

Sandsynlighedsteori og statistik

1. Introduktion

Allan Leck Jensen

alj@ece.au.dk

Dagens program

- Kapitel 1. Introduktion
 - Introduktion til kurset og underviseren
 - Lave og bearbejde et datasæt
 - Hvad er statistik og hvad bruges det til?
 - Statistikværktøjet R
 - Grundlæggende statistiske begreber
 - Wisdom of the crowds
- Introduktion til R i RStudio
- Kapitel 2. Præsentation af data
 - Deskriptorer
 - Grafisk præsentation

Om underviseren

- Allan Leck Jensen
- 60 år, bor i Viborg
- Gift, 4 børn, 1 barnebarn



Om underviseren

- Seniorforsker ved Operations Management, Institut for Elektro- og Computerteknologi
- Uddannelse:
 - Hovedfag matematik, Aalborg Universitet
 - Bifag biologi, Aarhus Universitet
 - PhD datalogi, Aalborg Universitet
- Ansættelser:
 - Statens Planteavlsforsøg (SP)
 - Danmarks JordbrugsForskning (DJF)
 - Aarhus Universitet, Det Jordbrugsvidenskabelige Fakultet (DJF)
 - Aarhus Universitet, Inst. for Ingeniørvidenskab (ENG)
 - Aarhus Universitet, Inst. for Elektro- og Computerteknologi (ECE)
- Kontorer på Katrinebjerg, AU Foulum og hjemme (kontakt med email)
- Undervisning:
 - Statistik for Ingeniører, fra 2015 (Diplom)
 - Programmering for Elektroteknologi (Python), fra 2019 (BSc EE)
 - Sandsynlighedsteori og Statistik, fra 2020 (BSc EE og CE).

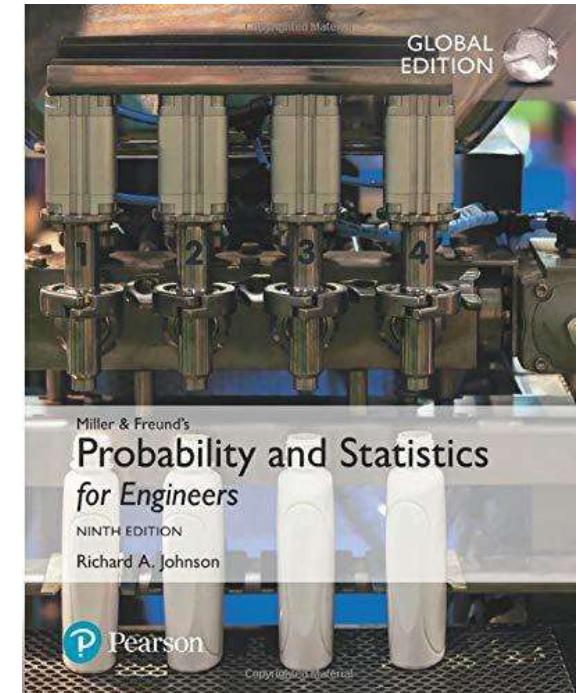
Kurset

- Litteratur:
 - Miller & Freund's Probability and Statistics for Engineers, 9th global edition, by Richard Johnson, Pearson 2018, ISBN-13: 978-1292176017
 - Kan købes i Stakbogladen (nu til 558.-):

Sandsynlighedsteori og statistik

Miller & Freund's Probability and Statistics for Engineers
Global Edition, 9th ed.
Pearson, 2017
ISBN: 9781292176017

620,00 kr.
558,00 kr.



- Man kan godt bestå uden bogen. Den er ikke fantastisk pædagogisk, men den har over 250 eksempler, og de fleste når vi ikke at tale om.

Emner fra bogen (pensum)

Emne		Litteratur	
Overordnet	Specifikt	Afsnit	Sider
1. Introduktion	Introduktion (selvstudium)	1.1-1.6	11-19
2. Præsentation af data	Diagrammer og deskriptorer	2.1-2.7	22-45
3. Sandsynlighedsteori	Beregning af sandsynlighed	3.1-3.7	56-87
4. Sandsynlighedsfordelinger	Diskrete stokastiske variable, bla. binomial- og poisson-fordeling	4.1-4.2, 4.4, 4.6-4.7	94-103, 107-114, 118-124
5. Sandsynlighedstætheder	Kontinuerte stokastiske variable, bla. normal- og eksponential-fordeling	5.1-5.2, 5.4-5.5, 5.7, 5.10, 5.12-5.13	134-147, 151-152, 155-157, 161-173, 180-183
6. Stikprøvefordeling	Populationer og stikprøver, fordeling af en stikprøves middelværdi og varians, den centrale grænseværdidisætning	6.1-6.4	193-210
7. Inferens om middelværdi	Inferens om middelværdi, konfidensinterval og hypotesetest	7.1-7.3, 7.5-7.8	223-232, 242-257
8. Sammenligning af to stikprøver	Hypotesetest for middelværdi af to stikprøver	8.1-8.5	266-286
9. Inferens om varians	Hypotesetest for varians af en og to stikprøver	9.1-9.3	290-298
10. Chi-i-anden test	Kontingenstabeller og goodness of fit	10.4-10.5	318-323
11. Regressionsanalyse	Lineær, ikke-lineær og multipel regression	11.1-11.7	327-381
12. Variansanalyse	ANOVA	12.1-12.2, 12.4	386-399, 410-413
13. Faktorielle eksperimenter	Eksperimenter med to og flere faktorer	13.1-13.2	425-438

Lektionsplan (14 lektioner)

