable

Dagsorden

1 Opsummering

2 Kontinuerte fordelinger

- Tæthedsfunktioner og fordelingsfunktioner
- Middelværdi, varians og kovarians

3 Vigtige kontinuerte fordelinger

- Den uniforme fordeling
- Normalfordelingen
- Log-normalfordelingen
- Eksponentialfordelingen

4 Regneregler for stokastiske variable

Opsummering: Uge 1 og 2

Vi ønsker at undersøge en population.

Populationen kan beskrives ved bl.a. positionsmål og spredningsmål, f.eks. gennemsnittet og variansen.

Man kan opstille et eksperiment og udtag en stikprøve for at estimere populationens parametre.

Men hvis vi vil regne på eksperimentet før det udføres, har vi brug for en matematisk model. Her kan vi definere en stokastisk variabel og tilknytte den en fordeling, der beskriver, hvorledes sandsynlighedsmassen fordeler sig over de reelle tal.

Opsummering: Uge 1 og 2

En sandsynlighedsfordeling er en teoretisk model, der beskriver, hvor sandsynligt det er, at den stokastiske variable antager forskellige værdier.

Vi skelner mellem diskrete og kontinuerte stokastiske variable (og fordelinger) baseret på tælleligheden af variablens værdimængde.

Man kan beskrive en sandsynlighedsfordeling igennem dens fordelingsfunktion eller dens tæthedsfunktion. Fordelingens funktionen er defineret ens for både diskrete og kontinuerte fordelinger, hvilket ikke er tilfældet for tæthedsfunktionen.

I den teoretiske model repræsenterer middelværdien af den stokastiske variabel $\mathbb{E}[X]$ det sande populationsgennemsnit og variansen af den stokastiske variabel $\mathbb{V}[X]$ den sande populationsvariанс.

Opsummering: Uge 1 og 2

For en diskret stokastisk variabel X med værdimængde A , defineres tæthedsfunktionen som

$$f(x) = \mathbb{P}(X = x)$$

og fordelingsfunktionen som

$$F(x) = \mathbb{P}(X \leq x) = \sum_{\{y \in A : y \leq x\}} \mathbb{P}(X = y) = \sum_{\{y \in A : y \leq x\}} f(y).$$

På engelsk kaldes tæthedsfunktionen ofte for *the probability mass function (pmf)* i det diskrete tilfælde.

Forventningsværdien (middelværdien) af X beregnes som

$$\mathbb{E}[X] = \sum_{x \in A} x \mathbb{P}(X = x) = \sum_{x \in A} xf(x).$$

Dagsorden

1 Opsummering

2 Kontinuerte fordelinger

- Tæthed- og fordelingsfunktioner
- Middelværdi, varians og kovarians

3 Vigtige kontinuerte fordelinger

- Den uniforme fordeling
- Normalfordelingen
- Log-normalfordelingen
- Eksponentialfordelingen

4 Regneregler for stokastiske variable

Tæthedsfunktionen, Definition 2.32

- Tæthedsfunktionen (density function/probability density function, pdf) for en stokastisk variabel betegnes med $f(x)$.

Tæthedsfunktionen, Definition 2.32

- Tæthedsfunktionen (density function/probability density function, pdf) for en stokastisk variabel betegnes med $f(x)$.
- Tæthedsfunktionen $f(x)$ siger noget om hyppigheden af udfaldsværdien x for den stokastiske variabel X .

Tæthedsfunktionen, Definition 2.32

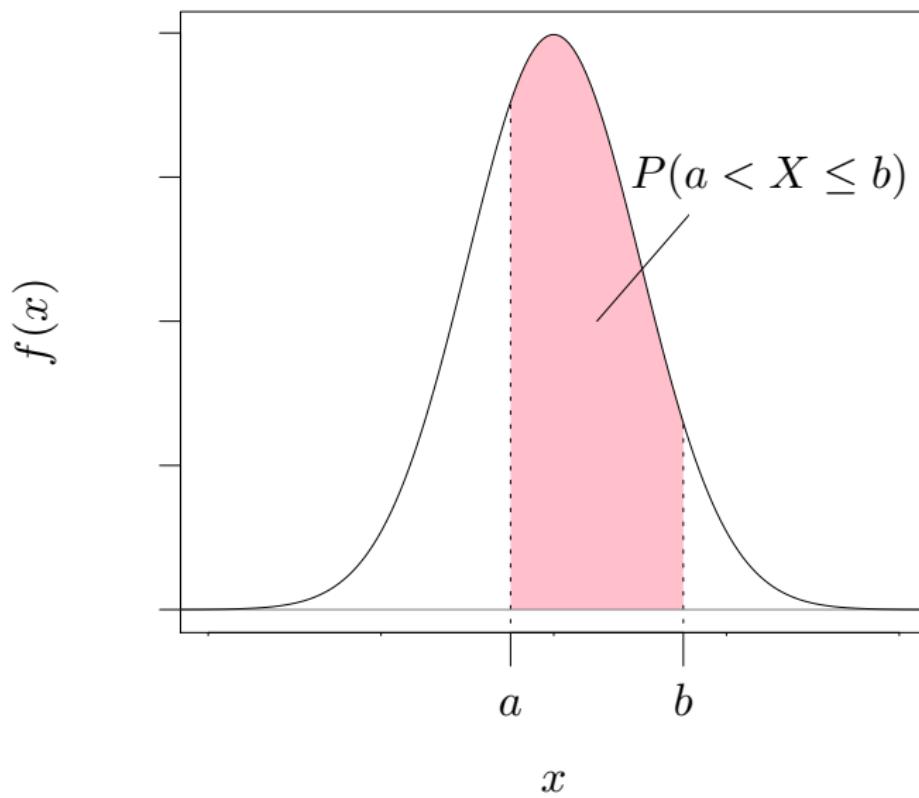
- Tæthedsfunktionen (density function/probability density function, pdf) for en stokastisk variabel betegnes med $f(x)$.
- Tæthedsfunktionen $f(x)$ siger noget om hyppigheden af udfaldsværdien x for den stokastiske variabel X .
- Tæthedsfunktionen for en kontinuert stokastisk variabel svarer *ikke* til sandsynligheden for at observere værdien x . Der gælder faktisk, $P(X = x) = 0$ for alle x .

Tæthedsfunktionen, Definition 2.32

- Tæthedsfunktionen (density function/probability density function, pdf) for en stokastisk variabel betegnes med $f(x)$.
- Tæthedsfunktionen $f(x)$ siger noget om hyppigheden af udfaldsværdien x for den stokastiske variabel X .
- Tæthedsfunktionen for en kontinuert stokastisk variabel svarer *ikke* til sandsynligheden for at observere værdien x . Der gælder faktisk, $P(X = x) = 0$ for alle x .
- Tæthedsfunktionen $f(x)$ hørende til fordelingen af en kontinuert stokastisk variabel opfylder at:

$$f(x) \geq 0 \text{ for alle } x \quad \text{og} \quad \int_{-\infty}^{\infty} f(x) dx = 1.$$

Tæthedsfunktionen



Fordelingsfunktionen, Definition 2.33

- Fordelingsfunktionen (distribution function/cumulative distribution function, cdf) hørende til en kontinuert stokastisk variabel benævnes med $F(x)$.

Fordelingsfunktionen, Definition 2.33

- Fordelingsfunktionen (distribution function/cumulative distribution function, cdf) hørende til en kontinuert stokastisk variabel benævnes med $F(x)$.
- Fordelingsfunktionen er defineret ved

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt.$$

Fordelingsfunktionen, Definition 2.33

- Fordelingsfunktionen (distribution function/cumulative distribution function, cdf) hørende til en kontinuert stokastisk variabel benævnes med $F(x)$.
- Fordelingsfunktionen er defineret ved

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt.$$

- Som følge af denne definition gælder, at

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

hvor fordelingsfunktionen er differentierbar.

Fordelingsfunktionen, Definition 2.33

- Fordelingsfunktionen (distribution function/cumulative distribution function, cdf) hørende til en kontinuert stokastisk variabel benævnes med $F(x)$.
- Fordelingsfunktionen er defineret ved

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt.$$

- Som følge af denne definition gælder, at

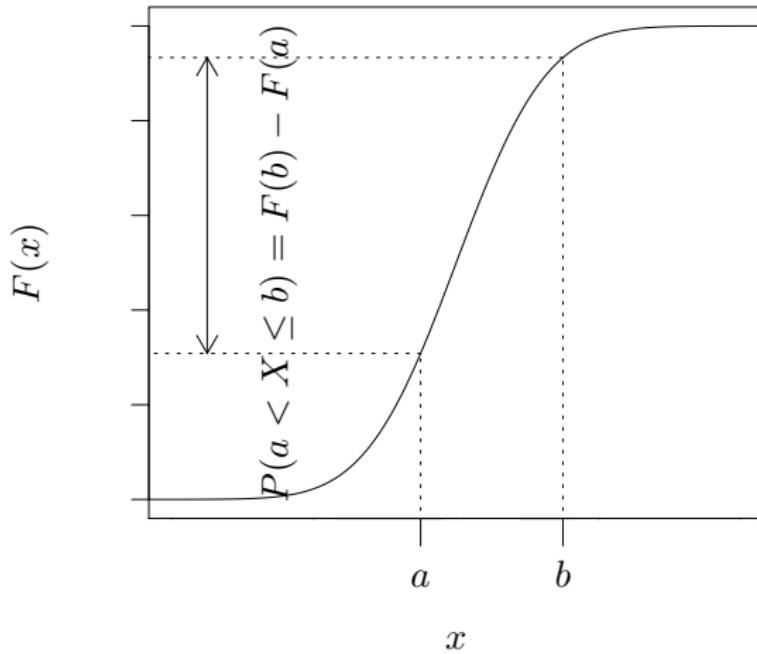
$$f(x) = F'(x),$$

hvor fordelingsfunktionen er differentierbar.

- Bemærk også, at:

$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x) dx.$$

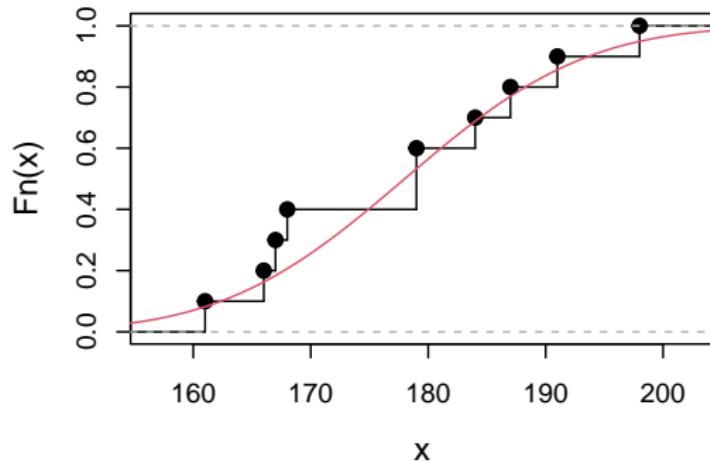
Fordelingsfunktionen



Empirisk fordelingsfunktion (ecdf)

```
# Empirical cdf for sample of height data from Chapter 1
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
plot(ecdf(x), verticals = TRUE, main = "")

# 'True cdf' for normal distribution (with sample mean and variance)
xp <- seq(0.9*min(x), 1.1*max(x), length = 100)
lines(xp, pnorm(xp, mean(x), sd(x)), col = 2)
```



Middelværdi (mean) af en kontinuert stokastisk variabel , Definition 2.34

Middelværdien af en kontinuert stokastisk variabel:

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

Middelværdi (mean) af en kontinuert stokastisk variabel , Definition 2.34

Middelværdien af en kontinuert stokastisk variabel:

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

Sammenlign med definitionen for en diskret stokastisk variabel:

$$\mu = \sum_{\text{alle } x} xf(x)$$

Varians af en kontinuert stokastisk variabel, Definition 2.34

Variansen af en kontinuert stokastisk variabel:

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

Varians af en kontinuert stokastisk variabel, Definition 2.34

Variansen af en kontinuert stokastisk variabel:

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

Sammenlign med definitionen for en diskret stokastisk variabel:

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

Kovarians, Definition 2.58

Kovariansen af to stokastisk variable:

Lad X og Y være to stokastiske variable. Kovariansen mellem X og Y er defineret som

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Kovariansen:

Hvis to stokastiske variable er *uafhængige*, så er kovariansen 0. *Det modsatte er ikke nødvendigvis tilfældet!*

Dagsorden

- 1 Opsummering
- 2 Kontinuerte fordelinger
 - Tæthed- og fordelingsfunktioner
 - Middelværdi, varians og kovarians
- 3 Vigtige kontinuerte fordelinger
 - Den uniforme fordeling
 - Normalfordelingen
 - Log-normalfordelingen
 - Eksponentialfordelingen
- 4 Regneregler for stokastiske variable

Vigtige kontinuerte fordelinger

Der findes en række statistiske fordelinger (både kontinuerte og diskrete), som kan bruges til at beskrive og analysere forskellige problemstillinger med

I dag ser vi nærmere på følgende kontinuerte fordelinger:

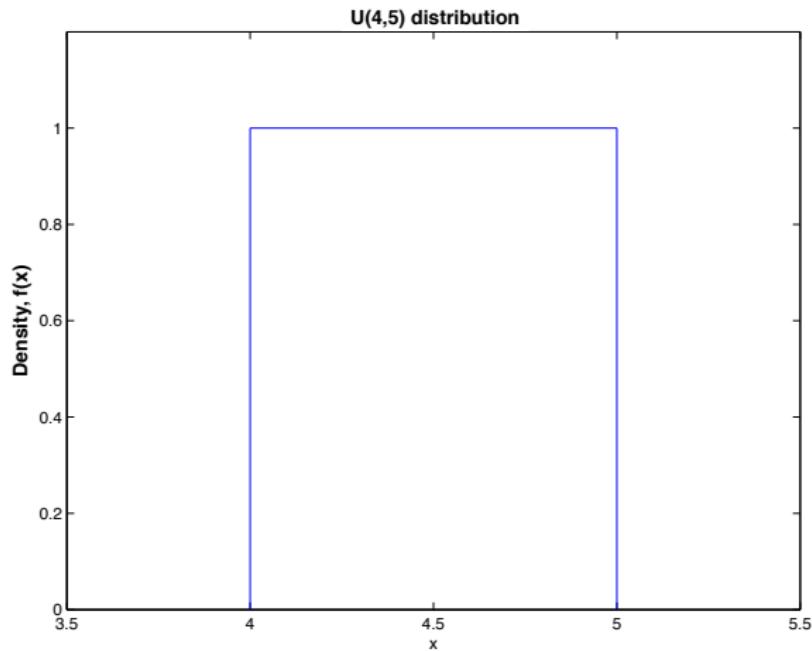
- Den uniforme fordeling
- Normalfordelingen
- Log-normalfordelingen
- Eksponentialfordelingen

Kontinuerte fordelinger in R

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

- d Tæthedsfunktion (density)
- p Fordelingsfunktion (probability)
- q Fraktilfunktion (quantile)
- r Tilfældighedsgenerator (random).

Tæthed for en uniform fordeling (eksempel)



Den uniforme fordeling, Def. 2.35 & sæt. 2.36

Notation:

$$X \sim U(\alpha, \beta)$$

Den uniforme fordeling, Def. 2.35 & sæt. 2.36

Notation:

$$X \sim U(\alpha, \beta)$$

Tæthedsfunktion:

$$f(x) = \frac{1}{\beta - \alpha} \text{ for } \alpha \leq x \leq \beta \text{ og ellers nul}$$

Den uniforme fordeling, Def. 2.35 & sæt. 2.36

Notation:

$$X \sim U(\alpha, \beta)$$

Tæthedsfunktion:

$$f(x) = \frac{1}{\beta - \alpha} \text{ for } \alpha \leq x \leq \beta \text{ og ellers nul}$$

Middelværdi:

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

Den uniforme fordeling, Def. 2.35 & sæt. 2.36

Notation:

$$X \sim U(\alpha, \beta)$$

Tæthedsfunktion:

$$f(x) = \frac{1}{\beta - \alpha} \text{ for } \alpha \leq x \leq \beta \text{ og ellers nul}$$

Middelværdi:

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

Varians:

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

Eksempel 1

Studerende på et statistikkursus ankommer til en forelæsning mellem 8.00 og 8.30. Det antages, at ankomsttiden kan beskrives ved en uniform fordeling.

Spørgsmål:

Hvad er sandsynligheden for, at en tilfældigt udvalgt studerende ankommer mellem 8:20 og 8:30?

Eksempel 1

Studerende på et statistikkursus ankommer til en forelæsning mellem 8.00 og 8.30. Det antages, at ankomsttiden kan beskrives ved en uniform fordeling.

Spørgsmål:

Hvad er sandsynligheden for, at en tilfældigt udvalgt studerende ankommer mellem 8:20 og 8:30?

Svar:

$$10/30 = 1/3$$

Lad $X \sim U(0,30)$ repræsentere "ankomsttiden" for en tilfældigt udvalgt studerende:

$$P(20 \leq X \leq 30) = P(X \leq 30) - P(X \leq 20) = 1 - 2/3 = 1/3$$

```
punif(30, 0, 30) - punif(20, 0, 30)
```

```
[1] 0.3333
```

Eksempel 1 (fortsat)

Spørgsmål:

Hvad er sandsynligheden for, at en tilfældigt udvalgt studerende ankommer efter 8:30?

Eksempel 1 (fortsat)

Spørgsmål:

Hvad er sandsynligheden for, at en tilfældigt udvalgt studerende ankommer efter 8:30?

Svar:

0

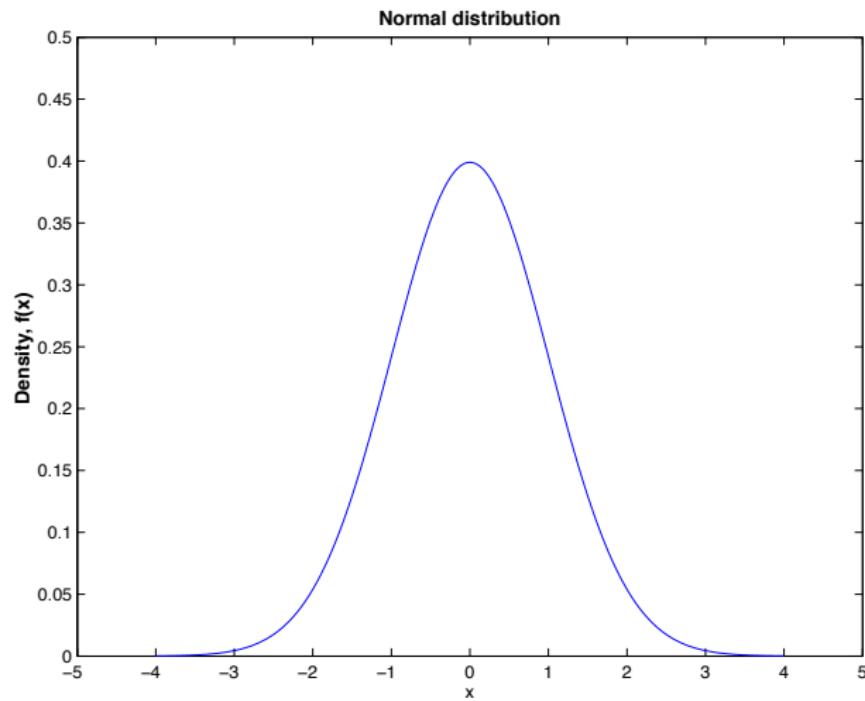
Lad $X \sim U(0, 30)$ repræsentere "ankomstiden" for en tilfældigt udvalgt studerende:

$$P(X > 30) = 1 - P(X \leq 30) = 1 - 1 = 0$$

```
1 - punif(30, 0, 30)
```

```
[1] 0
```

Tætheden for en normalfordeling (eksempel)



Normalfordelingen, Def. 2.37 & sæt. 2.38

Notation:

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

Normalfordelingen, Def. 2.37 & sæt. 2.38

Notation:

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

Tæthedsfunktion:

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

Normalfordelingen, Def. 2.37 & sæt. 2.38

Notation:

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

Tæthedsfunktion:

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

Middelværdi:

$$\mu$$

Normalfordelingen, Def. 2.37 & sæt. 2.38

Notation:

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

Tæthedsfunktion:

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

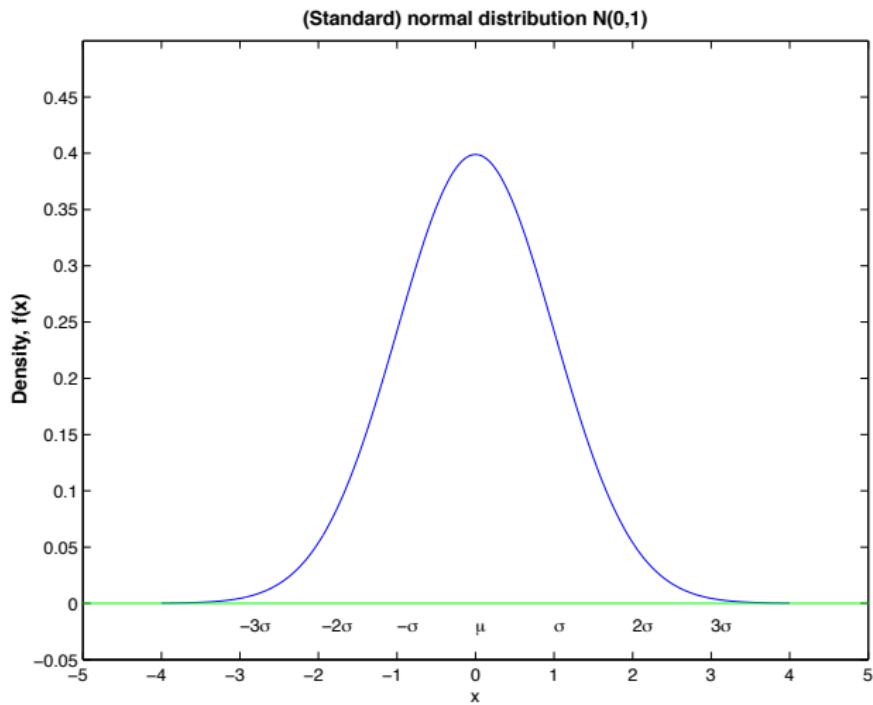
Middelværdi:

$$\mu$$

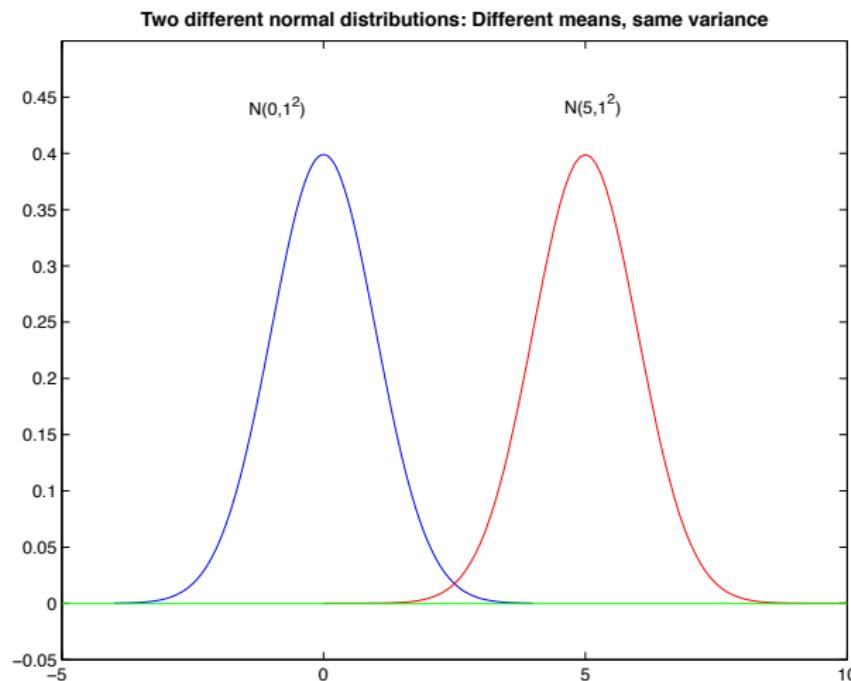
Varians:

$$\sigma^2$$

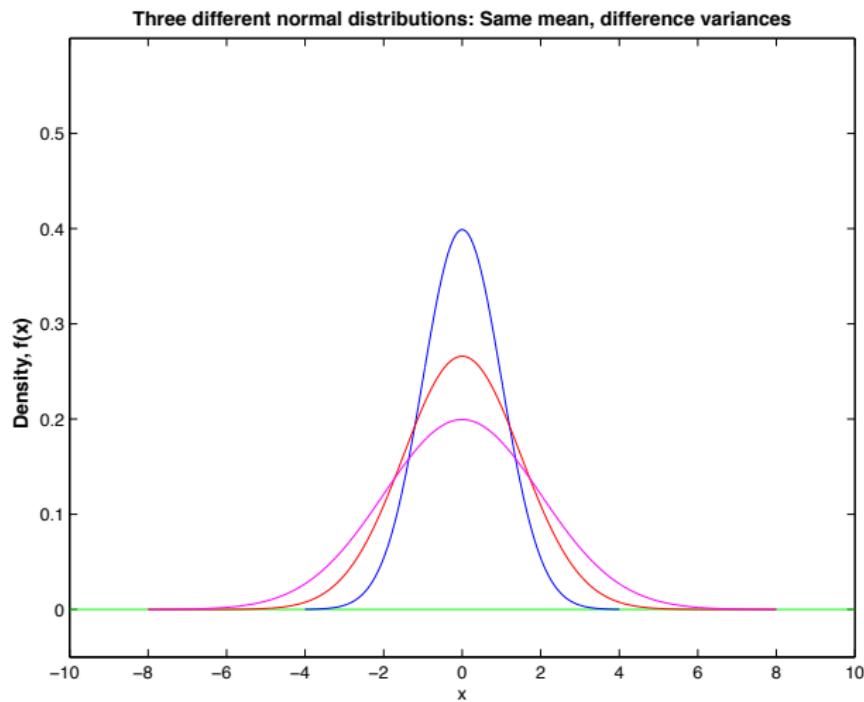
Tætheden for en standardnormalfordeling



Tætheder for to normalfordelinger (eksempel)



Tætheder for tre normalfordelinger (eksempel)



Standardnormalfordelingen

Standardnormalfordelingen:

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

Normalfordelingen med middelværdi 0 og varians 1.

Standardnormalfordelingen

Standardnormalfordelingen:

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

Normalfordelingen med middelværdi 0 og varians 1.

Standardisering:

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

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

Eksempel 2

Målefejl:

En given vægt har en målefejl (målt i gram), Z , som kan beskrives med en standardnormalfordeling,

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

Dvs. at den gennemsnitlige målefejl er $\mu = 0$ gram og standardafgivelsen er $\sigma = 1$ gram.
Antag, at vægten bruges til at veje et produkt.

Eksempel 2

Målefejl:

En given vægt har en målefejl (målt i gram), Z , som kan beskrives med en standardnormalfordeling,

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

Dvs. at den gennemsnitlige målefejl er $\mu = 0$ gram og standardafgivelsen er $\sigma = 1$ gram.
Antag, at vægten bruges til at veje et produkt.

Spørgsmål a):

Hvad er sandsynligheden for, at vægten giver et resultat, som er mindst 2 gram mindre end den sande vægt af produktet?

Eksempel 2

Målefejl:

En given vægt har en målefejl (målt i gram), Z , som kan beskrives med en standardnormalfordeling,

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

Dvs. at den gennemsnitlige målefejl er $\mu = 0$ gram og standardafgivelsen er $\sigma = 1$ gram.
Antag, at vægten bruges til at veje et produkt.

Spørgsmål a):

Hvad er sandsynligheden for, at vægten giver et resultat, som er mindst 2 gram mindre end den sande vægt af produktet?

Svar:

$$P(Z \leq -2) = 0.02275$$

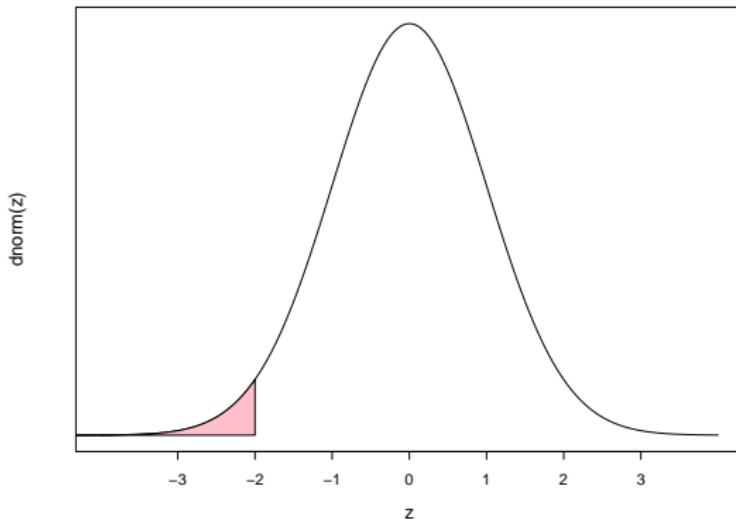
`pnorm(-2)`

Eksempel 2

Svar:

```
pnorm(-2)
```

```
[1] 0.02275
```



Eksempel 2

Spørgsmål b):

Hvad er sandsynligheden for, at vægten giver et resultat, som er mindst 2 gram større end den sande vægt af produktet?

Eksempel 2

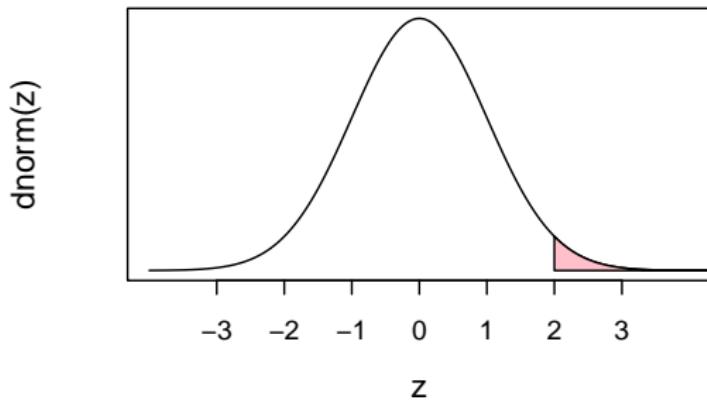
Spørgsmål b):

Hvad er sandsynligheden for, at vægten giver et resultat, som er mindst 2 gram større end den sande vægt af produktet?

Svar:

$$P(Z \geq 2) = 0.02275$$

```
1 - pnorm(2)
```



Eksempel 2

Spørgsmål c):

Hvad er sandsynligheden for, at vægten har en afvigelse på højst ± 1 gram?

Eksempel 2

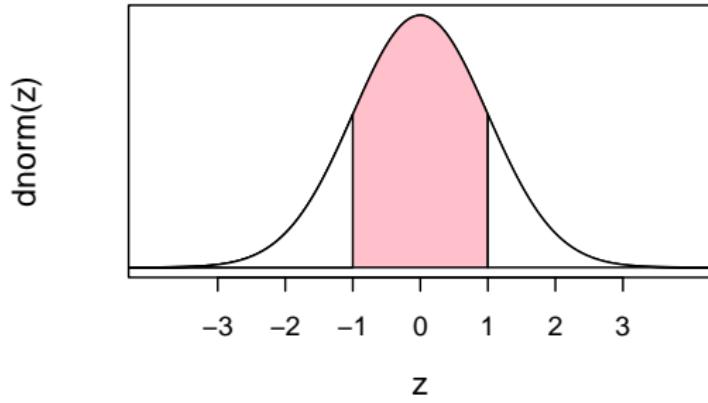
Spørgsmål c):

Hvad er sandsynligheden for, at vægten har en afvigelse på højst ± 1 gram?

Svar:

$$P(|Z| \leq 1) = P(-1 \leq Z \leq 1) = P(Z \leq 1) - P(Z \leq -1) = 0.683$$

```
pnorm(1) - pnorm(-1)
```



Eksempel 3

Indkomstfordeling:

Det antages, at folkeskolelæreres løn kan beskrives med en normalfordeling med middelværdi $\mu = 290$ (i 1000 DKK) og standardafvigelse $\sigma = 4$ (1000 DKK).

Eksempel 3

Indkomstfordeling:

Det antages, at folkeskolelæreres løn kan beskrives med en normalfordeling med middelværdi $\mu = 290$ (i 1000 DKK) og standardafvigelse $\sigma = 4$ (1000 DKK).

Spørgsmål a):

Hvad er sandsynligheden for, at en tilfældigt udvalgt lærer tjener mere end 300.000 kr.?

Eksempel 3

Spørgsmål a):

Hvad er sandsynligheden for, at en tilfældigt udvalgt lærer tjener mere end 300.000 kr?

Eksempel 3

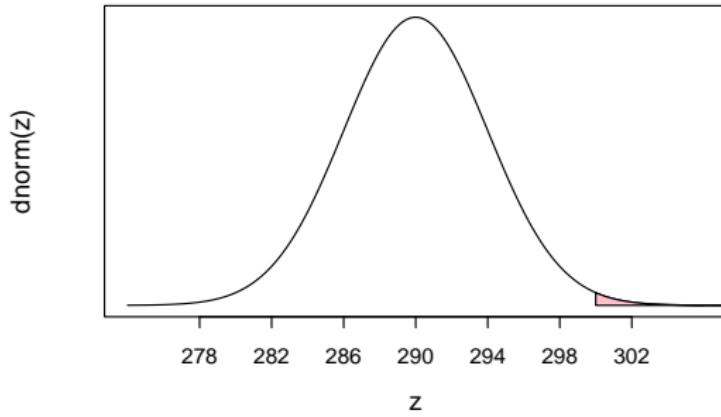
Spørgsmål a):

Hvad er sandsynligheden for, at en tilfældigt udvalgt lærer tjener mere end 300.000 kr?

Svar:

```
1 - pnorm(300, m = 290, s = 4)
```

```
[1] 0.00621
```



Eksempel 4

(Samme indkomstfordeling):

Det antages, at folkeskolelæreres løn kan beskrives med en normalfordeling med middelværdi $\mu = 290$ (i 1000 DKK) og standardafvigelse $\sigma = 4$ (1000 DKK).

Eksempel 4

(Samme indkomstfordeling):

Det antages, at folkeskolelæreres løn kan beskrives med en normalfordeling med middelværdi $\mu = 290$ (i 1000 DKK) og standardafvigelse $\sigma = 4$ (1000 DKK).

"Modsat" spørgsmål:

Specifér et løninterval (som er symmetrisk omkring middelværdien), som dækker 95% af lærernes lønninger.

Eksempel 4

(Samme indkomstfordeling):

Det antages, at folkeskolelæreres løn kan beskrives med en normalfordeling med middelværdi $\mu = 290$ (i 1000 DKK) og standardafvigelse $\sigma = 4$ (1000 DKK).

"Modsat" spørgsmål:

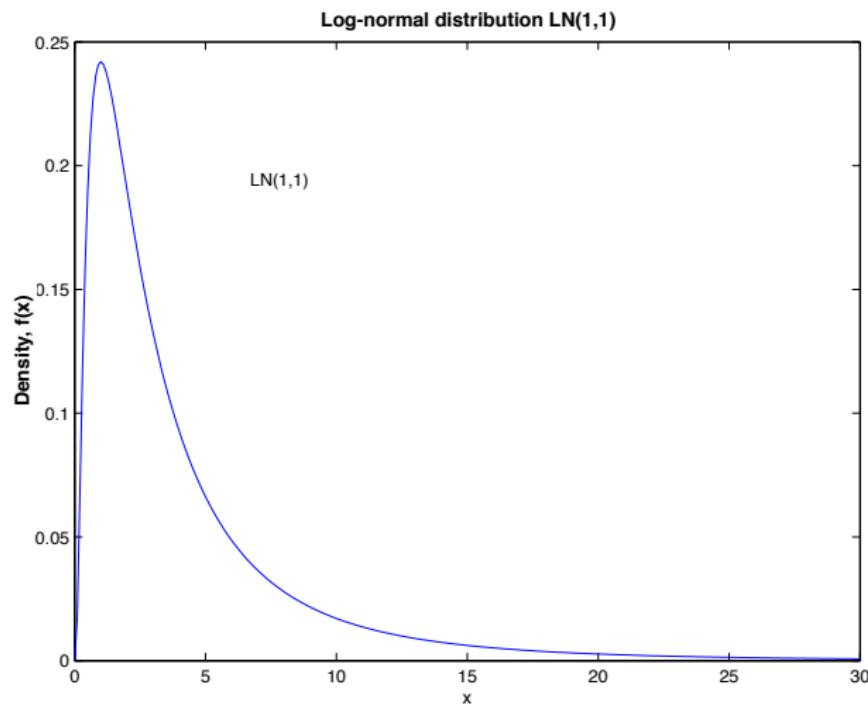
Specifér et løninterval (som er symmetrisk omkring middelværdien), som dækker 95% af lærernes lønninger.

Svar:

```
qnorm(c(0.025, 0.975), m = 290, s = 4)
```

```
[1] 282.2 297.8
```

Log-normalfordelingen



Log-normalfordelingen, Def. 2.46 & sæt. 2.47

Notation:

$$X \sim LN(\alpha, \beta^2) \text{ (hvor } \beta > 0)$$

Log-normalfordelingen, Def. 2.46 & sæt. 2.47

Notation:

$$X \sim LN(\alpha, \beta^2) \text{ (hvor } \beta > 0)$$

Tæthedsfunktion:

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

Log-normalfordelingen, Def. 2.46 & sæt. 2.47

Notation:

$$X \sim LN(\alpha, \beta^2) \text{ (hvor } \beta > 0)$$

Tæthedsfunktion:

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

Middelværdi:

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

Log-normalfordelingen, Def. 2.46 & sæt. 2.47

Notation:

$$X \sim LN(\alpha, \beta^2) \text{ (hvor } \beta > 0)$$

Tæthedsfunktion:

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

Middelværdi:

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

Varians:

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

Log-normalfordelingen

Log-normal- og normalfordelingen:

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

$$X = \ln(Y).$$

Her er X normalfordelt med middelværdi α og varians β^2 , dvs. $X \sim N(\alpha, \beta^2)$.

Log-normalfordelingen

Log-normal- og normalfordelingen:

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

$$X = \ln(Y).$$

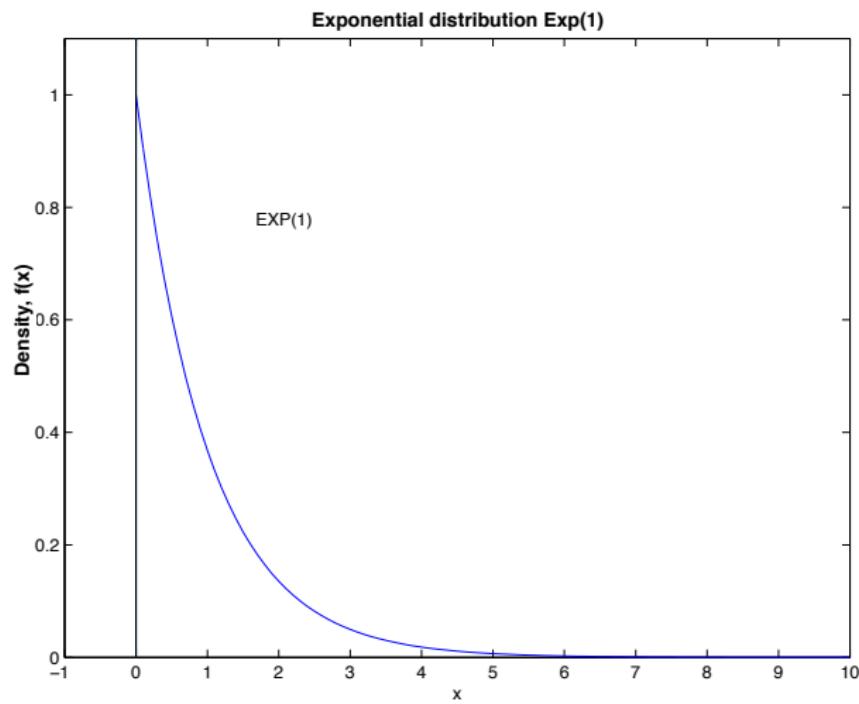
Her er X normalfordelt med middelværdi α og varians β^2 , dvs. $X \sim N(\alpha, \beta^2)$.

Ved at standardisere X igennem

$$Z = \frac{X - \mathbb{E}[X]}{\sqrt{\mathbb{V}[X]}} = \frac{\ln(Y) - \alpha}{\beta}$$

opnås en standardnormalfordelt variabel $Z \sim N(0, 1)$.

Eksponentiafordelingen



Eksponentialfordelingen, Def. 2.48 & sæt. 2.49

Notation:

$X \sim \text{Exp}(\lambda)$, hvor $\lambda > 0$.