Uge	Lokale 5106-110	
	Torsdage	12:15-16:00
35	01-09-2022	L1
36	08-09-2022	L2
37	15-09-2022	L3
38	22-09-2022	L4
39	29-09-2022	L5
40	06-10-2022	L6
41	13-10-2022	L7
42	20-10-2022	Efterårsferie
43	27-10-2022	L8
44	03-11-2022	L9
45	10-11-2022	L10
46	17-11-2022	L11
47	24-11-2022	L12
48	01-12-2022	L13
49	08-12-2022	L14
50	15-12-2022	Eksamensperiode

Statistik i Calculus Beta

- I har allerede lært noget sandsynlighedsteori og statistik (ifølge Calculus Beta compendium):

10 Diskrete stokastiske variable	136
10.1 Stokastiske eksperimenter og stokastiske variable	139
10.2 Lotterimodellen	143
10.3 Kombinatorik	146
10.4 Uafhængighed	151
11 Binomial-, Multinomial- og Poissonfordelingen	154
11.1 Binomialfordelingen	154
11.2 Multinomialfordelingen	157
11.3 Poissonfordelingen	161
12 Middelværdi, varians og kovarians – diskrete stokastiske variable	165
12.1 Middelværdi for diskrete stokastiske variable	165
12.1.1 Fortolkning af middelværdien: Store Tals Lov	168
12.2 Varians for diskrete stokastiske variable	168
12.3 Kovarians for diskrete stokastiske variable	170
13 Kontinuerte stokastiske variable	173
13.1 Kontinuerte stokastiske variable og tæthedsfunktioner	173
13.2 Eksempler på kontinuerte fordelinger	175
13.3 Fordelingsfunktioner	179
14 Middelværdi og varians — kontinuerte stokastiske variable	182
14.1 Middelværdi for kontinuerte stokastiske variable	182
14.2 Varians for kontinuerte stokastiske variable	185
15 Normal- og χ^2-fordelingen	188
15.1 Normalfordelingen	188
15.2 χ^2 -fordelingen	191

Eksamensoplysninger

- Fra kursuskataloget: <https://kursuskatalog.au.dk/da/course/114322>
- **Hjælpemidler:**
Alt er tilladt, bortset fra kommunikation med andre.

Eksamensoplysninger	SKRIFTLIG
Eksamensstid:	3 time(r)
Hjælpemidler:	Anviste

Bedømmelse:	7-trinsskala
Censurform:	ekstern censur

Bemærkninger
Hjælpemidler som computer, matematisk software, bøger, noter, pen og papir og internet er tilladt. Håndskrevne dele af eksamensbesvarelserne skal digitaliseres og vedhæftes eksamensbesvarelsen i "Digital eksamen". Læs mere om reglerne for "Digital eksamen" på Studieportalen.

Værktøjer

- Der er ingen krav til, hvilket statistikværktøj, I vil bruge
- Jeg er vant til at bruge MATLAB, men det må vi ikke pga. licens
- Excel og WordMat har statistik-funktionalitet, men er ikke tilstrækkelig
- Jeg har overvejet R og Python som alternativer. Jeg har valgt R, bl.a. fordi bogen bruger R til kodeeksempler
- *Jeg vil stærkt anbefale, at I også bruger R, så vi kan udveksle kode og erfaring. Der vil være løsningsforslag til opgaver i R*
- Jeg kan desuden anbefale at bruge RStudio som R editor (både R og RStudio kan installeres i gratis versioner)
- I kan bruge andet statistisk software, f.eks. Python, Mathcad, Maple, osv., hvis I vil. Men jeg kan ikke hjælpe!
- Desuden bruger vi **Brightspace** som kursusværktøj.

Installering af R og RStudio

- <https://cran.r-project.org/>
Jeg har version 4.1.3 til Windows
- <https://rstudio.com/products/rstudio/download/>
Jeg har RStudio Desktop 2022.07.1-544 til Windows

RStudio is becoming Posit in October. Learn more at posit.co ↗

Hvad er statistik (ifølge Wikipedia)?

- **Statistik** er en videnskabelig metode, hvormed man effektivt anvender numeriske **data**, som fx kan komme fra eksperimenter, spørgeskemaer eller registre
- Historisk set startede statistik med at være beskrivende, hvor fokus var at præsentere data grafisk, med tabeller og senere ved at regne **statistiske mål** som **gennemsnit**
- Moderne statistik omfatter at drage konklusioner om det generelle tilfælde (**hele populationen**) ud fra det enkelte tilfælde (en **stikprøve**). Det kan for eksempel være at bestemme **parametre** til **sandsynlighedsfordelingen** for populationen.
Dette kaldes **statistisk inferens**
- Et andet eksempel kunne være at bestemme, om der er **forskel** på to populationer (eksempelvis en behandlet gruppe og en placebo-gruppe).

Introduktion til statistik

- Lave et datasæt
- Beskrive datasættet ved grundlæggende statistiske mål
- Tale om hvad statistik er, og hvad det bruges til.

Datasæt: De studerendes gæt

- *Hvad er dybden af dette lokale, målt i hele millimeter?*
- Udfyld formular med dit gæt
- Kun 1 gæt per person!
- Diskutter *ikke* med andre!
- Den, der kommer tættest på den korrekte værdi, vinder en præmie



Gæt lokalets dybde - ECEStat 2022

Hvor langt er der fra væg til væg i millimeter?

allanleck@gmail.com (not shared) Switch accounts 

*Required

Navn: *

Your answer

Studienummer: *

Your answer

Gæt (svaret skal være et helt antal millimeter): *

Your answer

Submit **Clear form**

- Link til formularen er på Brightspace under K1.

Statistik kan give overraskende resultater

Ingeniøren

 Privatlivspolitik | Log

Nyheder | Blogs | Debat | Jobfinder | Avisarkiv | Kursusguide | Events | Insights

TENDENS SMART CITIES | SELVLÆRENDE ROBOTTER | FREMTIDENS FØDEVARER | FOKUS SELVKØRENDE

Cykelstier medfører flere ulykker

Antallet af uheld stiger, når der bygges cykelstier, viser undersøgelser fra Trafikforskningsgruppen på Aalborg Universitet.

Af Birgitte Marfelt 22. aug 2005 kl. 00:00



Når der bliver bygget nye cykelstier, kommer flere cyklister til skade. Den paradoksale konklusion kommer en ny undersøgelse fra Aalborg Universitet frem til.

Trafiksikkerhedsmæssigt er det overraskende, at cykelstier øger cyklisternes risiko. Flere cykelstier har hidtil været feltråbet fra både Rådet for Større Færdselssikkerhed og Dansk Cyklist Forbund.

- Hvem cyklede på strækningerne før og efter cykelstierne kom? Hvor mange?

F Professor tvivler på o årsagssammenhæng mellem A omskæring og autisme 9.

Rikke Gjøl Mansø, Berlingske Nyhedsbureau Fredag den 9. januar 2015, 07:28

**N Selvom et nyt forskningsprojekt viser, at drenge, der omskæres, har højere risiko
fo for at udvikle autisme, kalder professor det tvivlsomt. Han efterspørger
yderligere forskning.**

VDrenge, der omskæres, har 46 procent højere risiko for at udvikle autisme, inden de fylder ti år, end andre drenge. Sådan lyder konklusionen i et nyt forskningsprojekt fra Statens Serumssinsitut (SSI) ifølge Jyllands-Posten. Men man skal passe på at konkludere for meget, advarer Carsten Obel, der er professor i mental børnesundhed ved Aarhus Universitet:

»Jeg er overbevist om, at de to ting statistisk hænger sammen, men det er ikke det, vi taler om. Det interessante er jo, om der er en årsagssammenhæng. Og der skal mere til,« siger han til Berlingske Nyhedsbureau og uddyber:

»Der mangler i hvert fald noget, før vi kan formidle videre til folk, at omskæring og autisme skulle hænge årsagsmæssigt sammen. Når man finder sådan en sammenhæng, skal man gå videre for at undersøge sammenhængen og ikke bare konkludere definitivt på den.«

TForskeren bag forskningsprojektet, professor Morten Frisch fra SSI, der er kendt som en højlydt modstander af omskæring, foreslår, at det kan være den voldsomme smerteoplevelse, som en omskæring kan være, der giver øget risiko for autisme.

Ilad Før man melder noget ud, skal der kontrolleres for sociale forskelle, mener Carsten Obel:

»Jeg tror, man skal være forsiktig med at konkludere. Man skal eksempelvis overveje, om de mennesker, der lader deres børn omskære, adskiller sig fra de mennesker, der ikke gør det. Så det ikke er det at omskære, der gør udslaget, men en indikator for, hvem der lader deres børn omskære,« siger han og understreger, at autisme er en af de mest komplicerede sygdomme.

»Det er vigtigt at undersøge sammenhængen og følge den statistiske sammenhæng til dørs for at få undersøgt, om der også er en årsagssammenhæng.«

Sammenhænge i data

- Der er forskel på statistisk sammenhæng og årsagssammenhæng
- Eksempel:
Spørgeundersøgelse i et gymnasium i 2005 om holdning til ny EU-forfatning

Linje	Ja-sigere	Nej-sigere	Ja-pct
Sproglige	49	82	37 %
Matematikere	88	88	50 %

Konklusion: Matematikere er mere positive overfor EU

- Forkert!**

Køn	Linje	Ja-sigere	Nej-sigere	Ja-pct
Piger	Sproglige	34	74	31 %
Piger	Matematikere	30	60	33 %
Drenge	Sproglige	15	8	65 %
Drenge	Matematikere	58	28	67 %

Det er *drenge*, der er mere positive overfor EU. Køn var en *skjult variabel*.

Intimbarberede udsatte for kønssygdomme

STI: Sexually Transmitted Infections

ORIGINAL ARTICLE

Correlation between pubic hair grooming and STIs: results from a nationally representative probability sample

E Charles Osterberg,^{1,2} Thomas W Gaither,¹ Mohannad A Awad,¹ Matthew D Truesdale,¹ Isabel Allen,³ Siobhan Sutcliffe,⁴ Benjamin N Breyer^{1,3}

► Additional material is published online only. To view please visit the journal online (<http://dx.doi.org/10.1136/setrans-2016-052687>).

¹Department of Urology, University of California—San Francisco, San Francisco, California, USA

²Department of Surgery, University of Texas—Dell Medical School, Austin, Texas, USA

³Department of Biostatistics and Epidemiology, University of California—San Francisco, San Francisco, California, USA

⁴Division of Public Health Sciences, Department of Surgery, Washington University

ABSTRACT

Objective STIs are the most common infections among adults. Concurrently, pubic hair grooming is prevalent. Small-scale studies have demonstrated a relationship between pubic hair grooming and STIs. We aim to examine this relationship in a large sample of men and women.