Tæthedsfunktion:

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

Middelværdi:

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

Varians:

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

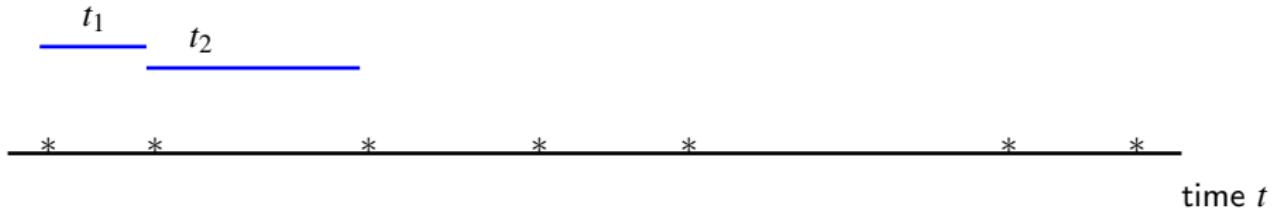
Eksponentiafordelingen

- Eksponentiafordelingen er et specialtilfælde af *gammafordelingen*.
- Eksponentiafordelingen anvendes f.eks. til at beskrive levetider og ventetider.
- Eksponentiafordelingen kan bruges til at beskrive (vente)tiden mellem hændelser i en poissonproces.

Sammenhæng mellem eksponential- og poissonfordelingen

Poisson: Diskrete hændelser pr. enhed

Eksponential: Kontinuert afstand mellem hændelser



Eksempel 5

Kø-model: Poissonproces

Tiden mellem kundeankomster på et posthus er eksponentiafordelt med middelværdi $\mu = 2$ minutter.

Eksempel 5

Kø-model: Poissonproces

Tiden mellem kundeankomster på et posthus er eksponentiafordelt med middelværdi $\mu = 2$ minutter.

Spørgsmål:

En kunde er netop ankommet. Hvad er sandsynligheden for, at der ikke kommer flere kunder indenfor en periode på 2 minutter?

Eksempel 5

Kø-model: Poissonproces

Tiden mellem kundeankomster på et posthus er eksponentiafordelt med middelværdi $\mu = 2$ minutter.

Spørgsmål:

En kunde er netop ankommet. Hvad er sandsynligheden for, at der ikke kommer flere kunder indenfor en periode på 2 minutter?

Svar:

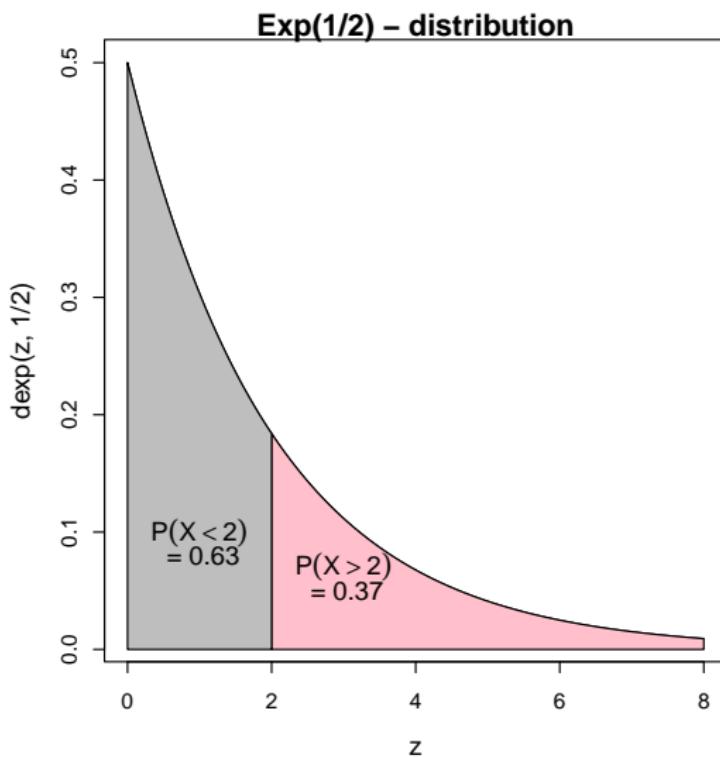
Lad $X \sim \text{Exp}(1/2)$ repræsentere ventetiden indtil ankomsten af den næste kunde.

$$P(X > 2) = 1 - P(X \leq 2)$$

```
1 - pexp(2, rate = 1/2)
```

```
[1] 0.3679
```

Eksempel 5



Eksempel 6

Spørgsmål:

En kunde er netop ankommet.

Brug poissonfordelingen til at beregne sandsynligheden for, at der ikke kommer flere kunder inden for de næste to minutter.

Eksempel 6

Spørgsmål:

En kunde er netop ankommet.

Brug poissonfordelingen til at beregne sandsynligheden for, at der ikke kommer flere kunder inden for de næste to minutter.

Svar:

$$\lambda_{2\text{min}} = 1, P(X = 0) = \frac{e^{-1}}{1!} 1^0 = e^{-1}$$

```
dpois(0, 1)
```

```
[1] 0.3679
```

```
exp(-1)
```

```
[1] 0.3679
```

Dagsorden

- 1 Opsummering
- 2 Kontinuerte fordelinger
 - Tæthed- og fordelingsfunktioner
 - Middelværdi, varians og kovarians
- 3 Vigtige kontinuerte fordelinger
 - Den uniforme fordeling
 - Normalfordelingen
 - Log-normalfordelingen
 - Eksponentialfordelingen
- 4 Regneregler for stokastiske variable

Regneregler for stokastiske variable

Disse regneregler gælder både for kontinuerte og diskrete stokastiske variable!

Lad X være en stokastisk variabel, medens a og b er konstanter.

Regneregler for stokastiske variable

Disse regneregler gælder både for kontinuerte og diskrete stokastiske variable!

Lad X være en stokastisk variabel, medens a og b er konstanter.

Middelværdi-regel:

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

Regneregler for stokastiske variable

Disse regneregler gælder både for kontinuerte og diskrete stokastiske variable!

Lad X være en stokastisk variabel, medens a og b er konstanter.

Middelværdi-regel:

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

Varians-regel:

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

Eksempel 7

Lad X være en stokastisk variabel med middelværdi 4 og varians 6.

Spørgsmål:

Beregn middelværdien og variansen af $Y = -3X + 2$.

Eksempel 7

Lad X være en stokastisk variabel med middelværdi 4 og varians 6.

Spørgsmål:

Beregn middelværdien og variansen af $Y = -3X + 2$.

Svar:

$$\mathbb{E}(Y) = -3\mathbb{E}(X) + 2 = -3 \cdot 4 + 2 = -10$$

$$\text{Var}(Y) = (-3)^2 \text{Var}(X) = 9 \cdot 6 = 54$$

Regneregler for stokastiske variable

Lad X_1, \dots, X_n være *uafhængige* stokastiske variable.

Regneregler for stokastiske variable

Lad X_1, \dots, X_n være *uafhængige* stokastiske variable.

Middelværdi-regel:

$$\begin{aligned} & \mathbb{E}(a_1X_1 + a_2X_2 + \cdots + a_nX_n) \\ &= a_1\mathbb{E}(X_1) + a_2\mathbb{E}(X_2) + \cdots + a_n\mathbb{E}(X_n) \end{aligned}$$

Regneregler for stokastiske variable

Lad X_1, \dots, X_n være *uafhængige* stokastiske variable.

Middelværdi-regel:

$$\begin{aligned} & \mathbb{E}(a_1X_1 + a_2X_2 + \cdots + a_nX_n) \\ &= a_1\mathbb{E}(X_1) + a_2\mathbb{E}(X_2) + \cdots + a_n\mathbb{E}(X_n) \end{aligned}$$

Varians-regel:

$$\begin{aligned} & \text{Var}(a_1X_1 + a_2X_2 + \cdots + a_nX_n) \\ &= a_1^2\text{Var}(X_1) + \cdots + a_n^2\text{Var}(X_n) \end{aligned}$$

Eksempel 8

Planlægning for flyselskab

Vægten af én passager på en flytur, X , antages at være normalfordelt så $X \sim N(70, 10^2)$.

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

Eksempel 8

Planlægning for flyselskab

Vægten af én passager på en flytur, X , antages at være normalfordelt så $X \sim N(70, 10^2)$.

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

Spørgsmål:

Beregn sandsynligheden for, at flyet bliver overlastet.

Eksempel 8

Planlægning for flyselskab

Vægten af én passager på en flytur, X , antages at være normalfordelt så $X \sim N(70, 10^2)$.

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

Spørgsmål:

Beregn sandsynligheden for, at flyet bliver overlastet.

Hvad er den samlede passagervægt Y på en afgang?

Eksempel 8

Planlægning for flyselskab

Vægten af én passager på en flytur, X , antages at være normalfordelt så $X \sim N(70, 10^2)$.

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

Spørgsmål:

Beregn sandsynligheden for, at flyet bliver overlastet.

Hvad er den samlede passagervægt Y på en afgang?

Hvad er Y ?

IKKE: $Y = 55 \cdot X$

Eksempel 8

Hvad er den samlede passagervægt Y ?

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

Eksempel 8

Hvad er den samlede passagervægt Y ?

$Y = \sum_{i=1}^{55} X_i$, hvor $X_i \sim N(70, 10^2)$ (som antages at være uafhængige)

Middelværdi og varians af Y :

$$\mathbb{E}(Y) = \sum_{i=1}^{55} \mathbb{E}(X_i) = \sum_{i=1}^{55} 70 = 55 \cdot 70 = 3850$$

$$\text{Var}(Y) = \sum_{i=1}^{55} \text{Var}(X_i) = \sum_{i=1}^{55} 100 = 55 \cdot 100 = 5500$$

Eksempel 8

Hvad er den samlede passagervægt Y ?

$Y = \sum_{i=1}^{55} X_i$, hvor $X_i \sim N(70, 10^2)$ (som antages at være uafhængige)

Middelværdi og varians af Y :

$$\mathbb{E}(Y) = \sum_{i=1}^{55} \mathbb{E}(X_i) = \sum_{i=1}^{55} 70 = 55 \cdot 70 = 3850$$

$$\text{Var}(Y) = \sum_{i=1}^{55} \text{Var}(X_i) = \sum_{i=1}^{55} 100 = 55 \cdot 100 = 5500$$

Y er normalfordelt, så vi kan finde $P(Y > 4000)$ ved:

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

[1] 0.02156

Eksempel 8 - FORKERT analyse

Hvad er Y ?

IKKE: $Y = 55 \cdot X$

Eksempel 8 - FORKERT analyse

Hvad er Y ?

IKKE: $Y = 55 \cdot X$

Middelværdi og varians af FORKERT Y :

$$\mathbb{E}(Y) = 55 \cdot 70 = 3850$$

$$\text{Var}(Y) = 55^2 \text{Var}(X) = 55^2 \cdot 100 = 550^2$$

Eksempel 8 - FORKERT analyse

Hvad er Y ?

IKKE: $Y = 55 \cdot X$

Middelværdi og varians af FORKERT Y :

$$\mathbb{E}(Y) = 55 \cdot 70 = 3850$$

$$\text{Var}(Y) = 55^2 \text{Var}(X) = 55^2 \cdot 100 = 550^2$$

Det FORKERTE Y er også normalfordelt. Her finder vi $P(Y > 4000)$ med FORKERT Y :

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

```
[1] 0.3925
```

Eksempel 8 - FORKERT analyse

Hvad er Y ?

IKKE: $Y = 55 \cdot X$

Middelværdi og varians af FORKERT Y :

$$\mathbb{E}(Y) = 55 \cdot 70 = 3850$$

$$\text{Var}(Y) = 55^2 \text{Var}(X) = 55^2 \cdot 100 = 550^2$$

Det FORKERTE Y er også normalfordelt. Her finder vi $P(Y > 4000)$ med FORKERT Y :

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

```
[1] 0.3925
```

Konsekvens af forkert beregning:

MANGE spilde penge for flyselskabet!!!

Dagsorden

- 1 Opsummering
- 2 Kontinuerte fordelinger
 - Tæthed- og fordelingsfunktioner
 - Middelværdi, varians og kovarians
- 3 Vigtige kontinuerte fordelinger
 - Den uniforme fordeling
 - Normalfordelingen
 - Log-normalfordelingen
 - Eksponentialfordelingen
- 4 Regneregler for stokastiske variable

02402 Statistik (Polyteknisk grundlag)

Uge 4: Konfidensintervaller

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

Dagsorden

- 1 Opsummering
- 2 Introduktion og eksempel
- 3 Fordelingen for gennemsnittet
 - t -fordelingen
- 4 Konfidensintervallet for μ
 - Eksempel: Højder
- 5 Statistisk sprogbrug og den formelle ramme
- 6 Ikke-normale data
- 7 Formel fortolkning af konfidensintervallet
- 8 Konfidensinterval for varians og spredning

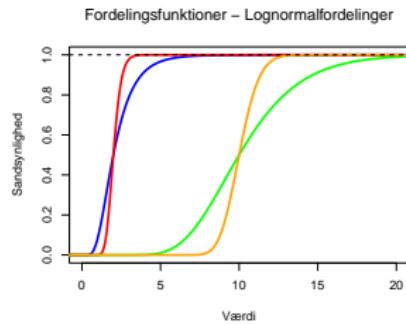
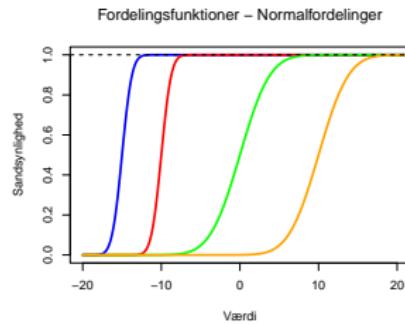
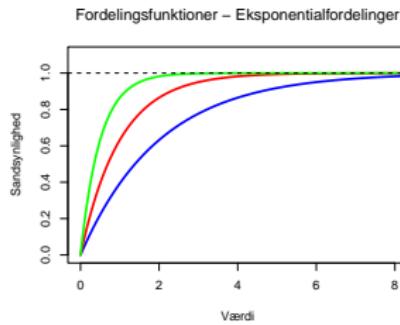
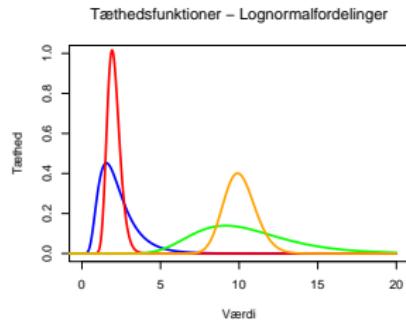
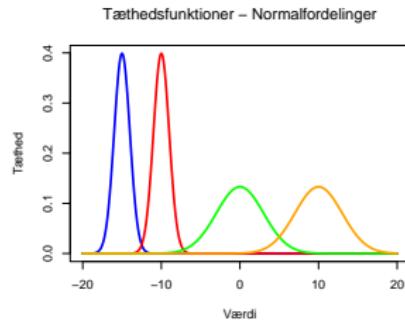
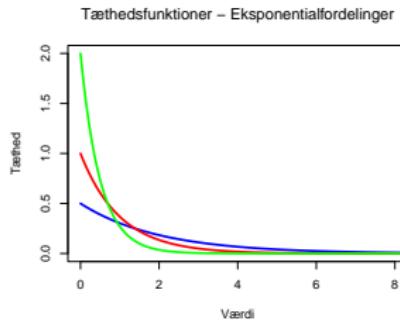
Dagsorden

- 1 Opsummering
- 2 Introduktion og eksempel
- 3 Fordelingen for gennemsnittet
 - t -fordelingen
- 4 Konfidensintervallet for μ
 - Eksempel: Højder
- 5 Statistisk sprogbrug og den formelle ramme
- 6 Ikke-normale data
- 7 Formel fortolkning af konfidensintervallet
- 8 Konfidensinterval for varians og spredning

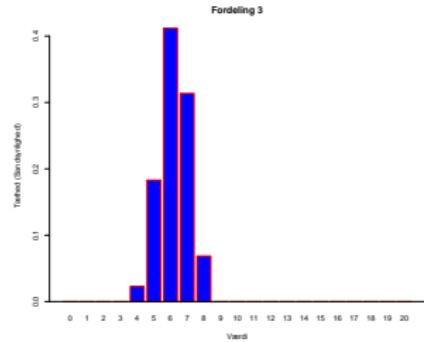
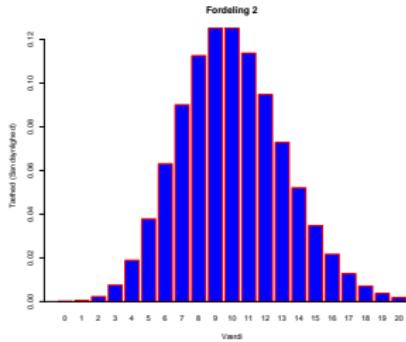
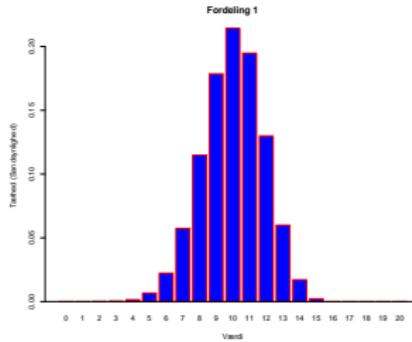
Læringsmål fra de første uger

- Beregne og fortolke simple statistiske størrelser, herunder gennemsnit, spredning, varians, median, fraktiler og korrelation
- Anvende enkle grafiske eksplorative teknikker
- Identificere og beskrive sandsynlighedsfordelinger som Poisson-, binomial-, eksponential- og normalfordelingen

Sidste uge: Kontinuerte fordelinger



Spørgsmål



Spørgsmål 1

- A) Bin(15,2/3) B) HG(14,8,18) C) Pois(10)

Spørgsmål 2

- A) $\mathbb{E}[X] = 56/9$, $\mathbb{V}[X] \approx 0.81$ B) $\mathbb{E}[X] = 10$, $\mathbb{V}[X] = 10$ C) $\mathbb{E}[X] = 10$, $\mathbb{V}[X] = 10/3$

Spørgsmål 3

- A) $\mathbb{P}(X = 8) \approx \mathbb{P}(X = 11)$ B) $\mathbb{P}(X = 8) < 0.1$ C) $\mathbb{P}(X \leq 5) < 0.01$

Tips fra underviserne og hjælpelærerne

- Brug bogen og dias
- Bogen har en formelsamling
- Prøv at løse problemer med blyant og papir før I bruger R

Dagsorden

- 1 Opsummering
- 2 Introduktion og eksempel
- 3 Fordelingen for gennemsnittet
 - t -fordelingen
- 4 Konfidensintervallet for μ
 - Eksempel: Højder
- 5 Statistisk sprogbrug og den formelle ramme
- 6 Ikke-normale data
- 7 Formel fortolkning af konfidensintervallet
- 8 Konfidensinterval for varians og spredning

Eksempel - Højde af 10 studerende:

Stikprøve, $n = 10$:

168 161 167 179 184 166 198 187 191 179

Eksempel - Højde af 10 studerende:

Stikprøve, $n = 10$:

168 161 167 179 184 166 198 187 191 179

Stikprøvegennemsnit og
-standardafvigelse:

$$\bar{x} = 178$$

$$s = 12.21$$

Eksempel - Højde af 10 studerende:

Stikprøve, $n = 10$:

168 161 167 179 184 166 198 187 191 179

Stikprøvegennemsnit og
-standardafvigelse:

$$\bar{x} = 178$$

$$s = 12.21$$

Estimater for populationens
middelværdi og standardafvigelse:

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

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

Eksempel - Højde af 10 studerende:

Stikprøve, $n = 10$:

168 161 167 179 184 166 198 187 191 179

Stikprøvegennemsnit og
-standardafvigelse:

$$\bar{x} = 178$$

$$s = 12.21$$

NYT: Konfidensinterval for μ :

$$[169.3; 186.7]$$

Estimater for populationens
middelværdi og standardafvigelse:

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

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

NYT: Konfidensinterval for σ :

$$[8.4; 22.3]$$

Dagsorden

- 1 Opsummering
- 2 Introduktion og eksempel
- 3 Fordelingen for gennemsnittet
 - t -fordelingen
- 4 Konfidensintervallet for μ
 - Eksempel: Højder
- 5 Statistisk sprogbrug og den formelle ramme
- 6 Ikke-normale data
- 7 Formel fortolkning af konfidensintervallet
- 8 Konfidensinterval for varians og spredning

(Empirisk) fordeling af stikprøvegennemsnittet

```

# 'Sand' middelværdi og standardafvigelse
mu <- 178
sigma <- 12

# Stikprøvestørrelsen
n <- 10

# Simuler normalfordelte  $X_i$  for  $n = 10$ 
x <- rnorm(n = n, mean = mu, sd = sigma)
x

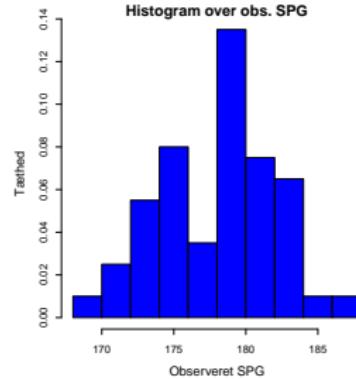
# Empirisk tæthed
hist(x, prob = TRUE, col = 'blue')
# Stikprøvegennemsnit
mean(x)

# Gentag eksperimentet (100 gange)
mat <- replicate(100, rnorm(n = n, mean = mu, sd = sigma))

# Udregn gennemsnit for hver stikprøve
xbar <- apply(mat, 2, mean)
xbar

# Fordelingen af stikprøvegennemsnittene (vist til højre)
hist(xbar, prob = TRUE, col = 'blue')
# Gns. og varians af stikprøvegennemsnittene
mean(xbar)
var(xbar)

```



Sætning 3.3: Fordeling for stikprøvegennemsnittet af normalfordelte variable

(Stikprøve-)fordelingen for \bar{X} :

Antag at X_1, \dots, X_n er uafhængige og ensfordelte (i.i.d) stokastiske variable, $X_i \sim N(\mu, \sigma^2), i = 1, \dots, n$, så:

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

Middelværdien og variansen følger af regneregler

Middelværdien af \bar{X} (Sætning 2.56):

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

Middelværdien og variansen følger af regneregler

Middelværdien af \bar{X} (Sætning 2.56):

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

Variansen for \bar{X} (Sætning 2.56):

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

Middelværdien og variansen følger af regneregler

Middelværdien af \bar{X} (Sætning 2.56):

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

Variansen for \bar{X} (Sætning 2.56):

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

Normaliteten af \bar{X} (Sætning 2.40):

Fra denne sætning følger, at \bar{X} er normalfordelt med middelværdi μ og varians σ^2/n .

Fordelingen af fejlen ($\bar{X} - \mu$)

Spredningen af \bar{X}

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Fordelingen af fejlen ($\bar{X} - \mu$)

Spredningen af \bar{X}

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Spredningen af $(\bar{X} - \mu)$

$$\sigma_{(\bar{X} - \mu)} = \frac{\sigma}{\sqrt{n}}$$

Standardiseret version af de samme ting, Sætning 3.4:

Fordelingen for den *standardiserede fejl*, vi begår:

Antag at X_1, \dots, X_n er uafhængige og ensfordelte (i.i.d.) stokastiske variable $X_i \sim N(\mu, \sigma^2)$, hvor $i = 1, \dots, n$, så:

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

Dvs. at det standardiserede stikprøvegennemsnit (Z) følger en standardnormalfordeling.

Praktisk problem i alt dette!

Hvordan skal resultaterne fra de foregående slides omsættes til et konkret interval for μ ?

Problemet: Populationsspredningen σ indgår i alle formlerne.

Praktisk problem i alt dette!

Hvordan skal resultaterne fra de foregående slides omsættes til et konkret interval for μ ?

Problemet: Populationsspredningen σ indgår i alle formlerne.

Oplagt løsning:

Anvend estimatet s i stedet for σ i formlerne!

Praktisk problem i alt dette!

Hvordan skal resultaterne fra de foregående slides omsættes til et konkret interval for μ ?

Problemet: Populationsspredningen σ indgår i alle formlerne.

Oplagt løsning:

Anvend estimatet s i stedet for σ i formlerne!

MEN:

Så bryder den givne teori faktisk sammen!

Praktisk problem i alt dette!

Hvordan skal resultaterne fra de foregående slides omsættes til et konkret interval for μ ?

Problemet: Populationsspredningen σ indgår i alle formlerne.

Oplagt løsning:

Anvend estimatet s i stedet for σ i formlerne!

MEN:

Så bryder den givne teori faktisk sammen!

HELDIGVIS:

Findes der en udvidet teori, der kan klare det!

Sætning 3.5: Mere anvendeligt resultat: (kopi af sætning 2.49)

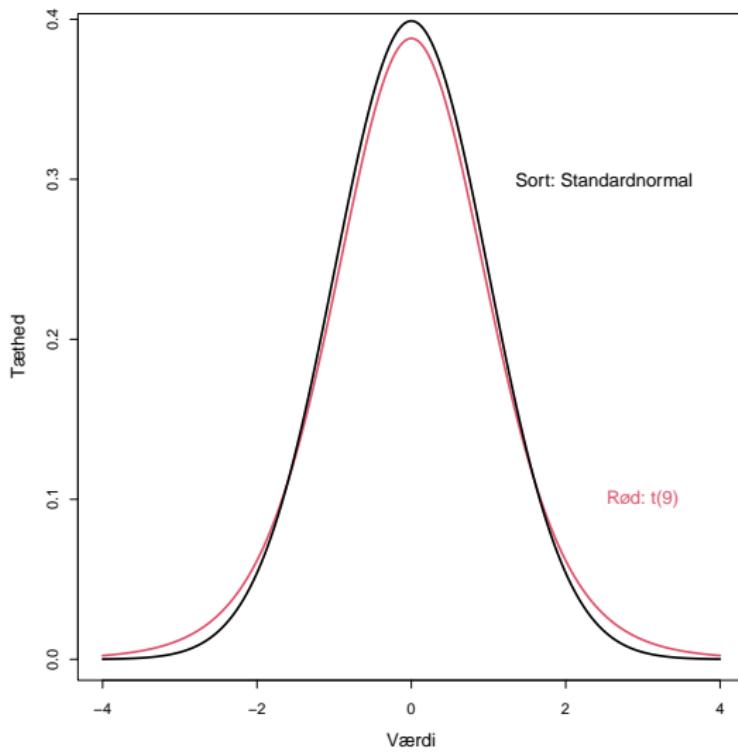
t-fordelingen tager højde for usikkerheden i at bruge stikprøvevariansen:

Antag at X_1, \dots, X_n er uafhængige og ensfordelte (i.i.d.) stokastiske variable, hvor $X_i \sim N(\mu, \sigma^2)$ og $i = 1, \dots, n$, så er:

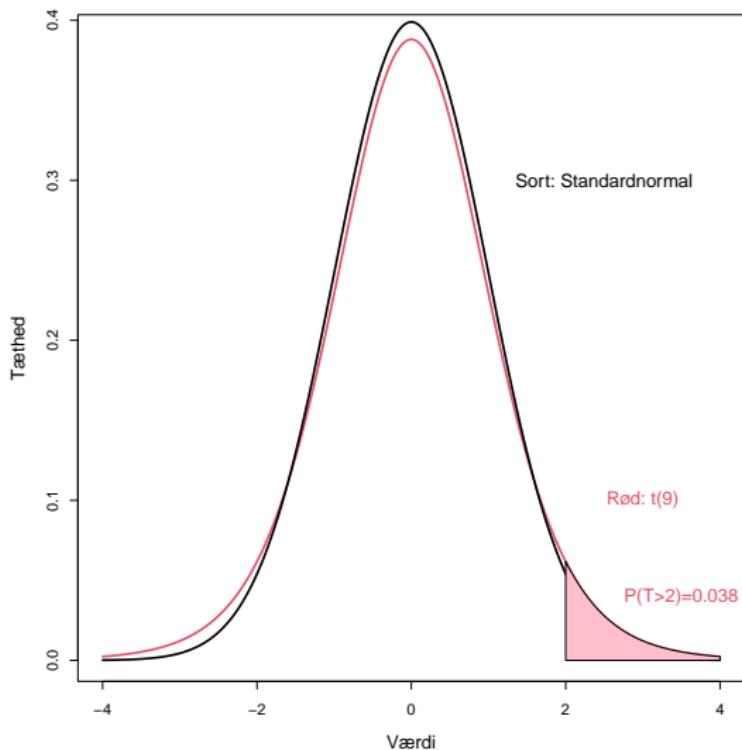
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

dvs. T følger en *t*-fordeling med $n-1$ frihedsgrader (degrees of freedom, df).

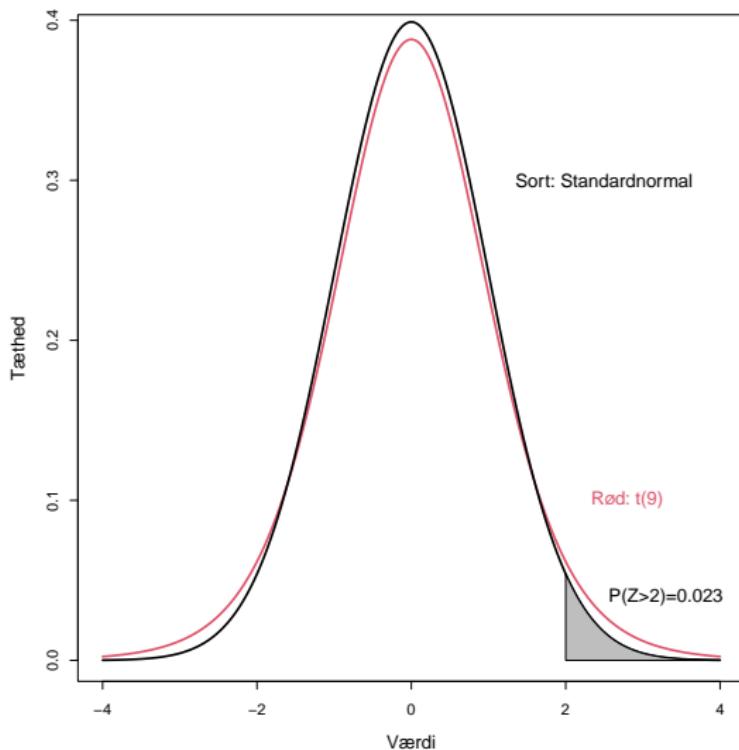
t-fordelingen med 9 frihedsgrader ($n = 10$):



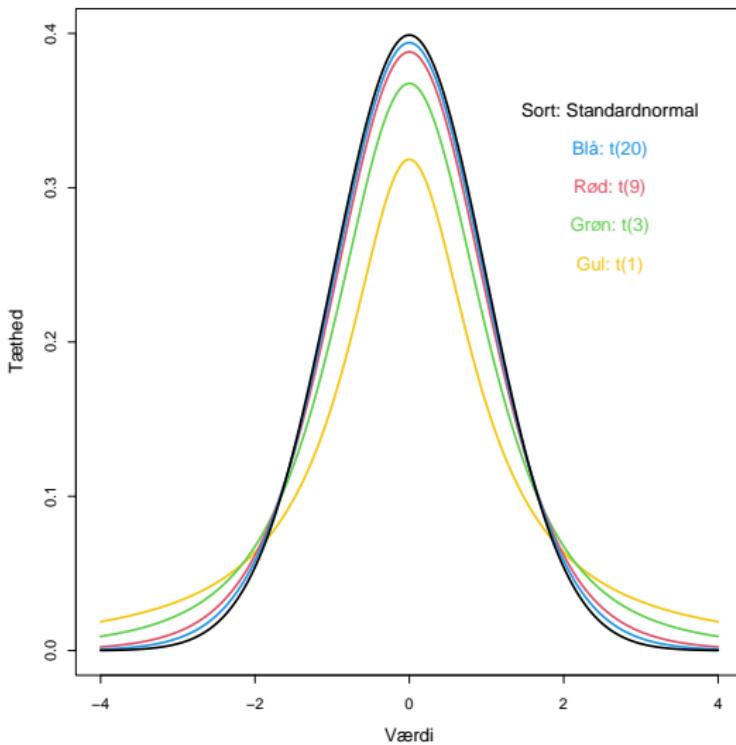
t-fordelingen med 9 frihedsgrader og standardnormalfordelingen:



t-fordelingen med 9 frihedsgrader og standardnormalfordelingen:



Forskellige *t*-fordelinger:



Dagsorden

- 1 Opsummering
- 2 Introduktion og eksempel
- 3 Fordelingen for gennemsnittet
 - t -fordelingen
- 4 Konfidensintervallet for μ
 - Eksempel: Højder
- 5 Statistisk sprogbrug og den formelle ramme
- 6 Ikke-normale data
- 7 Formel fortolkning af konfidensintervallet
- 8 Konfidensinterval for varians og spredning

Metodeboks 3.9: Konfidensinterval for μ

Brug den rigtige t -fordeling til at lave konfidensintervallet:

For en stikprøve x_1, \dots, x_n er $100(1 - \alpha)\%$ konfidensintervallet for μ givet ved:

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

hvor $t_{1-\alpha/2}$ er $100(1 - \alpha/2)\%$ fraktilen i t -fordelingen med $n - 1$ frihedsgrader.

Metodeboks 3.9: Konfidensinterval for μ

Brug den rigtige t -fordeling til at lave konfidensintervallet:

For en stikprøve x_1, \dots, x_n er $100(1 - \alpha)\%$ konfidensintervallet for μ givet ved:

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

hvor $t_{1-\alpha/2}$ er $100(1 - \alpha/2)\%$ fraktilen i t -fordelingen med $n - 1$ frihedsgrader.

Mest almindeligt med $\alpha = 0.05$:

Oftest bruger man 95%-konfidensintervallet:

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

Her kaldes $(1 - \alpha)$ for konfidensniveauet og α for signifikansniveauet.

Højde-eksempel

```
## 0.975-fraktilen i t(9)-fordelingen (n=10):  
qt(0.975, 9)
```

[1] 2.262

Dette giver os, at $t_{0.975} = 2.26$.

Resultatet fra metodeboks 3.9:

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

som udregnes til

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

Højde-eksempel, 99% Konfidensintervallet (CI)

```
qt(0.995, 9)
```

[1] 3.25

Dette giver resultatet $t_{0.995} = 3.25$.

I dette tilfælde fås

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

som giver

$$178 \pm 12.55 = [165.5; 190.5].$$

En R-funktion, der kan gøre det hele (og mere til):

```
x <- c(168,161,167,179,184,166,198,187,191,179)
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:
## 165.5 190.5
## sample estimates:
## mean of x
##      178
```

Dagsorden

- 1 Opsummering
- 2 Introduktion og eksempel
- 3 Fordelingen for gennemsnittet
 - t -fordelingen
- 4 Konfidensintervallet for μ
 - Eksempel: Højder
- 5 Statistisk sprogbrug og den formelle ramme
- 6 Ikke-normale data
- 7 Formel fortolkning af konfidensintervallet
- 8 Konfidensinterval for varians og spredning

Den formelle ramme for *statistisk inferens*

Fra kapitel 1 i boget:

- 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**)

Den formelle ramme for *statistisk inferens*

Fra kapitel 1 i boget:

- 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:

- μ og σ er *parametre*, som beskriver populationen
- \bar{x} er *estimatet* for μ (konkret udfaldsværdi)
- \bar{X} er *estimatoren* for μ (nu set som stokastisk variabel)
- Begrebet *teststørrelse (statistic)* er en fællesbetegnelse for begge

Den formelle ramme for *statistisk inferens* - Eksempel

Fra kapitel 1 i boget: Modificeret højde-eksempel

Vi måler højden for 10 tilfældige personer i Danmark.

Den formelle ramme for *statistisk inferens* - Eksempel

Fra kapitel 1 i boget: Modificeret højde-eksempel

Vi måler højden for 10 tilfældige personer i Danmark.

Stikprøven:

De 10 observationer: x_1, \dots, x_{10} .

Den formelle ramme for *statistisk inferens* - Eksempel

Fra kapitel 1 i boget: Modificeret højde-eksempel

Vi måler højden for 10 tilfældige personer i Danmark.

Stikprøven:

De 10 observationer: x_1, \dots, x_{10} .

Populationen:

Højderne for alle mennesker i Danmark.

Den formelle ramme for *statistisk inferens* - Eksempel

Fra kapitel 1 i boget: Modificeret højde-eksempel

Vi måler højden for 10 tilfældige personer i Danmark.

Stikprøven:

De 10 observationer: x_1, \dots, x_{10} .

Populationen:

Højderne for alle mennesker i Danmark.

Observationsenheden:

Én person.

Statistisk inferens: Læring fra data

Læring fra data:

Man ønsker at udlede parameterværdierne for den underliggende population.

Statistisk inferens: Læring fra data

Læring fra data:

Man ønsker at udlede parameterværdierne for den underliggende population.

Vigtigt i den forbindelse:

Stikprøven skal på meningsfuld vis være *repræsentativ* for en veldefineret population.

Statistisk inferens: Læring fra data

Læring fra data:

Man ønsker at udlede parameterværdierne for den underliggende population.

Vigtigt i den forbindelse:

Stikprøven skal på meningsfuld vis være *repræsentativ* for en veldefineret population.

Hvordan sikrer man det?

F.eks. ved at sikre, at stikprøven er fuldstændig *tilfældigt udtaget*.

Tilfældig stikprøveudtagning (random sampling)

Definition 3.12 :

- En tilfældig stikprøve fra en (uendelig) population: De stokastiske variable X_1, X_2, \dots, X_n udgør en tilfældig stikprøve af størrelse n fra den uendelige population, hvis:
 - ① Alle de stokastiske variable har samme fordeling
 - ② De n stokastiske variable er uafhængige

Tilfældig stikprøveudtagning (random sampling)

Definition 3.12 :

- En tilfældig stikprøve fra en (uendelig) population: De stokastiske variable X_1, X_2, \dots, X_n udgør en tilfældig stikprøve af størrelse n fra den uendelige population, hvis:
 - ① Alle de stokastiske variable har samme fordeling
 - ② De n stokastiske variable er uafhængige

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)

Dagsorden

- 1 Opsummering
- 2 Introduktion og eksempel
- 3 Fordelingen for gennemsnittet
 - t -fordelingen
- 4 Konfidensintervallet for μ
 - Eksempel: Højder
- 5 Statistisk sprogbrug og den formelle ramme
- 6 **Ikke-normale data**
- 7 Formel fortolkning af konfidensintervallet
- 8 Konfidensinterval for varians og spredning

Sætning 3.14: Den centrale grænseværdidisætning (CLT)

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

Lad \bar{X} være gennemsnittet for en tilfældigt udtrukket stikprøve af størrelse n taget fra en population med middelværdi μ og varians σ^2 . Så gælder det, at fordelingen for

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

tilnærmer sig standardnormalfordelingen, $N(0, 1^2)$, når $n \rightarrow \infty$.

Sætning 3.14: Den centrale grænseværdidisætning (CLT)

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

Lad \bar{X} være gennemsnittet for en tilfældigt udtrukket stikprøve af størrelse n taget fra en population med middelværdi μ og varians σ^2 . Så gælder det, at fordelingen for

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

tilnærmer sig standardnormalfordelingen, $N(0, 1^2)$, når $n \rightarrow \infty$.