Design We conducted a probability survey of US residents aged 18–65 years. The survey ascertained self-reported pubic hair grooming practices, sexual behaviours and STI history. We defined extreme grooming as removal of all pubic hair more than 11 times per year and high-frequency grooming as daily/weekly trimming. Cutaneous STIs included herpes, human papillomavirus, syphilis and molluscum. Secretory STIs included gonorrhoea, chlamydia and HIV. We analysed lice separately.

by bacterial or viral STIs, such as human papillomavirus (HPV) and molluscum contagiosum.⁶ This hypothesis is supported by a small-scale report of increased molluscum contagiosum acquisition among groomers.⁷ On the other hand, grooming removes the amount and length of pubic hair, which may reduce the risk of acquiring other sexually transmitted pathogens, such as pubic lice. This hypothesis is also supported by a small-scale report.⁸ Finally, as pubic hair grooming is correlated with an increased number of lifetime sexual partners and is viewed as a preparatory act to sexual engagement,^{1 4 9 10} it may also serve as a marker of increased STI risk. Irrespective of the underlying mechanism—whether a causal relation or statistical association—understanding the possible link between pubic hair grooming and STI

Statistik har et dårligt ry



Statistik er som en gadelygte. Ikke særligt oplysende, men god at støtte sig til. - Robert Storm Petersen

- **Statistik er ...**
- den videnskab der kan bevise alt undtagen nytten af statistik.
- *Evan Esar*
- kunsten med præcise termer at fastslå det man ikke ved.
- *William Kruskal*
- indsamling, analyse og tolkning af numeriske data på en sådan måde, at de kan forstås af computere og misforstås af alle andre.
- *Anonym*
- som en bikini: den viser noget interessant og skjuler noget væsentligt.
- *Peter von Zahn m.fl..*

Andre har sagt om statistik ...

- Jeg har kun tiltro til en statistik, når jeg selv har forfalsket den.
- *Winston Churchill*
- Der findes tre slags løgne: almindelige løgne, forbandede løgne og statistik.
- *Benjamin Disraeli m.fl.*
- Jeg kan bevise alt ved hjælp af statistik, bortset fra sandheden.
- *Georg Canning*
- Med statistik kan man bevise alt - også det modsatte.
- *James Callaghan.*

Dataindsamling og databehandling

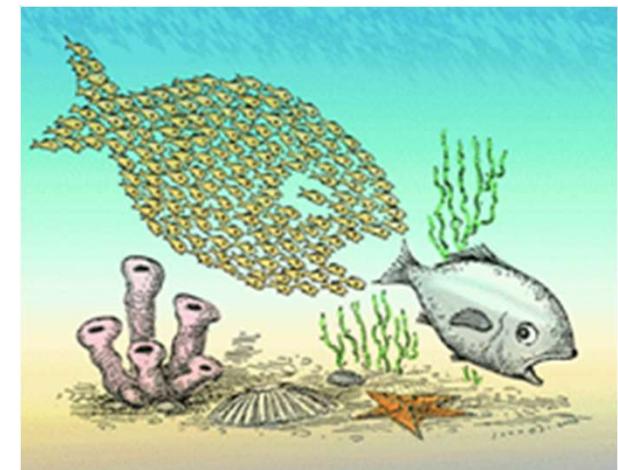
- Vi laver et datasæt, som består af jeres gæt på værdien
- Hver persons gæt er en observation i datasættet
- Vi kan beskrive datasættet med statistiske deskriptorer, f.eks.:
 - Median (engelsk: *median*)
 - Middelværdi (engelsk: *average, mean*)
 - Varians (engelsk: *variance*)
 - Spredning / standardafvigelse (engelsk: *standard deviation*)
- Vi kan beskrive datasættet grafisk, f.eks.:
 - Stolpediagram (engelsk: *bar chart*).

And the winner is ...



Wisdom of the Crowds

- Dyrskue i Plymouth 1906
- Konkurrence om at gætte vægten på en okse
- 787 personer deltog i konkurrencen, men ingen ramte rigtigt ([1198](#) pund)
- En af deltagerne, statistikeren Francis Galton (fætter til Charles Darwin), opdagede at:
 - *medianen* for observationerne ([1207](#) pund) var mindre end 1% fra den korrekte værdi
 - *middelværdien* ([1197](#) pund) ramte næsten præcis rigtigt
- Morale: Ofte er vi klogere i flok end individuelt
- **Wisdom of the Crowds:**
En forsamlings individuelle gæt kan modelleres som en sandsynlighedsfordeling, hvis middelværdi er tæt på den sande værdi for det, der skal gættes.



Hjemmeopgave

- Hent datasættet over jeres gæt på Brightspace (Filer under K1)
- Indlæs datasættet i RStudio
- Brug R til at bestemme:
 - Antal observationer
 - Minimumværdi
 - Maksimumværdi
 - Median
 - Middelværdi
 - Varians
 - Standardafvigelse.

Sandsynlighedsteori og statistik

Kapitel 2. Organisering og præsentation af data (afsnit 2.1-2.7)

Allan Leck Jensen
alj@ece.au.dk

Præsentation af data

Hvorfor?

- Forsøge at forstå data
- Forsøge at beskrive data
- "Listen to the data"
- Få hurtigt overblik – vurdere om det er relevant at lave yderligere statistisk analyse

0.3	1.0	0.8	1.1	1.3	1.1	2.4
2.9	1.6	1.3	0.4	0.7	1.5	1.3
2.1	1.9	1.0	1.8	1.1	2.2	1.9
0.7	0.4	1.5	0.7	2.4	2.4	2.8
2.7	1.1	1.2	1.1	1.1	0.9	1.3

Hvordan?

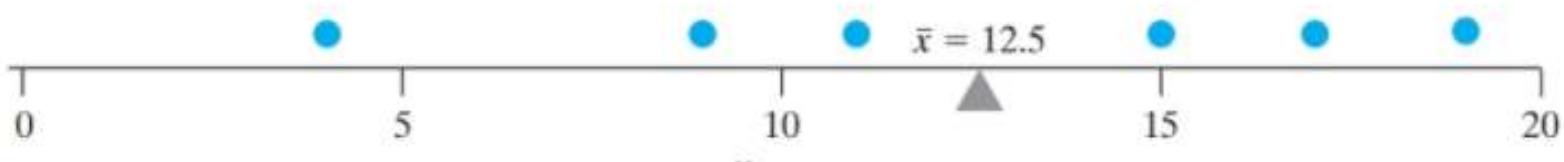
- Beregne deskriptorer
- Præsentere data grafisk.