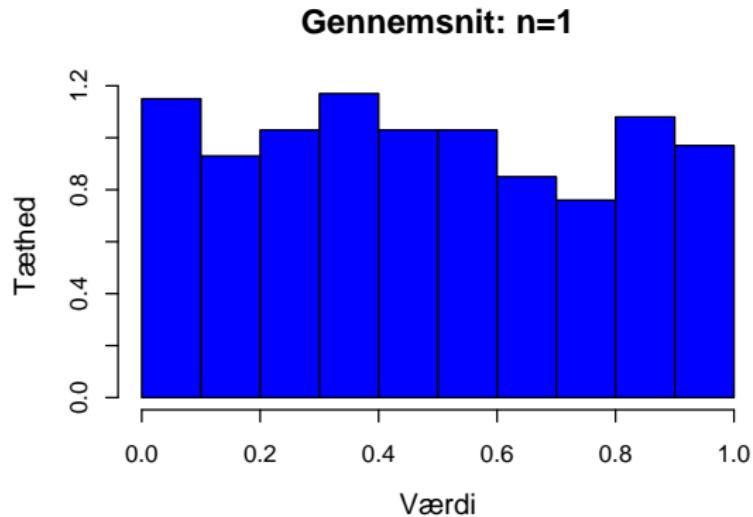
Dvs., hvis n er stor nok, kan vi (tilnærmelsesvist) antage:

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

Engelsk: *Central Limit Theorem* (CLT)

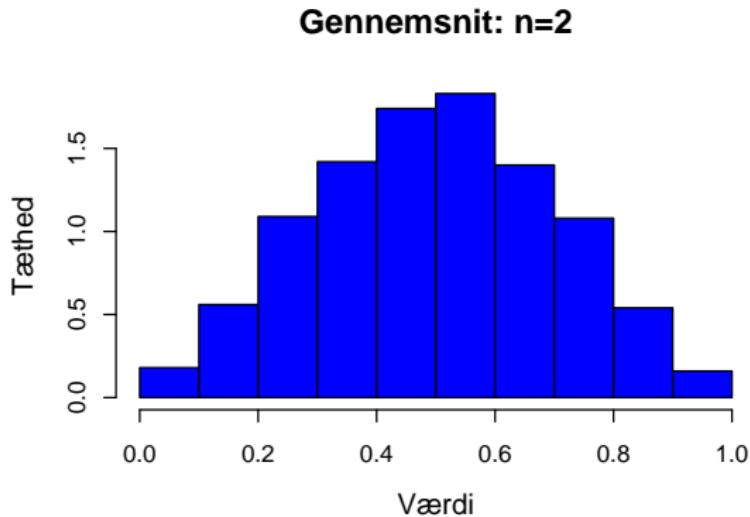
CLT in action - gennemsnit af uniformt fordelte variable

```
n=1  
k=1000  
u=matrix(runif(k*n),ncol=n)  
hist(apply(u,1,mean), col="blue", main="Gennemsnit: n=1", xlab="Værdi",ylab="Tæthed",freq=FALSE)
```



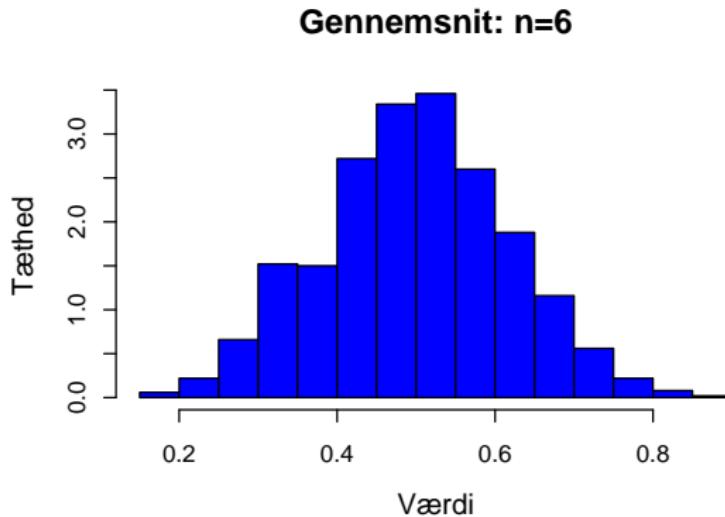
CLT in action - gennemsnit af uniformt fordelte variable

```
n=2  
k=1000  
u=matrix(runif(k*n),ncol=n)  
hist(apply(u,1,mean), col="blue", main="Gennemsnit: n=2", xlab="Værdi",ylab="Tæthed",freq=FALSE)
```



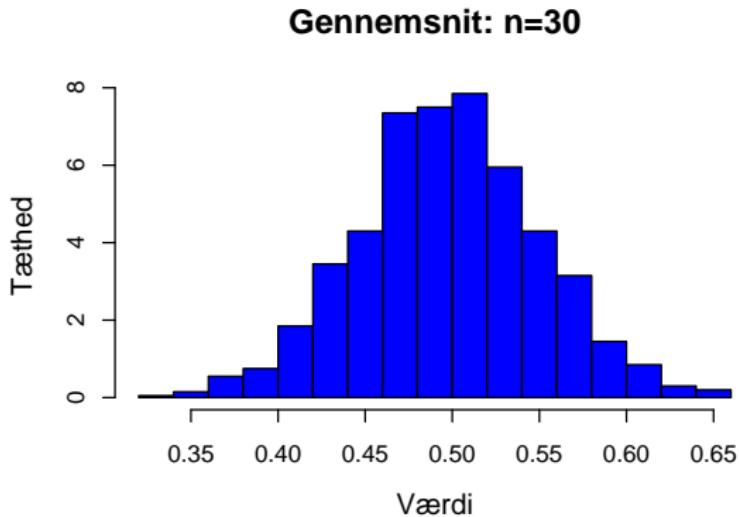
CLT in action - gennemsnit af uniformt fordelte variable

```
n=6  
k=1000  
u=matrix(runif(k*n),ncol=n)  
hist(apply(u,1,mean), col="blue", main="Gennemsnit: n=6", xlab="Værdi",ylab="Tæthed",freq=FALSE)
```



CLT in action - gennemsnit af uniformt fordelte variable

```
n=30  
k=1000  
u=matrix(runif(k*n),ncol=n)  
hist(apply(u,1,mean), col="blue", main="Gennemsnit: n=30", xlab="Værdi",ylab="Tæthed",freq=FALSE,nclass=15)
```



Konsekvens af den centrale grænseværdidisætning:

Konfidensintervallet for μ gælder også for ikke-normale data:

Man kan bruge konfidensintervaller baseret på t -fordelingen i stort set alle situationer, blot n er "stør nok".

Konsekvens af den centrale grænseværdidisætning:

Konfidensintervallet for μ gælder også for ikke-normale data:

Man kan bruge konfidensintervaller baseret på t -fordelingen i stort set alle situationer, blot n er "stør nok".

Hvornår er n "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.

Dagsorden

- 1 Opsummering
- 2 Introduktion og eksempel
- 3 Fordelingen for gennemsnittet
 - t -fordelingen
- 4 Konfidensintervallet for μ
 - Eksempel: Højder
- 5 Statistisk sprogbrug og den formelle ramme
- 6 Ikke-normale data
- 7 Formel fortolkning af konfidensintervallet**
- 8 Konfidensinterval for varians og spredning

'Repeated sampling' fortolkning

I det lange løb fanger vi den sande værdi i 95% af tilfældene:

Konfidensintervallet vil variere i både bredde (s) og position (\bar{x}), hvis man gentager sit studie.

'Repeated sampling' fortolkning

I det lange løb fanger vi den sande værdi i 95% af tilfældene:

Konfidensintervallet vil variere i både bredde (s) og position (\bar{x}), hvis man gentager sit studie.

Mere formelt udtrykt:

$$P\left(\frac{|\bar{X} - \mu|}{S/\sqrt{n}} < t_{0.975}\right) = 0.95,$$

'Repeated sampling' fortolkning

I det lange løb fanger vi den sande værdi i 95% af tilfældene:

Konfidensintervallet vil variere i både bredde (s) og position (\bar{x}), hvis man gentager sit studie.

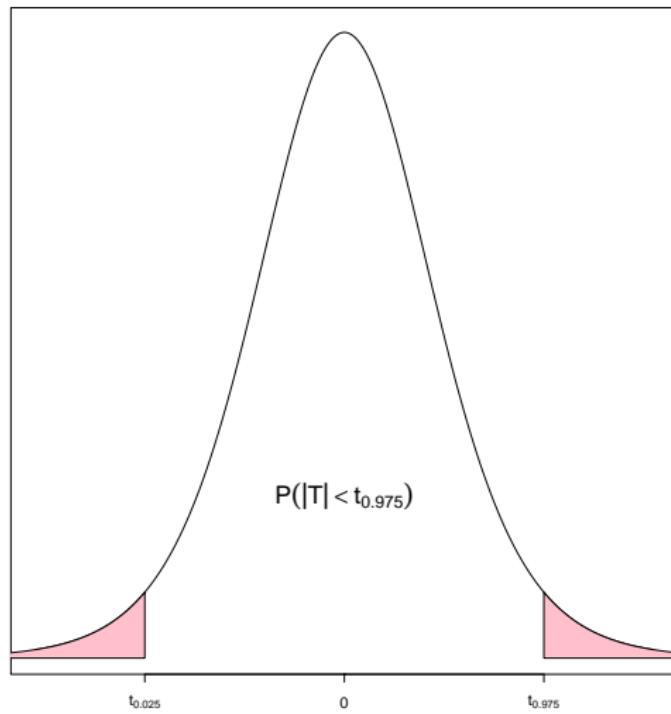
Mere formelt udtrykt:

$$P\left(\frac{|\bar{X} - \mu|}{S/\sqrt{n}} < t_{0.975}\right) = 0.95,$$

som er ækvivalent med:

$$P\left(\bar{X} - t_{0.975} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{0.975} \frac{S}{\sqrt{n}}\right) = 0.95.$$

'Repeated sampling' fortolkning



Dagsorden

- 1 Opsummering
- 2 Introduktion og eksempel
- 3 Fordelingen for gennemsnittet
 - t -fordelingen
- 4 Konfidensintervallet for μ
 - Eksempel: Højder
- 5 Statistisk sprogbrug og den formelle ramme
- 6 Ikke-normale data
- 7 Formel fortolkning af konfidensintervallet
- 8 Konfidensinterval for varians og spredning

Motiverende eksempel

Produktion af tabletter:

I produktionen af tabletter blandes et aktivt stof med et pulver, hvorefter blandingen formes til tabletter. Vi producerer altså pulverblanding og deraf pillerne. Det er vigtigt, at blandingen er så homogen (ensartet) som mulig, således at tabletternes styrke er ens.

Vi betragter en blanding af det aktive stof og fyldpulver, hvoraf vi vil producere en stor mængde tabletter.

Vi ønsker, at koncentrationen af det aktive stof i tabletterne skal være 1 mg/g med den mindst mulige spredning. En tilfældig stikprøve udtages, hvor vi mäter koncentrationen af det aktive stof (i mg/g). Vi antager endvidere, at vores målinger følger en normalfordeling.

Fordelingen for stikprøvevariansen, sætning 2.81

Stikprøven defineres som (X_1, \dots, X_n) , hvor X_i (for $i = 1, \dots, n$) repræsenterer den i'te måling af koncentrationen, som her antages at følge en normal(μ, σ^2)fordeling. Vi antager yderligere, at stikprøven er repræsentativ (variablene er uafhængige og ensfordelte).

Stikprøvevariansen (varianseestimatet) følger en χ^2 -fordeling:

Lad

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

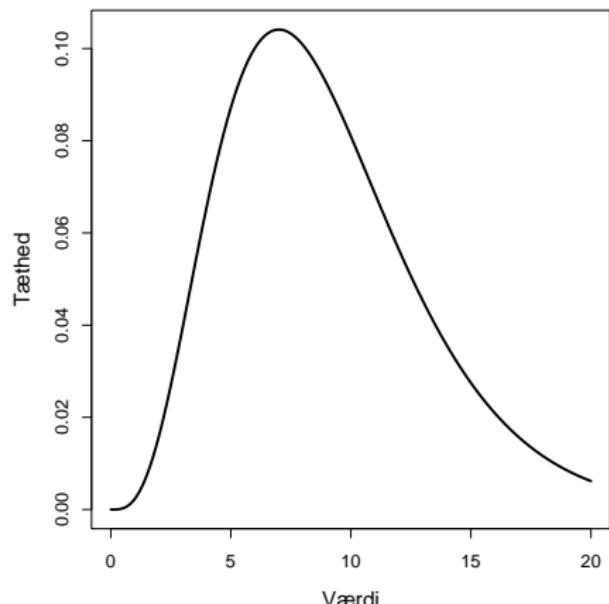
Så gælder at:

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

følger en χ^2 -fordelt med $v = n - 1$ frihedsgrader.

χ^2 -fordelingen med $v = 9$ frihedsgrader (degrees of freedom)

```
x <- seq(0, 20, by = 0.1)
plot(x, dchisq(x, df = 9), type = "l", ylab="Tæthed", xlab="Værdi", lwd=2)
```



Metode 3.19: Konfidensintervaller for variansen og spredningen

Lad $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \dots, n$ være uafhængige (og ensfordelte).

Variansen:

Et $100(1 - \alpha)\%$ konfidensinterval for variansen σ^2 er givet ved:

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

hvor fraktilerne kommer fra en χ^2 -fordeling med $v = n - 1$ frihedsgrader.

Metode 3.19: Konfidensintervaller for variansen og spredningen

Lad $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \dots, n$ være uafhængige (og ensfordelte).

Variansen:

Et $100(1 - \alpha)\%$ konfidensinterval for variansen σ^2 er givet ved:

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

hvor fraktilerne kommer fra en χ^2 -fordeling med $v = n - 1$ frihedsgrader.

Standardafvigelsen:

Et $100(1 - \alpha)\%$ konfidensinterval for standardafvigelsen σ er:

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

Eksempel

Data:

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

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

Eksempel

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 χ^2 -fraktilerne (19 frihedsgrader):

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

```
qchisq(c(0.025, 0.975), df = 19)
```

```
[1] 8.907 32.852
```

Eksempel

Så konfidensintervallet for variansen σ^2 bliver:

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

Eksempel

Så konfidensintervallet for variansen σ^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 spredningen σ bliver:

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

Højdeeksempel

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

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

```
qchisq(c(0.025, 0.975), df = 9)
```

```
[1] 2.70 19.02
```

Højdeeksempel

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

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

```
qchisq(c(0.025, 0.975), df = 9)
```

```
[1] 2.70 19.02
```

Så konfidensintervallet for højdens standardafvigelse σ 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 - Resultater:

Stikprøve, $n = 10$:

168 161 167 179 184 166 198 187 191 179

Gennemsnit og standardafvigelse for stikprøven:

$$\bar{x} = 178$$

$$s = 12.21$$

NYT: Konfidensinterval for μ :

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

Estimater for populationsgennemsnit og standardafvigelse:

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

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

NYT: Konfidensinterval for σ :

$$[8.4; 22.3]$$

Dagsorden

- 1 Opsummering
- 2 Introduktion og eksempel
- 3 Fordelingen for gennemsnittet
 - t -fordelingen
- 4 Konfidensintervallet for μ
 - Eksempel: Højder
- 5 Statistisk sprogbrug og den formelle ramme
- 6 Ikke-normale data
- 7 Formel fortolkning af konfidensintervallet
- 8 Konfidensinterval for varians og spredning

02402 Statistik (Polyteknisk grundlag)

Uge 5: Hypotesetest

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

Dagsorden

- 1 Opsummering
- 2 Motiverende eksempel – sovemedicin
- 3 t -test med en stikprøve
- 4 Kritiske værdier og konfidensintervaller
- 5 Hypotesetest – generelt
 - Den alternative hypotese (Modhypotesen)
 - Den generelle metode
 - Fejlslutninger ved hypotesetest!
- 6 Modelkontrol: Normalfordelingsantagelsen
 - Q-Q plot for normalfordelingen
 - Transformation mod normalitet

Dagsorden

- 1 Opsummering
- 2 Motiverende eksempel – sovemedicin
- 3 t -test med en stikprøve
- 4 Kritiske værdier og konfidensintervaller
- 5 Hypotesetest – generelt
 - Den alternative hypotese (Modhypotesen)
 - Den generelle metode
 - Fejlslutninger ved hypotesetest!
- 6 Modelkontrol: Normalfordelingsantagelsen
 - Q-Q plot for normalfordelingen
 - Transformation mod normalitet

De forrige uger

Vi vil undersøge en population ved at udføre et eksperiment og udtagte en repræsentativ stikprøve.

Vi definerer en stokastisk variabel $X : S \rightarrow \mathbb{R}$, som afbilder eksperimentets udfald til numeriske værdier. Den stokastiske variabel repræsenterer eksperimentets udfaldsværdi før det udføres.

Vi kan så formulere en statistisk model ved at tilknytte den stokastiske variabel en sandsynlighedsfordeling.

Vi vil så betragte en stikprøvefunktion $g : \mathbb{R}^n \rightarrow \mathbb{R}$. Det kunne eksempelvis være stikprøvegennemsnittet eller stikprøvevariansen.

- Hvis funktionen g bruges til at estimere en ukendt populationsparameter θ , kaldes $g(X_1, \dots, X_n)$ en estimator for θ og $g(x_1, \dots, x_n)$ et estimat for θ .
- Hvis funktionen g bruges til hypotesetest, kaldes $g(X_1, \dots, X_n)$ en teststørrelse og $g(x_1, \dots, x_n)$ den observerede teststørrelse.

Sidste uge

Sidste uge omhandlede en type intervalestimatorer kaldet konfidensintervaller.

Lad $X = (X_1, \dots, X_n)$ være en stikprøve med ensfordelte, uafhængige variable. Et γ -konfidensinterval for populationsparameteren θ er givet ved $[u(X), v(X)]$ sådan, at

$$\mathbb{P}(u(X) \leq \theta \leq v(X)) = \gamma.$$

Her er u og v stikprøvefunktioner, der afhænger af fordelingen for de stokastiske variable.

Sidste uge

Konfidensintervallerne for gennemsnittet/middelværdien (μ) var baseret på hovedresultaterne nedenfor:

Lad (X_1, \dots, X_n) være en stikprøve med ensfordelte, uafhængige variable.

- Hvis $X_i \sim N(\mu, \sigma^2)$, hvor variansen er kendt:

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

- Hvis $X_i \sim N(\mu, \sigma^2)$, hvor variansen er ukendt og estimeres med S^2 :

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim t(n-1).$$

Sidste uge

Den centrale grænseværdidisætning

Lad X følge en vilkårlig fordeling med $\mathbb{E}[X_i] = \mu$ og $\mathbb{V}[X_i] = \sigma^2$. Hvis n er stor ($n \geq 30$), så vil både Z og T defineret som

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \quad \text{og} \quad T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}}$$

følge en standardnormalfordeling.

Siden t -fordelingen konvergerer til en standardnormalfordeling for $n \rightarrow \infty$, kan man benytte begge fordelinger til at konstruere konfidensintervaller.

I dette kursus baseres konfidensintervaller for middelværdien altid på en t -fordeling, jf. metode 3.9.

Sidste uge

Et $(1 - \alpha)$ -konfidensinterval for μ fås ved følgende beregninger:

Da $T \sim t(n - 1)$ må $\mathbb{P}(t_{\alpha/2} \leq T \leq t_{1-\alpha/2}) = 1 - \alpha$, hvor t_p er p -fraktilen i en t -fordeling med $n - 1$ frihedsgrader. Det gælder endvidere, at

$$\begin{aligned}\mathbb{P}(t_{\alpha/2} \leq T \leq t_{1-\alpha/2}) &= \mathbb{P}\left(t_{\alpha/2} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{1-\alpha/2}\right) \\ &= \mathbb{P}\left(t_{\alpha/2} \frac{S}{\sqrt{n}} \leq \bar{X} - \mu \leq t_{1-\alpha/2} \frac{S}{\sqrt{n}}\right) \\ &= \mathbb{P}\left(\bar{X} - t_{\alpha/2} \frac{S}{\sqrt{n}} \geq \mu \geq \bar{X} - t_{1-\alpha/2} \frac{S}{\sqrt{n}}\right).\end{aligned}$$

Da t -fordelingen er symmetrisk omkring nul gælder, at $t_p = -t_{1-p}$. Derfor kan dette omskrives til:

$$1 - \alpha = \mathbb{P}\left(\bar{X} + t_{1-\alpha/2} \frac{S}{\sqrt{n}} \geq \mu \geq \bar{X} - t_{1-\alpha/2} \frac{S}{\sqrt{n}}\right).$$

Derfor bliver $(1 - \alpha)$ -konfidensintervallet for μ : $\left[\bar{X} - t_{1-\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2} \frac{S}{\sqrt{n}}\right]$.

Sidste uge

Et $(1 - \alpha)$ -konfidensinterval for σ^2 , når $X_i \sim N(\mu, \sigma^2)$, fås ved følgende beregninger:

Lad

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

Så vil Y følge en χ^2 -fordeling med $n-1$ frihedsgrader. Derfor gælder, at

$$\mathbb{P}\left(\chi_{\alpha/2}^2 \leq Y \leq \chi_{1-\alpha/2}^2\right) = 1 - \alpha,$$

hvor χ_p^2 er p -fraktilen i en χ^2 -fordeling med $n-1$ frihedsgrader. Bemærk så, at

$$\begin{aligned} \mathbb{P}\left(\chi_{\alpha/2}^2 \leq Y \leq \chi_{1-\alpha/2}^2\right) &= \mathbb{P}\left(\chi_{\alpha/2}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{1-\alpha/2}^2\right) \\ &= \mathbb{P}\left(\frac{1}{\chi_{\alpha/2}^2} \geq \frac{\sigma^2}{(n-1)S^2} \geq \frac{1}{\chi_{1-\alpha/2}^2}\right) \\ &= \mathbb{P}\left(\frac{(n-1)S^2}{\chi_{\alpha/2}^2} \geq \sigma^2 \geq \frac{(n-1)S^2}{\chi_{1-\alpha/2}^2}\right) \end{aligned}$$

Generelt for i dag

For at undersøge en hypotese gør vi følgende:

- ① Vi definerer og afgrænser en population
- ② Vi formulerer en statistisk model
- ③ Vi opstiller et eksperiment og udtager en repræsentativ stikprøve
- ④ Vi beregner en eller flere teststørrelser
- ⑤ Vi sammenholder den teoretiske model med de foreliggende observationer

Vi kan så udtale os om hypotesen i forhold til vores antagelser og den foreliggende data.

Vi arbejder efter den videnskabelige metode.

Dagsorden

- 1 Opsummering
- 2 Motiverende eksempel – sovemedicin
- 3 t -test med en stikprøve
- 4 Kritiske værdier og konfidensintervaller
- 5 Hypotesetest – generelt
 - Den alternative hypotese (Modhypotesen)
 - Den generelle metode
 - Fejlslutninger ved hypotesetest!
- 6 Modelkontrol: Normalfordelingsantagelsen
 - Q-Q plot for normalfordelingen
 - Transformation mod normalitet

Eksempel – sovemedicin

Forskel på sovemedicin

I et studie er man interesseret i at sammenligne 2 sovemidler, 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 med $n = 10$:

Person	Forskel
1	1.2
2	2.4
3	1.3
4	1.3
5	0.9
6	1.0
7	1.8
8	0.8
9	4.6
10	1.4

Eksempel – sovemedicin

Forskel på sovemedicin

I et studie er man interesseret i at sammenligne 2 sovemidler, 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 med $n = 10$:

Person	Forskel
1	1.2
2	2.4
3	1.3
4	1.3
5	0.9
6	1.0
7	1.8
8	0.8
9	4.6
10	1.4

$$\bar{x} = 1.67 \text{ (gennemsnit)}$$
$$s = 1.13 \text{ (standardafvigelse)}$$

Eksempel – sovemedicin

Hypotesen om ingen forskel:

$$H_0 : \mu = 0$$

hvor μ er den gennemsnitlige forskel i søvnslængde ("effekten").

Eksempel – sovemedicin

Hypotesen om ingen forskel:

$$H_0 : \mu = 0$$

hvor μ er den gennemsnitlige forskel i søvnslængde ("effekten").

Stikprøvegennemsnit og
-standardafvigelse:

$$\bar{x} = 1.670 = \hat{\mu}$$

$$s = 1.13 = \hat{\sigma}$$

Eksempel – sovemedicin

Hypotesen om ingen forskel:

$$H_0 : \mu = 0$$

hvor μ er den gennemsnitlige forskel i søvnslængde ("effekten").

Stikprøvegennemsnit og -standardafvigelse:

$$\bar{x} = 1.670 = \hat{\mu}$$

$$s = 1.13 = \hat{\sigma}$$

NYT: ***p*-værdi**

$$p = 0.00117$$

(Udregnet under antagelsen, at H_0 er sand).

Er data i overensstemmelse med nulhypotesen H_0 ?

Data: $\bar{x} = 1.67$, $H_0 : \mu = 0$

NYT: **Konklusion**

Vi **forkaster** H_0 og konkluderer, at der er en **signifikant** forskel på effekten af middel B sammenlignet med middel A.

Dagsorden

- 1 Opsummering
- 2 Motiverende eksempel – sovemedicin
- 3 ***t*-test med en stikprøve**
- 4 Kritiske værdier og konfidensintervaller
- 5 Hypotesetest – generelt
 - Den alternative hypotese (Modhypotesen)
 - Den generelle metode
 - Fejlslutninger ved hypotesetest!
- 6 Modelkontrol: Normalfordelingsantagelsen
 - Q-Q plot for normalfordelingen
 - Transformation mod normalitet

Metode 3.23: Test med en stikprøve

En *t*-test med en stikprøve undersøger om populationsmiddelværdien afviger signifikant fra værdien μ_0 :

For en (kvantitativ) situation med **én stikprøve**, er *p*-værdien givet ved:

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

hvor T følger en *t*-fordeling med $(n - 1)$ frihedsgrader.

Den observerede værdi af teststørrelsen, som skal udregnes, er

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

hvor μ_0 er værdien af μ under nulhypotesen:

$$H_0 : \mu = \mu_0.$$

Definition og fortolkning af p-værdien (generelt)

p-værdien udtrykker evidens mod nulhypotesen – Tabel 3.1:

$p < 0.001$	Meget stærk evidens imod H_0
$0.001 \leq p < 0.01$	Stærk evidens imod H_0
$0.01 \leq p < 0.05$	Nogen evidens imod H_0
$0.05 \leq p < 0.1$	Svag evidens imod H_0
$p \geq 0.1$	Meget svag eller ingen evidens imod H_0

Definition 3.22 af p-værdien:

p-værdien er sandsynligheden for at observere en teststørrelse som er **mindst lige så ekstrem** som den observerede testværdi. Denne sandsynlighed udregnes under antagelse om, at nulhypotesen er sand.

Eksempel – sovemedicin

Hypotesen om ingen forskel:

$$H_0 : \mu = 0$$

hvor μ er den gennemsnitlige forskel i søvnslængde.

Eksempel – sovemedicin

Hypotesen om ingen forskel:

$$H_0 : \mu = 0$$

hvor μ er den gennemsnitlige forskel i søvnslængde.

Udregn testværdien:

$$t_{\text{obs}} = \frac{1.67 - 0}{1.13/\sqrt{10}} = 4.67$$

Eksempel – sovemedicin

Hypotesen om ingen forskel:

$$H_0: \mu = 0$$

hvor μ er den gennemsnitlige forskel i søvnslængde.

Beregn p -værdien:

Udregn testværdien:

$$t_{\text{obs}} = \frac{1.67 - 0}{1.13/\sqrt{10}} = 4.67$$

$$2P(T > 4.67) = 0.00117$$

```
2 * (1 - pt(4.67, df = 9))
```

Fortolkningen af p -værdien ud fra Tabel 3.1:

Der er stærk evidens imod nulhypotesen.

Eksempel – sovemedicin: Manuelt i R

```
# Indlæs 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) # Stikprøvestørrelsen (Antal observationer)

# Udregn 'tobs' - den observerede teststørrelse/testværdi
tobs <- (mean(x) - 0) / (sd(x) / sqrt(n))

# Udregn p-værdien
# (med den relevante t-fordeling):
2 * (1 - pt(abs(tobs), df = n-1))

## [1] 0.001166
```

Eksempel – sovemedicin: Automatisk i R

```
t.test(x)

##
##  One Sample t-test
##
## data: x
## t = 4.7, df = 9, p-value = 0.001
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.8613 2.4787
## sample estimates:
## mean of x
##      1.67
```

Definition af hypotesetest og signifikans (generelt)

Definition 3.24: Hypotesetest

Vi siger, at vi *udfører en hypotesetest*, når vi vælger at afvise eller acceptere en nulhypotese ud fra data.

En nulhypotese *afvises* på et α -signifikansniveau, hvis den observerede data giver anledning til en p -værdi mindre end *signifikansniveauet* α , der er valgt på forhånd.

Ellers siges nulhypotesen at være '*accepteret*'. Det er mere korrekt (langt at foretrække) at sige, at nulhypotesen ikke kan afvises.

Definition 3.29: Statistisk signifikans

En effekt siges at være (*statistisk*) *signifikant*, hvis p -værdien er mindre end signifikansniveauet α .

Oftest bruges $\alpha = 0.05$.

Eksempel – sovemedicin

Med $\alpha = 0.05$ kan vi konkludere følgende:

Idet p -værdien er mindre end α , **forkaster** vi nulhypotesen.

Eksempel – sovemedicin

Med $\alpha = 0.05$ kan vi konkludere følgende:

Idet p -værdien er mindre end α , **forkaster** vi nulhypotesen.

Og:

Vi har påvist en **signifikant** forskel på effekten af middel B sammenlignet med middel A. (Og dermed, at B virker bedre end A).

Dagsorden

- 1 Opsummering
- 2 Motiverende eksempel – sovemedicin
- 3 t -test med en stikprøve
- 4 Kritiske værdier og konfidensintervaller
- 5 Hypotesetest – generelt
 - Den alternative hypotese (Modhypotesen)
 - Den generelle metode
 - Fejlslutninger ved hypotesetest!
- 6 Modelkontrol: Normalfordelingsantagelsen
 - Q-Q plot for normalfordelingen
 - Transformation mod normalitet

Kritiske værdier

Man kan også udføre hypotesetest ved brug af kritiske værdier, som er tærskelværdier for observerede teststørrelser.

Definition 3.31 - Kritiske værdier for t -testen:

$(1 - \alpha)$ kritiske værdier for den dobbeltsidet t -test med en stikprøve er $(\alpha/2)$ - og $(1 - \alpha/2)$ -fraktilerne i t -fordelingen med $n - 1$ frihedsgrader:

$$t_{\alpha/2} \text{ og } t_{1-\alpha/2}.$$

Kritiske værdier

Man kan også udføre hypotesetest ved brug af kritiske værdier, som er tærskelværdier for observerede teststørrelser.

Definition 3.31 - Kritiske værdier for t -testen:

$(1 - \alpha)$ kritiske værdier for den dobbeltsidet t -test med en stikprøve er $(\alpha/2)$ - og $(1 - \alpha/2)$ -fraktilerne i t -fordelingen med $n - 1$ frihedsgrader:

$$t_{\alpha/2} \text{ og } t_{1-\alpha/2}.$$

Metode 3.32: t -test med en stikprøve ved brug af kritiske værdier

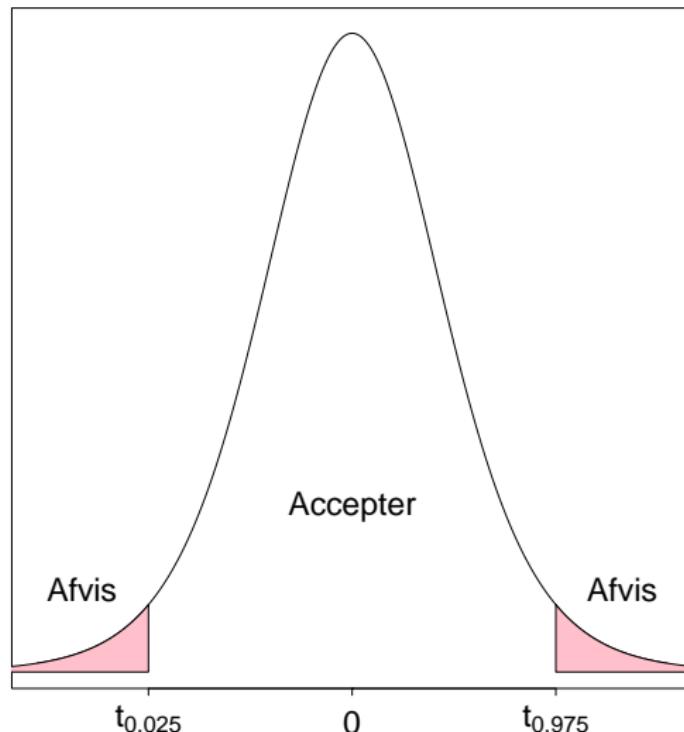
En nulhypotese **afvises** på et α -signifikansniveau, hvis den observerede teststørrelse er mere ekstrem end de kritiske værdier, dvs. hvis

$$t_{\text{obs}} < t_{\alpha/2} \text{ eller } t_{1-\alpha/2} < t_{\text{obs}} \quad (\text{alt. } |t_{\text{obs}}| > t_{1-\alpha/2}).$$

Ellers **accepteres** nulhypotesen (Ellers kan nulhypotesen ikke afvises).

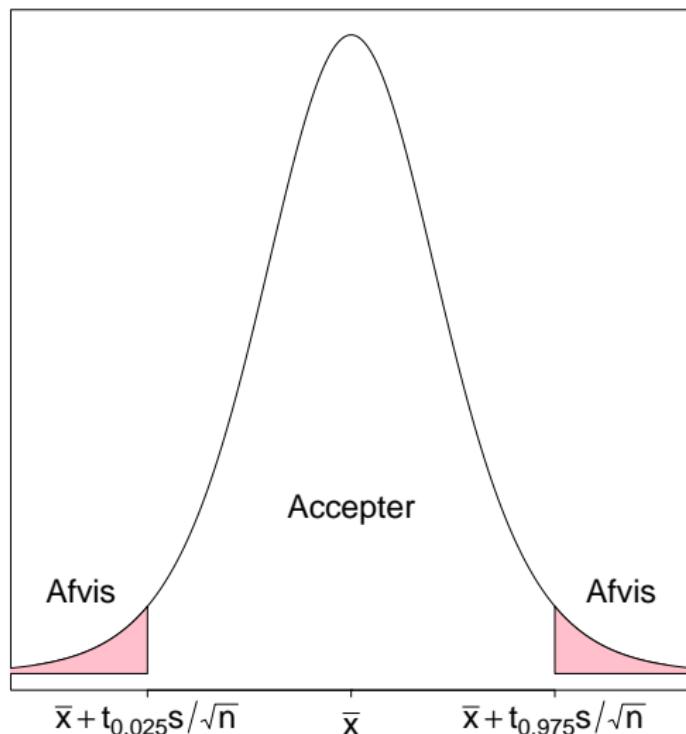
Kritiske værdier og hypotesetest

Acceptområdet består af de værdier af t_{obs} , som ikke er for langt væk fra 0 (standardiseret skala):



Kritiske værdier og hypotesetest

Acceptområdet består af de værdier af μ , som ikke er for langt væk fra stikprøvegennemsnittet (oprindelige skala):



Kritiske værdier, konfidensintervaller og hypotesetest

Man kan også udføre hypotesetest med konfidensintervaller.

Sætning 3.33: Konfidensintervaller i hypotesetest

Vi betragter et $(1 - \alpha)$ -konfidensinterval for μ :

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

Konfidensintervallet svarer til acceptområdet for H_0 , når man tester hypotesen (imod en dobbeltsiden modhypotese)

$$H_0 : \mu = \mu_0.$$

Kritiske værdier, konfidensintervaller og hypotesetest

Man kan også udføre hypotesetest med konfidensintervaller.

Sætning 3.33: Konfidensintervaller i hypotesetest

Vi betragter et $(1 - \alpha)$ -konfidensinterval for μ :

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

Konfidensintervallet svarer til acceptområdet for H_0 , når man tester hypotesen (imod en dobbeltsiden modhypotese)

$$H_0 : \mu = \mu_0.$$

(Ny) fortolkning af konfidensintervallet:

Konfidensintervallet indeholder de værdier, som vil blive accepteret i hypotesestesten på baggrund af den observerede data.

Bevis:

Bemærkning 3.34

Et μ_0 inden for konfidensintervallet opfylder, at

$$\mu_0 \in \left[\bar{x} - t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}} \right] \Leftrightarrow |\bar{x} - \mu_0| < t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}},$$

hvilket er ækvivalent med

$$\frac{|\bar{x} - \mu_0|}{\frac{s}{\sqrt{n}}} < t_{1-\alpha/2}.$$

Dette er ensbetydende med

$$|t_{\text{obs}}| < t_{1-\alpha/2},$$

hvilket netop siger, at μ_0 accepteres, idet t_{obs} er inden for de kritiske værdier.

Dagsorden

- 1 Opsummering
- 2 Motiverende eksempel – sovemedicin
- 3 t -test med en stikprøve
- 4 Kritiske værdier og konfidensintervaller
- 5 Hypotesetest – generelt
 - Den alternative hypotese (Modhypotesen)
 - Den generelle metode
 - Fejlslutninger ved hypotesetest!
- 6 Modelkontrol: Normalfordelingsantagelsen
 - Q-Q plot for normalfordelingen
 - Transformation mod normalitet

Den alternative hypotese (Modhypotesen)

Indtil nu har det været underforstået, at testen er todørs (dobbeltodørs) (non-directional)

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

Den alternative hypotese (Modhypotesen)

Indtil nu har det været underforstået, at testen er todosidet (dobbeltosidet): (non-directional)

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

Der kan være andre situationer, f.eks. ensidet (= directional) modhypoteser:

Alternativet til $H_0 : \mu = \mu_0$ er $H_1 : \mu < \mu_0$.

Den alternative hypotese (Modhypotesen)

Indtil nu har det været underforstået, at testen er todosidet (dobbeltosidet): (non-directional)

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

Der kan være andre situationer, f.eks. ensidet (= directional) modhypoteser:

Alternativet til $H_0 : \mu = \mu_0$ er $H_1 : \mu < \mu_0$.

Vi holder os til den tosidet test (non-directional) i dette kursus!

Trin i en hypotesetest – Et overblik

Helt generelt består en hypotesetest af følgende trin:

- 1 Formulér nulhypotesen (og modhypotesen) og vælg et signifikansniveau α (vælg "risikoniveauet").
- 2 Udregn værdien af teststørrelsen ud fra de observerede data.
- 3 Udregn p-værdien ud fra teststørrelsen holdt op imod den rette fordeling.
- 4 Sammenlign p -værdien med signifikansniveauet α og konkludér.

Alternativt, konkludér ud fra de relevante kritiske værdier eller det relevante konfidensinterval.

Fejlslutninger ved hypotesetests

Der findes to slags fejl (dog kun én af gangen)

Type I: Afvisning af H_0 , når H_0 er sand.

Type II: Ikke afvisning/godkendelse af H_0 , når H_1 er sand.

Risikoen for de to typer fejl kaldes sædvanligvis:

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

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

Type I fejl kaldes en falsk-positiv, medens en type II fejl kaldes en falsk-negativ. Endvidere kaldes sandsynligheden $1 - \beta$ nogle gange teststyrken (power) eller den statistiske følsomhed (sensitivity).

Retssalsanalogi

En person står stillet for en domstol:

En person bliver stillet for en domstol under en specifik anklage.

Nul- og modhypotesen (den alternative hypotese) er:

H_0 : Personen er uskyldig.

H_1 : Personen er skyldig.

Retssalsanalogi

En person står stillet for en domstol:

En person bliver stillet for en domstol under en specifik anklage.

Nul- og modhypotesen (den alternative hypotese) er:

H_0 : Personen er uskyldig.

H_1 : Personen er skyldig.

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

Sagt på en anden måde:

Accept af en nulhypotese er ikke et statistisk bevis for, at nulhypotesen er sand!

Fejlslutninger ved hypotesetest

Sætning 3.39: Signifikansniveauet er risikoen for at begå en Type I fejl

Signifikansniveauet α i hypotesetests er risikoen for en Type I fejl:

$$P(\text{Type I fejl}) = P(\text{Afvisning af } H_0 \text{ når } H_0 \text{ er sand}) = \alpha$$

Mindre $\alpha =$ større β (og omvendt)

Fejlslutninger ved hypotesetest

Sætning 3.39: Signifikansniveauet er risikoen for at begå en Type I fejl

Signifikansniveauet α i hypotesetests er risikoen for en Type I fejl:

$$P(\text{Type I fejl}) = P(\text{Afvisning af } H_0 \text{ når } H_0 \text{ er sand}) = \alpha$$

To mulige sandheder mod to mulige konklusioner:

	Afviser H_0	Afviser ikke H_0
H_0 er sand	Type I fejl (α)	Korrekt accept af H_0
H_0 er falsk	Korrekt afvisning af H_0	Type II fejl (β)

Mindre $\alpha =$ større β (og omvendt)

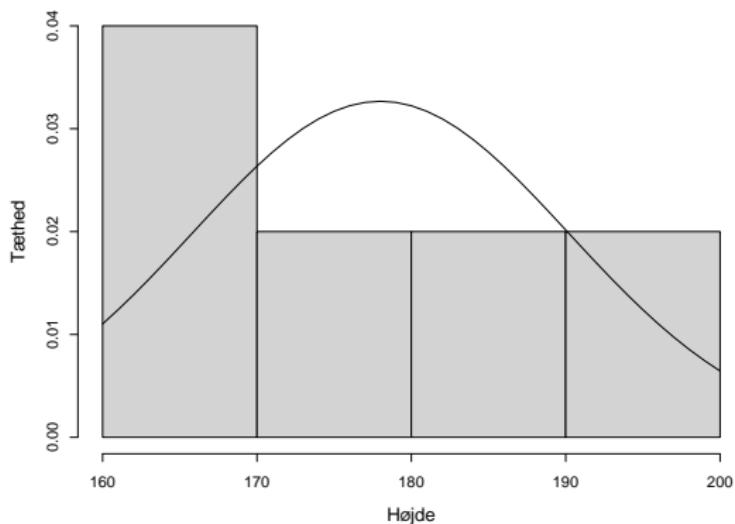
Dagsorden

- 1 Opsummering
- 2 Motiverende eksempel – sovemedicin
- 3 t -test med en stikprøve
- 4 Kritiske værdier og konfidensintervaller
- 5 Hypotesetest – generelt
 - Den alternative hypotese (Modhypotesen)
 - Den generelle metode
 - Fejlslutninger ved hypotesetest!
- 6 Modelkontrol: Normalfordelingsantagelsen
 - Q-Q plot for normalfordelingen
 - Transformation mod normalitet

Eksempel – Højde på studerende

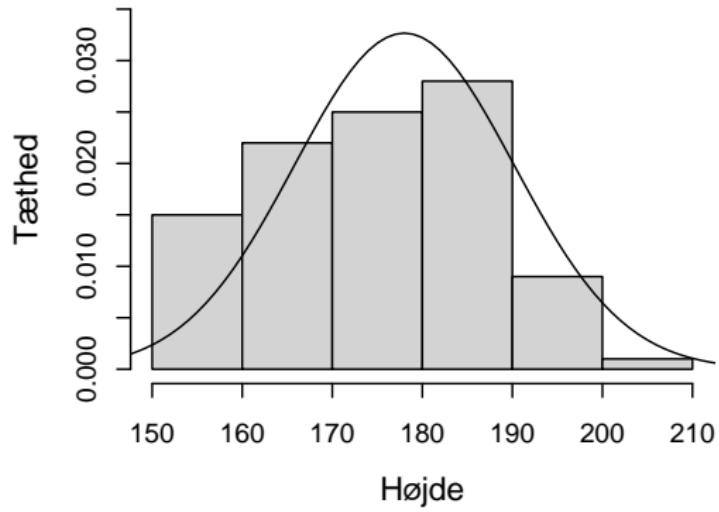
```
# Data - Højde målt i cm
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)

# Histogram af data sammen med en normal tæthedsfunktion
hist(x, xlab = "Højde", main = "", freq = FALSE, ylab="Tæthed")
lines(seq(160, 200, 1), dnorm(seq(160, 200, 1), mean(x), sd(x)))
```



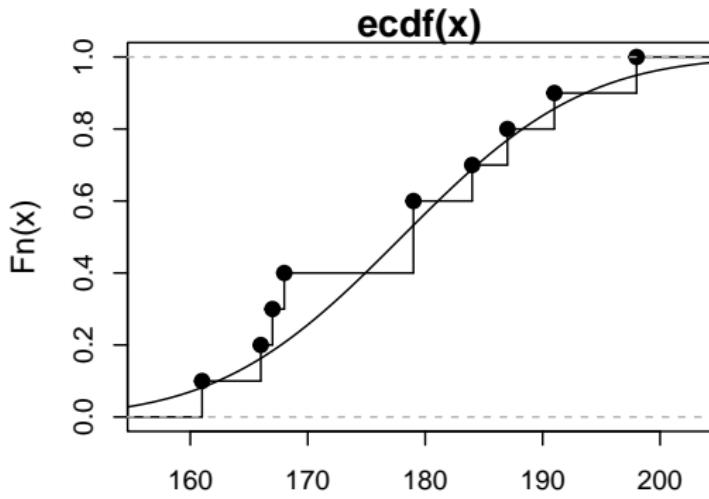
Eksempel – 100 observationer fra en normalfordeling

```
# Histogram over simuleret data fra en normalfordeling  
xr <- rnorm(100, mean(x), sd(x))  
hist(xr,xlab="Højde",main="",freq=F,ylab="Tæthed",ylim=c(0, 0.035))  
lines(seq(130, 230, 1), dnorm(seq(130, 230, 1), mean(x), sd(x)))
```



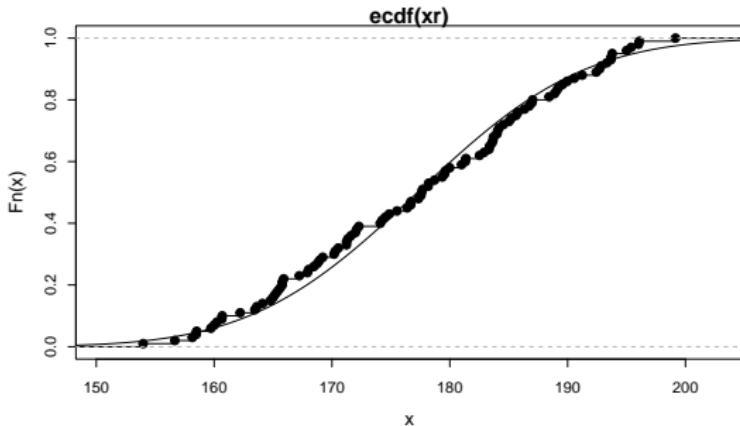
Eksempel – Højde på studerende: ECDF

```
# Empirisk fordelingsfunktion for data  
# sammen med en normal fordelingsfunktion  
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)))
```



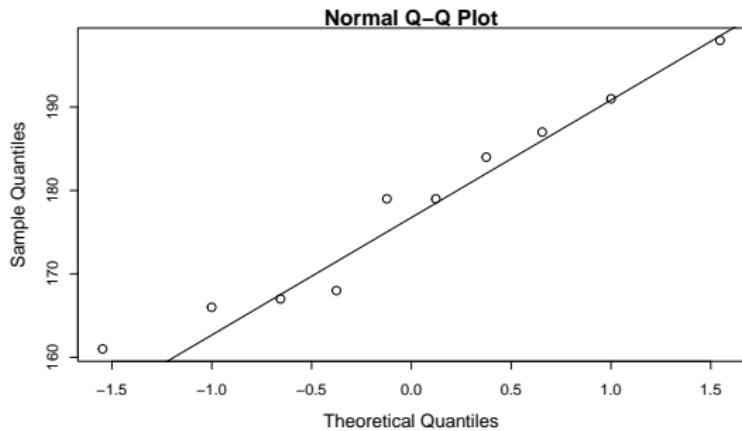
Eksempel – 100 observationer fra en normalford.: ECDF

```
# Empirisk fordelingsfunktion for simuleret normalfordeling
# (n = 100) sammen med en normal fordelingsfunktion
xr <- rnorm(100, mean(x), sd(x))
plot(ecdf(xr), verticals = TRUE)
xp <- seq(0.9*min(xr), 1.1*max(xr), length.out = 100)
lines(xp, pnorm(xp, mean(xr), sd(xr)))
```



Eksempel – Højde på studerende – Q-Q plot

```
# Normal Q-Q plot for de studerendes højder  
qqnorm(x)  
qqline(x)
```



Q-Q plot for normalfordelingen

Metode 3.42 - Formel definition

De sorterede observationer, $x_{(1)}, \dots, x_{(n)}$ plottes mod de teoretiske fraktiler i normalfordelingen. Der findes forskellige definitioner af fraktilerne:

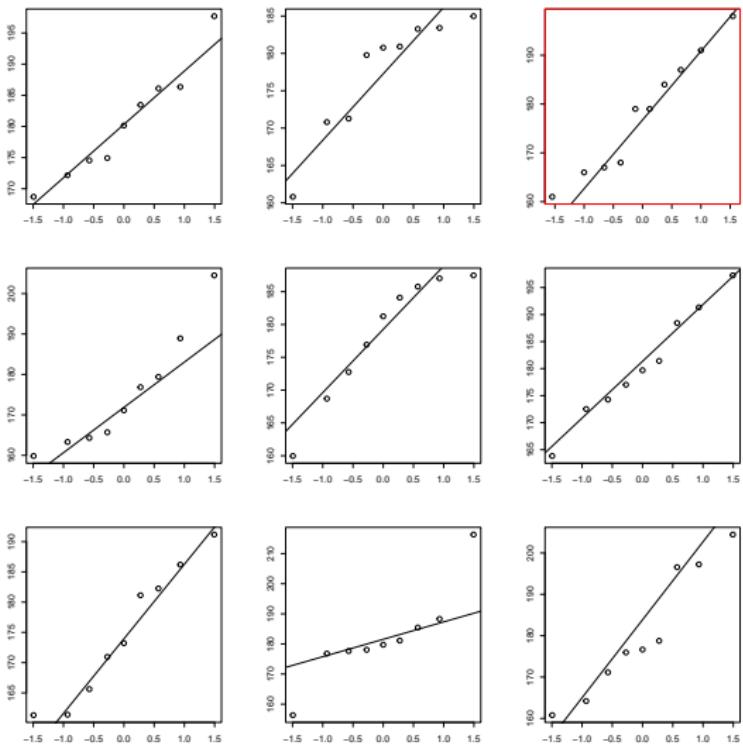
- I R, når $n > 10$:

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

- I R, når $n \leq 10$:

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

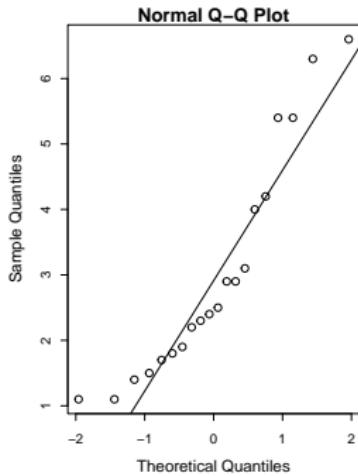
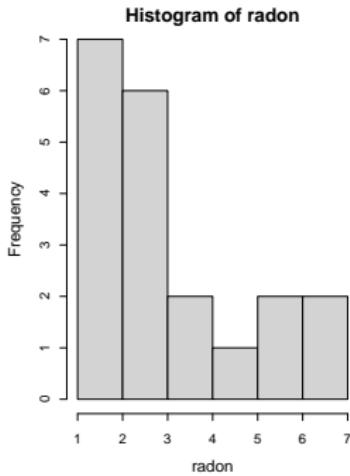
Eksempel – Højde på studerende: Sammenligning med simulerede data



Eksempel – Radon data

```
## Indlæs 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)

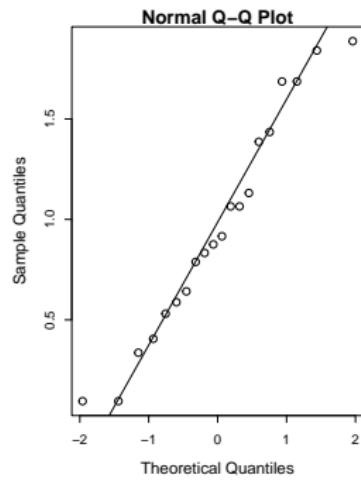
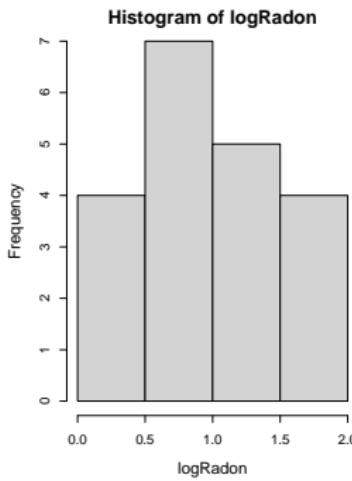
## Histogram og Q-Q plot af data
par(mfrow = c(1,2))
hist(radon)
qqnorm(radon)
qqline(radon)
```



Eksempel – Radon data: Log-transformation

```
# Log-transformation af data
logRadon<-log(radon)

## Histogram og Q-Q plot af den transformerede data
par(mfrow = c(1,2))
hist(logRadon)
qqnorm(logRadon)
qqline(logRadon)
```



Dagsorden

- 1 Opsummering
- 2 Motiverende eksempel – sovemedicin
- 3 t -test med en stikprøve
- 4 Kritiske værdier og konfidensintervaller
- 5 Hypotesetest – generelt
 - Den alternative hypotese (Modhypotesen)
 - Den generelle metode
 - Fejlslutninger ved hypotesetest!
- 6 Modelkontrol: Normalfordelingsantagelsen
 - Q-Q plot for normalfordelingen
 - Transformation mod normalitet

02402 Statistik (Polyteknisk grundlag)

Uge 6: Analyser med to stikprøver

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

Dagsorden

- 1 Opsummering
- 2 Motiverende eksempel: Ernæringsstudie
- 3 t -test med to uparrede stikprøver
- 4 Konfidensintervallet for forskellen i middelværdi
- 5 Overlappende konfidensintervaller?
- 6 t -test med to parrede stikprøver (parret t -test)
- 7 Normalfordelingesantagelserne
- 8 Styrke og stikprøvestørrelse – Forsøgsdesign
 - Krav til præcision
 - Styrke og stikprøvestørrelse – En stikprøve
 - Styrke og stikprøvestørrelse – To stikprøver
- 9 t -test med sammenvejet varians – Et alternativ

Hjælp os med at forbedre kurset:

Brug 5 minutter på at udfylde kursets
midtvejsevaluering på DTU Inside!

Dagsorden

- 1 Opsummering
- 2 Motiverende eksempel: Ernæringsstudie
- 3 t -test med to uparrede stikprøver
- 4 Konfidensintervallet for forskellen i middelværdi
- 5 Overlappende konfidensintervaller?
- 6 t -test med to parrede stikprøver (parret t -test)
- 7 Normalfordelingesantagelserne
- 8 Styrke og stikprøvestørrelse – Forsøgsdesign
 - Krav til præcision
 - Styrke og stikprøvestørrelse – En stikprøve
 - Styrke og stikprøvestørrelse – To stikprøver
- 9 t -test med sammenvejet varians – Et alternativ

Hypotesetest

Med udgangspunkt i et forskningsspørgsmål opstilles hypotesetestens grundrammer:

- ① Definition og afgrænsning af populationen
- ② Formulering af nulhypotesen (og modhypotesen)
- ③ Fastsættelse af signifikansniveauet (og teststyrken)

Baseret på kravene til hypotesestesten og en statistisk model opstilles et eksperiment (forsøg), hvorfra der udtages en eller flere repræsentative stikprøver:

- *Beregning af de nødvendige stikprøvestørrelser (Hvis muligt)*
- Undersøgelse af modellens antagelser

På baggrund af observationerne og den statistiske model udvælges en passende test. Testen kan evalueres på flere måder ved at sammenligne:

- Et konfidensinterval med parameterværdien under nulhypotesen
- En teststørrelse med kritiske værdier
- En p -værdi med signifikansniveauet

Undersøgelse af modellens fordelingsantagelse

Man udleder teststørrelser på baggrund af nogle fordelingsantagelser. Man kan undersøge om fordelingsantagelserne er opfyldt på flere måder, bl.a. ved brug af:

- Histogrammer: Man grupperer observationerne og sammenligner gruppehyppighederne med en teoretisk tæthedsfunktion baseret på estimerede parameterværdier.
- Den empiriske fordelingsfunktion: Man tegner den empiriske fordelingsfunktion og sammenligner den med en teoretisk fordelingsfunktion baseret på estimerede parameterværdier.
- QQ-plot: Man sammenholder fraktiler fra den empiriske og en teoretisk fordeling.

Man kan sammenholde med plots baseret på simulerede data fra den teoretiske fordeling.

QQ-plot

Man sammenligner fraktiler fra den empiriske fordeling med fraktiler fra den teoretiske fordeling. Hvis fordelingerne er ens, vil fraktilerne ligge på linjen $y = x$. Et QQ-plot kan bl.a. bruges til at vurdere:

- Fordelingens spredning (haler)
- Fordelingens skævhed (skewness)

Man kan teste mod alle fordelinger, men fraktilerne er kun unikt bestemt for absolut kontinuerte fordelinger.

Et normal QQ-plot er et QQ-plot, hvor den teoretiske fordeling er en normalfordeling. Normalfordelingen har nogle specielle egenskaber, som medfører, at man blot kan sammenligne med fraktiler fra en standardnormalfordeling. Få andre fordelinger, som f.eks. eksponentialfordelingen, har lignende egenskaber.

QQ-plot

Normalfordeling

Lad $X \sim N(\mu, \sigma^2)$ med fraktiler q_p for $p \in (0, 1)$ og lad $Y \sim N(0, 1^2)$ med fraktiler z_p for $p \in (0, 1)$. Så har man, at

$$p = \mathbb{P}(X \leq q_p) = \mathbb{P}(\sigma X^* + \mu \leq q_p) = \mathbb{P}\left(X^* \leq \frac{q_p - \mu}{\sigma}\right).$$

Da X^* og Y har samme fordeling, må $z_p = (q_p - \mu)/\sigma$.

Eksponentialfordeling

Lad $X \sim \text{Eks}(\lambda)$ med fraktiler q_p for $p \in (0, 1)$ og lad $Y \sim \text{Eks}(1)$ med fraktiler f_p for $p \in (0, 1)$. Så har man, at

$$p = \mathbb{P}(X \leq q_p) = \mathbb{P}(\lambda X \leq \lambda q_p).$$

Da λX og Y har samme fordeling, må $f_p = \lambda q_p$.

Links om QQ-plot

To gode links med diskussioner om normal QQ-plot:

[https://stats.stackexchange.com/questions/101274/
how-to-interpret-a-qq-plot](https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot)

[https://stats.stackexchange.com/questions/22258/
what-is-the-use-of-the-line-produced-by-qqline-in-r](https://stats.stackexchange.com/questions/22258/what-is-the-use-of-the-line-produced-by-qqline-in-r)

Hentet 2. oktober 2023.

Dagsorden

- 1 Opsummering
- 2 Motiverende eksempel: Ernæringsstudie
- 3 t -test med to uparrede stikprøver
- 4 Konfidensintervallet for forskellen i middelværdi
- 5 Overlappende konfidensintervaller?
- 6 t -test med to parrede stikprøver (parret t -test)
- 7 Normalfordelingesantagelserne
- 8 Styrke og stikprøvestørrelse – Forsøgsdesign
 - Krav til præcision
 - Styrke og stikprøvestørrelse – En stikprøve
 - Styrke og stikprøvestørrelse – To stikprøver
- 9 t -test med sammenvejet varians – Et alternativ

Motiverende eksempel: Ernæringsstudie

Forskel på energiforbrug?

I et ernæringsstudie ønsker man at undersøge, om der er en forskel i energiforbruget for forskellige typer (moderat fysisk krævende) arbejde.

I studiet har 9 sygeplejersker fra hospital A og 9 (andre) sygeplejersker fra hospital B fået målt deres energiforbrug. Målingerne ses i følgende tabel (i enheden megajoule MJ):

Stikprøve fra hvert 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: Ernæringsstudie

Hypotesen om ingen forskel (i det gennemsnitlige energiforbrug) ønskes undersøgt:

$$H_0 : \mu_A = \mu_B$$

Eksempel: Ernæringsstudie

Hypotesen om ingen forskel (i det gennemsnitlige energiforbrug) ønskes undersøgt:

$$H_0 : \mu_A = \mu_B$$

Stikprøvegennemsnit og
-standardafvigelser:

$$\hat{\mu}_A = \bar{x}_A = 8.293 \quad (s_A = 1.428)$$

$$\hat{\mu}_B = \bar{x}_B = 10.298 \quad (s_B = 1.398)$$

Eksempel: Ernæringsstudie

Hypotesen om ingen forskel (i det gennemsnitlige energiforbrug) ønskes undersøgt:

$$H_0 : \mu_A = \mu_B$$

Stikprøvegennemsnit og -standardafvigelser:

$$\hat{\mu}_A = \bar{x}_A = 8.293 \quad (s_A = 1.428)$$

$$\hat{\mu}_B = \bar{x}_B = 10.298 \quad (s_B = 1.398)$$

Er data i overensstemmelse med nulhypotesen H_0 ?

Data: $\bar{x}_B - \bar{x}_A = 2.005$

Nulhypotese: $H_0 : \mu_B - \mu_A = 0$

Eksempel: Ernæringsstudie

Hypotesen om ingen forskel (i det gennemsnitlige energiforbrug) ønskes undersøgt:

$$H_0 : \mu_A = \mu_B$$

Stikprøvegennemsnit og -standardafvigelser:

$$\hat{\mu}_A = \bar{x}_A = 8.293 \quad (s_A = 1.428)$$

$$\hat{\mu}_B = \bar{x}_B = 10.298 \quad (s_B = 1.398)$$

Er data i overensstemmelse med nulhypotesen H_0 ?

Data: $\bar{x}_B - \bar{x}_A = 2.005$

Nulhypote: $H_0 : \mu_B - \mu_A = 0$

NYT: *p*-værdi for forskel:

$$p = 0.0083$$

(Beregnet under antagelsen, at H_0 er sand.)

NYT: Konfidensinterval for forskellen:

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

Dagsorden

- 1 Opsummering
- 2 Motiverende eksempel: Ernæringsstudie
- 3 ***t*-test med to uparrede stikprøver**
- 4 Konfidensintervallet for forskellen i middelværdi
- 5 Overlappende konfidensintervaller?
- 6 *t*-test med to parrede stikprøver (parret *t*-test)
- 7 Normalfordelingesantagelserne
- 8 Styrke og stikprøvestørrelse – Forsøgsdesign
 - Krav til præcision
 - Styrke og stikprøvestørrelse – En stikprøve
 - Styrke og stikprøvestørrelse – To stikprøver
- 9 *t*-test med sammenvejet varians – Et alternativ

Metode 3.49: Teststørrelsen i en (Welch) *t*-test med to uparrede stikprøver

Forudsætninger:

Testen gælder, når begge stikprøver er store, eller når begge stikprøver kommer fra normalfordelte populationer.

Beregning af den observerede teststørrelse:

Vi betragter følgende nulhypotese om forskellen i middelværdi mellem to *uafhængige* og *uparrede* stikprøver: (Bemærk fejl i bogen)

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

$$H_0: \delta = \delta_0,$$

så er den observerede teststørrelse:

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

Sætning 3.50: Fordelingen af teststørrelsen

Teststørrelsen er (tilnærmelsesvis) *t*-fordelt:

Under nulhypotesen er teststørrelsen:

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

Den følger tilnærmelsesvis en *t*-fordeling med v frihedsgrader, hvor

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

hvis de to populationer er normalfordelte eller stikprøvestørrelserne er tilstrækkelig store.

Eksempel: Ernæringsstudie

Hypotesen om ingen forskel ønskes undersøgt:

$$H_0 : \delta = \mu_B - \mu_A = 0$$

mod det tosidede alternativ:

$$H_1 : \delta = \mu_B - \mu_A \neq 0.$$

Først beregnes t_{obs} og V :

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

og

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

Eksempel: Ernæringsstudie

Dernæst findes p -værdien:

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

```
## Eksempel - Ernæringsstudie: P(T > 3.01)
1 - pt(3.01, df = 15.99)

## [1] 0.004154
```

Eksempel: Ernæringsstudie

Dernæst findes p -værdien:

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

```
## Eksempel - Ernæringsstudie: P(T > 3.01)
1 - pt(3.01, df = 15.99)

## [1] 0.004154
```

Vurdér evidensen (Tabel 3.1):

Der er stærk evidens imod nulhypotesen.

Eksempel: Ernæringsstudie

Dernæst findes *p*-værdien:

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

```
## Eksempel - Ernæringsstudie: P(T > 3.01)
1 - pt(3.01, df = 15.99)

## [1] 0.004154
```

Vurdér evidensen (Tabel 3.1):

Der er stærk evidens imod nulhypotesen.

Konklusion baseret på et 5%-signifikansniveau ($\alpha = 0.05$):

Vi forkaster nulhypotesen. Der er signifikant forskel på de to grupper – sygeplejersker på Hospital B kan siges at have et større (middel)energiforbrug end sygeplejesker på Hospital A.

Dagsorden

- 1 Opsummering
- 2 Motiverende eksempel: Ernæringsstudie
- 3 t -test med to uparrede stikprøver
- 4 Konfidensintervallet for forskellen i middelværdi
- 5 Overlappende konfidensintervaller?
- 6 t -test med to parrede stikprøver (parret t -test)
- 7 Normalfordelingesantagelserne
- 8 Styrke og stikprøvestørrelse – Forsøgsdesign
 - Krav til præcision
 - Styrke og stikprøvestørrelse – En stikprøve
 - Styrke og stikprøvestørrelse – To stikprøver
- 9 t -test med sammenvejet varians – Et alternativ

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

Konfidensintervallet for forskellen i middelværdi:

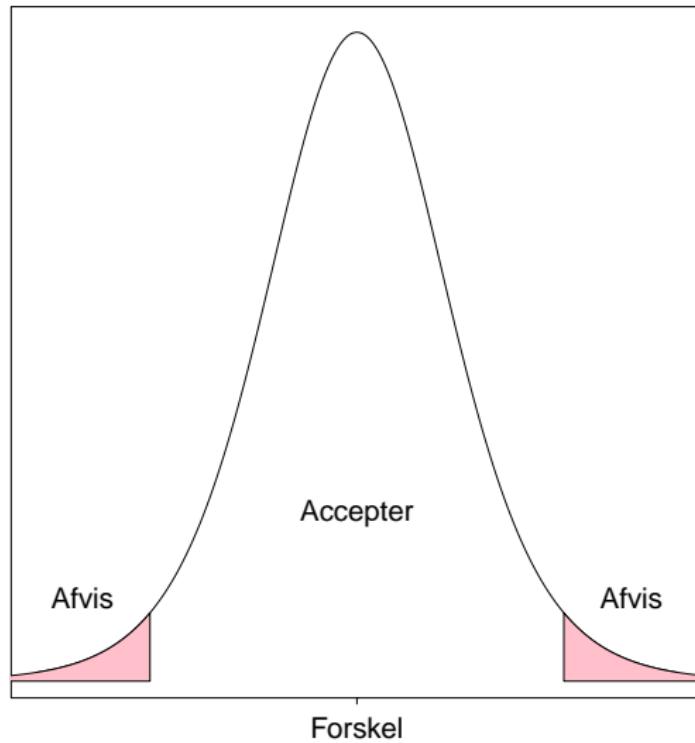
For to stikprøver (x_1, \dots, x_{n_1}) og (y_1, \dots, y_{n_2}) er
 $(1 - \alpha)$ -konfidensintervallet for $\mu_1 - \mu_2$ givet ved:

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

hvor $t_{1-\alpha/2}$ er $(1 - \alpha/2)$ -fraktilen i t -fordelingen med v frihedsgrader (givet i sætning 3.50).

Konfidensinterval og hypotesetest (Repetition)

Acceptområdet er de mulige værdier for $\mu_1 - \mu_2$, som ikke ligger for langt væk fra data:



Eksempel: Ernæringsstudie

Lad os finde 95%-konfidensintervallet for $\mu_B - \mu_A$. Med $v = 15.99$ er den relevante t -fraktil givet ved

$$t_{0.975} = 2.120,$$

så konfidensintervallet bliver

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

Udregnet giver dette:

$$[0.59; 3.42].$$

Eksempel: Ernæringsstudie – Det hele i R:

```
# Indlæs data
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)

# Udfør t-test med to uparrede stikprøver
t.test(xB, xA)

##
##  Welch Two Sample t-test
##
## data: xB and xA
## t = 3, df = 16, p-value = 0.008
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.5923 3.4166
## sample estimates:
## mean of x mean of y
##      10.298      8.293
```

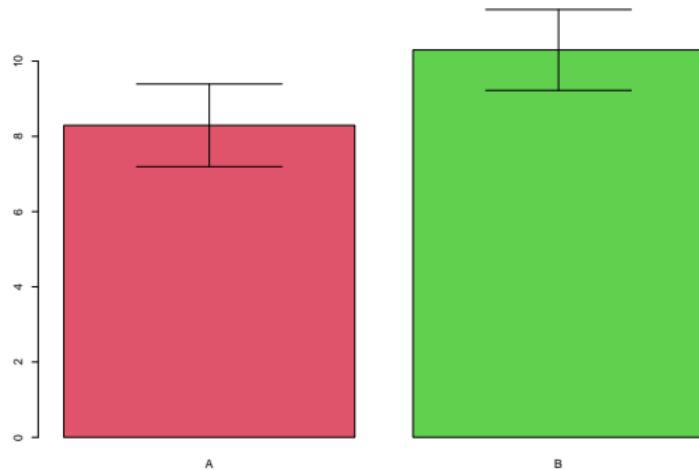
Dagsorden

- 1 Opsummering
- 2 Motiverende eksempel: Ernæringsstudie
- 3 t -test med to uparrede stikprøver
- 4 Konfidensintervallet for forskellen i middelværdi
- 5 Overlappende konfidensintervaller?
 - 6 t -test med to parrede stikprøver (parret t -test)
 - 7 Normalfordelingesantagelserne
 - 8 Styrke og stikprøvestørrelse – Forsøgsdesign
 - Krav til præcision
 - Styrke og stikprøvestørrelse – En stikprøve
 - Styrke og stikprøvestørrelse – To stikprøver
 - 9 t -test med sammenvejet varians – Et alternativ

Eksempel: Ernæringsstudie – Præsentation af resultater

Søjlediagrammer med *fejlbjælker* ses ofte:

Et sjøldiagram med nogle fejlbjælker (error bars): Herunder vises 95%-konfidensintervallerne for hver gruppe:



Vær varsom med at bruge "overlappende konfidensintervaller"

Man bruger den forkerte variation til at vurdere forskellen:

$$\sigma_{(\bar{X}_A - \bar{X}_B)} \neq \sigma_{\bar{X}_A} + \sigma_{\bar{X}_B}$$

$$\text{Var}(\bar{X}_A - \bar{X}_B) = \text{Var}(\bar{X}_A) + \text{Var}(\bar{X}_B)$$

Antag at de to standardafvigelser er 3 og 4: Summen er 7, men $\sqrt{3^2 + 4^2} = 5$.

Vær varsom med at bruge "overlappende konfidensintervaller"

Man bruger den forkerte variation til at vurdere forskellen:

$$\sigma_{(\bar{X}_A - \bar{X}_B)} \neq \sigma_{\bar{X}_A} + \sigma_{\bar{X}_B}$$

$$\text{Var}(\bar{X}_A - \bar{X}_B) = \text{Var}(\bar{X}_A) + \text{Var}(\bar{X}_B)$$

Antag at de to standardafvigelser er 3 og 4: Summen er 7, men $\sqrt{3^2 + 4^2} = 5$.

Det korrekte forhold mellem standardafvigelserne er således:

$$\sigma_{(\bar{X}_A - \bar{X}_B)} < \sigma_{\bar{X}_A} + \sigma_{\bar{X}_B}.$$

Vær varsom med at bruge "overlappende konfidensintervaller"

Bemærkning 3.59 – Regel for brug af "overlappende konfidensintervaller":

Når to konfidensintervaller IKKE overlapper: De to grupper er signifikant forskellige.

Når to konfidensintervaller overlapper: Ingen konklusion kan drages uden at undersøge konfidensintervallet for *forskellen* mellem grupperne.

Dagsorden

- 1 Opsummering
- 2 Motiverende eksempel: Ernæringsstudie
- 3 *t*-test med to uparrede stikprøver
- 4 Konfidensintervallet for forskellen i middelværdi
- 5 Overlappende konfidensintervaller?
- 6 ***t*-test med to parrede stikprøver (parret *t*-test)**
- 7 Normalfordelingesantagelserne
- 8 Styrke og stikprøvestørrelse – Forsøgsdesign
 - Krav til præcision
 - Styrke og stikprøvestørrelse – En stikprøve
 - Styrke og stikprøvestørrelse – To stikprøver
- 9 *t*-test med sammenvejet varians – Et alternativ

Motiverende eksempel: Sovemedicin

Forskel på sovemedicin?

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

Stikprøve med $n = 10$:

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

Motiverende eksempel: Sovemedicin

Forskelse på sovemedicin?

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

Stikprøve med $n = 10$:

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	$\bar{x} = 1.67$
6	+3.4	+4.4	+1.0	$s = 1.13$
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	

Analyse af to parrede stikprøver: Parret *t*-test

```
# Indlæs data
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)

# Udregn forskellene
dif = x2 - x1

# Udfør en parret t-test
t.test(dif)

##
##  One Sample t-test
##
## data: dif
## t = 4.7, df = 9, p-value = 0.001
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.8613 2.4787
## sample estimates:
## mean of x
##          1.67
```

Analyse af to parrede stikprøver: Parret *t*-test

```
# En anden måde at udføre en parret t-test
t.test(x2, x1, paired = TRUE)

##
##  Paired t-test
##
## data: x2 and x1
## t = 4.7, df = 9, p-value = 0.001
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.8613 2.4787
## sample estimates:
## mean of the differences
##                           1.67
```

Forsøgsopsætning: Parrede og uafhængige stikprøver

Fuldstændigt tilfældigt (uafhængige stikprøver):

Vi har 20 patienter, som tilfældigt fordeles på to grupper (normalt lige mange i hver gruppe). Dvs. at der er forskellige (uafhængige) patienter i de to grupper.

Parrede observationer (afhængige stikprøver):

Vi har 10 patienter, som alle får begge behandlinger (typisk med noget tid imellem og med tilfældig rækkefølge af behandlingerne).

Dvs. de samme patienter fremgår i de to grupper.

Eksempel: Sovemedicin – FORKERT analyse

```
# Forkert analyse (t-test for to uparrede stikprøver)
t.test(x1, x2)

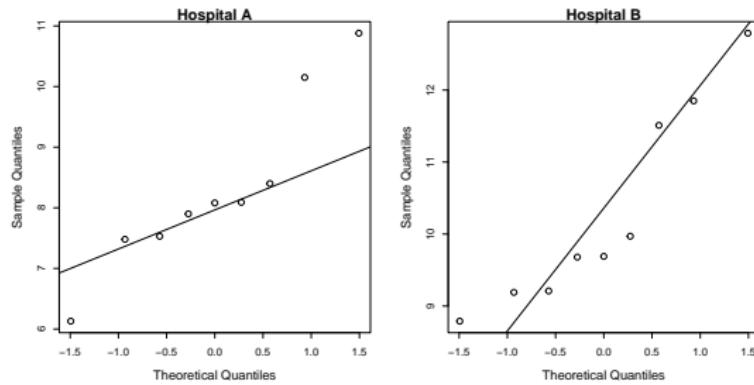
##
##  Welch Two Sample t-test
##
## data: x1 and x2
## t = -1.9, df = 18, p-value = 0.07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -3.4854 0.1454
## sample estimates:
## mean of x mean of y
##      0.66      2.33
```

Dagsorden

- 1 Opsummering
- 2 Motiverende eksempel: Ernæringsstudie
- 3 t -test med to uparrede stikprøver
- 4 Konfidensintervallet for forskellen i middelværdi
- 5 Overlappende konfidensintervaller?
- 6 t -test med to parrede stikprøver (parret t -test)
- 7 **Normalfordelingesantagelserne**
- 8 Styrke og stikprøvestørrelse – Forsøgsdesign
 - Krav til præcision
 - Styrke og stikprøvestørrelse – En stikprøve
 - Styrke og stikprøvestørrelse – To stikprøver
- 9 t -test med sammenvejet varians – Et alternativ

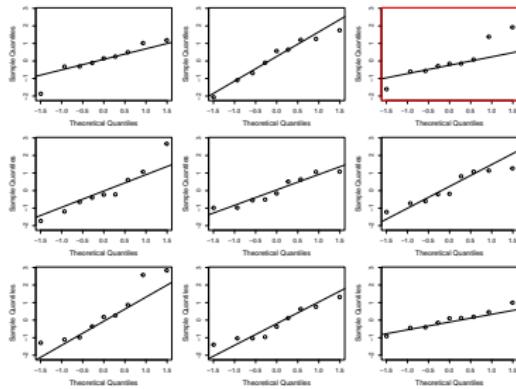
Eksempel: QQ-plot for hver af stikprøverne:

```
# QQ plot for hver stikprøve  
qqnorm(xA, main = "Hospital A")  
qqline(xA)  
qqnorm(xB, main = "Hospital B")  
qqline(xB)
```



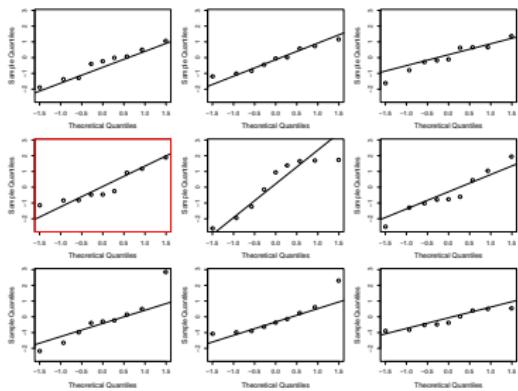
Eksempel: Sammenligning med simulerede data, Hospital A

```
# Flere simulerede QQ plot
require(MESS)
fitA <- lm(xA ~ 1)
qqnorm.wally <- function(x, y, ...) { qqnorm(y, ...); qqline(y, ...) }
wallyplot(fitA, FUN = qqnorm.wally, main = "")
```



Eksempel: Sammenligning med simulerede data, Hospital B

```
# Flere simulerede QQ plot
fitB <- lm(xB ~ 1)
qqnorm.wally <- function(x, y, ...) { qqnorm(y, ...); qqline(y, ...) }
wallyplot(fitB, FUN = qqnorm.wally, main = "")
```



Dagsorden

- 1 Opsummering
- 2 Motiverende eksempel: Ernæringsstudie
- 3 t -test med to uparrede stikprøver
- 4 Konfidensintervallet for forskellen i middelværdi
- 5 Overlappende konfidensintervaller?
- 6 t -test med to parrede stikprøver (parret t -test)
- 7 Normalfordelingesantagelserne
- 8 Styrke og stikprøvestørrelse – Forsøgsdesign
 - Krav til præcision
 - Styrke og stikprøvestørrelse – En stikprøve
 - Styrke og stikprøvestørrelse – To stikprøver
- 9 t -test med sammenvejet varians – Et alternativ

Forsøgsplanlægning med krav til præcisionen

Fejlmarginen (margin of error - ME) er defineret som

$$t_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Metode 3.63: Stikprøvestørrelse for konfidensintervallet baseret på en stikprøve

Hvis σ er kendt, eller vurderet til at være en bestemt værdi, så kan vi beregne den stikprøvestørrelse, som kræves for at opnå en given fejlmargin, med sandsynlighed $1 - \alpha$.

$$n = \left(\frac{z_{1-\alpha/2} \cdot \sigma}{ME} \right)^2.$$

Eksempel: Højde på studerende

Stikprøvemiddelværdien
og -standardafvigelsen:

$$\bar{x} = 178$$

$$s = 12.21$$

Estimater for
populationens middelværdi
og standardafvigelse:

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

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

Eksempel: Højde på studerende

Stikprøvemiddelværdien
og -standardafvigelsen:

$$\bar{x} = 178$$

$$s = 12.21$$

Estimater for
populationens middelværdi
og standardafvigelse:

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

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

Hvis vi ønsker en fejlmargin på 3 cm på et
5%-signifikansniveau, hvor stor skal stikprøvestørrelsen n
så være?

$$n = \left(\frac{1.96 \cdot 12.21}{3} \right)^2 = 63.64 \approx 64.$$

Forsøgsplanlægning: Styrke

Hvad er styrken af et fremtidigt eksperiment:

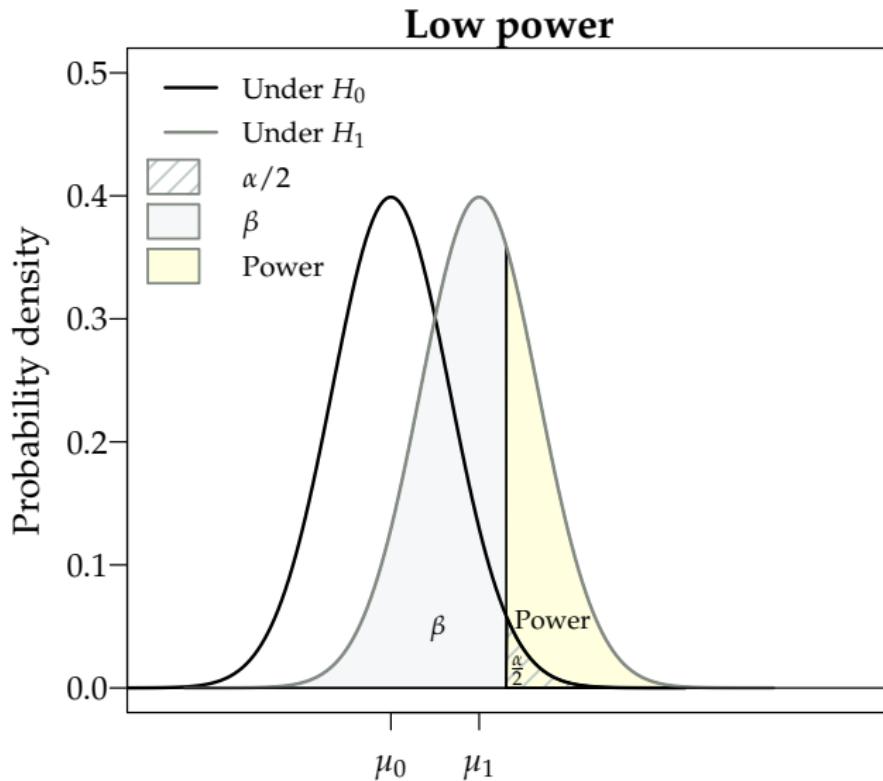
- Sandsynligheden for at detektere en (påstået) effekt.
- $P(H_0 \text{ afvises})$ når H_1 er sand.
- Sandsynligheden for en korrekt afvisning af H_0 .
- MEN: En nulhypotese kan være forkert på mange måder!
- I praksis: Brug en scenarie-baseret tilgang
 - F.eks. "Hvis $\mu = 86$, hvor sikkert vil mit forsøg være i stand til at detektere dette?"
 - F.eks. "Hvis $\mu = 84$, hvor sikkert vil mit forsøg være i stand til at detektere dette?"

Forsøgsplanlægning: Styrke

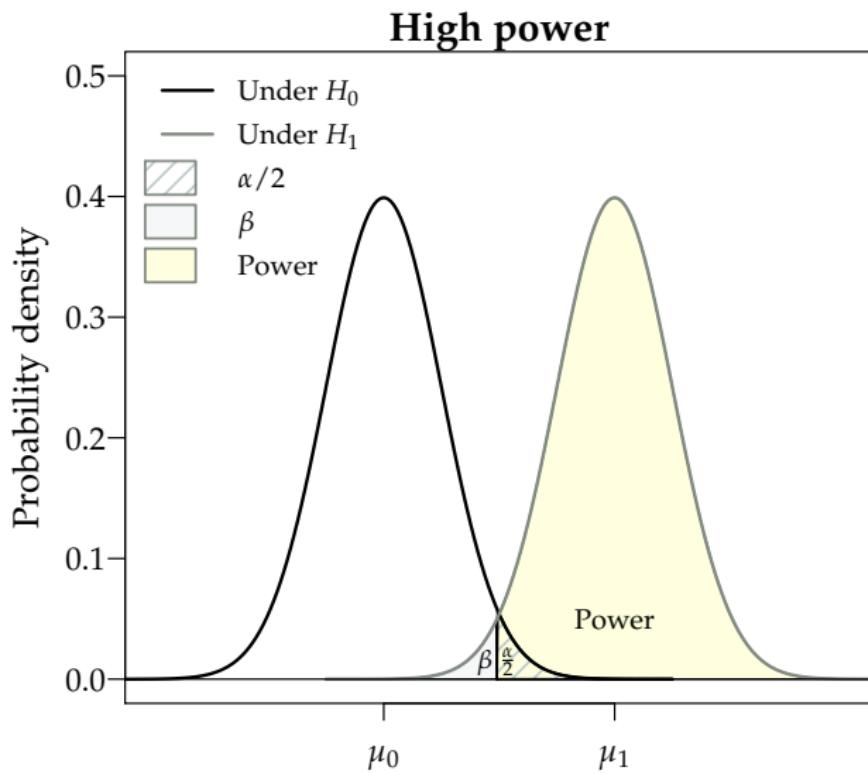
Hvis vi kender (eller antager) fire ud af de fem følgende størrelser, så kan vi finde den manglende:

- Stikprøvestørrelsen (sample size), n .
- Signifikansniveauet, α , som vi tester på.
- Forskellen i middelværdi (effekt-størrelsen), $\mu_0 - \mu_1$.
- Populationsstandardafvigelsen, σ .
- Styrken (power), $1 - \beta$.