De vigtigste deskriptorer

- Median
(engelsk: **Median**)
- Middelværdi
(engelsk: **Mean** eller **Average**)
- Varians
(engelsk: **Variance**)
- Standardafvigelse (også kaldet spredning)
(engelsk: **Standard deviation**)
- Variationskoefficient
(engelsk: **Coefficient of variation**).

Middelværdi μ ('my')

- Datasæt med n observationer: x_1, x_2, \dots, x_n
Middelværdi: $\mu = \frac{1}{n} \sum_{i=1}^n x_i$
- Middelværdien kan opfattes som observationernes balancepunkt



- Hvis data kun kan antage k forskellige værdier, så kan det være lettere at beregne middelværdien som

$$\mu = \frac{1}{n} \sum_{i=1}^k h_i \cdot x_i$$

hvor h_i er antal gange, som x_i er observeret.

Således er

$$\sum_{i=1}^k h_i = n .$$

Varians σ^2 ('sigma i anden')

- Varians er den gennemsnitlige, kvadrerede afstand til middelværdien:

Varians:
$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

- Varians kan også beregnes således:

Varians:
$$\sigma^2 = \frac{1}{n} \left(\sum_{i=1}^n x_i^2 \right) - \mu^2 .$$

Varians σ^2

Beregningsmetode (fordel ved manuel beregning):

$$\begin{aligned}\sigma^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \\&= \frac{1}{n} \sum_{i=1}^n (x_i^2 + \mu^2 - 2\mu x_i) \\&= \frac{1}{n} \sum_{i=1}^n x_i^2 + \frac{1}{n} \sum_{i=1}^n \mu^2 - \frac{1}{n} \sum_{i=1}^n 2\mu x_i \\&= \frac{1}{n} \sum_{i=1}^n x_i^2 + \mu^2 \frac{1}{n} \sum_{i=1}^n 1 - 2\mu \frac{1}{n} \sum_{i=1}^n x_i \\&= \frac{1}{n} \sum_{i=1}^n x_i^2 + \mu^2 - 2\mu^2 \\&= \frac{1}{n} \sum_{i=1}^n x_i^2 - \mu^2 .\end{aligned}$$

Standardafvigelse / spredning σ ('sigma')

- Standardafvigelsen er kvadratroden af variansen. Dermed har standardafvigelsen samme skala (måleenhed) som data

Standardafvigelse:

$$\begin{aligned}\sigma &= \sqrt{\sigma^2} \\ &= \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - \mu^2}\end{aligned}$$

- Standardafvigelsen opfattes som et mål for observationernes gennemsnitlige afstand fra middelværdien.

Eksempel: Højde på 25 elever i A klasse

167	183	188	172	167
162	182	173	160	173
170	160	166	163	179
184	183	169	151	182
162	167	172	169	186

- Middelværdi:

$$\mu = \frac{1}{25} (167 + 162 + 170 + \dots + 182 + 186) = 171.6 \text{ [cm]}$$

- Varians:

$$\sigma^2 = \frac{1}{25} (167^2 + 162^2 + 170^2 + \dots + 186^2) - 171.6^2$$

$$\sigma^2 = 29535.68 - 29446.56 = 89.12 \text{ [cm}^2\text{]}$$

- Standardafvigelse:

$$\sigma = \sqrt{89.12} = 9.44 \text{ [cm]} .$$

Eksempel: Højde på 25 elever i A klasse

- B klassen har også 25 elever:
 - 10 elever på 171 cm og
 - 15 elever på 172 cm

- Middelværdi:

$$\mu = \frac{1}{25} (10 \cdot 171 + 15 \cdot 172) = 171.6 \text{ [cm]}$$

- Varians:

$$\begin{aligned}\sigma^2 &= \frac{1}{25} (10 \cdot 171^2 + 15 \cdot 172^2) - 171.6^2 \\ \sigma^2 &= 29446.80 - 29446.56 = 0.24 \text{ [cm}^2\text{]}\end{aligned}$$

- Standardafvigelse:

$$\sigma = \sqrt{0.24} = 0.49 \text{ [cm]}$$

De to klasser har samme gennemsnitshøjde, men B-klassen har meget mindre standardafvigelse end A-klassen.

Vi kan bruge middelværdi og standardafvigelse til at beskrive de to datasæt.

Population eller stikprøve

- I eksemplet med højden af 25 elever fra A klassen har vi opfattet dem som hele ‘populationen’, vi var interesserede i
- Alternativt kan man opfatte data som en **stikprøve**, som vi ønsker at bruge til at udtales om hele **populationen** af tilsvarende elever
- Vi kender ikke middelværdi, varians og standardafvigelse for alle danske gymnasieelever, men vi kan estimere det ved en repræsentativ stikprøve, f.eks. A klassen
- I det tilfælde beregnes variansen lidt anderledes:

	Population	Stikprøve
Middelværdi	$\mu = \frac{1}{n} \sum_{i=1}^n x_i$	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
Varians	$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$	$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
Standardafvigelse	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

- Ved at dele med $n - 1$ i stedet for n får vi et bedre estimat for σ^2 .