Eksempel med lav styrke



Eksempel med høj styrke



Forsøgsplanlægning: Stikprøvestørrelsen n

Det store spørgsmål: Hvor stort skal n være?

Vi skal have nok observationer til at kunne detektere en relevant effekt med høj styrke $1 - \beta$ (typisk mindst 80%):

Forsøgsplanlægning: Stikprøvestørrelsen n

Det store spørgsmål: Hvor stort skal n være?

Vi skal have nok observationer til at kunne detektere en relevant effekt med høj styrke $1 - \beta$ (typisk mindst 80%):

Metode 3.65: Formel for stikprøvestørrelse med en stikprøve

For en t-test med en stikprøve, hvor α , β og σ er givet:

$$n = \left(\sigma \frac{z_{1-\beta} + z_{1-\alpha/2}}{\mu_0 - \mu_1} \right)^2.$$

Her er $\mu_0 - \mu_1$ den forskel i middelværdi, som vi ønsker at måle, medens $z_{1-\beta}$ og $z_{1-\alpha/2}$ er fraktiler i standardnormalfordelingen.

Eksempel: Styrken når $n = 40$

```
# Udregning af styrken (en stikprøve)
power.t.test(n = 40, delta = 4, sd = 12.21, 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
```

Eksempel: Stikprøvestørrelsen når styken skal være 80%

```
# Udregning af stikprøvestørrelsen (en stikprøve)
power.t.test(power = .80, delta = 4, sd = 12.21, 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
```

Styrke og stikprøvestørrelse: To stikprøver

Find styrken af en test, som kan detektere en forskel/effektstørrelse på 2 på et 5%-signifikansniveau, når $\sigma = 1$ og $n = 10$:

```
# Udregning af styrken (to stikprøver)
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: To stikprøver

Find stikprøvestørrelsen, hvis en test med en styrke på 90% skal kunne detektere en forskel/effektstørrelse på 2, når $\sigma = 1$:

```
# Udregning af stikprøvestørrelsen (to stikprøver)
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
```

Styrke og stikprøvestørrelse: To stikprøver

Hvilken effektstørrelse δ kan detekteres på et 5%-signifikansniveau i en test med en styrke på 90%, når $\sigma = 1$ og $n = 10$:

```
## Udregning af effektstørrelsen (to stikprøver)
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
```

Dagsorden

- 1 Opsummering
- 2 Motiverende eksempel: Ernæringsstudie
- 3 *t*-test med to uparrede stikprøver
- 4 Konfidensintervallet for forskellen i middelværdi
- 5 Overlappende konfidensintervaller?
- 6 *t*-test med to parrede stikprøver (parret *t*-test)
- 7 Normalfordelingesantagelserne
- 8 Styrke og stikprøvestørrelse – Forsøgsdesign
 - Krav til præcision
 - Styrke og stikprøvestørrelse – En stikprøve
 - Styrke og stikprøvestørrelse – To stikprøver
- 9 *t*-test med sammenvejet varians – Et alternativ

t-test med sammenvejet varians for to uparrede stikprøver

Det *sammenvejede* (pooled) variansenestimat (her antages $\sigma_1^2 = \sigma_2^2$)

Metode 3.52

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

Teststørrelsen i en *t*-test med sammenvejet varians, Metode 3.53

Betrugter vi nulhypotesen om forskellen i middelværdi mellem to *uafhængige* stikprøver:

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

$$H_0 : \delta = \delta_0,$$

så er teststørrelsen i en *t*-test med den sammenvejede varians givet ved:

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

Sætning 3.54: Fordelingen af teststørrelsen

Resultatet

Teststørrelsen i en *t*-test med sammenvejet varians:

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

følger under nulhypotesen (og under antagelsen $\sigma_1^2 = \sigma_2^2$) en *t*-fordeling med $n_1 + n_2 - 2$ frihedsgrader, hvis de to populationer er normalfordelte.

Vi bruger altid "Welch"-versionen

Nogenlunde (idiot)sikkert at bruge Welch-versionen altid:

- Hvis $s_1^2 = s_2^2$, så er de to test ens. Hvis det er tilfældet, så foretrækker vi ikke nødvendigvis testen med den sammenvejede varians, da antagelsen om ens varianser kan være højest tvivlsom.
- Kun hvis de to varianser er meget forskellige, kan det ske, at de to test giver meget forskellige resultater. Hvis varianserne virker meget forskellige, brydes antagelsen om ens varianser formodentligt.
- I tilfælde med en lille stikprøvestørrelse i mindst en af grupperne, kan testen med sammenvejet varians give en højere styrke (under antagelse om ens varianser). I disse tilfælde er Welch-versionen en "forsiktig" tilgang.

Dagsorden

- ① Opsummering
- ② Motiverende eksempel: Ernæringsstudie
- ③ *t*-test med to uparrede stikprøver
- ④ Konfidensintervallet for forskellen i middelværdi
- ⑤ Overlappende konfidensintervaller?
- ⑥ *t*-test med to parrede stikprøver (parret *t*-test)
- ⑦ Normalfordelingesantagelserne
- ⑧ Styrke og stikprøvestørrelse – Forsøgsdesign
 - Krav til præcision
 - Styrke og stikprøvestørrelse – En stikprøve
 - Styrke og stikprøvestørrelse – To stikprøver
- ⑨ *t*-test med sammenvejet varians – Et alternativ

02402 Statistik (Polyteknisk grundlag)

Uge 7: Simulation og bootstrapping

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

Dagsorden

- 1 Introduktion til simulation
 - Eksempel: Areal af plader
- 2 Fejlophobningslove
- 3 Parametrisk bootstrapping
 - Introduktion til bootstrapping
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver
- 4 Ikke-parametrisk bootstrapping
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver

Dagsorden

- ① Introduktion til simulation
 - Eksempel: Areal af plader
- ② Fejlophobningslove
- ③ Parametrisk bootstrapping
 - Introduktion til bootstrapping
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver
- ④ Ikke-parametrisk bootstrapping
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver

Motivation

- Mange relevante stikprøbefunktioner har komplikerede fordelinger. Det kunne f.eks. være:
 - Medianen
 - Fraktiler
 - Den interkvartile variationsbredde (IQR)
 - Variationskoefficienten (coefficient of variation)
 - Ikke-lineære funktioner af en eller flere variable
 - Variansen (el. spredningen)
- Vi mangler værktøjer, når antagelserne for vores test ikke er opfyldte.

Motivation

- Mange relevante stikprøvefunktioner har komplikerede fordelinger. Det kunne f.eks. være:
 - Medianen
 - Fraktiler
 - Den interkvartile variationsbredde (IQR)
 - Variationskoefficienten (coefficient of variation)
 - Ikke-lineære funktioner af en eller flere variable
 - Variansen (el. spredningen)
- Vi mangler værktøjer, når antagelserne for vores test ikke er opfyldte.
- **Løsning:** Simulation og bootstrapping – R er et super værktøj til dette!

Simulation

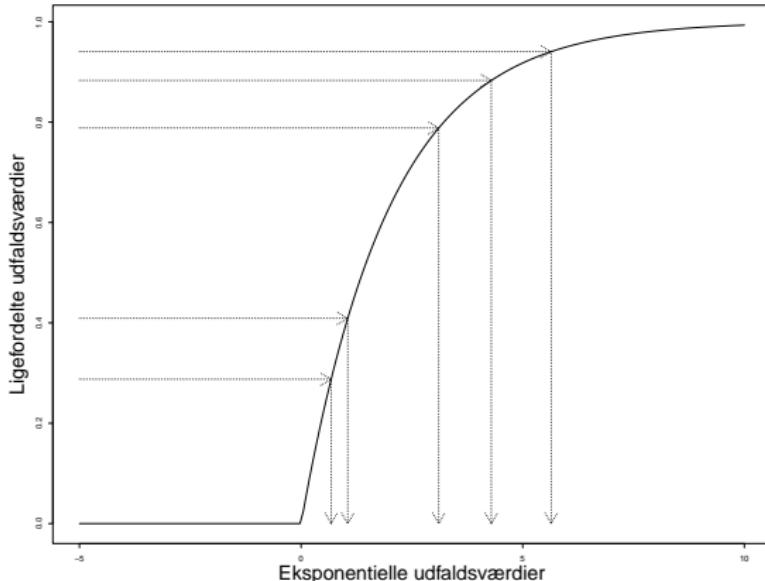
- (Pseudo)-tilfældige tal genereret af en computer.
- En *tilfældighedsgenerator* er en algoritme, der kan generere en talfølge af tilsyneladende tilfældige tal.
- Algoritmen kræver en "start" kaldet et *seed*.
- Man kan simulere fra (næsten) alle fordelinger igennem den uniforme fordeling ved at benytte følgende resultat:

Sætning 2.51: Alle fordelinger kan "fremskaffes" fra den uniforme fordeling

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

Eksempel: Eksponentialfordelingen med $\lambda = 0.5$

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



I praksis i R

Mange fordelinger er gjort klar til simulering, for eksempel:

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

Eksempel: Areal af plader

En virksomhed producerer rektangulære plader.

Længden af pladerne (i meter), X , antages at kunne beskrives ved normalfordelingen $N(2, 0.01^2)$, medens bredden af pladerne (i meter), Y , antages at kunne beskrives ved normalfordelingen $N(3, 0.02^2)$. Man kan antage, at pladernes længder og bredder er uafhængige.

Man er interesseret i arealet, A , som jo så givet ved $A = XY$.

- Hvad er middelarealet?
- Hvad er spredningen i arealet fra plade til plade?
- Hvor ofte har sådanne plader et areal, der afviger mere end 0.1 m^2 fra de angivne 6 m^2 ?
- Sandsynligheder for andre hændelser.
- Generelt: Hvad er fordelingen for den stokastiske variabel A ?

Eksempel: Areal af plader – løsning ved simulation

```
k = 10000 # Antal simulationer  
X = rnorm(k, 2, 0.01)  
Y = rnorm(k, 3, 0.02)  
A = X*Y  
  
mean(A)
```

```
[1] 6
```

```
var(A)
```

```
[1] 0.002458
```

```
mean(abs(A - 6) > 0.1)
```

```
[1] 0.0439
```

Dagsorden

1 Introduktion til simulation

- Eksempel: Areal af plader

2 Fejlophobningslove

3 Parametrisk bootstrapping

- Introduktion til bootstrapping
- Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
- Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver

4 Ikke-parametrisk bootstrapping

- Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
- Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver

Fejlophobningslove (propagation of error)

Man ønsker at finde:

$$\sigma_{f(X_1, \dots, X_n)}^2 = \text{Var}(f(X_1, \dots, X_n))$$

Fejlophobningslove (propagation of error)

Man ønsker at finde:

$$\sigma_{f(X_1, \dots, X_n)}^2 = \text{Var}(f(X_1, \dots, X_n))$$

Lineærkombination af uafhængige variable:

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

Fejlophobningslove (propagation of error)

Man ønsker at finde:

$$\sigma_{f(X_1, \dots, X_n)}^2 = \text{Var}(f(X_1, \dots, X_n))$$

Lineærkombination af uafhængige variable:

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

Metode 4.3: For ikke-lineære funktioner af uafhængige variable X_1, \dots, X_n :

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

Eksempel: Areal af plader (fortsat)

Vi brugte simulation i den første del af eksemplet.

Nu er vi givet to konkrete målinger for X og Y , $x = 2.00$ m og $y = 3.00$ m:

Hvad er variansen af $A = XY$ beregnet med fejlophobningsloven?

Eksempel: Areal af plader (fortsat)

Varianserne er:

$$\sigma_1^2 = \text{Var}(X) = 0.01^2 \text{ og } \sigma_2^2 = \text{Var}(Y) = 0.02^2.$$

Eksempel: Areal af plader (fortsat)

Varianserne er:

$$\sigma_1^2 = \text{Var}(X) = 0.01^2 \text{ og } \sigma_2^2 = \text{Var}(Y) = 0.02^2.$$

Funktionen og dens partielt afledte er:

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

Eksempel: Areal af plader (fortsat)

Varianserne er:

$$\sigma_1^2 = \text{Var}(X) = 0.01^2 \text{ og } \sigma_2^2 = \text{Var}(Y) = 0.02^2.$$

Funktionen og dens partielt afledte er:

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

Så resultatet bliver:

$$\begin{aligned} \text{Var}(A) &\approx \left(\frac{\partial f}{\partial x}\right)^2 \sigma_1^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_2^2 \\ &= y^2 \sigma_1^2 + x^2 \sigma_2^2 \\ &= 3.00^2 \cdot 0.01^2 + 2.00^2 \cdot 0.02^2 \\ &= 0.0025. \end{aligned}$$

Fejlophobning – ved simulation

Metode 4.4: Fejlophobning ved simulation

Antag at vi har (faktiske) målinger x_1, \dots, x_n med kendte/antagede (estimerede) varianser $\sigma_1^2, \dots, \sigma_n^2$.

- ① Simulér k udfaldsværdier af alle n målinger fra de antagne fordelinger, f.eks. $X_i^{(j)} \sim N(x_i, \sigma_i^2)$, $j = 1, \dots, k$, $i = 1, \dots, n$.
- ② Udregn standardafvigelsen som den observerede standardafvigelse af de k simulerede værdier af $f(X_1^{(j)}, \dots, X_n^{(j)})$:

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

hvor

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

Eksempel: Areal af plader (fortsat)

Faktisk kan vi i dette eksempel udlede variansen for A teoretisk:

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

Eksempel: Areal af plader (fortsat)

Tre forskellige tilgange:

- ① Simulation
- ② Den approksimative *fejlophobningslov*
- ③ Teoretisk udledning

Eksempel: Areal af plader (fortsat)

Tre forskellige tilgange:

- ① Simulation
- ② Den approksimative *fejlophobningslov*
- ③ Teoretisk udledning

Simulationstilgangen har nogle vigtige fordele:

- ① Nem måde at beregne andre størrelser end blot standardafvigelsen (de teoretiske udledninger kan være meget komplicerede sammenlignet med variansen).
- ② Nem måde at bruge andre fordelinger end normalfordelingen, hvis vi tror, at det bedre beskriver virkeligheden.
- ③ Afhænger ikke af en lineær tilnærmelse af den underliggende ikke-lineære funktion (i modsætning til fejlophobningsloven).

Dagsorden

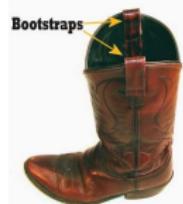
- 1 Introduktion til simulation
 - Eksempel: Areal af plader
- 2 Fejlophobningslove
- 3 Parametrisk bootstrapping
 - Introduktion til bootstrapping
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver
- 4 Ikke-parametrisk bootstrapping
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver

Bootstrapping

Bootstrap = Støvlestrop

Bootstrapping findes i to versioner:

- ① Parametrisk bootstrap: Simulér gentagne stikprøver fra den antagede (og estimerede) fordeling.
- ② Ikke-parametrisk bootstrap: Simulér gentagne stikprøver direkte fra data.



<https://en.wikipedia.org/wiki/Bootstrapping#Etymology>

Eksempel: Konfidensinterval for middelværdien i en eksponentiaffordeling

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.

Eksempel: Konfidensinterval for middelværdien i en eksponentiafordeling

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.

Vi estimerer middelværdien og intensiteten ud fra data:

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

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.

Vi estimerer middelværdien og intensiteten ud fra data:

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

Fordelingsantagelse:

Ventetiderne kommer fra en eksponentialfordeling.

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.

Vi estimerer middelværdien og intensiteten ud fra data:

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

Fordelingsantagelse:

Ventetiderne kommer fra en eksponentialfordeling.

Hvad er konfidensintervallet for μ ?

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

Eksempel: Konfidensinterval for middelværdien i en eksponentiaffordeling

```
# Antal simulationer
k <- 100000

# Simulerer 10 observationer med den 'rigtige' intensitet k gange
sim_samples <- replicate(k, rexp(10, 1/26.08))

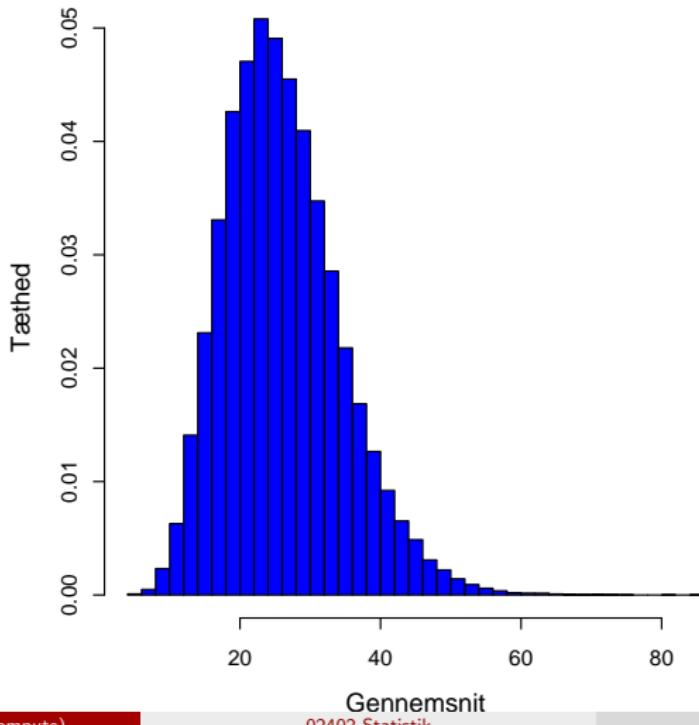
# Udregner gennemsnittet af de 10 simulerede observationer k gange
sim_means <- apply(sim_samples, 2, mean)

# Finder relevante fraktiler i fordelingen af de k simulerede gennemsnit
quantile(sim_means, c(0.025, 0.975))

## 2.5% 97.5%
## 12.59 44.63
```

Eksempel: Histogram

```
# Histogram over simulerede gennemsnit  
hist(sim_means, col = "blue", nclass = 30, main = "", prob = TRUE, xlab = "Gennemsnit", ylab = "Tæthed")
```



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.

Eksempel: Konfidensinterval for medianen i en eksponentiafordeling

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.

Vi estimerer medianen og middelværdien ud fra data:

$$q_{0.5} = 21.4 \text{ og } \hat{\mu} = \bar{x} = 26.08.$$

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.

Vi estimerer medianen og middelværdien ud fra data:

$$q_{0.5} = 21.4 \text{ og } \hat{\mu} = \bar{x} = 26.08.$$

Fordelingsantagelse:

Ventetiderne kommer fra en eksponentialfordeling.

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.

Vi estimerer medianen og middelværdien ud fra data:

$$q_{0.5} = 21.4 \text{ og } \hat{\mu} = \bar{x} = 26.08.$$

Fordelingsantagelse:

Ventetiderne kommer fra en eksponentialfordeling.

Hvad er konfidensintervallet for medianen?

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

Eksempel: Konfidensinterval for medianen i en eksponentiaffordeling

```
# Antal simulationer
k <- 100000

# Simulerer 10 observationer med den 'rigtige' intensitet k gange
sim_samples <- replicate(k, rexp(10, 1/26.08))

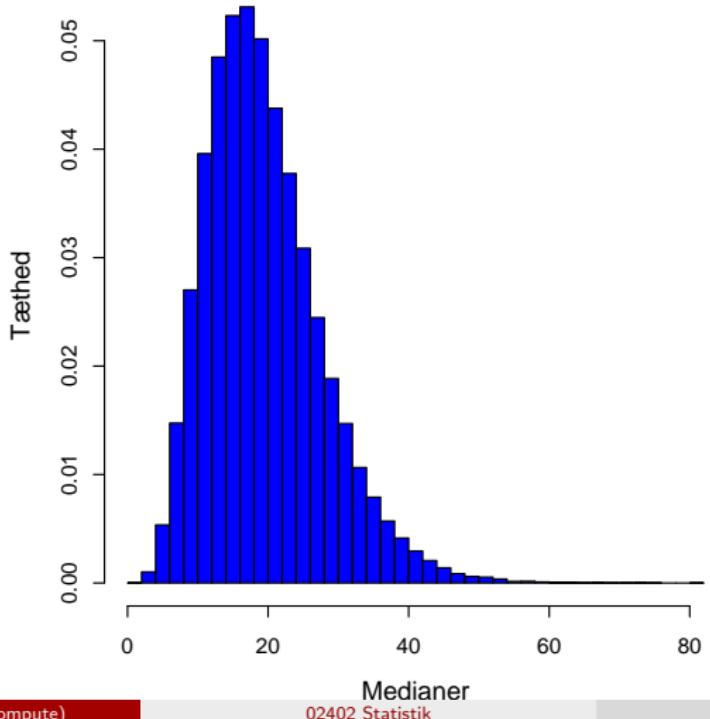
# Udregner medianen af de 10 simulerede observationer k gange
sim_medians <- apply(sim_samples, 2, median)

# Finder relevante fraktiler i fordelingen af de k simulerede medianer
quantile(sim_medians, c(0.025, 0.975))

##    2.5% 97.5%
## 7.038 38.465
```

Eksempel: Histogram

```
# Make histogram of simulated medians  
hist(sim_medians, col = "blue", nclass = 30, main = "", prob = TRUE, xlab = "Medianer", ylab = "Tæthed")
```



Konfidensinterval for en vilkårlig stikprøvefunktion (inkl. μ)

Metode 4.7: Konfidensinterval for en vilkårlig stikprøvefunktion θ ved parametrisk bootstrap

Antag at vi har faktiske observationer x_1, \dots, x_n , og at disse kommer fra en sandsynlighedsfordeling (med tæthed) f .

- ① Simulér k stikprøver af n observationer fra den antagede fordeling f , hvor middelværdien er lig \bar{x} . ^a
- ② Udregn estimatet $\hat{\theta}$ for hver af de k stikprøver, $\hat{\theta}_1^*, \dots, \hat{\theta}_k^*$.
- ③ Find $\alpha/2$ - og $(1 - \alpha/2)$ -fraktilerne i $\hat{\theta}_1^*, \dots, \hat{\theta}_k^*$, $q_{\alpha/2}^*$ og $q_{1-\alpha/2}^*$, så vi får et $(1 - \alpha)$ -konfidensinterval: $[q_{\alpha/2}^*, q_{1-\alpha/2}^*]$

^aAndre parametre/størrelser i fordelingen skal også matche data bedst muligt.

Nogle fordelinger har mere end en parameter, f.eks. har log-normalfordelingen to parametre. Mere generelt bør man anvende den såkaldte *maximum likelihood* tilgang.

Eksempel: 99% KI for Q_3 i en normalfordeling

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

# Definerer en Q3-funktion
Q3 <- function(x){ quantile(x, 0.75) }

# Antal simulationer
k <- 100000

# Simulerer 10 observationer med 'rige' parametre k gange
sim_samples <- replicate(k, rnorm(n, mean(x), sd(x)))

# Udregner Q3 af de 10 simulerede observationer k gange
simQ3s <- apply(sim_samples, 2, Q3)

# Finder relevante fraktiler i fordelingen af de k simulerede Q3
quantile(simQ3s, c(0.005, 0.995))

## 0.5% 99.5%
## 172.8 198.0
```

Konfidensinterval for en vilkårlig stikprøvefunktion (sammenligning) $\theta_1 - \theta_2$ (inkl. $\mu_1 - \mu_2$) fra to stikprøver

Metode 4.10: Konfidensinterval for en vilkårlig sammenligning $\theta_1 - \theta_2$ baseret på to stikprøver ved parametrisk bootstrap:

Antag at vi har faktiske observationer x_1, \dots, x_n , og at disse kommer fra sandsynlighedsfordelinger f_1 og f_2 . (Fordelingerne antages uafhængige)

- ① Simulér k grupper af 2 stikprøver med hhv. n_1 og n_2 observationer fra de antagede fordelinger, hvor middelværdierne er hhv. $\hat{\mu}_1 = \bar{x}$ og $\hat{\mu}_2 = \bar{y}$.
- ② Udregn forskellen mellem stikprøvefunktionerne i hver af de k stikprøver: $\hat{\theta}_{x1}^* - \hat{\theta}_{y1}^*, \dots, \hat{\theta}_{xk}^* - \hat{\theta}_{yk}^*$.
- ③ Find $\alpha/2$ - og $(1 - \alpha/2)$ -fraktilerne i disse, $q_{\alpha/2}^*$ og $q_{1-\alpha/2}^*$, så vi får et $(1 - \alpha)$ -konfidensinterval: $[q_{\alpha/2}^*, q_{1-\alpha/2}^*]$

Eksempel: Konfidensinterval for forskellen mellem middelværdierne i to eksponentialfordelinger

```
# Dag 1 data
x <- c(32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0)
n1 <- length(x)

# Dag 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)
n2 <- length(y)
```

Eksempel: Konfidensinterval for forskellen mellem middelværdierne i to eksponentialfordelinger

```
# Antal simulationer
k <- 100000

# Simulerer k par af stikprøver med hhu. n1 = 10 and n2 = 12 observationer
# fra eksponentialfordelinger med de "rigtige" intensiteter.

simX_samples <- replicate(k, rexp(n1, 1/mean(x)))
simY_samples <- replicate(k, rexp(n2, 1/mean(y)))

# Udregner forskellen mellem de simulerede middelværdier k gange
sim_dif_means <- apply(simX_samples, 2, mean) -
  apply(simY_samples, 2, mean)

# Finder relevante fraktiler i fordelingen af de k simulerede forskelle
quantile(sim_dif_means, c(0.025, 0.975))

##    2.5%  97.5%
## -40.74  14.12
```

Parametrisk bootstrapping: Et overblik

Vi antager en eller anden fordeling!

To metoder med konfidensintervaller bliver givet:

	Med en SP	Med to SP'er
Vilkårlig stikprøvefunktion	Metode 4.7	Metode 4.10

Dagsorden

- 1 Introduktion til simulation
 - Eksempel: Areal af plader
- 2 Fejlophobningslove
- 3 Parametrisk bootstrapping
 - Introduktion til bootstrapping
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver
- 4 Ikke-parametrisk bootstrapping
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver

Ikke-parametrisk bootstrapping: Et overblik

Vi antager *ikke* noget om fordelinger!

To metoder med konfidensintervaller bliver givet:

	Med en SP	Med to SP'er
Vilkårlig stikprøvefunktion	Metode 4.15	Metode 4.17

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 middelværdierne før og efter! Er der sket nogen ændring i gennemsnitsforbruget?

Eksempel: Kvinders cigaretforbrug

En parret test, *men* data er tydeligvis ikke normalfordelt!

```
# Data
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)

# Udregner forskellene
dif <- x1-x2
dif

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

# Udregner gennemsnitsforskellen
mean(dif)

## [1] 5.273
```

Eksempel: Kvinders cigaretforbrug – Ikke-parametrisk bootstrapping

```
t(replicate(5, sample(dif, replace = TRUE)))  
  
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]  
## [1,]    -2    0    9   22    0   -1    0   -2    0     3     0  
## [2,]    13    3   -2   -1   -2    7   13   -4   -2    -1     5  
## [3,]     9   -4    5   -4    5    3   -4   13    3     0    22  
## [4,]    -1   22   -2   -1   13    6   -4    0     0    -1    22  
## [5,]     9   -2   13    6    9   22    0   -1     7     7    -1
```

Eksempel: Kvinders cigaretforbrug – Resultater

Lad os finde et 95%-konfidensinterval for *middelændringen* i cigaretforbruget.

```
k = 100000
sim_samples = replicate(k, sample(dif, replace = TRUE))
sim_means = apply(sim_samples, 2, mean)
quantile(sim_means, c(0.025,0.975))

##  2.5% 97.5%
## 1.364 9.818
```

Konfidensinterval for en vilkårlig stikprøvefunktion θ (inkl. μ) fra en stikprøve

Metode 4.15: Konfidensinterval for en vilkårlig stikprøvefunktion θ ved ikke-parametrisk bootstrapping

Antag at vi har observeret x_1, \dots, x_n .

- ① Simulér k stikprøver af størrelse n ved tilfældig trækning (med tilbagelægning) fra de observerede/tilgængelige data.
- ② Udregn estimatet $\hat{\theta}$ for hver af de k stikprøver: $\hat{\theta}_1^*, \dots, \hat{\theta}_k^*$.
- ③ Find $\alpha/2$ - og $(1 - \alpha/2)$ -fraktilerne for disse, $q_{\alpha/2}^*$ og $q_{1-\alpha/2}^*$, så vi får et $(1 - \alpha)$ -konfidensinterval: $[q_{\alpha/2}^*, q_{1-\alpha/2}^*]$

Eksempel: Kvinders cigaretforbrug

Lad os finde et 95%-konfidensinterval for *medianændringen* i cigaretforbruget i eksemplet fra før.

```
k = 100000
sim_samples = replicate(k, sample(dif, replace = TRUE))
sim_medians = apply(sim_samples, 2, median)
quantile(sim_medians, c(0.025,0.975))

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

Eksempel: Tandsundhed og spædbørns brug af flaske

I et studie undersøgtes det om børn, der som spæde havde fået mælk fra flaske, 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
N	9	N	10	J	16
J	14	N	8	J	14
J	15	N	6	J	9
N	10	J	12	N	12
N	12	J	13	J	12
N	6	N	20		
J	19	J	13		

Eksempel: Tandsundhed og spædbørns brug af flaske – 95%-konfidensinterval for $\mu_1 - \mu_2$

```
# Indlæser data
x <- c(9, 10, 12, 6, 10, 8, 6, 20, 12)
y <- c(14, 15, 19, 12, 13, 13, 16, 14, 9, 12)

# 95% KI: gns. forskel ved ikke-parametrisk bootstrapping
k <- 100000

simx_samples <- replicate(k, sample(x, replace = TRUE))
simy_samples <- replicate(k, sample(y, replace = TRUE))
sim_mean_difs <- apply(simx_samples, 2, mean)-
                        apply(simy_samples, 2, mean)
quantile(sim_mean_difs, c(0.025, 0.975))

##      2.5%    97.5%
## -6.2111 -0.1111
```

Konfidensinterval for $\theta_1 - \theta_2$ (inkl. $\mu_1 - \mu_2$) ved ikke-parametrisk bootstrapping fra to stikprøver

Metode 4.17: Konfidensinterval for $\theta_1 - \theta_2$ ved ikke-parametrisk bootstrapping fra to stikprøver

Antag at vi har observationer x_1, \dots, x_n og y_1, \dots, y_n .

- ① Udtag k par bootstrap-stikprøver med hhv. n_1 og n_2 observationer fra de respektive stikprøver (ved tilfæld trækning med tilbagelægning).
- ② Udregn forskellen mellem estimaterne i hver af de k par bootstrap-stikprøver:
$$\hat{\theta}_{x1}^* - \hat{\theta}_{y1}^*, \dots, \hat{\theta}_{xk}^* - \hat{\theta}_{yk}^*.$$
- ③ Find $\alpha/2$ - og $(1 - \alpha/2)$ - fraktilerne i disse, $q_{\alpha/2}^*$ og $q_{1-\alpha/2}^*$, så vi får et $(1 - \alpha)$ -konfidensinterval:
$$\left[q_{\alpha/2}^*, q_{1-\alpha/2}^* \right]$$

Eksempel: Tandsundhed og spædbørns brug af flaske – Et 99%-konfidensinterval for median-forskellen

```
k <- 100000  
simx_samples <- replicate(k, sample(x, replace = TRUE))  
simy_samples <- replicate(k, sample(y, replace = TRUE))  
sim_median_difs <- apply(simx_samples, 2, median)-  
                      apply(simy_samples, 2, median)  
quantile(sim_median_difs, c(0.005,0.995))  
  
## 0.5% 99.5%  
##      -8       0
```

Bootstrapping: Et overblik

Vi har set 4 ikke så forskellige metode-bokse

- ① Med eller uden fordelingsantagelse (parametrisk eller ikke-parametrisk)
- ② Analyser med en eller to stikprøver (en eller to grupper)

Bootstrapping: Et overblik

Vi har set 4 ikke så forskellige metode-bokse

- ① Med eller uden fordelingsantagelse (parametrisk eller ikke-parametrisk)
- ② Analyser med en eller to stikprøver (en eller to grupper)

Bemærk:

Middelværdier (means) er også inkluderet i *vilkårlige stikprøvefunktioner* (other features). dvs. disse metoder kan også anvendes for andre analyser end for middelværdier!

Bootstrapping: Et overblik

Vi har set 4 ikke så forskellige metode-bokse

- ① Med eller uden fordelingsantagelse (parametrisk eller ikke-parametrisk)
- ② Analyser med en eller to stikprøver (en eller to grupper)

Bemærk:

Middelværdier (means) er også inkluderet i *vilkårlige stikprøvefunktioner* (other features). dvs. disse metoder kan også anvendes for andre analyser end for middelværdier!

Hypotesetest også muligt

Vi kan udføre hypotesetest ved at kigge på konfidensintervallerne!

Dagsorden

- 1 Introduktion til simulation
 - Eksempel: Areal af plader
- 2 Fejlphobningslove
- 3 Parametrisk bootstrapping
 - Introduktion til bootstrapping
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver
- 4 Ikke-parametrisk bootstrapping
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på en stikprøve
 - Konfidensinterval for en vilkårlig stikprøvefunktion baseret på to stikprøver

02402 Statistik (Polyteknisk grundlag)

Uge 8: Simpel lineær regression

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

Dagsorden

- ① Opsummering
- ② Eksempel: Højde og vægt
- ③ Lineære regressionsmodeller
- ④ Mindste kvadraters metode (Least squares)
- ⑤ Statistik og lineær regression
- ⑥ Hypotesetest og konfidensintervaller for β_0 og β_1
- ⑦ Konfidens- og prædiktionsintervaller
- ⑧ Outputtet fra summary
- ⑨ Korrelation
- ⑩ Modelkontrol - Analyse af residualer

Dagsorden

- 1 Opsummering
- 2 Eksempel: Højde og vægt
- 3 Lineære regressionsmodeller
- 4 Mindste kvadraters metode (Least squares)
- 5 Statistik og lineær regression
- 6 Hypotesetest og konfidensintervaller for β_0 og β_1
- 7 Konfidens- og prædiktionsintervaller
- 8 Outputtet fra summary
- 9 Korrelation
- 10 Modelkontrol - Analyse af residualer

Opsummering af kursets første halvdel

Sandsynlighedsregning

- Fordelinger som statistiske modeller
- Regneregler til udledninger

Hypotesetest med en eller to stikprøver

- Konfidensintervaller
- Kritiske værdier
- p -værdier

Bootstrapping

- Parametrisk bootstrapping (m. antagelser)
- Ikke-parametrisk bootstrapping (u. antagelser)

Reformulering og modeller

Lad $Y \sim N(\mu, \sigma^2)$. Man kan så opskrive en model med en middelværdi og støj:

$$Y = \mu + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

Estimation af populationsmiddelværdien svarer altså til at estimere en modelparameter.

Stikprøvegennemsnittet \bar{Y} er en middelret (unbiased) estimator for μ :

$$\mathbb{E} [\bar{Y}] = \mu.$$

Hvis variansen er kendt:

$$\frac{\bar{Y} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1^2).$$

Hvis variansen ikke er kendt (og skal estimeres med S^2):

$$\frac{\bar{Y} - \mu}{\sqrt{S^2/n}} \sim t(n-1).$$

Reformulering og modeller

Lad $X = \mu_1 + \varepsilon_1$, hvor $\varepsilon_1 \sim N(0, \sigma_1^2)$, og lad $Y = \mu_2 + \varepsilon_2$, hvor $\varepsilon_2 \sim N(0, \sigma_2^2)$.

Hypotesetest med en stikprøve

$$H_0 : \mu_1 = \mu_0,$$

$$H_1 : \mu_1 \neq \mu_0.$$

Hypotesetest med to stikprøver - Ikke parret

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

$$H_1 : \mu_1 - \mu_2 \neq \delta.$$

Hypotesetest med to stikprøver - Parret

Ny model: $Z = X - Y = \mu + \varepsilon$, hvor $\varepsilon \sim N(0, \sigma^2)$.

$$H_0 : \mu = \mu_1 - \mu_2 = \mu_0,$$

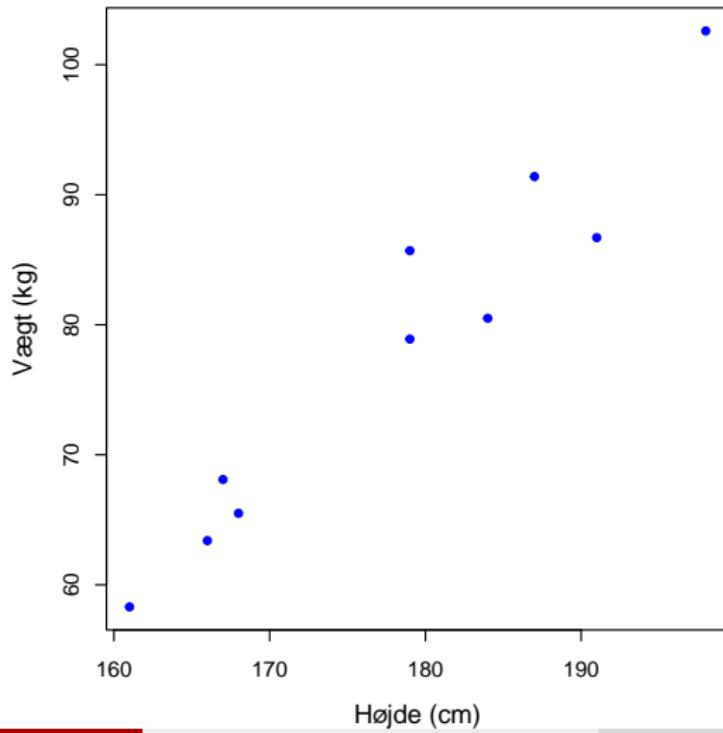
$$H_1 : \mu = \mu_1 - \mu_2 \neq \mu_0.$$

Dagsorden

- 1 Opsummering
- 2 Eksempel: Højde og vægt
- 3 Lineære regressionsmodeller
- 4 Mindste kvadraters metode (Least squares)
- 5 Statistik og lineær regression
- 6 Hypotesetest og konfidensintervaller for β_0 og β_1
- 7 Konfidens- og prædiktionsintervaller
- 8 Outputtet fra summary
- 9 Korrelation
- 10 Modelkontrol - Analyse af residualer

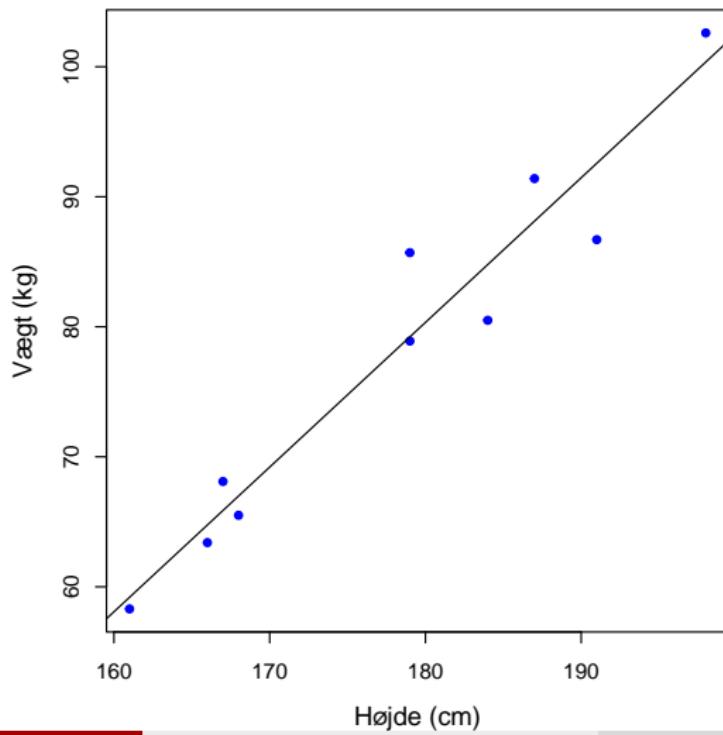
Eksempel: Højde og vægt

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



Eksempel: Højde og vægt

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



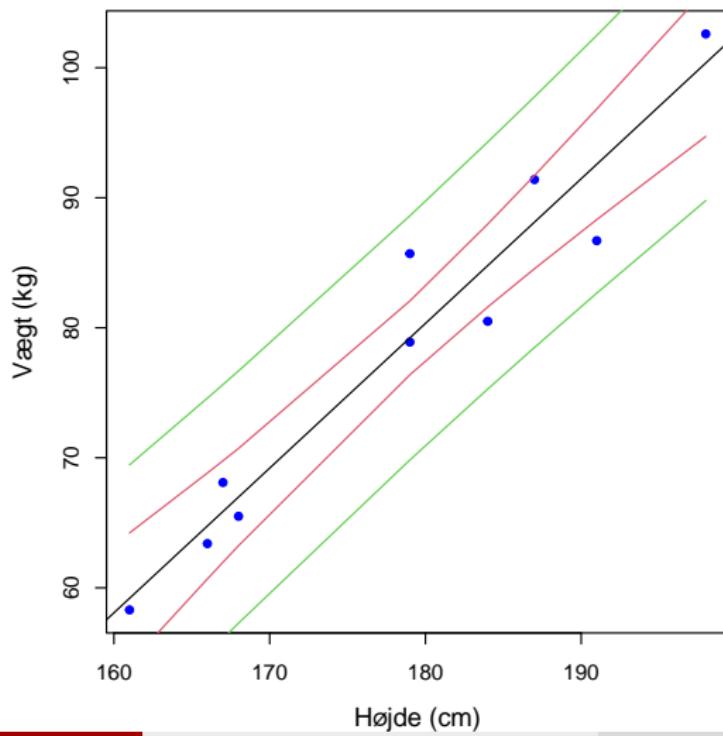
Højde (x_i)	168	161	167	179	184	166	198	187	191	179
Vægt (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9