Variationskoefficient

- En standardafvigelse på 5 er lille, hvis middelværdien for datasættet er 1000. Men hvis middelværdien er 2, så er det en stor standardafvigelse
- **Variationskoefficient** (*Coefficient of Variation, CV*) er et standardiseret mål for den relative standardafvigelse:

$$CV = \frac{\sigma}{\mu} \cdot 100\%$$

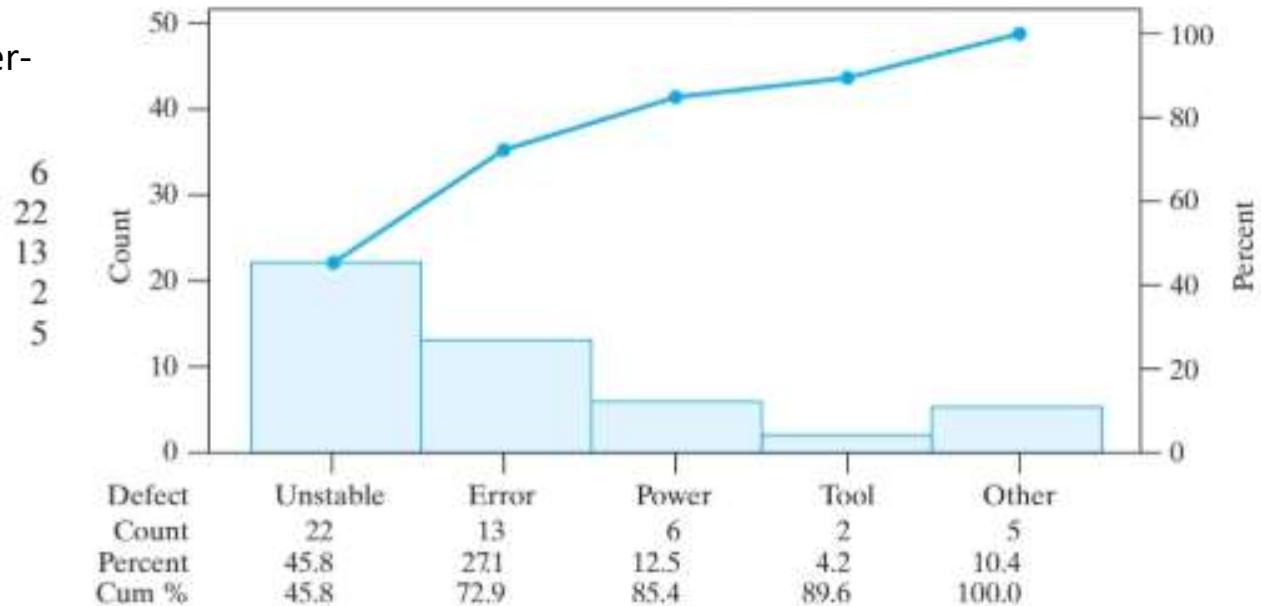
- For klasse A er $CV_A = \frac{9.44}{171.6} \cdot 100 \% = 5.5 \%$
- For klasse B er $CV_B = \frac{0.49}{171.6} \cdot 100 \% = 0.3 \%$
- Variationskoefficienten er et mål for præcisionen i data. Da det er standardiseret og dimensionsløst kan CV for forskellige datasæt sammenlignes.

Grafisk præsentation

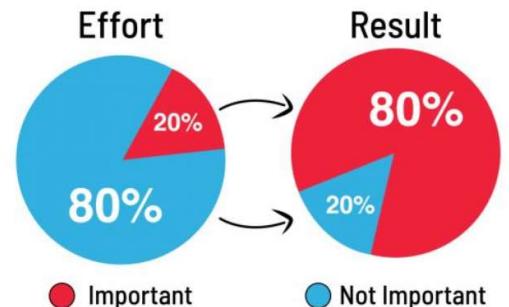
- M&F starter med at præsentere et **Pareto diagram**:

Årsager til nedbrud af en computer-styret drejebænk:

power fluctuations
controller not stable
operator error
worn tool not replaced
other