```
summary(lm(y ~ x))
```

```
##
## 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: 3.88 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
```

Eksempel: Højde og vægt

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

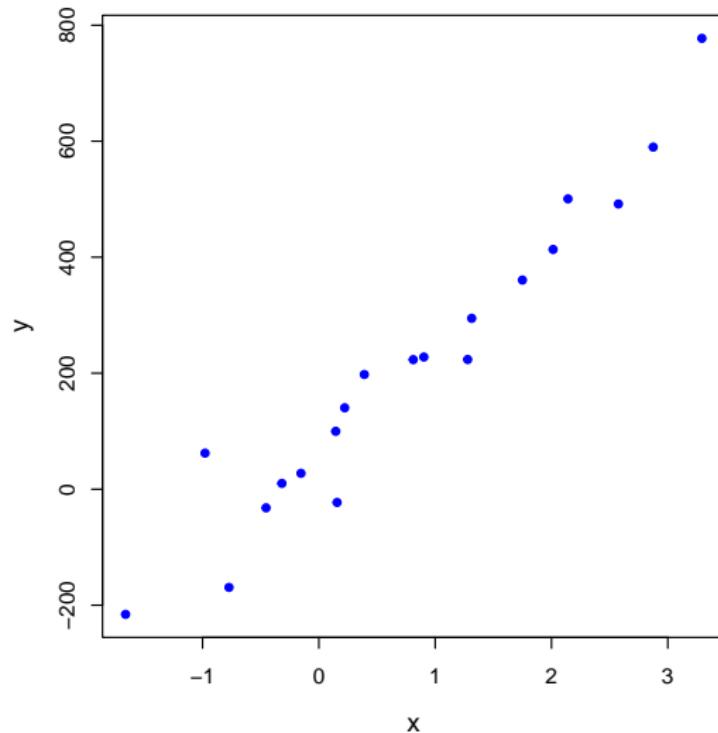


Dagsorden

- 1 Opsummering
- 2 Eksempel: Højde og vægt
- 3 Lineære regressionsmodeller
- 4 Mindste kvadraters metode (Least squares)
- 5 Statistik og lineær regression
- 6 Hypotesetest og konfidensintervaller for β_0 og β_1
- 7 Konfidens- og prædiktionsintervaller
- 8 Outputtet fra summary
- 9 Korrelation
- 10 Modelkontrol - Analyse af residualer

Et scatterplot

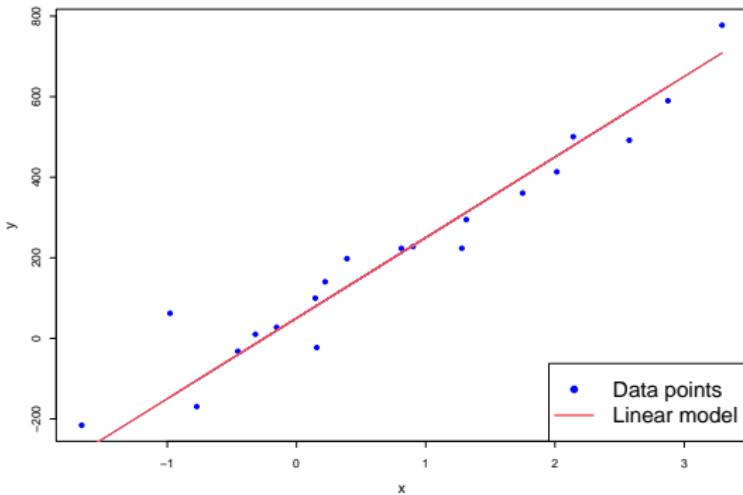
- Vi har n par datapunkter (x_i, y_i) .



En lineær model

Hvis datapunkterne ligger på en linje, kan sammenhængen mellem x - og y -værdierne beskrives ved ligningen:

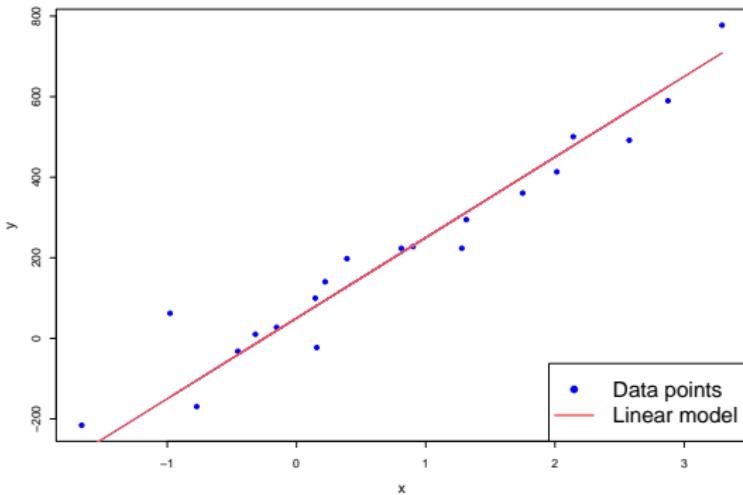
$$y_i = \beta_0 + \beta_1 x_i.$$



En lineær model

Hvis datapunkterne ligger på en linje, kan sammenhængen mellem x - og y -værdierne beskrives ved ligningen:

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



- Vi mangler en beskrivelse af den *tilfældige variation*.

Den lineære regressionsmodel

- Den *lineære regressionsmodel*:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, \dots, n).$$

- Y_i er en *afhængige variabel* (dependent/outcome variable) - En stokastisk variabel.
- x_i er en *forklarende variabel* (independent/predictor/regressor/explanatory variable, covariate) - En deterministisk værdi.
- ε_i er en *støj/afvigelse/fejl* (error) - En stokastisk variabel.
- Vi antager, at fejlene ε_i ($i = 1, \dots, n$) er *uafhængige og ensfordelte* (i.i.d.) med $\varepsilon_i \sim N(0, \sigma^2)$.

Den lineære regressionsmodel

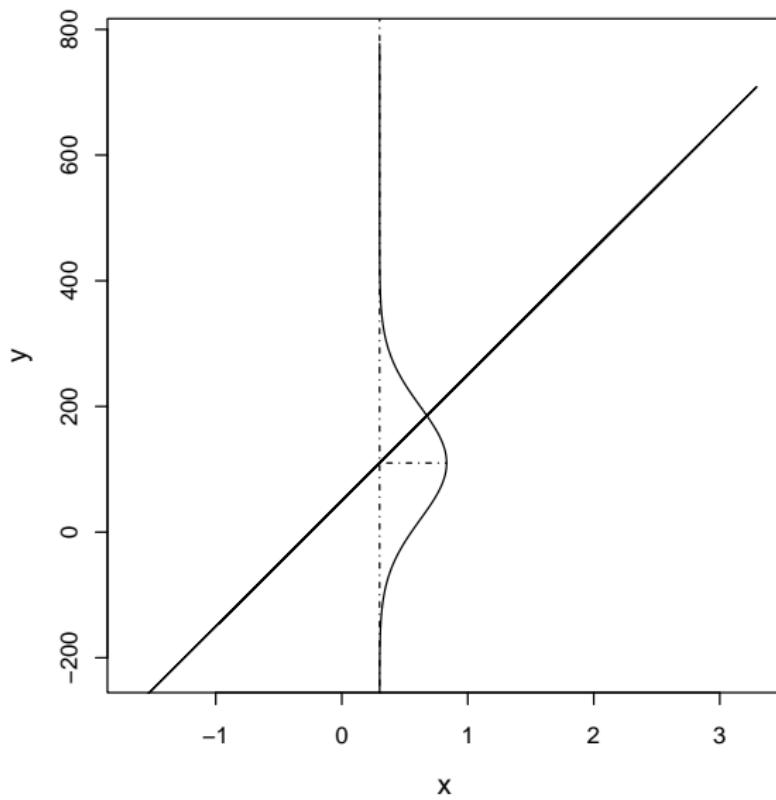
- Den *lineære regressionsmodel*:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, \dots, n).$$

- Y_i er en *afhængige variabel* (dependent/outcome variable) - En stokastisk variabel.
- x_i er en *forklarende variabel* (independent/predictor/regressor/explanatory variable, covariate) - En deterministisk værdi.
- ε_i er en *støj/afvigelse/fejl* (error) - En stokastisk variabel.
- Vi antager, at fejlene ε_i ($i = 1, \dots, n$) er *uafhængige og ensfordelte* (i.i.d.) med $\varepsilon_i \sim N(0, \sigma^2)$.

Overvej: *Hvilken slags fordeling følger Y_i ? Er Y_i 'erne ensfordelte?*

Illustration af den statistiske model



Dagsorden

- 1 Opsummering
- 2 Eksempel: Højde og vægt
- 3 Lineære regressionsmodeller
- 4 Mindste kvadraters metode (Least squares)
- 5 Statistik og lineær regression
- 6 Hypotesetest og konfidensintervaller for β_0 og β_1
- 7 Konfidens- og prædiktionsintervaller
- 8 Outputtet fra summary
- 9 Korrelation
- 10 Modelkontrol - Analyse af residualer

Mindste kvadraters metode

- Vi ønsker at estimere parametrene β_0 og β_1 .

Mindste kvadraters metode

- Vi ønsker at estimere parametrene β_0 og β_1 .
- God ide: Lad os minimere variansen af residualerne/afvigelsen (σ^2).

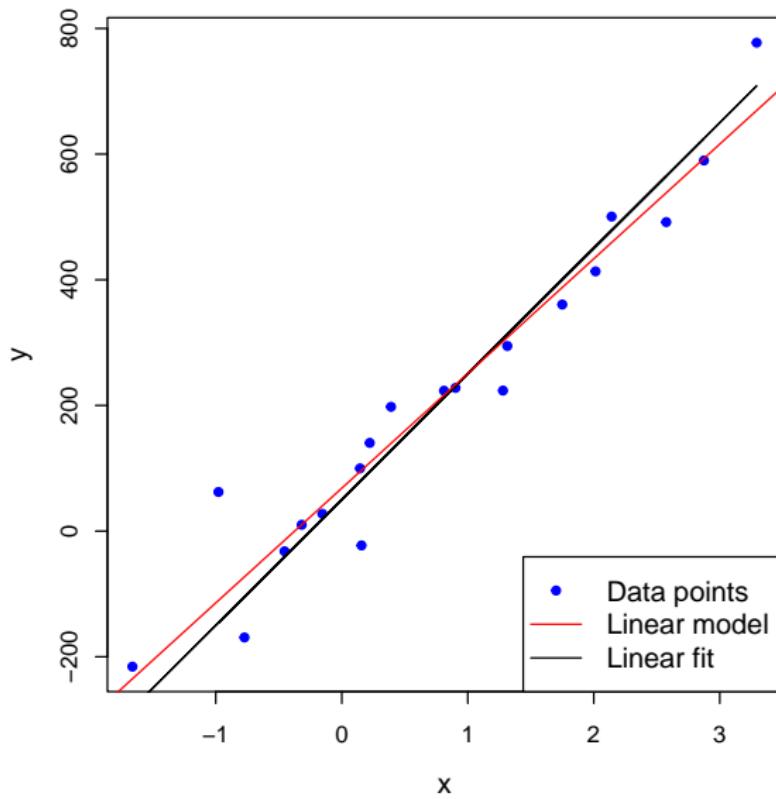
Mindste kvadraters metode

- Vi ønsker at estimere parametrene β_0 og β_1 .
- God ide: Lad os minimere variansen af residualerne/afvigelsen (σ^2).
- Vi minimerer summen af de kvadrerede afvigelser (Residual Sum of Squares, RSS):

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

Dvs. at vi vælger $\hat{\beta}_0$ og $\hat{\beta}_1$, sådan at de minimerer RSS.

Illustration af model, data og fit



'Least squares'-estimatorer

Sætning 5.4 (her for estimatorer, som i bogen)

'Least squares'-estimatorerne for β_0 og β_1 er givet ved:

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

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

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

'Least squares'-estimator

Sætning 5.4 (her for estimator)

'Least squares'-estimaterne for β_0 og β_1 er givet ved:

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

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

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

Eksempel i R

```
set.seed(100)

# Simuler værdier for x
x <- runif(n = 20, min = -2, max = 4)

# Simuler værdier for y
beta0 <- 50; beta1 <- 200; sigma <- 90
y <- beta0 + beta1 * x + rnorm(n = length(x), mean = 0, sd = sigma)

# Scatter plot af y mod x
plot(x, y)

# Find LS estimerater (Sætning 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 finde LS estimerater
lm(y ~ x)

# Plot den bedste rette linje
abline(lm(y ~ x), col="red")
```

Dagsorden

- 1 Opsummering
- 2 Eksempel: Højde og vægt
- 3 Lineære regressionsmodeller
- 4 Mindste kvadraters metode (Least squares)
- 5 Statistik og lineær regression
- 6 Hypotesetest og konfidensintervaller for β_0 og β_1
- 7 Konfidens- og prædiktionsintervaller
- 8 Outputtet fra summary
- 9 Korrelation
- 10 Modelkontrol - Analyse af residualer

Variation i parameterestimaterne

Vi udtager en ny stikprøve

Vil estimaterne af $\hat{\beta}_0$ and $\hat{\beta}_1$ så blive de samme?

Variation i parameterestimaterne

Vi udtager en ny stikprøve

Vil estimaterne af $\hat{\beta}_0$ and $\hat{\beta}_1$ så blive de samme?

Nej - Der er en variation!

En ny stikprøve giver anledning til nye realiseringer af estimatorerne, dvs. nye estimerter.

Variation i parameterestimaterne

Vi udtager en ny stikprøve

Vil estimaterne af $\hat{\beta}_0$ and $\hat{\beta}_1$ så blive de samme?

Nej - Der er en variation!

En ny stikprøve giver anledning til nye realiseringer af estimatorerne, dvs. nye estimer.

Hvad er fordelingerne af parameterestimatorerne?

Vi skal kende dem for at lave hypotesetest mv.

Variation i parameterestimaterne

Vi udtager en ny stikprøve

Vil estimaterne af $\hat{\beta}_0$ and $\hat{\beta}_1$ så blive de samme?

Nej - Der er en variation!

En ny stikprøve giver anledning til nye realiseringer af estimatorerne, dvs. nye estimer.

Hvad er fordelingerne af parameterestimatorerne?

Vi skal kende dem for at lave hypotesetest mv.

Simulation kan hjælpe os! R kan give os en intuitiv ide!

Fordelingerne af $\hat{\beta}_0$ og $\hat{\beta}_1$

Estimatorerne $\hat{\beta}_0$ og $\hat{\beta}_1$ er normalfordelte med varianserne:

Sætning 5.8 (første del)

$$V[\hat{\beta}_0] = \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{S_{xx}},$$

$$V[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}},$$

$$Cov[\hat{\beta}_0, \hat{\beta}_1] = -\frac{\bar{x}\sigma^2}{S_{xx}}.$$

Kovariansen $Cov[\hat{\beta}_0, \hat{\beta}_1]$ gør vi ikke mere ud af her.

Standardafvigelserne for $\hat{\beta}_0$ og $\hat{\beta}_1$

Sætning 5.8 (anden del)

Da σ^2 er ukendt, benyttes det *centrale estimat for σ^2* :

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}.$$

Vi estimerer altså variansen (standardafvigelsen) for fejlen og derved også varianserne (standardafvigelserne) for estimatorerne. Vi benævner disse $\hat{\sigma}_{\hat{\beta}_0}^2$ og $\hat{\sigma}_{\hat{\beta}_1}^2$.

Standardafvigelserne for $\hat{\beta}_0$ og $\hat{\beta}_1$

Sætning 5.8 (anden del)

Da σ^2 er ukendt, benyttes det *centrale estimat for σ^2* :

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}.$$

Vi estimerer altså variansen (standardafvigelsen) for fejlen og derved også varianserne (standardafvigelserne) for estimatorerne. Vi benævner disse $\hat{\sigma}_{\beta_0}^2$ og $\hat{\sigma}_{\beta_1}^2$.

Man får følgende estimerater af standardafvigelserne for $\hat{\beta}_0$ og $\hat{\beta}_1$:

$$\hat{\sigma}_{\beta_0} = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}, \quad \hat{\sigma}_{\beta_1} = \hat{\sigma} \sqrt{\frac{1}{S_{xx}}}.$$

Dagsorden

- 1 Opsummering
- 2 Eksempel: Højde og vægt
- 3 Lineære regressionsmodeller
- 4 Mindste kvadraters metode (Least squares)
- 5 Statistik og lineær regression
- 6 Hypotesetest og konfidensintervaller for β_0 og β_1
- 7 Konfidens- og prædiktionsintervaller
- 8 Outputtet fra summary
- 9 Korrelation
- 10 Modelkontrol - Analyse af residualer

Hypotesetest for β_0 og β_1

Vi kan udføre hypotesetest for parametrene i en lineær regressionsmodel:

$$H_{0,i} : \beta_i = \beta_{0,i},$$

$$H_{1,i} : \beta_i \neq \beta_{0,i}.$$

Hypotesetest for β_0 og β_1

Vi kan udføre hypotesetest for parametrene i en lineær regressionsmodel:

$$H_{0,i} : \beta_i = \beta_{0,i},$$

$$H_{1,i} : \beta_i \neq \beta_{0,i}.$$

Sætning 5.12

Under nulhypoteserne ($\beta_0 = \beta_{0,0}$ og $\beta_1 = \beta_{0,1}$) er teststørrelserne

$$T_{\beta_0} = \frac{\hat{\beta}_0 - \beta_{0,0}}{\hat{\sigma}_{\beta_0}}, \quad T_{\beta_1} = \frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}},$$

t -fordelte med $n - 2$ frihedsgrader.

Hypotest for β_0 og β_1

- Se eksempel 5.13 for et eksempel på en hypotesetest.
- Test om parametrene er signifikant forskellige fra 0:

$$H_{0,i} : \beta_i = 0, \quad H_{1,i} : \beta_i \neq 0.$$

```
# Indlæs data
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
y <- c(65.5, 58.3, 68.1, 85.7, 80.5, 63.4, 102.6, 91.4, 86.7, 78.9)

# Fit model til data
fit <- lm(y ~ x)

# Find teststørrelser og p-værdier
summary(fit)
```

Konfidensintervaller for β_0 og β_1

Metode 5.15

$(1 - \alpha)$ konfidensintervaller for β_0 og β_1 er givet ved:

$$\hat{\beta}_0 \pm t_{1-\alpha/2} \hat{\sigma}_{\beta_0},$$

$$\hat{\beta}_1 \pm t_{1-\alpha/2} \hat{\sigma}_{\beta_1},$$

hvor $t_{1-\alpha/2}$ er $(1 - \alpha/2)$ -fraktilen i t -fordelingen med $n - 2$ frihedsgrader.

- Husk at $\hat{\sigma}_{\beta_0}$ og $\hat{\sigma}_{\beta_1}$ kan findes med ligninger 5-43 og 5-44.
- I R kan $\hat{\sigma}_{\beta_0}$ og $\hat{\sigma}_{\beta_1}$ aflæses under "Std. Error" fra `summary(fit)`.

Dagsorden

- 1 Opsummering
- 2 Eksempel: Højde og vægt
- 3 Lineære regressionsmodeller
- 4 Mindste kvadraters metode (Least squares)
- 5 Statistik og lineær regression
- 6 Hypotesetest og konfidensintervaller for β_0 og β_1
- 7 Konfidens- og prædiktionsintervaller**
- 8 Outputtet fra summary
- 9 Korrelation
- 10 Modelkontrol - Analyse af residualer

Metode 5.18: Konfidensinterval for regressionslinjen

En simpel lineær regressionsmodel kan skrives som

$$Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

eller

$$Y \sim N(\mu(x), \sigma^2),$$

hvor $\mu(x) = \beta_0 + \beta_1 x$.

For $x = x_0$ kan vi derfor finde et konfidensinterval for $\mu(x_0) = \beta_0 + \beta_1 x_0$.

($1 - \alpha$)-konfidensintervallet for regressionslinjen i $x = x_0$ (for $\mu(x_0)$) kan findes ved:

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2} \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}.$$

Metode 5.18: Prædiktionsinterval for en ny observation

- Vi ønsker et prædiktionsinterval for en ny observation Y_0 for $x = x_0$.
- $(1 - \alpha)$ -prædiktionsintervallet for en ny observation Y_0 for $x = x_0$ kan findes ved:

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}.$$

- Prædiktionsintervallet vil $100(1 - \alpha)\%$ af gangene indeholde det observerede y_0 .
- For fastholdt α er prædiktionsintervallet større end konfidensintervallet.

Eksempel: Konfidensinterval for linjen

```
# Simuler 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), sd = sigma)

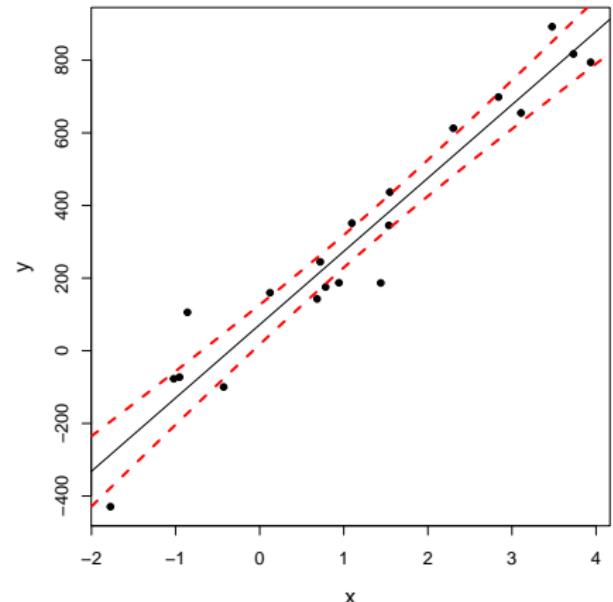
# Brug lm() til at fitte modellen
fit <- lm(y ~ x)

# Lav x-værdier til plot
xval <- seq(from = -2, to = 6, length.out = 100)

# Brug predict-funktionen
CI <- predict(fit, newdata = data.frame(x = xval),
              interval = "confidence",
              level = 0.95)

# Tjek værdierne
head(CI)

# Plot data, regressionslinjen og intervallerne
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: Prædiktionsinterval for observation

```
# Simuler 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), sd = sigma)

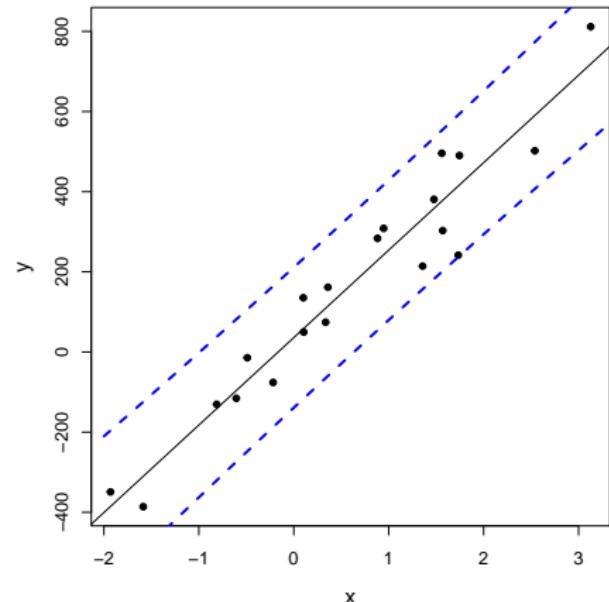
# Brug lm() til at fitte modellen
fit <- lm(y ~ x)

# Lav x-værdier til plot
xval <- seq(from = -2, to = 6, length.out = 100)

# Brug predict-funktionen
PI <- predict(fit, newdata = data.frame(x = xval),
              interval = "prediction",
              level = 0.95)

# Tjek værdierne
head(PI)

# Plot data, regressionslinjen og intervallerne
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)
```



Konfidens- og prædiktionsintervaller

- Konfidensintervallet angiver usikkerheden for *regressionslinjen*.
- Prædiktionsintervallet angiver usikkerheden for en *ny observation*.
- Prædiktionsintervallet kan aldrig blive mindre end den tilfældige variation i data (altså den fra fejlleddet ε).

Dagsorden

- 1 Opsummering
- 2 Eksempel: Højde og vægt
- 3 Lineære regressionsmodeller
- 4 Mindste kvadraters metode (Least squares)
- 5 Statistik og lineær regression
- 6 Hypotesetest og konfidensintervaller for β_0 og β_1
- 7 Konfidens- og prædiktionsintervaller
- 8 Outputtet fra summary**
- 9 Korrelation
- 10 Modelkontrol - Analyse af residualer

Hvad får vi ud af `summary()` i R?

```
summary(fit)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -172.5  -67.2   16.7   58.9  119.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.9       20.0    1.79    0.09 .
## x           218.3      14.0   15.56   7e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 80.9 on 18 degrees of freedom
## Multiple R-squared:  0.931, Adjusted R-squared:  0.927
## F-statistic: 242 on 1 and 18 DF,  p-value: 6.97e-12
```

Summary(lm(y~x))

• Residuals: Min 1Q Median 3Q Max

Summary(lm(y~x))

- Residuals: Min 1Q Median 3Q Max

Residualernes: Minimum, 1. kvartil, Median, 3. kvartil, Maximum

Summary(lm(y~x))

- Residuals: Min 1Q Median 3Q Max

Residualernes: Minimum, 1. kvartil, Median, 3. kvartil, Maximum

- Coefficients:

Estimate	Std. Error	t value	Pr(> t)	"stars"
----------	------------	---------	----------	---------

Summary(lm(y~x))

- Residuals: Min 1Q Median 3Q Max

Residualernes: Minimum, 1. kvartil, Median, 3. kvartil, Maximum

- Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	"stars"
--	----------	------------	---------	----------	---------

Koefficienternes:

$\hat{\beta}_i$	$\hat{\sigma}_{\beta_i}$	t_{obs}	p -værdi
-----------------	--------------------------	------------------	------------

- Testen er $H_{0,i} : \beta_i = 0$ mod $H_{1,i} : \beta_i \neq 0$.
- Stjernerne er sat efter p -værdien.

Summary(lm(y~x))

- Residuals: Min 1Q Median 3Q Max

Residualernes: Minimum, 1. kvartil, Median, 3. kvartil, Maximum

- Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	"stars"
--	----------	------------	---------	----------	---------

Koefficienternes:

$\hat{\beta}_i$	$\hat{\sigma}_{\beta_i}$	t_{obs}	p -værdi
-----------------	--------------------------	------------------	------------

- Testen er $H_{0,i} : \beta_i = 0$ mod $H_{1,i} : \beta_i \neq 0$.
- Stjernerne er sat efter p -værdien.
- Residual standard error: XXX on XXX degrees of freedom

Summary(lm(y~x))

- Residuals: Min 1Q Median 3Q Max

Residualernes: Minimum, 1. kvartil, Median, 3. kvartil, Maximum

- Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	"stars"
--	----------	------------	---------	----------	---------

Koefficienternes:

$\hat{\beta}_i$	$\hat{\sigma}_{\beta_i}$	t_{obs}	p -værdi
-----------------	--------------------------	------------------	------------

- Testen er $H_{0,i} : \beta_i = 0$ mod $H_{1,i} : \beta_i \neq 0$.
- Stjernerne er sat efter p -værdien.
- Residual standard error: XXX on XXX degrees of freedom
 $\varepsilon_i \sim N(0, \sigma^2)$: outputtet viser $\hat{\sigma}$ og antallet af frihedsgrader v brugt i hypotesetest samt konfidens- og prædiktionsintervaller

Summary(lm(y~x))

- Residuals: Min 1Q Median 3Q Max

Residualernes: Minimum, 1. kvartil, Median, 3. kvartil, Maximum

- Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	"stars"
--	----------	------------	---------	----------	---------

Koefficienternes:

$\hat{\beta}_i$	$\hat{\sigma}_{\beta_i}$	t_{obs}	p -værdi
-----------------	--------------------------	------------------	------------

- Testen er $H_{0,i} : \beta_i = 0$ mod $H_{1,i} : \beta_i \neq 0$.
- Stjernerne er sat efter p -værdien.
- Residual standard error: XXX on XXX degrees of freedom
 $\varepsilon_i \sim N(0, \sigma^2)$: outputtet viser $\hat{\sigma}$ og antallet af frihedsgrader v brugt i hypotesetest samt konfidens- og prædiktionsintervaller
- Multiple R-squared: XXX

Summary(lm(y~x))

- Residuals: Min 1Q Median 3Q Max

Residualernes: Minimum, 1. kvartil, Median, 3. kvartil, Maximum

- Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	"stars"
--	----------	------------	---------	----------	---------

Koefficienternes:

$\hat{\beta}_i$	$\hat{\sigma}_{\beta_i}$	t_{obs}	p -værdi
-----------------	--------------------------	------------------	------------

- Testen er $H_{0,i} : \beta_i = 0$ mod $H_{1,i} : \beta_i \neq 0$.
- Stjernerne er sat efter p -værdien.
- Residual standard error: XXX on XXX degrees of freedom
 $\varepsilon_i \sim N(0, \sigma^2)$: outputtet viser $\hat{\sigma}$ og antallet af frihedsgrader v brugt i hypotesetest samt konfidens- og prædiktionsintervaller
- Multiple R-squared: XXX
 Forklaret varians R^2

Summary(lm(y~x))

- Residuals: Min 1Q Median 3Q Max

Residualernes: Minimum, 1. kvartil, Median, 3. kvartil, Maximum

- Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	"stars"
--	----------	------------	---------	----------	---------

Koefficienternes:

$\hat{\beta}_i$	$\hat{\sigma}_{\beta_i}$	t_{obs}	p -værdi
-----------------	--------------------------	------------------	------------

- Testen er $H_{0,i} : \beta_i = 0$ mod $H_{1,i} : \beta_i \neq 0$.
- Stjernerne er sat efter p -værdien.
- Residual standard error: XXX on XXX degrees of freedom
 $\epsilon_i \sim N(0, \sigma^2)$: outputtet viser $\hat{\sigma}$ og antallet af frihedsgrader v brugt i hypotesetest samt konfidens- og prædiktionsintervaller
- Multiple R-squared: XXX
 Forklaret varians R^2
- Resten bruger vi ikke kurset.

Dagsorden

- 1 Opsummering
- 2 Eksempel: Højde og vægt
- 3 Lineære regressionsmodeller
- 4 Mindste kvadraters metode (Least squares)
- 5 Statistik og lineær regression
- 6 Hypotesetest og konfidensintervaller for β_0 og β_1
- 7 Konfidens- og prædiktionsintervaller
- 8 Outputtet fra summary
- 9 Korrelation
- 10 Modelkontrol - Analyse af residualer

Forklaret varians og korrelation

- Den forklarede varians i en model er R^2 (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 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

- Andelen af den totale varians, der er forklaret med modellen.

Forklaret varians og korrelation

- Korrelationen ρ er et mål for *lineær sammenhæng* mellem to stokastiske variable.
- Den estimerede (dvs. empiriske) korrelation opfylder

$$\hat{\rho} = R = \sqrt{R^2} sgn(\hat{\beta}_1),$$

hvor $sgn(\hat{\beta}_1)$ er -1 for $\hat{\beta}_1 \leq 0$ og 1 for $\hat{\beta}_1 > 0$

Forklaret varians og korrelation

- Korrelationen ρ er et mål for *lineær sammenhæng* mellem to stokastiske variable.
- Den estimerede (dvs. empiriske) korrelation opfylder

$$\hat{\rho} = R = \sqrt{R^2} \operatorname{sgn}(\hat{\beta}_1),$$

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.

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 β_1 er hældningen i den simple lineære regressionsmodel.

Eksempel: Korrelation og R^2 for højde/vægt data

```
# Indlæs data

x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
y <- c(65.5, 58.3, 68.1, 85.7, 80.5, 63.4, 102.6, 91.4, 86.7, 78.9)

# Fit modellen
fit <- lm(y ~ x)

# Scatter plot af data mod regressionslinjen
plot(x,y, xlab = "Height", ylab = "Weight")
abline(fit, col="red")

# Summary
summary(fit)

# Korrelationen mellem X og Y
cor(x,y)

# Den kuadrerede korrelation er "Multiple R-squared"
cor(x,y)^2
```

Dagsorden

- 1 Opsummering
- 2 Eksempel: Højde og vægt
- 3 Lineære regressionsmodeller
- 4 Mindste kvadraters metode (Least squares)
- 5 Statistik og lineær regression
- 6 Hypotesetest og konfidensintervaller for β_0 og β_1
- 7 Konfidens- og prædiktionsintervaller
- 8 Outputtet fra summary
- 9 Korrelation
- 10 Modelkontrol - Analyse af residualer

Residualanalyse

Metode 5.28

- Undersøg normalitetsantagelse med et qq-plot.
- Undersøg evt. systematiske afvigelser ved at plotte residualerne (e_i) som en funktion af de fittede værdier (\hat{y}_i).

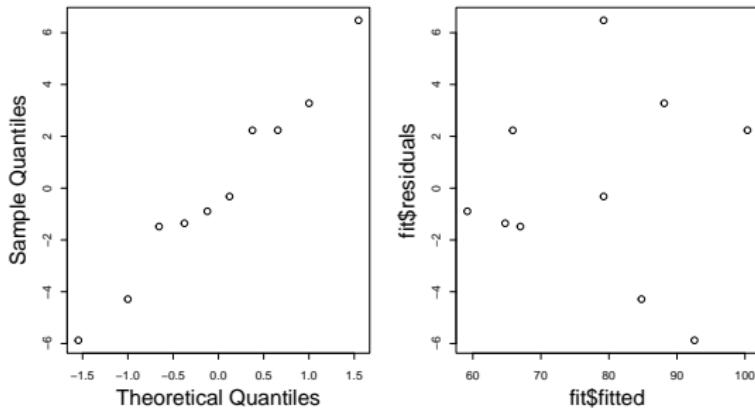
(Metode 5.29)

- Er uafhængighedsantagelsen rimelig?

Modelkontrol i R

```
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
y <- c(65.5, 58.3, 68.1, 85.7, 80.5, 63.4, 102.6, 91.4, 86.7, 78.9)
fit <- lm(y ~ x)

par(mfrow = c(1, 2))
qqnorm(fit$residuals, main = "", cex.lab = 1.5)
plot(fit$fitted, fit$residuals, cex.lab = 1.5)
```



Dagsorden

- 1 Opsummering
- 2 Eksempel: Højde og vægt
- 3 Lineære regressionsmodeller
- 4 Mindste kvadraters metode (Least squares)
- 5 Statistik og lineær regression
- 6 Hypotesetest og konfidensintervaller for β_0 og β_1
- 7 Konfidens- og prædiktionsintervaller
- 8 Outputtet fra summary
- 9 Korrelation
- 10 Modelkontrol - Analyse af residualer

02402 Statistik (Polyteknisk grundlag)

Uge 9: Multipel lineær regression

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

Dagsorden

- 1 Opsummering
- 2 Multipel lineær regression
- 3 Modeludvælgelse (Model selection)
- 4 Modelkontrol (Analyse af residualerne)
- 5 Kurvelinearitet
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet
- 8 Den 'samlede' regressionsmetode

Dagsorden

- 1 Opsummering
- 2 Multipel lineær regression
- 3 Modeludvælgelse (Model selection)
- 4 Modelkontrol (Analyse af residualerne)
- 5 Kurvilinearitet
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet
- 8 Den 'samlede' regressionsmetode

Statistiske modeller

Gennemsnittet

$$Y_i = \mu + \varepsilon_i.$$

Simpel lineær regression

$$\begin{aligned} Y_i &= \mu_i + \varepsilon_i, \\ \mu_i &= \beta_0 + \beta_1 x_i. \end{aligned}$$

Multipel lineær regression

$$\begin{aligned} Y_i &= \mu_i + \varepsilon_i, \\ \mu_i &= \beta_0 + \sum_{j=1}^p \beta_j x_{i,j}. \end{aligned}$$

Fejlene er uafhængige og følger en $N(0, \sigma^2)$ -fordeling.

Terminologi

Fejl: Forskel mellem en sand værdi og en observation

$$\varepsilon = Y - \mu.$$

Residual: Forskel mellem en prædikteret (fittet) værdi og en observation

$$e = Y - \hat{Y}.$$

I en simpel lineær regression har man f.eks.

$$\mu = \beta_0 + \beta_1 x,$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Vi kan sige, at residualerne estimerer fejlene.

Estimation

Modelparametrene kan estimeres med mindste kvadraters metode.

(I disse modeller er LS (Least squares) og ML (Maximum likelihood) estimatorerne de samme)

Estimatorerne $(\hat{\beta}_0, \hat{\beta}_1)$ findes som løsningen til et minimeringsproblem. De vælges således, at kvadraterne af residualerne minimeres:

$$(\hat{\beta}_0, \hat{\beta}_1) = \operatorname{argmin} \text{RSS}(a, b),$$

hvor RSS (Residual Sum of Squares) er defineret som:

$$\text{RSS}(a, b) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - [a + bx_i])^2.$$

Eksempel - Indledning

Baggrund

En bilfabrikant lover, at en bestemt model kan køre mindst 20 km per liter diesel ved bykørsel.

Indsamling af data

Man ønsker at undersøge, hvorvidt påstanden er korrekt, hvorfor man har kørt 25 ture af varierende længde. Efter hver tur har man målt rutens længde og brændstofforbruget.

Model

Man opstiller en lineær regressions model under de sædvanlige antagelser:

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

Eksempel - Indledning

Baggrund

En bilfabrikant lover, at en bestemt model kan køre mindst 20 km per liter diesel ved bykørsel.

Indsamling af data

Man ønsker at undersøge, hvorvidt påstanden er korrekt, hvorfor man har kørt 25 ture af varierende længde. Efter hver tur har man målt rutens længde og brændstofferbruget.

Model

Man opstiller en lineær regressions model under de sædvanlige antagelser:

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

Spørgsmål

Hvad repræsenterer de forskellige variable?

Er forudsætningerne opfyldt?

Hvilken nulhypotese ønsker man at undersøge?

Eksempel - Hypotesetest

Vi ønsker at teste nulhypotesen

$$H_0 : \beta_1 = 0.05$$

mod en tosidet modhypoteze på et 5% signifikansniveau.

Hvilke tal skal vi bruge til testen?

Hvad bliver den observerede teststørrelse?

Hvilken fordeling skal teststørrelsen sammenlignes mod?

Eksempel - Hypotesetest

Vi ønsker at teste nulhypotesen

$$H_0 : \beta_1 = 0.05$$

mod en tosidet modhypoteze på et 5% signifikansniveau.

Hvilke tal skal vi bruge til testen?

$$\hat{\beta}_1 = 0.0527, \quad \hat{\sigma}_{\beta_1} = 0.0015.$$

Hvad bliver den observerede teststørrelse?

Hvilken fordeling skal teststørrelsen sammenlignes mod?

Eksempel - Hypotesetest

Vi ønsker at teste nulhypotesen

$$H_0 : \beta_1 = 0.05$$

mod en tosidet modhypoteze på et 5% signifikansniveau.

Hvilke tal skal vi bruge til testen?

$$\hat{\beta}_1 = 0.0527, \quad \hat{\sigma}_{\beta_1} = 0.0015.$$

Hvad bliver den observerede teststørrelse?

$$T_{\beta_1} = \frac{0.0527 - 0.05}{0.0015} = \frac{0.0027}{0.0015} = \frac{27}{15} = 1.8.$$

Hvilken fordeling skal teststørrelsen sammenlignes mod?

Eksempel - Hypotesetest

Vi ønsker at teste nulhypotesen

$$H_0 : \beta_1 = 0.05$$

mod en tosidet modhypoteze på et 5% signifikansniveau.