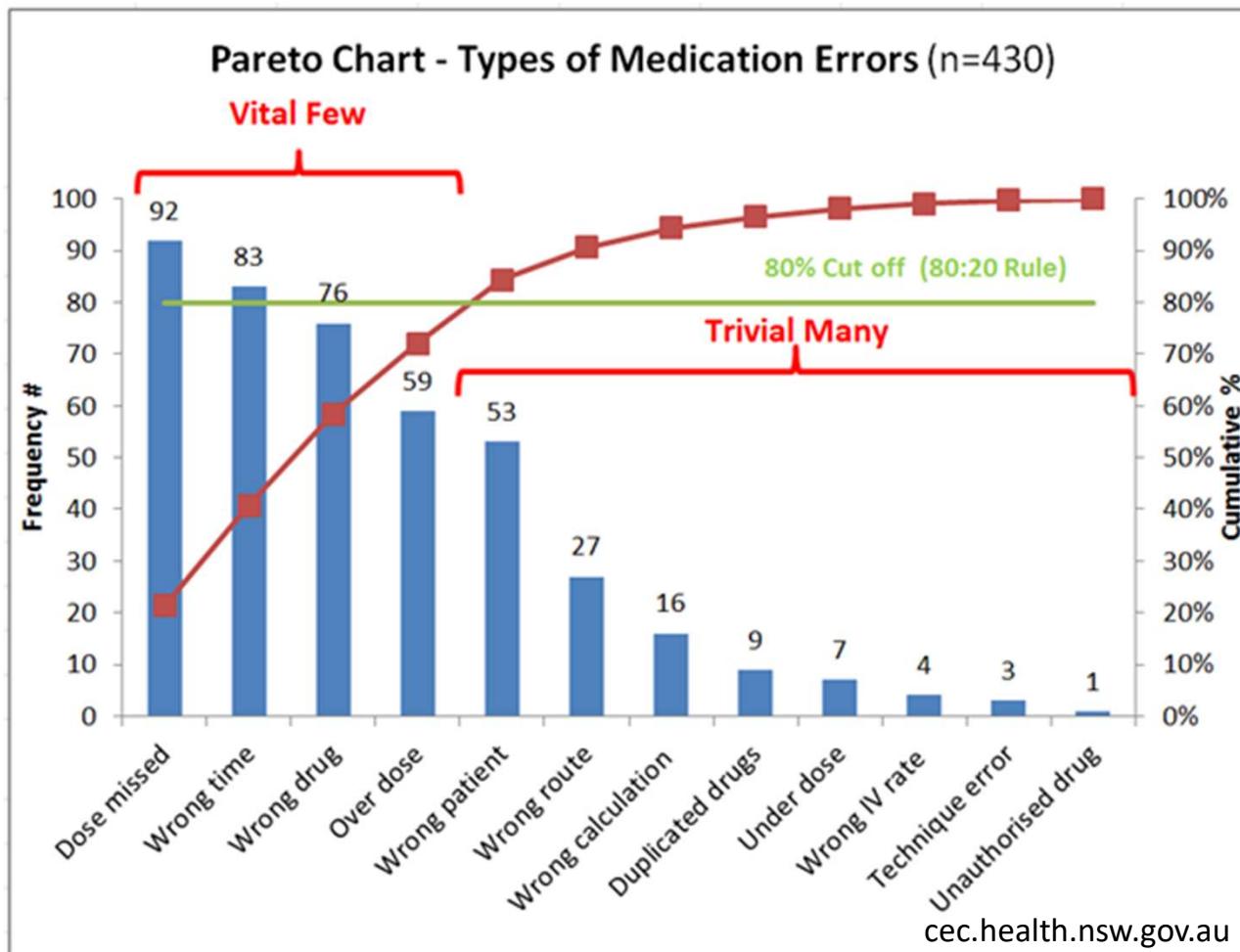


- Pareto diagrammet består af to elementer:
 - Søjlediagram over årsagerne, typisk sorteret efter størrelse (venstre akse)
 - Kurve over den kumulerede procentvise andel af nedbrud (højre akse)
- Pareto's regel (80/20 reglen): 80 % af effekterne kommer fra 20 % af årsagerne (f.eks. 20 % af de Covid-19 smittede er skyld i 80 % af smittespredningen)



Eksempel på Pareto-reglen fra medicin

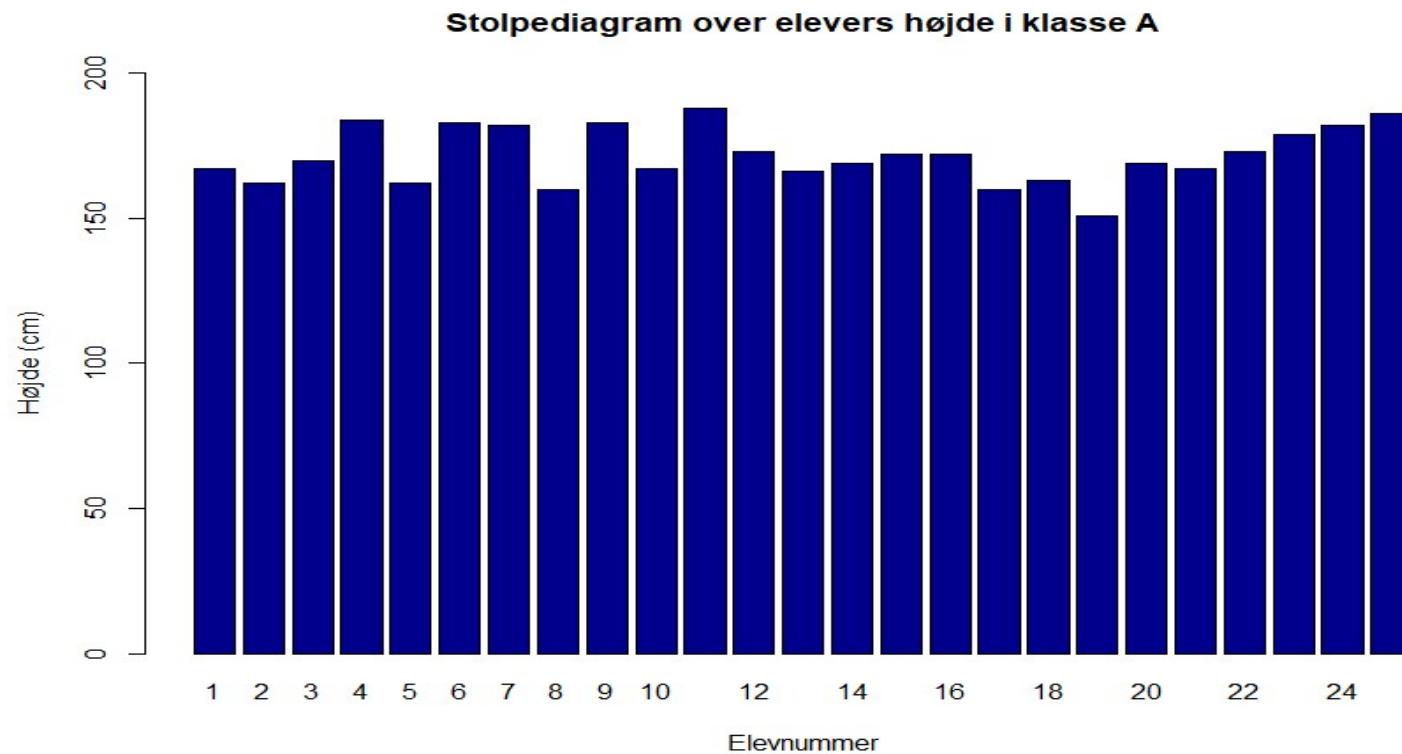
- $310/430 = 72\%$ af medicineringsfejl skyldes $4/12 = 33\%$ af årsagerne
- Hvis man fokuserer på de ‘vital few’ årsager kan mange fejl undgås
- *Vi kommer ikke til at bruge Pareto diagrammer.*



Stolpediagram

- **Stolpediagram** over observationerne af A-klassens elevers højde ‘råt’ i observeret rækkefølge
- Funktion i R: barplot().