Hvilke tal skal vi bruge til testen?

$$\hat{\beta}_1 = 0.0527, \quad \hat{\sigma}_{\beta_1} = 0.0015.$$

Hvad bliver den observerede teststørrelse?

$$T_{\beta_1} = \frac{0.0527 - 0.05}{0.0015} = \frac{0.0027}{0.0015} = \frac{27}{15} = 1.8.$$

Hvilken fordeling skal teststørrelsen sammenlignes mod?
En t -fordeling med 23 frihedsgrader.

Eksempel: Ozonkoncentration

Vi har et sæt af sammenhængende målinger af: logaritmen af ozonkoncentration ($\log(\text{ppb})$), 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: Ozonkoncentration

```
## See info about data
?airquality
## Copy the data
Air <- airquality
## Remove rows with at least one NA value
Air <- na.omit(Air)

## Remove one outlier
Air <- Air[-which(Air$Ozone == 1), ]

## Check the empirical density
hist(Air$Ozone, probability=TRUE, xlab="Ozon", main="")

## Concentrations are positive and very skewed, let's
## log-transform right away:
## (although really one could wait and check residuals from models)
Air$logOzone <- log(Air$Ozone)
## Bedre epdf?
hist(Air$logOzone, probability=TRUE, xlab="log Ozone", main="")

## Make a time variable (R timeclass, see ?POSIXct)
Air$t <- ISOdate(1973, Air$Month, Air$Day)
## Keep only some of the columns
Air <- Air[,c(7,4,3,2,8)]
## New names of the columns
names(Air) <- c("logOzone","temperature","wind","radiation","t")

## What's in Air?
str(Air)
Air
head(Air)
tail(Air)

## Typically one would begin with a pairs plot
pairs(Air, panel = panel.smooth, main = "airquality data")
```

Fit modellen i R

```
#####
## See the relation between ozone and temperature
plot(Air$temperature, Air$logOzone, xlab="Temperature", ylab="Ozon")

## Correlation
cor(Air$logOzone, Air$temperature)

## Fit a simple linear regression model
summary(lm(logOzone ~ temperature, data=Air))

## Add a vector with random values, is there a significant linear relation?
## ONLY for ILLUSTRATION purposes
Air$noise <- rnorm(nrow(Air))
plot(Air$logOzone, Air$noise, xlab="Noise", ylab="Ozon")
cor(Air$logOzone, Air$noise)
summary(lm(logOzone ~ noise, data=Air))
```

Alternativer

Vi kan også lave en simpel lineær regressionsmodel med de to andre forklarende variable:

```
#####
## With each of the other two independent variables

## Simple linear regression model with the wind speed
plot(Air$logOzone, Air$wind, xlab="logOzone", ylab="Wind speed")
cor(Air$logOzone, Air$wind)
summary(lm(logOzone ~ wind, data=Air))

## Simple linear regression model with the radiation
plot(Air$logOzone, Air$radiation, xlab="logOzone", ylab="Radiation")
cor(Air$logOzone, Air$radiation)
summary(lm(logOzone ~ radiation, data=Air))
```

Dagsorden

- 1 Opsummering
- 2 Multipel lineær regression
- 3 Modeludvælgelse (Model selection)
- 4 Modelkontrol (Analyse af residualerne)
- 5 Kurvelinearitet
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet
- 8 Den 'samlede' regressionsmetode

Multipel lineær regression

En udvidelse af den simple lineære regressionsmodel, hvor flere *forklarende/uafhængige* variable inkluderes.

- I en multipel lineær regression med p *forklarende* variable benævnes de deterministiske variable x_1, x_2, \dots, x_p .
- Vi modellerer en *lineær sammenhæng* mellem Y og x_1, x_2, \dots, x_p , ved en regressionsmodel på formen

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i,$$

hvor fejlene er uafhængige og ensfordelte med $\varepsilon_i \sim N(0, \sigma^2)$.

Estimation

Estimation og prædiktion udføres ligesom i den simple lineære regressionsmodel.

- Parameterestimaterne findes ved at minimere RSS:

$$(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p) = \operatorname{argmin} \sum_{i=1}^n \text{RSS}(b_0, b_1, \dots, b_p)$$

Bemærk:

$$\operatorname{argmin} \sum_{i=1}^n \text{RSS}(b_0, b_1, \dots, b_p) = \operatorname{argmin} \sum_{i=1}^n e_i^2 = \operatorname{argmin} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

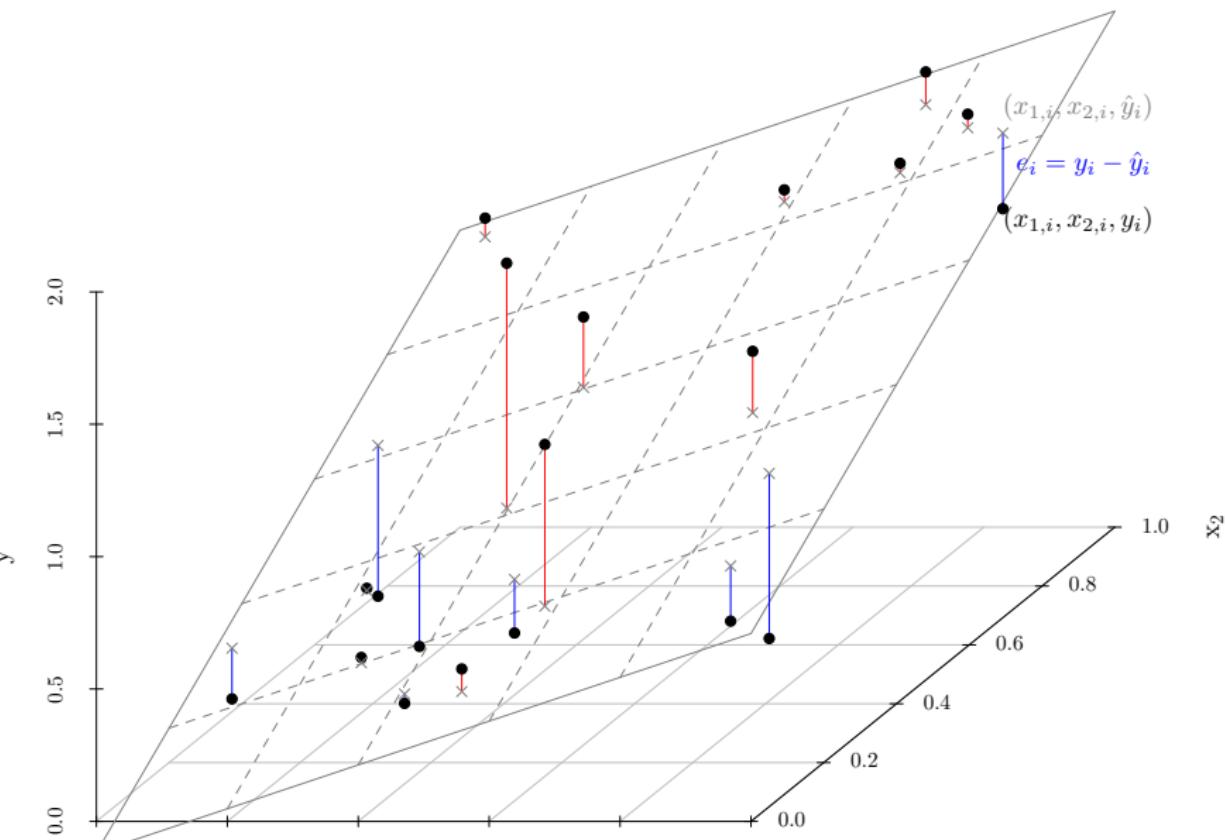
- De prædikterede (fittede) værdier findes ved:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_p x_{i,p},$$

- og residualerne er så:

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

Mindste kvadraters metode



Vigtige resultater

- Bemærkning 6.6: Find $\hat{\beta}_i$ og $\hat{\sigma}_{\beta_i}$ fra R-outputtet (summary(myfit))

Vigtige resultater

- Bemærkning 6.6: Find $\hat{\beta}_i$ og $\hat{\sigma}_{\beta_i}$ fra R-outputtet (summary(myfit))
- Sætning 6.2: t -fordelingen kan bruges til inferens for modelparametre.

Vigtige resultater

- Bemærkning 6.6: Find $\hat{\beta}_i$ og $\hat{\sigma}_{\beta_i}$ fra R-outputtet (`summary(myfit)`)
- Sætning 6.2: t -fordelingen kan bruges til inferens for modelparametre.
- Metoder 6.4 og 6.5: Hypotesetest og konfidensintervaller fra R-outputtet.

Vigtige resultater

- Bemærkning 6.6: Find $\hat{\beta}_i$ og $\hat{\sigma}_{\beta_i}$ fra R-outputtet (`summary(myfit)`)
- Sætning 6.2: t -fordelingen kan bruges til inferens for modelparametre.
- Metoder 6.4 og 6.5: Hypotesetest og konfidensintervaller fra R-outputtet.
- Altsammen: **Samme som for simpel lineær regression!**

Vigtige resultater

- Bemærkning 6.6: Find $\hat{\beta}_i$ og $\hat{\sigma}_{\beta_i}$ fra R-outputtet (`summary(myfit)`)
- Sætning 6.2: t -fordelingen kan bruges til inferens for modelparametre.
- Metoder 6.4 og 6.5: Hypotesetest og konfidensintervaller fra R-outputtet.
- Altsammen: **Samme som for simpel lineær regression!**
- (I Afsnit 6.6 af bogen: Matrix-baseret tilgang med eksplisitive formler.
Ikke pensum i kursus 02402)

R-outputtet

```
summary(lm(logOzone ~ temperature + wind + radiation, data=Air))

##
## Call:
## lm(formula = logOzone ~ temperature + wind + radiation, data = Air)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.0203 -0.3150 -0.0094  0.3230  1.1223 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 0.261436  0.520496   0.50    0.62    
## temperature 0.044457  0.005678   7.83 3.9e-12 ***
## wind        -0.069283  0.014514  -4.77 5.8e-06 ***
## radiation   0.002190  0.000516   4.25 4.6e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.467 on 106 degrees of freedom
## Multiple R-squared:  0.674, Adjusted R-squared:  0.664 
## F-statistic: 72.9 on 3 and 106 DF,  p-value: <2e-16
```

- Læs estimerter, usikkerheder osv. fra outputtet.

Fortolkning af parametre (Bemærkning 6.14)

Hvad er $\hat{\beta}_i$ udtryk for?

- Den forventede ændring i y når x_i ændres én enhed.

Fortolkning af parametre (Bemærkning 6.14)

Hvad er $\hat{\beta}_i$ udtryk for?

- Den forventede ændring i y når x_i ændres én enhed.
- Effekten af x_i givet de øvrige variable.
- Effekten af x_i korrigert for de øvrige variables effekt.
- Effekten af x_i "når de andre variable er uændret".

Fortolkning af parametre (Bemærkning 6.14)

Hvad er $\hat{\beta}_i$ udtryk for?

- Den forventede ændring i y når x_i ændres én enhed.
- Effekten af x_i givet de øvrige variable.
- Effekten af x_i korrigert for de øvrige variables effekt.
- Effekten af x_i "når de andre variable er uændret".
- Afhænger af hvad der ellers i modellen!
- Generelt: IKKE en kausal effekt/interventionseffekt!

Dagsorden

- 1 Opsummering
- 2 Multipel lineær regression
- 3 Modeludvælgelse (Model selection)
- 4 Modelkontrol (Analyse af residualerne)
- 5 Kurvilinearitet
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet
- 8 Den 'samlede' regressionsmetode

Modeludvidelse (forward selection)

- *Ikke inkluderet i bogen*
- Start med en *simpel lineær regressionsmodel* med en signifikant forklarende variabel
- *Udvid modellen* med andre forklarende variable én ad gangen
- *Stop* når der ikke er flere signifikante udvidelser

```
#####
## Extend the model

## Forward selection:
## Add wind to the model
summary(lm(logOzone ~ temperature + wind, data=Air))
## Add radiation to the model
summary(lm(logOzone ~ temperature + wind + radiation, data=Air))
```

Modelreduktion (backward selection)

- Beskrevet i bogen under sektion 6.5
- Start med den fulde model
- Fjern den "mindst signifikante" variabel
- Stop når alle tilbageværende parametre er signifikante

```
#####
## Backward selection

## Fit the full model
summary(lm(logOzone ~ temperature + wind + radiation + noise, data=Air))
## Remove the most non-significant input, are all now significant?
summary(lm(logOzone ~ temperature + wind + radiation, data=Air))
```

Modeludvælgelse

- Der er ikke nogen sikker metode til at finde den bedste model!
- Det kræver subjektive beslutninger at udvælge en model.
- Forskellige procedurer, enten forward eller backward selection (eller begge), afhænger af forholdene.
- Der findes statistiske metoder og test til at sammenligne modeller.
- Her i kurset er kun backward selection beskrevet.

Dagsorden

- 1 Opsummering
- 2 Multipel lineær regression
- 3 Modeludvælgelse (Model selection)
- 4 Modelkontrol (Analyse af residualerne)
- 5 Kurvilinearitet
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet
- 8 Den 'samlede' regressionsmetode

Modelkontrol (Analyse af residualerne)

- Modelkontrol: Analysér residualerne for at tjekke om antagelserne er opfyldt.
- Samme antagelser som for den simple lineære model.

Antagelsen om normalfordelte residualer

- Brug normal QQ-plot:

```
#####
## Assumption of normal distributed residuals

## Save the selected fit
fitSel <- lm(logOzone ~ temperature + wind + radiation, data=Air)

## qq-normalplot
qqnorm(fitSel$residuals)
qqline(fitSel$residuals)
```

Antagelse om ensfordelte residualer

Vi kigger efter varianshomogenitet og systematiske tendenser.

- Plot residualerne (e_i) mod de prædikterede (fittede) værdier: (\hat{y}_i)

```
#####
## Plot the residuals vs. predicted values

plot(fitSel$fitted.values, fitSel$residuals, xlab="Predicted values",
      ylab="Residuals")
```

- Plot residualerne mod de forklarende variable:

```
#####
## Plot the residuals vs. the independent variables

par(mfrow=c(1,3))
plot(Air$temperature, fitSel$residuals, xlab="Temperature")
plot(Air$wind, fitSel$residuals, xlab="Wind speed")
plot(Air$radiation, fitSel$residuals, xlab="Radiation")
```

Dagsorden

- 1 Opsummering
- 2 Multipel lineær regression
- 3 Modeludvælgelse (Model selection)
- 4 Modelkontrol (Analyse af residualerne)
- 5 Kurvilinearitet
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet
- 8 Den 'samlede' regressionsmetode

En kurvelineær model

Regressionsmodeller til ikke-lineær data baseret på Taylorudviklinger.

Hvis vi vil benytte en model af typen

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,$$

så kan vi bruge en multipel lineær regressionsmodel

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i,$$

hvor

$$x_{i,1} = x_i, \quad x_{i,2} = x_i^2$$

og bruge de samme metoder som for multipel lineær regression.

Udvid ozonmodellen med passende kurvelineær regression

```
#####
## Extend the ozone model with appropriate curvilinear regression

## Make the squared wind speed
Air$WindSq <- Air$wind^2
## Add it to the model
fitWindSq <- lm(logOzone ~ temperature + wind + windSq + radiation, data=Air)
summary(fitWindSq)

## Equivalently for the temperature
Air$temperature2 <- Air$temperature^2
## Add it
fitTemperatureSq <- lm(logOzone ~ temperature + temperature2 + wind + radiation, data=Air)
summary(fitTemperatureSq)

## Equivalently for the radiation
Air$radiation2 <- Air$radiation^2
## Add it
fitRadiationSq <- lm(logOzone ~ temperature + wind + radiation + radiation2, data=Air)
summary(fitRadiationSq)

## Which one was best?
## One could try to extend the model further
fitWindSqTemperaturSq <- lm(logOzone ~ temperature + temperature2 + wind + windSq + radiation, data=Air)
summary(fitWindSqTemperaturSq)

## Model validation
qqnorm(fitWindSq$residuals)
qqline(fitWindSq$residuals)
plot(fitWindSq$residuals, fitWindSq$fitted.values, pch=19)
```

Dagsorden

- 1 Opsummering
- 2 Multipel lineær regression
- 3 Modeludvælgelse (Model selection)
- 4 Modelkontrol (Analyse af residualerne)
- 5 Kurvilinearitet
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet
- 8 Den 'samlede' regressionsmetode

Konfidens- og prædiktionsintervaller - Metode 6.9:

Som for simpel lineær regression (i princippet).

```
#####
## Confidence and prediction intervals for the curvilinear model

## Generate a new data.frame with constant temperature and radiation, but with varying wind speed
wind<-seq(1,20.3,by=0.1)
AirForPred <- data.frame(temperature=mean(Air$temperature), wind=wind,
                           windSq=wind^2, radiation=mean(Air$radiation))

## Calculate confidence and prediction intervals (actually bands)
CI <- predict(fitWindSq, newdata=AirForPred, interval="confidence", level=0.95)
PI <- predict(fitWindSq, newdata=AirForPred, interval="prediction", level=0.95)

## Plot them
plot(wind, CI[, "fit"], ylim=range(CI,PI), type="l",
      main=paste("At temperature =",format(mean(Air$temperature),digits=3),
                 "and radiation =", format(mean(Air$radiation),digits=3)))
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("Prediction", "95% confidence band", "95% prediction band"), lty=c(1,2,2), col=1:3)
```

Dagsorden

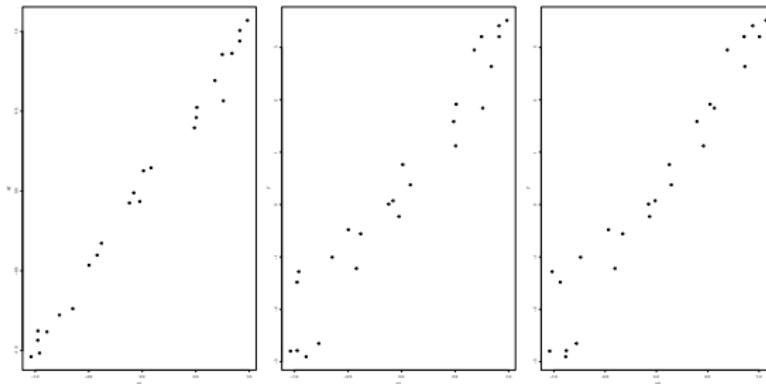
- 1 Opsummering
- 2 Multipel lineær regression
- 3 Modeludvælgelse (Model selection)
- 4 Modelkontrol (Analyse af residualerne)
- 5 Kurvilinearitet
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet**
- 8 Den 'samlede' regressionsmetode

Kollinearitet

- Hvis to (eller flere) forklarende variable har en perfekt lineær sammenhæng, så kan vi ikke afgøre, hvilken som er forklarende.
- Også et problem hvis sammenhængen er tæt på lineær.
- Relateret til konceptet "confounders".
- Med to meget korrelerede x -variable:
 - *Sammen* kan det være at ingen af dem har en "unik" effekt.
 - *Separat* kan de have en stor effekt.

Kollinearitet – Eksempel

To meget korrelerede forklarende variable x_1 og x_2 og responsvariabel y .



Kollinearitet – Eksempel

```
#####
L <- lm(y ~ x1 + x2)
summary(L)

##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -0.7951 -0.3723  0.0038  0.3546  1.2247 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.376     0.109    3.44   0.0023 **  
## x1          0.709     1.535    0.46   0.6485    
## x2          2.167     1.523    1.42   0.1688    
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 0.534 on 22 degrees of freedom
## Multiple R-squared:  0.941, Adjusted R-squared:  0.936 
## F-statistic: 175 on 2 and 22 DF, p-value: 3.05e-14
```

Kollinearitet – Konklusion

- Svært at separere effekter af kollinære variable
- Ingen nem løsning på kollinearitet
- Et fornuftigt design af eksperimentet kan hjælpe

Kollinearitet – Konklusion

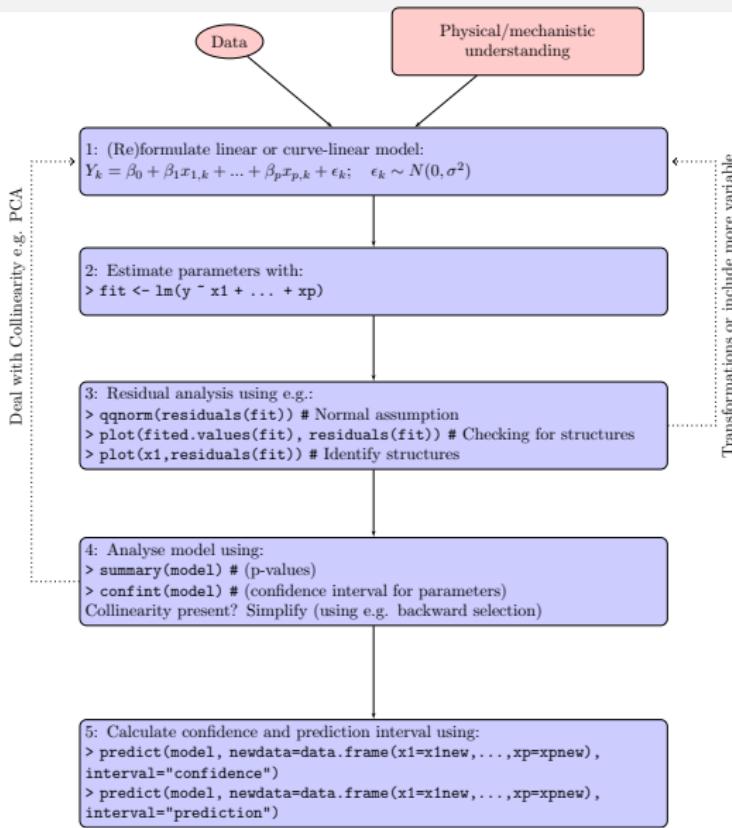
- Svært at separere effekter af kollinære variable
- Ingen nem løsning på kollinearitet
- Et fornuftigt design af eksperimentet kan hjælpe

Det er vigtigt, hvordan man
designer sit eksperiment!

Dagsorden

- 1 Opsummering
- 2 Multipel lineær regression
- 3 Modeludvælgelse (Model selection)
- 4 Modelkontrol (Analyse af residualerne)
- 5 Kurvilinearitet
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet
- 8 Den 'samlede' regressionsmetode

Metode 6.16



Dagsorden

- 1 Opsummering
- 2 Multipel lineær regression
- 3 Modeludvælgelse (Model selection)
- 4 Modelkontrol (Analyse af residualerne)
- 5 Kurvelinearitet
- 6 Konfidens- og prædiktionsintervaller
- 7 Kollinearitet
- 8 Den 'samlede' regressionsmetode

02402: Introduktion til Statistik

Forelæsning 10: Inferens for andele (forholdstal)

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Overblik

- 1 Introduktion
- 2 Konfidensinterval for én andel
 - Stikprøvestørrelse, forsøgsplanlægning
- 3 Hypotesetest for én andel
- 4 Konfindensinterval og hypotesetest for to andele
- 5 Hypotesetest for flere andele
- 6 Statistik for antalstabeller

Overview

1 Introduktion

2 Konfidensinterval for én andel

- Stikprøvestørrelse, forsøgsplanlægning

3 Hypotesetest for én andel

4 Konfindensinterval og hypotesetest for to andele

5 Hypotesetest for flere andele

6 Statistik for antalstabeller

Forskellige analyse/data-situationer i 02402

Middelværdi for kvantitative data

- Hypotesetest/CI for én middelværdi (én stikprøve) (kapitel 3)
- Hypotesetest/CI for to middelværdier (to stikprøver) (kapitel 3)
- Hypotesetest/CI for flere middelværdier (K stikprøver) (kapitel 8, næste uge)

Forskellige analyse/data-situationer i 02402

Middelværdi for kvantitative data

- Hypotesetest/CI for én middelværdi (én stikprøve) (kapitel 3)
- Hypotesetest/CI for to middelværdier (to stikprøver) (kapitel 3)
- Hypotesetest/CI for flere middelværdier (K stikprøver) (kapitel 8, næste uge)

I dag: Andele (forholdstal, engelsk: *proportions*)

- Hypotestest/CI for én andel
- Hypotestest/CI for to andele
- Hypotesetest for flere andele
- Hypotesetest test for “multi-kategoriske” andele/tabeller

Estimation af andele

- Estimation af en andel (sandsynlighed) fås ved at observere antallet af gange (x) en hændelse har indtruffet ud af n (uafhængige) forsøg:

$$\hat{p} = \frac{x}{n}$$

Bemærk:

- $\hat{p} \in [0; 1]$.
- \hat{p} er en stokastisk variabel. Gentagelser af forsøget kan give forskellige udfald.

Overview

- 1 Introduktion
- 2 Konfidensinterval for én andel
 - Stikprøvestørrelse, forsøgsplanlægning
- 3 Hypotesetest for én andel
- 4 Konfindensinterval og hypotesetest for to andele
- 5 Hypotesetest for flere andele
- 6 Statistik for antalstabeller

Konfidensinterval for én andel

Method 7.3

Hvis stikprøven er **stor**, så er $(1 - \alpha)100\%$ konfidensintervallet for p givet ved:

$$\frac{x}{n} - z_{1-\alpha/2} \sqrt{\frac{\frac{x}{n}(1 - \frac{x}{n})}{n}} < p < \frac{x}{n} + z_{1-\alpha/2} \sqrt{\frac{\frac{x}{n}(1 - \frac{x}{n})}{n}}$$

Konfidensinterval for én andel

Method 7.3

Hvis stikprøven er **stor**, så er $(1 - \alpha)100\%$ konfidensintervallet for p givet ved:

$$\frac{x}{n} - z_{1-\alpha/2} \sqrt{\frac{\frac{x}{n}(1 - \frac{x}{n})}{n}} < p < \frac{x}{n} + z_{1-\alpha/2} \sqrt{\frac{\frac{x}{n}(1 - \frac{x}{n})}{n}}$$

Hvordan?

Følger af at approksimere binomialfordelingen med normalfordelingen.

Konfidensinterval for én andel

Method 7.3

Hvis stikprøven er **stor**, så er $(1 - \alpha)100\%$ konfidensintervallet for p givet ved:

$$\frac{x}{n} - z_{1-\alpha/2} \sqrt{\frac{\frac{x}{n}(1 - \frac{x}{n})}{n}} < p < \frac{x}{n} + z_{1-\alpha/2} \sqrt{\frac{\frac{x}{n}(1 - \frac{x}{n})}{n}}$$

Hvordan?

Følger af at approksimere binomialfordelingen med normalfordelingen.

Tommelfingerregel

Antag $X \sim \text{binom}(n, p)$. Normalfordelingen er en god tilnærmelse til binomialfordelingen hvis np og $n(1 - p)$ (forventet antall successer og fiaskoer) begge er større 15.

Konfidensinterval for én andel

Middelværdi og varians i binomialfordelingen, kapitel 2.21

$$\text{E}(X) = np$$

$$\text{Var}(X) = np(1 - p)$$

Altså fås:

$$\text{E}(\hat{p}) = \text{E}\left(\frac{X}{n}\right) = \frac{np}{n} = p$$

$$\text{Var}(\hat{p}) = \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2} \text{Var}(X) = \frac{p(1-p)}{n}$$

Eksempel 1

Venstrehåndede:

$p = \text{Andelen af venstrehåndede i Danmark}$

og/eller:

Kvindelige ingeniørstuderende:

$p = \text{Andelen af kvindelige ingeniørstuderende}$

Example 1

Venstrehåndede (observeret: $x = 10$ ud af $n = 100$):

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{10/100(1-10/100)}{100}} = 0.03$$

$$0.10 \pm 1.96 \cdot 0.03 = 0.10 \pm 0.059 = [0.041, 0.159]$$

Example 1

Venstrehåndede (observeret: $x = 10$ ud af $n = 100$):

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{10/100(1-10/100)}{100}} = 0.03$$

$$0.10 \pm 1.96 \cdot 0.03 = 0.10 \pm 0.059 = [0.041, 0.159]$$

Bedre "small sample"-metode – "plus 2"-tilgangen (Remark 7.7)

Anvend samme formel på $\tilde{x} = 10 + 2 = 12$ og $\tilde{n} = 100 + 2 + 2 = 104$:

$$\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}} = \sqrt{\frac{12/104(1-12/104)}{104}} = 0.031$$

$$0.115 \pm 1.96 \cdot 0.031 = 0.115 \pm 0.061 = [0.054, 0.177]$$

Fejlmarginen (Margin of Error (ME))

Fejlmarginen

ved et $(1 - \alpha)100\%$ konfidensniveau er:

$$ME = z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

hvor vi estimerer p med $\hat{p} = \frac{x}{n}$.

Fejlmarginen:

- svarer til den halve bredde af $(1 - \alpha)100\%$ konfidensintervallet.
- Beskriver den “mindst ønskede præcision” på estimatet \hat{p} .

Præcision og stikprøvestørrelse

Forsøgsplanlægning:

Hvor stor skal stikprøvestørrelsen være for at opnå en given præcision?

Metode 7.13

Ønskes en given fejlmargin, ME, med $(1 - \alpha)100\%$ konfidens, da kræves følgende stikprøvestørrelse:

$$n = p(1 - p) \left(\frac{z_{1-\alpha/2}}{ME} \right)^2$$

Præcision og stikprøvestørrelse

Method 7.13

Ønskes en given fejlmargen, ME, med $(1 - \alpha)100\%$ konfidens, men hvor vi *ikke* har et fornuftigt gæt på p , da kræves følgende stikprøvestørrelse:

$$n = \frac{1}{4} \left(\frac{z_{1-\alpha/2}}{ME} \right)^2$$

thi "worst case" er $p = \frac{1}{2}$.

Eksempel 1 – fortsat

Venstrehåndthed:

Antag at vi ønsker $ME = 0.01$ (hvor $\alpha = 0.05$) – hvad skal n så være?

Eksempel 1 – fortsat

Venstrehåndthed:

Antag at vi ønsker $ME = 0.01$ (hvor $\alpha = 0.05$) – hvad skal n så være?

Antag $p \approx 0.10$:

$$n = 0.1 \cdot 0.9 \left(\frac{1.96}{0.01} \right)^2 = 3467.4 \approx 3468$$

Eksempel 1 – fortsat

Venstrehåndthed:

Antag at vi ønsker $ME = 0.01$ (hvor $\alpha = 0.05$) – hvad skal n så 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 hvad p er:

$$n = \frac{1}{4} \left(\frac{1.96}{0.01} \right)^2 = 9604$$

Overview

- 1 Introduktion
- 2 Konfidensinterval for én andel
 - Stikprøvestørrelse, forsøgsplanlægning
- 3 Hypotesetest for én andel
- 4 Konfindensinterval og hypotesetest for to andele
- 5 Hypotesetest for flere andele
- 6 Statistik for antalstabeller

Trin i et hypotesetest – overblik (repetition!)

- ① Opstil hypotesen og vælg signifikansniveau α
- ② Beregn teststørrelse
- ③ Beregn p -værdi ud fra teststørrelsen og den relevante fordeling. Sammenlign p -værdien med signifikansniveauet α og konkludér.
- ④ (Alternativt: lav konklusion ud fra kritiske værdier).

Hypotesetest for én andel

Vi betragter en nul- og alternativ hypotese for én andel p :

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

Som sædvanligt enten accepteres H_0 eller afvises H_0 .

Hypotesetest: teststørrelsen

Theorem 7.10 og Method 7.11

Hvis 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 stokastiske variabel Z (tilnærmelsesvis) følger en standard normalfordeling, dvs. $Z \sim N(0, 1^2)$.

Hypotesetest: p -værdi og konklusion (Method 7.11)

Find p -værdien (evidens imod nulhypotesen):

- $2P(Z > |z_{\text{obs}}|)$

Test ved brug af kritisk værdi:

Vi afviser nulhypotesen hvis $z_{\text{obs}} < -z_{1-\alpha/2}$ eller
 $z_{\text{obs}} > z_{1-\alpha/2}$.

Eksempel 1 – fortsat

Er halvdelen af alle danskere venstrehåndede?

$$H_0 : p = 0.5, \quad H_1 : p \neq 0.5$$

Eksempel 1 – fortsat

Er halvdelen af alle danskere venstrehåndede?

$$H_0 : p = 0.5, \quad H_1 : p \neq 0.5$$

Teststørrelse:

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}} = \frac{10 - 100 \cdot 0.5}{\sqrt{100 \cdot 0.5(1 - 0.5)}} = -8$$

Eksempel 1 – fortsat

Er halvdelen af alle danskere venstrehåndede?

$$H_0 : p = 0.5, \quad H_1 : p \neq 0.5$$

Teststørrelse:

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}} = \frac{10 - 100 \cdot 0.5}{\sqrt{100 \cdot 0.5(1 - 0.5)}} = -8$$

p-værdi:

$$2 \cdot P(Z > 8) = 1.2 \cdot 10^{-15}$$

Der er meget stærk evidens mod nulhypotesen.

Eksempel 1 – fortsat

Hypotesetest in R

```
prop.test(10, 100, p = 0.5, correct = FALSE)

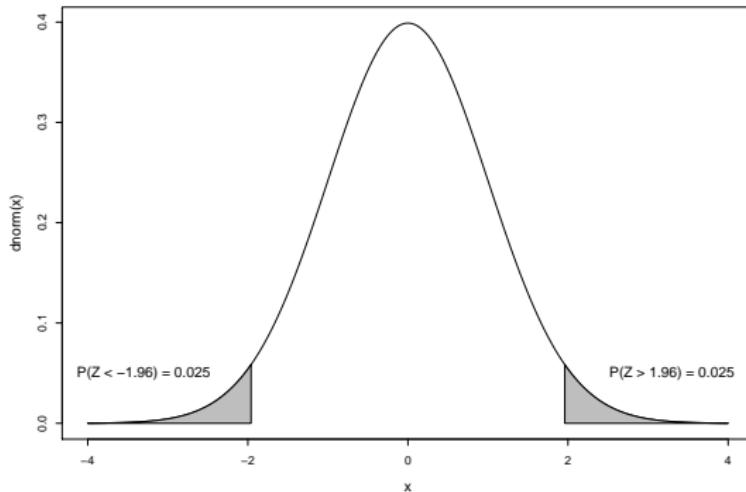
##
## 1-sample proportions test without continuity correction
##
## data: 10 out of 100, null probability 0.5
## X-squared = 64, df = 1, p-value = 1e-15
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
## 0.05523 0.17437
## sample estimates:
## p
## 0.1
```

Eksempel 1 – fortsat

Ved brug af kritiske værdier:

$$z_{0.975} = 1.96$$

Da $z_{\text{obs}} = -8$ er (meget) mindre end -1.96 så afvises hypotesen.



Overview

- 1 Introduktion
- 2 Konfidensinterval for én andel
 - Stikprøvestørrelse, forsøgsplanlægning
- 3 Hypotesetest for én andel
- 4 Konfindensinterval og hypotesetest for to andele
- 5 Hypotesetest for flere andele
- 6 Statistik for antalstabeller

Konfidensinterval for forskellen 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

$$\hat{\sigma}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

Tommelfingerregel

Både $n_i p_i \geq 10$ og $n_i(1 - p_i) \geq 10$ for $i = 1, 2$.

Hypotesetest for forskellen på to andele, Method 7.18

Two sample proportions hypothesis test

Såfremt man ønsker at sammenligne to andele (her vist for et tosidet alternativ)

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

skal vi bruge 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 tilstrækkeligt store stikprøver:

Brug standardnormalfordelingen igen!

Eksempel 2

Er der en sammenhæng mellem brugen af p-piller og risikoen for blodprop i hjertet?

I et studie (USA, 1975) undersøgtes sammenhængen mellem p-piller og risikoen for blodprop i hjertet.

	Blodprop	Ikke blodprop
p-piller	23	34
Ikke p-piller	35	132

Undersøg om der er sammenhæng mellem brug af p-piller og risiko for blodprop i hjertet. Anvend signifikansniveau $\alpha = 5\%$.

Example 2

I et studie (USA, 1975) undersøgtes sammenhængen mellem p-piller og risikoen for blodprop i hjertet.

	Blodprop	Ikke blodprop
p-piller	23	34
Ikke p-piller	35	132

Estimater i hver stikprøve

$$\hat{p}_1 = \frac{23}{57} = 0.4035, \quad \hat{p}_2 = \frac{35}{167} = 0.2096$$

Fælles estimat:

$$\hat{p} = \frac{23 + 35}{57 + 167} = \frac{58}{224} = 0.2589$$

Eksempel 2 – fortsat

prop.test: test om to andele er ens i R

```
# Read data table into R
pill.study <- matrix(c(23, 34, 35, 132),
                      ncol = 2, byrow = TRUE)
colnames(pill.study) <- c("Blood Clot", "No Clot")
rownames(pill.study) <- c("Pill", "No pill")
pill.study

# Test whether probabilities are equal for the two groups
prop.test(pill.study, correct = FALSE)
```

Eksempel 2 – fortsat

prop.test: test om to andele er ens i R

```
##          Blood Clot No Clot
## Pill           23      34
## No pill       35     132
```

```
##
## 2-sample test for equality of proportions without continuity
## correction
##
## data: pill.study
## X-squared = 8.3, df = 1, p-value = 0.004
## alternative hypothesis: two.sided
## 95 percent confidence interval:
## 0.05239 0.33546
## sample estimates:
## prop 1 prop 2
## 0.4035 0.2096
```

Overview

- 1 Introduktion
- 2 Konfidensinterval for én andel
 - Stikprøvestørrelse, forsøgsplanlægning
- 3 Hypotesetest for én andel
- 4 Konfindensinterval og hypotesetest for to andele
- 5 **Hypotesetest for flere andele**
- 6 Statistik for antalstabeller

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 den alternative hypotese om at disse andele ikke er ens.

Hypotesetest for flere andele

Tabel af observerede antal for c stikprøver:

	Sample 1	Sample 2	...	Sample c	Total
Succes	x_1	x_2	...	x_c	x
Failure	$n_1 - x_1$	$n_2 - x_2$...	$n_c - x_c$	$n - x$
Total	n_1	n_2	...	n_c	n

Fælles (gennemsnitligt) estimat:

Under nulhypotesen er estimatet for p :

$$\hat{p} = \frac{x}{n}$$

Hypotesetest for flere andele

Fælles (gennemsnitligt) estimat:

Under nulhypotesen er estimatet for p :

$$\hat{p} = \frac{x}{n}$$

"Brug" dette fælles estimat i hver gruppe:

Hvis nulhypotesene er sand, så forventer vi at den j 'te gruppe har e_{1j} successer og e_{2j} fiaskoer, hvor

$$e_{1j} = n_j \cdot \hat{p} = \frac{n_j \cdot x}{n}$$

$$e_{2j} = n_j(1 - \hat{p}) = \frac{n_j \cdot (n - x)}{n}$$

Hypotesetest for flere andele

Make a table with the *expected* counts for the c samples:

e_{ij}	Sample 1	Sample 2	...	Sample c	Total
Succes	e_{11}	e_{12}	...	e_{1c}	x
Failure	e_{21}	e_{22}	...	e_{2c}	$n - x$
Total	n_1	n_2	...	n_c	n

Generel formel for beregning af forventede værdier i antalstabeller:

$$e_{ij} = \frac{(i\text{'th row total}) \cdot (j\text{'th column total})}{\text{total}}$$

Beregning af teststørrelsen - 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 det *observerede* antal i celle (i,j) og e_{ij} er det *forventede* antal i celle (i,j) .

Find p -værdien eller brug kritisk værdi – Method 7.20

Stikprøvefordeling for teststørrelse (under H_0):

χ^2 -fordeling med $(c - 1)$ frihedsgrader (approx.)

Kritisk værdi metode:

Hvis $\chi_{\text{obs}}^2 > \chi_{\alpha}^2(c - 1)$, så afvises nulhypotesen.

Tommelfingerregel for om testen er valid:

Alle forventede værdier $e_{ij} \geq 5$.

Eksempel 2 – fortsat

De *observerede* værdier o_{ij}

Observed	Blodprop	Ikke blodprop
p-piller	23	34
Ikke p-piller	35	132

Eksempel 2 – fortsat

Beregn de *forventede* værdier e_{ij}

Expected	Blodprop	Ikke blodprop	Total
p-piller			57
Ikke p-piller			167
Total	58	166	224

Eksempel 2 – fortsat

Brug “reglen” for forventede værdier fire gange, e.g.:

$$e_{22} = \frac{167 \cdot 166}{224} = 123.76$$

De *forventede* værdier e_{ij} :

Expected	Blodprop	Ikke blodprop	Total
p-piller	14.76	42.24	57
Ikke p-piller	43.24	123.76	167
Total	58	166	224

Eksempel 2 – fortsat

Teststørrelsen (husk at inkludere alle celler):

$$\chi_{\text{obs}}^2 = \frac{(23 - 14.76)^2}{14.76} + \frac{(34 - 42.24)^2}{42.24} + \frac{(35 - 43.24)^2}{43.24} + \frac{(132 - 123.76)^2}{123.76}$$
$$= 8.33$$

Kritisk værdi:

```
qchisq(0.95, 1)
```

```
[1] 3.841
```

Konklusion:

Vi afviser nulhypotesen – der er altså en signifikant højere risiko for blodpropper i gruppen, som indtager p-piller.

Eksempel 2 – fortsat

chisq.test for at teste om to forhold er ens i R.

```
# Test whether probabilities are equal for the two groups
chisq.test(pill.study, correct = FALSE)

##
## Pearson's Chi-squared test
##
## data: pill.study
## X-squared = 8.3, df = 1, p-value = 0.004

# Expected values
chisq.test(pill.study, correct = FALSE)$expected

##          Blood Clot No Clot
## Pill        14.76   42.24
## No pill    43.24  123.76
```

Overview

- 1 Introduktion
- 2 Konfidensinterval for én andel
 - Stikprøvestørrelse, forsøgsplanlægning
- 3 Hypotesetest for én andel
- 4 Konfindensinterval og hypotesetest for to andele
- 5 Hypotesetest for flere andele
- 6 Statistik for antalstabeller

Eksempel 3: Analyse af en antalstabel

En 3×3 -tabel – 3 stikprøver, 3-kategori udfald

	4 weeks bef	2 weeks bef	1 week bef
Candidate I	79	91	93
Candidate II	84	66	60
Undecided	37	43	47
	$n_1 = 200$	$n_2 = 200$	$n_3 = 200$

Er stemmefordelingen ens?

$$H_0 : p_{i1} = p_{i2} = p_{i3}, \quad i = 1, 2, 3.$$

En anden slags antalstabbel

En 3×3 -tabel – 1 stikprøve, to 3-kategori variable:

	bad	average	good
bad	23	60	29
average	28	79	60
good	9	49	63

Er der uafhængighed mellem inddelingskriterier?

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}$$

Beregning af teststørrelse – uanset typen af tabel 7.22

I en antalstabel med r rækker og c søjler, da er 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 det *forventede antal* i celle (i,j) (under hypotesen).

Generel formel for beregning af forventede værdier i antalstabeller:

$$e_{ij} = \frac{(i\text{'th row total}) \cdot (j\text{'th column total})}{\text{total}}$$

Find p -værdi eller brug kritisk værdi - Method 7.22

Stikprøvefordeling for test-størrelse under H_0 :

χ^2 -fordeling med $(r - 1)(c - 1)$ frihedsgrader.

Kritisk værdi-metode:

Såfremt $\chi_{\text{obs}}^2 > \chi_{\alpha}^2$ med $(r - 1)(c - 1)$ frihedsgrader, da forkastes nulhypotesen.

Tommelfingerregel for validitet af test:

Alle forventede værdier $e_{ij} \geq 5$.

Eksempel 3 – fortsat

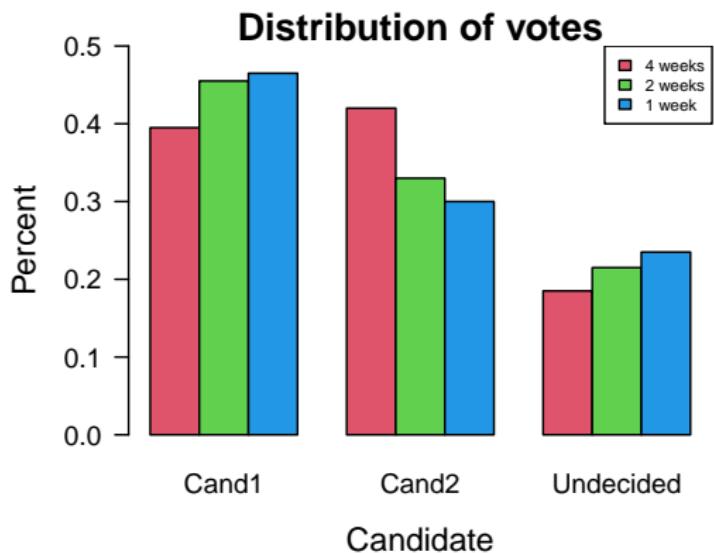
chisq.test for antalstabeller

```
# Read data 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")

# Show column percentages
prop.table(poll, 2)

##          4 weeks 2 weeks 1 week
## Cand1      0.395   0.455   0.465
## Cand2      0.420   0.330   0.300
## Undecided  0.185   0.215   0.235
```

Eksempel 3 – fortsat



Eksempel 3 – fortsat

```
# Testing for same distribution in the three populations
chisq.test(poll, correct = FALSE)

##
## Pearson's Chi-squared test
##
## data: poll
## X-squared = 7, df = 4, p-value = 0.1

# Expected values
chisq.test(poll, correct = FALSE)$expected

##          4 weeks 2 weeks 1 week
## Cand1      87.67   87.67   87.67
## Cand2      70.00   70.00   70.00
## Undecided  42.33   42.33   42.33
```

Overview

- 1 Introduktion
- 2 Konfidensinterval for én andel
 - Stikprøvestørrelse, forsøgsplanlægning
- 3 Hypotesetest for én andel
- 4 Konfindensinterval og hypotesetest for to andele
- 5 Hypotesetest for flere andele
- 6 Statistik for antalstabeller

02402 Statistik (Polyteknisk grundlag)

Uge 11: Ensidet variansanalyse - ANOVA

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

Dagsorden

- 1 Introduktion
- 2 Model og hypoteser
- 3 Beregning: Variansdekomposition og ANOVA-tabellen
- 4 Hypotesetest (F-test)
- 5 Variabilitet og sammenhæng med t-testen for to stikprøver
- 6 Post hoc sammenligninger
- 7 Modelkontrol
- 8 Et gennemregnet eksempel – fra bogen

Variansanalyse - ANOVA

"ANalysis Of VAriance" (ANOVA) blev introduceret af R.A. Fisher for ca. 100 år siden som en systematisk måde at analysere grupper på og har siden da været helt centrale i statistik og anvendelser deraf.

- I dag: Et inddelingskriterium (ensidet ANOVA)
- Næste uge: To inddelingskriterier (tosidet ANOVA)
- Inddelingskriterium = **faktor**
- Første faktor kaldes typisk *treatment*, anden faktor *block*

Dagsorden

- 1 Introduktion
- 2 Model og hypoteser
- 3 Beregning: Variansdekomposition og ANOVA-tabellen
- 4 Hypotesetest (F-test)
- 5 Variabilitet og sammenhæng med t-testen for to stikprøver
- 6 Post hoc sammenligninger
- 7 Modelkontrol
- 8 Et gennemregnet eksempel – fra bogen

Ensidet 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 (i middelværdien) på grupperne A, B og C?

Variansanalyse (ANOVA) kan anvendes til analysen, såfremt observationerne i hver gruppe kan antages at være normalfordelte.

Envejs variansanalyse – eksempel i R

```
# Indlæs data
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)

# Definer (treatment) grupper
treatm <- factor(c(1, 1, 1, 1,
                     2, 2, 2, 2,
                     3, 3, 3, 3))

# Plot data mod grupperne
par(mfrow = c(1,2))
plot(y ~ as.numeric(treatm), xlab = "Gruppe (Treatment)", ylab = "Værdi")
boxplot(y ~ treatm, xlab = "Gruppe (Treatment)", ylab = "Værdi")
```

Dagsorden

- 1 Introduktion
- 2 Model og hypoteser**
- 3 Beregning: Variansdekomposition og ANOVA-tabellen
- 4 Hypotesetest (F-test)
- 5 Variabilitet og sammenhæng med t-testen for to stikprøver
- 6 Post hoc sammenligninger
- 7 Modelkontrol
- 8 Et gennemregnet eksempel – fra bogen

Ensidet variansanalyse - Model

- Modellen kan opskrives som

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

hvor det antages ε_{ij} er i.i.d. med

$$\varepsilon_{ij} \sim N(0, \sigma^2).$$

- μ er den samlede middelværdi
- α_i angiver effekten af gruppe (treatment) i
- Y_{ij} er måling j i gruppe i (j går fra 1 til n_i)

Ensidet variansanalyse - Hypotesetest

- Vi vil nu sammenligne (flere end to) middelværdier ($\mu + \alpha_i$) i modellen

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$

- Nulhypotesen er givet ved:

$$H_0 : \alpha_i = 0 \quad \text{for alle } i.$$

- Modhypotesen (alternativhypotesen) er givet ved:

$$H_1 : \alpha_i \neq 0 \quad \text{for mindst et } i.$$

Dagsorden

- 1 Introduktion
- 2 Model og hypoteser
- 3 Beregning: Variansdekomposition og ANOVA-tabellen
- 4 Hypotesetest (F-test)
- 5 Variabilitet og sammenhæng med t-testen for to stikprøver
- 6 Post hoc sammenligninger
- 7 Modelkontrol
- 8 Et gennemregnet eksempel – fra bogen

Ensidet variansanalyse - Dekomposition og ANOVA-tabellen

- Med modellen

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

kan den totale variation i data opspaltes:

$$SST = SS(Tr) + SSE.$$

- 'Ensidet' hentyder til, at der kun er én faktor i forsøget (med k niveauer).
- Metoden kaldes variansanalyse, fordi testningen foregår ved at sammenligne varianser.

Formler for kvadratafvigelsessummer

- Den samlede variation

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

Formler for kvadratafvigelsessummer

- Den samlede variation

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

- Variation inden for grupperne (Variation tilbage efter model, dvs. af residualerne)

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

Formler for kvadratafvigelsessummer

- Den samlede variation

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

- Variation inden for grupperne (Variation tilbage efter model, dvs. af residualerne)

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

- Variation mellem grupperne (Variation forklaret af modellen)

$$SS(Tr) = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2$$

Ensidet variansanalyse - Parameterestimater

- $\hat{\mu} = \bar{y}$
- $\hat{\alpha}_i = \bar{y}_i - \bar{y}$
- $\hat{\sigma}^2 = MSE = \frac{SSE}{n-k}$

```
# Samlet gennemsnit  
mean(y)  
  
## [1] 5.233  
  
# Gruppegennemsnit  
tapply(y, treatm, mean)  
  
##      1      2      3  
## 3.025 5.900 6.775  
  
# SSE: Brug anova(..)
```

Variansanalyseskema

<i>Source of variation</i>	Deg. of freedom	Sums of squares	Mean sum of squares
<i>Treatment</i>	$k - 1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$
<i>Residual</i>	$n - k$	SSE	$MSE = \frac{SSE}{n-k}$
<i>Total</i>	$n - 1$	SST	

```
# Ensidet ANOVA med anova() og lm()
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
```

Dagsorden

- 1 Introduktion
- 2 Model og hypoteser
- 3 Beregning: Variansdekomposition og ANOVA-tabellen
- 4 Hypotesetest (F-test)
- 5 Variabilitet og sammenhæng med t-testen for to stikprøver
- 6 Post hoc sammenligninger
- 7 Modelkontrol
- 8 Et gennemregnet eksempel – fra bogen

Envejs variansanalyse - F-test

- Vi har (Sætning 8.2)

$$SST = SS(Tr) + SSE$$

- Herfra kan udlede teststørrelsen:

$$F = \frac{SS(Tr)/(k-1)}{SSE/(n-k)} = \frac{MS(Tr)}{MSE},$$

hvor

- k er antal nivauer af faktoren,
- n er antal observationer.

- Vælg et signifikansniveau α og beregn teststørrelsen F .
- Sammenlign teststørrelsen med $(1 - \alpha)$ -fraktilen i F -fordelingen:

$$F \sim F(k-1, n-k) \text{ (Sætning 8.6)}$$

F-fordelingen og F-testen

```
# Under H0:  
  
# Antal grupper  
k <- 3  
  
# Antal observationer  
n <- 12  
  
# Talrække plot  
xseq <- seq(0, 10, by = 0.1)  
  
# Plot tætheden for F-fordelingen  
plot(xseq, df(xseq, df1 = k-1, df2 = n-k), type = "l", xlab = "x", ylab = "f(x)")  
  
# Plot kritiske værdier for 5%-signifikansniveauet  
cr <- qf(0.95, df1 = k-1, df2 = n-k)  
abline(v = cr, col = "red")
```

Variansanalyseskema

<i>Source of variation</i>	Deg. of freedom	Sums of squares	Mean sum of squares	Test-statistic F	<i>p</i> -value
<i>treatment</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			

```
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
```

Ensidet ANOVA F-test "i hånden"

```
k <- 3; n <- 12 # Antal grupper og observationer
```

```
# Samlet variation: SST
```

```
SST <- sum((y - mean(y))^2)
```

```
# Variation af residualerne (inden for grupperne): SSE
```

```
y1 <- y[1:4]; y2 <- y[5:8]; y3 <- y[9:12]
```

```
SSE <- sum((y1 - mean(y1))^2) +  
       sum((y2 - mean(y2))^2) +  
       sum((y3 - mean(y3))^2)
```

```
# Variation forklaret af modellen/grupperingen (mellem grupperne): SS(Tr)
```

```
SSTr <- SST - SSE
```

```
# Teststørrelsen
```

```
Fobs <- (SSTr/(k-1)) / (SSE/(n-k))
```

```
# P-værdien
```

```
1 - pf(Fobs, df1 = k-1, df2 = n-k)
```

Dagsorden

- 1 Introduktion
- 2 Model og hypoteser
- 3 Beregning: Variansdekomposition og ANOVA-tabellen
- 4 Hypotesetest (F-test)
- 5 **Variabilitet og sammenhæng med t-testen for to stikprøver**
- 6 Post hoc sammenligninger
- 7 Modelkontrol
- 8 Et gennemregnet eksempel – fra bogen

Variabilitet og sammenhæng med t-testen for to stikprøver (Sætning 8.4)

Residualkvadratafvigelsessummen, SSE , divideret med $n - k$, også kaldet middelkvadratafvigelsen $MSE = SSE/(n - k)$, er et vægtet gennemsnit af stikprøvevarianserne for grupperne:

$$MSE = \frac{SSE}{n - k} = \frac{(n_1 - 1)s_1^2 + \cdots + (n_k - 1)s_k^2}{n - k},$$

$$s_i^2 = \frac{1}{n_i - 1} \sum_{i=1}^{n_i} (y_{ij} - \bar{y}_i)^2.$$

KUN når $k = 2$: (jf. Metode 3.52)

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

$$F_{\text{obs}} = t_{\text{obs}}^2,$$

hvor t_{obs} er den sammenvejede t-teststørrelse fra Metode 3.52 og 3.53.

Dagsorden

- 1 Introduktion
- 2 Model og hypoteser
- 3 Beregning: Variansdekomposition og ANOVA-tabellen
- 4 Hypotesetest (F-test)
- 5 Variabilitet og sammenhæng med t-testen for to stikprøver
- 6 Post hoc sammenligninger**
- 7 Modelkontrol
- 8 Et gennemregnet eksempel – fra bogen

Post hoc konfidensinterval – Metode 8.9

- En enkelt *forudplanlagt* sammenligning af forskellen på behandling i og j findes ved:

$$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2} \sqrt{\frac{SSE}{n-k} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)},$$

hvor $t_{1-\alpha/2}$ er fra t -fordelingen med $n - k$ frihedsgrader.

- Bemærk de færre frihedsgrader, da der estimeres flere parametre i beregningen af $MSE = SSE/(n - k) = s_p^2$ (det sammenvejede varianseestimat)
- Hvis alle $M = k(k - 1)/2$ kombinationer af parvise konfidensintervaller udregnes, så brug formlen M gange, men hver gang med $\alpha_{\text{Bonferroni}} = \alpha/M$.

Post hoc parvis hypotesetest – Metode 8.10

- For en enkelt *forudplanlagt* hypotesetest

$$H_0 : \mu_i = \mu_j, \quad H_1 : \mu_i \neq \mu_j$$

på niveau α , benyttes teststørrelsen

$$t_{\text{obs}} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}}$$

og p -værdien

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

hvor t -fordelingen med $n - k$ frihedsgrader anvendes.

- Hvis alle $M = k(k - 1)/2$ kombinationer af parvise hypotesetest udføres, så bruges det korrigerede signifikansniveau $\alpha_{\text{Bonferroni}} = \alpha/M$.

Dagsorden

- 1 Introduktion
- 2 Model og hypoteser
- 3 Beregning: Variansdekomposition og ANOVA-tabellen
- 4 Hypotesetest (F-test)
- 5 Variabilitet og sammenhæng med t-testen for to stikprøver
- 6 Post hoc sammenligninger
- 7 Modelkontrol**
- 8 Et gennemregnet eksempel – fra bogen

Varianshomogenitet

Se på box-plottet om spredningen ser (meget) forskellig ud for hver gruppe

```
# Tjek antagelsen om varianshomogenitet  
plot(treatm, y)
```

Normalfordelingsantagelsen

Se normalfordelings-QQ-plottet af residualerne:

```
# Tjek antagelsen om normalfordeling
fit1 <- lm(y ~ treatm)
qqnorm(fit1$residuals)
qqline(fit1$residuals)
```

Dagsorden

- 1 Introduktion
- 2 Model og hypoteser
- 3 Beregning: Variansdekomposition og ANOVA-tabellen
- 4 Hypotesetest (F-test)
- 5 Variabilitet og sammenhæng med t-testen for to stikprøver
- 6 Post hoc sammenligninger
- 7 Modelkontrol
- 8 Et gennemregnet eksempel – fra bogen

Et gennemregnet eksempel – fra bogen

Introduction to Statistics

Agendas

eNotes

Course Material

Podcast

Forum

Quiz

Admin

Dokumentegenhælder...

8.2.5 A complete worked through example: plastic types for lamps

Example 8.17 Plastic types for lamps

On a lamp two plastic screens are to be mounted. It is essential that these plastic screens have a good impact strength. Therefore an experiment is carried out for 5 different types of plastic. 6 samples in each plastic type are tested. The strengths of these items are determined. The following measurement data was found (strength in kJ/m²):

Type of plastic				
I	II	III	IV	V
44.6	52.8	53.1	51.5	48.2
50.5	58.3	50.0	53.7	40.8
46.3	55.4	54.4	50.5	44.5
48.5	57.4	55.3	54.4	43.9
45.2	58.1	50.6	47.5	45.9
52.3	54.6	53.4	47.8	42.5

Dagsorden

- 1 Introduktion
- 2 Model og hypoteser
- 3 Beregning: Variansdekomposition og ANOVA-tabellen
- 4 Hypotesetest (F-test)
- 5 Variabilitet og sammenhæng med t-testen for to stikprøver
- 6 Post hoc sammenligninger
- 7 Modelkontrol
- 8 Et gennemregnet eksempel – fra bogen

02402 Statistik (Polyteknisk grundlag)

Uge 12: Tosidet variansanalyse - ANOVA

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

Dagsorden

- 1 Introduktion: Regneeksempel og TV-data fra B&O
- 2 Modellen
- 3 Beregning: Variansdekomposition og ANOVA-tabellen
- 4 Hypotesetest (F-test)
- 5 Post hoc sammenligninger
- 6 Modekontrol
- 7 Et gennemregnet eksempel fra bogen

Variansanalyse - ANOVA

"ANalysis Of VAriance" (ANOVA) blev introduceret af R.A. Fisher for ca. 100 år siden som en systematisk måde at analysere grupper på og har siden da været helt centrale i statistik og anvendelser deraf.

- Sidste uge: Et inddelingskriterium (ensidet ANOVA)
- Denne uge: To inddelingskriterier (tosidet ANOVA)
- Inddelingskriterium = **faktor**
- Første faktor kaldes typisk *treatment*, anden faktor *block*

Dagsorden

- 1 Introduktion: Regneeksempel og TV-data fra B&O
- 2 Modellen
- 3 Beregning: Variansdekomposition og ANOVA-tabellen
- 4 Hypotesetest (F-test)
- 5 Post hoc sammenligninger
- 6 Modekontrol
- 7 Et gennemregnet eksempel fra bogen

Udvikling af fjernsyn hos Bang & Olufsen

Lyd- og billedkvalitet vurderes af mennesker:



Bang & Olufsen data i R

```
# Get the B&O data from the lmerTest-package
library(lmerTest)
data(TVbo)

# Alle 8 dommere bedømte de 12 kombinationer 2 gange.
# Her er pointene for første bedømmelse af et bestemt
# billede på tre forskellige TV
TVbo_sub <- subset(TVbo, Picture == 1 & Repeat == 1)[, c(1, 2, 9)]
sharp <- matrix(TVbo_sub$Sharpness, nrow = 8, byrow = T)
colnames(sharp) <- c("TV3", "TV2", "TV1")
rownames(sharp) <- c("Person 1", "Person 2", "Person 3",
                      "Person 4", "Person 5", "Person 6",
                      "Person 7", "Person 8")
library(xtable)
xtable(sharp)
```

Bang & Olufsen data i R

	TV3	TV2	TV1
Person 1	9.30	4.70	6.60
Person 2	10.20	7.00	8.80
Person 3	11.50	9.50	8.00
Person 4	11.90	6.60	8.20
Person 5	10.70	4.20	5.40
Person 6	10.90	9.10	7.10
Person 7	8.50	5.00	6.30
Person 8	12.60	8.90	10.70

Tosidet variansanalyse - Eksempel

- Samme data som for den ensidet ANOVA, men nu videt det, at forsøget var inddelt i blokke:

	Gruppe A	Gruppe B	Gruppe C
Blok 1	2.8	5.5	5.8
Blok 2	3.6	6.3	8.3
Blok 3	3.4	6.1	6.9
Blok 4	2.3	5.7	6.1

- Tre grupper fordelt på fire blokke
- Tre behandlinger fordelt på fire personer

Tosidet variansanalyse - Eksempel

- Samme data som for den ensidet ANOVA, men nu videt det, at forsøget var inddelt i blokke:

	Gruppe A	Gruppe B	Gruppe C
Blok 1	2.8	5.5	5.8
Blok 2	3.6	6.3	8.3
Blok 3	3.4	6.1	6.9
Blok 4	2.3	5.7	6.1

- Tre grupper fordelt på fire blokke
- Tre behandlinger fordelt på fire personer
- Ensidet eller tosidet ANOVA
- Fuldstændigt randomiseret forsøg eller Randomiseret blokforsøg

Tosidet variansanalyse - Eksempel

- Samme data som for ensidet ANOVA, men nu vides det, at forsøget var inddelt i blokke:

	Gruppe A	Gruppe B	Gruppe C
Blok 1	2.8	5.5	5.8
Blok 2	3.6	6.3	8.3
Blok 3	3.4	6.1	6.9
Blok 4	2.3	5.7	6.1

- Er der forskel (i middelværdierne) på grupperne A, B og C?

Tosidet variansanalyse - Eksempel

- Samme data som for ensidet ANOVA, men nu vides det, at forsøget var inddelt i blokke:

	Gruppe A	Gruppe B	Gruppe C
Blok 1	2.8	5.5	5.8
Blok 2	3.6	6.3	8.3
Blok 3	3.4	6.1	6.9
Blok 4	2.3	5.7	6.1

- Er der forskel (i middelværdierne) på grupperne A, B og C?
- Variansanalyse (ANOVA) kan anvendes til analysen, såfremt observationerne i hver celle kan antages at være normalfordelte, eller hvis der er tilstrækkeligt mange observationer (CLT).

Eksemplet i R

```
# 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)

# Behandling (gruppe)
treatm <- factor(c(1, 1, 1, 1,
                     2, 2, 2, 2,
                     3, 3, 3, 3))

# Blok (person)
block <- factor(c(1, 2, 3, 4,
                  1, 2, 3, 4,
                  1, 2, 3, 4))

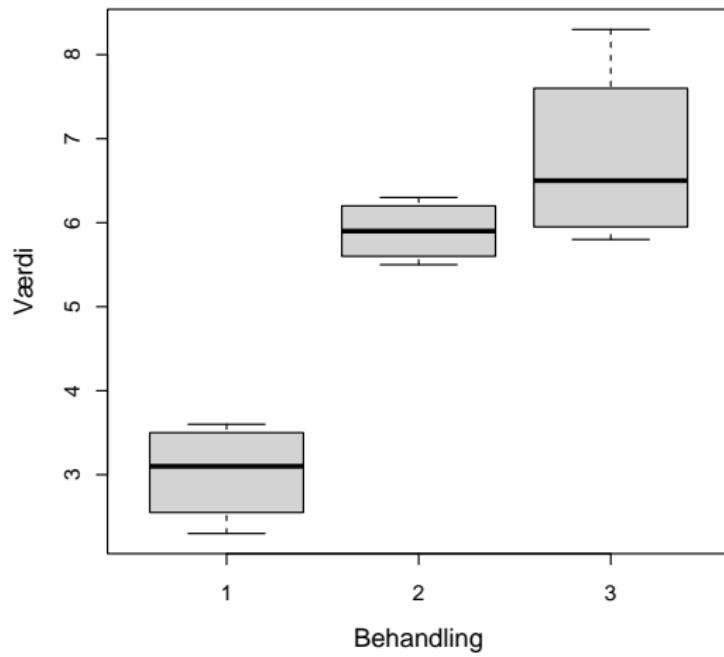
# Antal behandlinger og blokke
k <- length(unique(treatm))
l <- length(unique(block))

# Boxplot fordelt på behandlinger
plot(treatm, y, xlab = "Gruppe", ylab = "y")

# Boxplot fordelt på blokke
plot(block, y, xlab = "Blok", ylab="y")
```

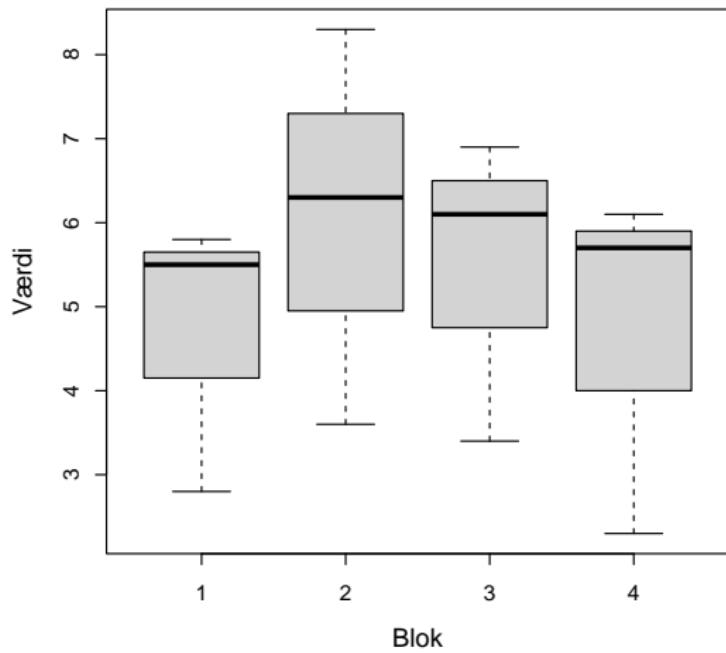
Eksemplet i R

```
# Boxplot fordelt på behandlinger  
plot(treatm, y, xlab = "Behandling", ylab = "Værdi")
```



Eksemplet i R

```
# Boxplot fordelt på blokke  
plot(block, y, xlab = "Blok", ylab="Værdi")
```



Dagsorden

- 1 Introduktion: Regneeksempel og TV-data fra B&O
- 2 **Modellen**
- 3 Beregning: Variansdekomposition og ANOVA-tabellen
- 4 Hypotesetest (F-test)
- 5 Post hoc sammenligninger
- 6 Modekontrol
- 7 Et gennemregnet eksempel fra bogen

Tosidet variansanalyse - Model

- Modellen:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij},$$

hvor fejlene er uafhængige og ensfordelte med

$$\varepsilon_{ij} \sim N(0, \sigma^2).$$

- μ er den samlede middelværdi
- α_i angiver effekten for behandling $i \in \{1, \dots, k\}$
- β_j angiver effekten for blok $j \in \{1, \dots, l\}$
- Der er k behandlinger og l blokke

Tosidet variansanalyse - Estimation

- Man beregner parameterestimaterne $\hat{\mu}$, $\hat{\alpha}_i$ og $\hat{\beta}_j$ ved:

$$\hat{\mu} = \bar{y} = \frac{1}{k \cdot l} \sum_{i=1}^k \sum_{j=1}^l y_{ij}$$

$$\hat{\alpha}_i = \left(\frac{1}{l} \sum_{j=1}^l y_{ij} \right) - \hat{\mu}$$

$$\hat{\beta}_j = \left(\frac{1}{k} \sum_{i=1}^k y_{ij} \right) - \hat{\mu}$$

```
# Samlet gennemsnit
mu_hat <- mean(y)

# Effekt af behandlinger
alpha_hat <- tapply(y, treatm, mean) - mu_hat

# Effekt af blokke
beta_hat <- tapply(y, block, mean) - mu_hat
```

Dagsorden

- 1 Introduktion: Regneeksempel og TV-data fra B&O
- 2 Modellen
- 3 Beregning: Variansdekomposition og ANOVA-tabellen
- 4 Hypotesetest (F-test)
- 5 Post hoc sammenligninger
- 6 Modekontrol
- 7 Et gennemregnet eksempel fra bogen

Tosidet variansanalyse - Dekomposition og variansanalyseskema - Sætning 8.20

- Med modellen

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2),$$

kan den totale variation i data opspaltes:

$$SST = SS(Tr) + SS(Bl) + SSE.$$

- 'Tosidet' hentyder til, at der er to faktorer i forsøget.
- Metoden kaldes variansanalyse, fordi testningen foregår ved at sammenligne varianser.

Formler for kvadratafvigelsessummer

- Den samlede variation (samme som for en ensidet analyse)

$$SST = \sum_{i=1}^k \sum_{j=1}^l (y_{ij} - \hat{\mu})^2$$

Formler for kvadratafvigelsessummer

- Den samlede variation (samme som for en ensidet analyse)

$$SST = \sum_{i=1}^k \sum_{j=1}^l (y_{ij} - \hat{\mu})^2$$

- Variation mellem behandlingerne/grupperne (Variation forklaret af *behandlingerne*)

$$SS(Tr) = l \cdot \sum_{i=1}^k (\bar{y}_{i \cdot} - \hat{\mu})^2 = l \cdot \sum_{i=1}^k \hat{\alpha}_i^2$$

Formler for kvadratafvigelsessummer

- Variation mellem blokkene/personerne (Variation forklaret af *blokkene*)

$$SS(Bl) = k \cdot \sum_{j=1}^l (\bar{y}_{\cdot j} - \hat{\mu})^2 = k \cdot \sum_{j=1}^l \hat{\beta}_j^2$$

- Variation af residualerne (Variation som modellen ikke forklarer)

$$SSE = \sum_{i=1}^k \sum_{j=1}^l (y_{ij} - \hat{y}_{ij})^2 = \sum_{i=1}^k \sum_{j=1}^l (y_{ij} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\mu})^2$$

Dagsorden

- 1 Introduktion: Regneeksempel og TV-data fra B&O
- 2 Modellen
- 3 Beregning: Variansdekomposition og ANOVA-tabellen
- 4 Hypotesetest (F-test)
- 5 Post hoc sammenligninger
- 6 Modekontrol
- 7 Et gennemregnet eksempel fra bogen

Tosidet variansanalyse - Hypoteze om forskellige behandlingseffekter - Sætning 8.22

- Man ønsker at sammenligne behandlingseffekterne (middelværdierne α_i) i modellen

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$

- Nulhypotesen om *ingen forskel/effekt mellem behandlingerne* kan formuleres som:

$$H_{0,Tr} : \quad \alpha_i = 0 \quad \text{for alle } i$$

$$H_{1,Tr} : \quad \alpha_i \neq 0 \quad \text{for mindst et } i$$

Tosidet variansanalyse - Hypoteze om forskellige behandlingseffekter - Sætning 8.22

- Man ønsker at sammenligne behandlingseffekterne (middelværdierne α_i) i modellen

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$

- Nulhypotesen om *ingen forskel/effekt mellem behandlingerne* kan formuleres som:

$$H_{0,Tr} : \quad \alpha_i = 0 \quad \text{for alle } i$$

$$H_{1,Tr} : \quad \alpha_i \neq 0 \quad \text{for mindst et } i$$

- Under $H_{0,Tr}$ gælder, at teststørrelsen

$$F_{Tr} = \frac{SS(Tr)/(k-1)}{SSE/((k-1)(l-1))}$$

er F -fordelt med $k-1$ og $(k-1)(l-1)$ frihedsgrader.

Tosidet variansanalyse - Hypoteze om forskellige blokeffekter - Sætning 8.22

- Man ønsker at sammenligne blokeffekterne (middelværdierne β_j) i modellen

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$

- Nulhypotesen om *ingen forskel/effekt mellem blokkene* kan formuleres som:

$$H_{0,Bl} : \beta_j = 0 \quad \text{for alle } j$$

$$H_{1,Bl} : \beta_j \neq 0 \quad \text{for mindst et } j$$

Tosidet variansanalyse - Hypoteze om forskellige blokeffekter - Sætning 8.22

- Man ønsker at sammenligne blokeffekterne (middelværdierne β_j) i modellen

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$

- Nulhypotesen om *ingen forskel/effekt mellem blokkene* kan formuleres som:

$$H_{0,Bl} : \beta_j = 0 \quad \text{for alle } j$$

$$H_{1,Bl} : \beta_j \neq 0 \quad \text{for mindst et } j$$

- Under $H_{0,Bl}$ gælder, at teststørrelsen

$$F_{Bl} = \frac{SS(Bl)/(l-1)}{SSE/((k-1)(l-1))}$$

er F -fordelt med $l-1$ og $(k-1)(l-1)$ frihedsgrader.

F-fordelingen og hypotesen for behandlinger

```
# Plot density of relevant F-distribution. Remember that this is "under H0"
# (computed as if H0 were true)
xseq <- seq(0, 10, by = 0.1)
plot(xseq, df(xseq, df1 = k-1, df2 = (k-1)*(l-1)), type = "l")

# Show critical value (5% signif. level) for test of treatment hypothesis
critical_value <- qf(0.95, df1 = k-1, df2 = (k-1)*(l-1))
abline(v = critical_value, col = "red")

# Compute value of the test statistic
(FTr <- (SSTr/(k-1)) / (SSE/((k-1)*(l-1)))) 

# Compute p-value for the test
1 - pf(FTr, df1 = k-1, df2 = (k-1)*(l-1))
```

F-fordelingen og hypotesen for blokke

```
# Plot density of relevant F-distribution. Remember that this is "under H0"
# (computed as if H0 were true)
xseq <- seq(0, 10, by = 0.1)
plot(xseq, df(xseq, df1 = l-1, df2 = (k-1)*(l-1)), type = "l")

# Show critical value (5% signif. level) for test of treatment hypothesis
critical_value <- qf(0.95, df1 = l-1, df2 = (k-1)*(l-1))
abline(v = critical_value, col = "red")

# Compute value of the test statistic
(FB1 <- (SSB1/(l-1)) / (SSE/((k-1)*(l-1)))) 

# Compute p-value for the test
1 - pf(FB1, df1 = l-1, df2 = (k-1)*(l-1))
```

Variansanalyseskema

<i>Source of variation</i>	Deg. of freedom	Sums of squares	Mean sum of squares	Test-statistic F	<i>p</i> -value
<i>Treatment</i>	$k - 1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F_{Tr} = \frac{MS(Tr)}{MSE}$	$P(F > F_{Tr})$
<i>Block</i>	$l - 1$	$SS(Bl)$	$MS(Bl) = \frac{SS(Bl)}{l-1}$	$F_{Bl} = \frac{MS(Bl)}{MSE}$	$P(F > F_{Bl})$
<i>Residual</i>	$(k - 1)(l - 1)$	SSE	$MSE = \frac{SSE}{(k-1)(l-1)}$		
<i>Total</i>	$n - 1$	SST			

```
anova(lm(y ~ treatm + block))

## Analysis of Variance Table
##
## Response: y
##             Df Sum Sq Mean Sq F value    Pr(>F)
## treatm      2 30.79   15.40   74.40 5.8e-05 ***
## block       3   3.95    1.32     6.37   0.027 *
## Residuals   6   1.24    0.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dagsorden

- 1 Introduktion: Regneeksempel og TV-data fra B&O
- 2 Modellen
- 3 Beregning: Variansdekomposition og ANOVA-tabellen
- 4 Hypotesetest (F-test)
- 5 Post hoc sammenligninger
- 6 Modekontrol
- 7 Et gennemregnet eksempel fra bogen

Post hoc konfidensintervaller

- Som ved ensidet ANOVA (brug metode 8.9 og 8.10), men skift $n - k$ frihedsgrader ud med $(k - 1)(l - 1)$ og brug MSE fra en tosidet ANOVA.
- Gøres med enten behandlinger eller blokke.

Post hoc konfidensintervaller

- Som ved ensidet ANOVA (brug metode 8.9 og 8.10), men skift $n - k$ frihedsgrader ud med $(k - 1)(l - 1)$ og brug MSE fra en tosidet ANOVA.
- Gøres med enten behandlinger eller blokke.
- En enkelt forudplanlagt sammenligning af forskellen på behandling i og j findes ved:

$$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2} \sqrt{\frac{SSE}{(k-1)(l-1)} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

hvor $t_{1-\alpha/2}$ er fra t -fordelingen med $(k - 1)(l - 1)$ frihedsgrader.

Post hoc konfidensintervaller

- Som ved ensidet ANOVA (brug metode 8.9 og 8.10), men skift $n - k$ frihedsgrader ud med $(k - 1)(l - 1)$ og brug MSE fra en tosidet ANOVA.
- Gøres med enten behandlinger eller blokke.
- En enkelt forudplanlagt sammenligning af forskellen på behandling i og j findes ved:

$$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2} \sqrt{\frac{SSE}{(k-1)(l-1)} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

hvor $t_{1-\alpha/2}$ er fra t -fordelingen med $(k - 1)(l - 1)$ frihedsgrader.

- Hvis M kombinationer af parvise konfidensintervaller udregnes, så brug formlen M gange, men hver gang med $\alpha_{\text{Bonferroni}} = \alpha/M$.

Post hoc parvis hypotesetest

- For en enkelt *forudplanlagt* hypotesetest

$$H_0 : \alpha_i = \alpha_j, \quad H_1 : \alpha_i \neq \alpha_j$$

på signifikansniveauet α beregnes teststørrelsen 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 -værdien ved:

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

hvor T følger en t -fordeling med $(k-1)(l-1)$ frihedsgrader.

Post hoc parvis hypotesetest

- For en enkelt *forudplanlagt* hypotesetest

$$H_0 : \alpha_i = \alpha_j, \quad H_1 : \alpha_i \neq \alpha_j$$

på signifikansniveauet α beregnes teststørrelsen 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 -værdien ved:

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

hvor T følger en t -fordeling med $(k-1)(l-1)$ frihedsgrader.

- Hvis M kombinationer af parvise konfidensintervaller udregnes, så bruges det korrigerede signifikansniveau: $\alpha_{\text{Bonferroni}} = \alpha/M$.

Dagsorden

- 1 Introduktion: Regneeksempel og TV-data fra B&O
- 2 Modellen
- 3 Beregning: Variansdekomposition og ANOVA-tabellen
- 4 Hypotesetest (F-test)
- 5 Post hoc sammenligninger
- 6 Modekontrol
- 7 Et gennemregnet eksempel fra bogen

Modelkontrol - Varianshomogenitet

Se på box plots om spredningen af *residualerne* ser ud til at afhænge af grupperne eller blokkene

```
# Estimer modellen
fit <- lm(y ~ treatm + block)

# Tjek box plots af residualerne
par(mfrow = c(1,2))
plot(treatm, fit$residuals, xlab = "Behandling")
plot(block, fit$residuals, xlab = "Blok")
```

Modelkontrol - Normalfordelingsantagelse

Se normalfordelings-QQ-plottet af residualerne:

```
# Normal QQ-plot for residualerne
qqnorm(fit$residuals)
qqline(fit$residuals)
```

Dagsorden

- 1 Introduktion: Regneeksempel og TV-data fra B&O
- 2 Modellen
- 3 Beregning: Variansdekomposition og ANOVA-tabellen
- 4 Hypotesetest (F-test)
- 5 Post hoc sammenligninger
- 6 Modekontrol
- 7 Et gennemregnet eksempel fra bogen

Et gennemregnet eksempel – fra bogen

The screenshot shows a navigation bar with the DTU logo, "Introduction to Statistics", and links for "Agendas", "eNotes", "Course Material", "Podcast", "Forum", "Quiz", and "Admin". On the right, there are links for "perbb", "Logout", and "arktosj". Below the navigation bar, a breadcrumb trail indicates the current page: "0.5.5 A complete worked through example: Car tires".

||| Example 8.26 Car tires

In a study of 3 different types of tires ("treatment") effect on the fuel economy, drives of 1000 km in 4 different cars ("blocks") were carried out. The results are listed in the following table in km/l.

	Car 1	Car 2	Car 3	Car 4	Mean
Tire 1	22.5	24.3	24.9	22.4	22.525
Tire 2	21.5	21.3	23.9	18.4	21.275
Tire 3	22.2	21.9	21.7	17.9	20.925
Mean	21.400	22.167	23.167	19.567	21.575

Let us analyse these data with a two-way ANOVA model, but first some explorative plotting:

Tosidet variansanalyse - Manglende eller flere observationer

Tosidet ANOVA i kurset her er meget "pæn" – vi har præcis én observation pr. række og søjle. Men i praksis:

- Mangler der ofte observationer i en gruppe.
- Man har mere end én observation nogle steder.
- Modellen kan nemt tilpasses. (R håndterer det meste)

	Gruppe A	Gruppe B	Gruppe C
Blok 1	2.8	5.5	5.8
Blok 2	NA	6.3	8.3
Blok 3	3.4	6.1	NA
Blok 4	2.3	5.7	6.1

Dagsorden

- 1 Introduktion: Regneeksempel og TV-data fra B&O
- 2 Modellen
- 3 Beregning: Variansdekomposition og ANOVA-tabellen
- 4 Hypotesetest (F-test)
- 5 Post hoc sammenligninger
- 6 Modekontrol
- 7 Et gennemregnet eksempel fra bogen