167	183	188	172	167
162	182	173	160	173
170	160	166	163	179
184	183	169	151	182
162	167	172	169	186

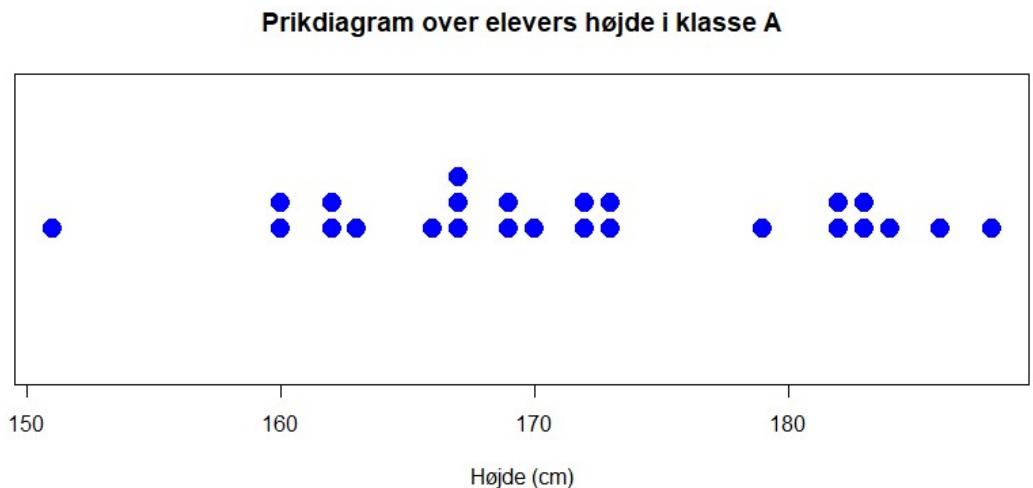


Prikdiagram

- Observationerne organiseres efter størrelse og hyppighed

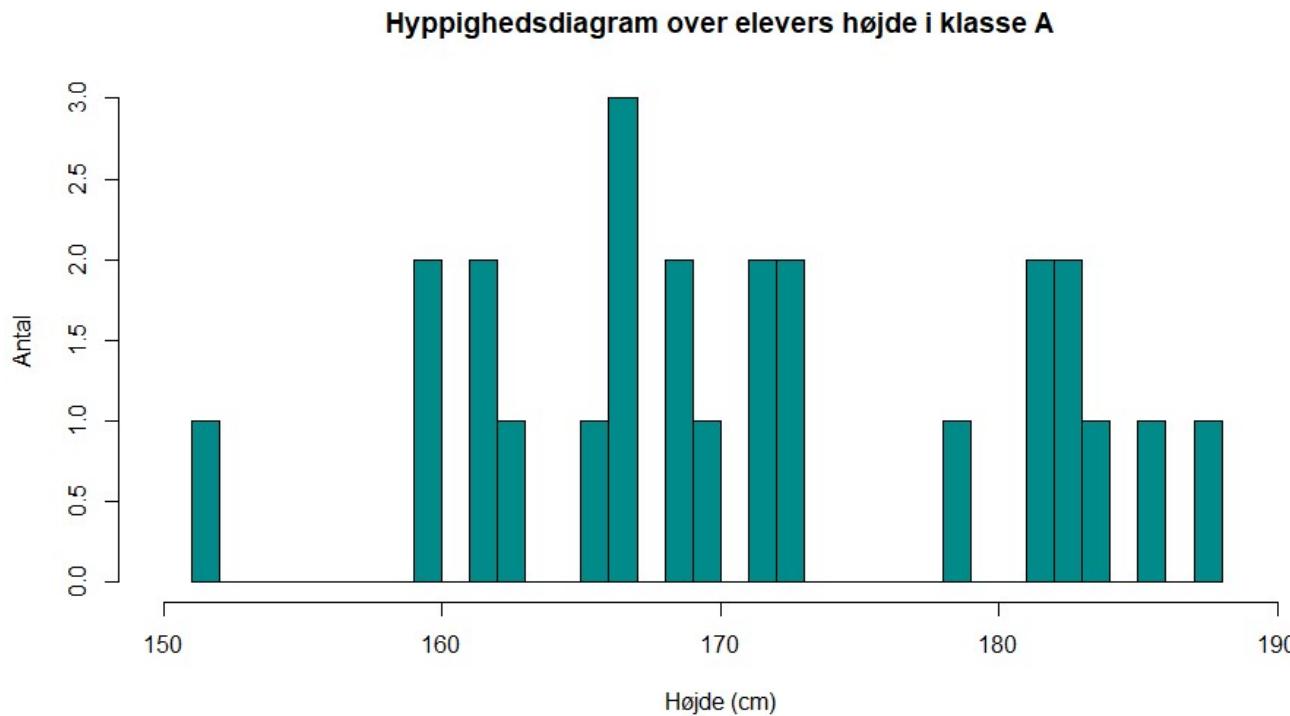
Observation	151	160	162	163	166	167	169	170	172	173	179	182	183	184	186	188
Hyppighed	1	2	2	1	1	3	2	1	2	2	1	2	2	1	1	1

- For hver observeret højde vises en prik for antal observationer af højden (fordel: data kan genfindes i diagrammet)
- Prikdiagrammer viser outliers og hyppige observationer i små datasæt
- Funktion til prikdiagram i R: `stripchart()`.



Stolpediagram over hyppigheder

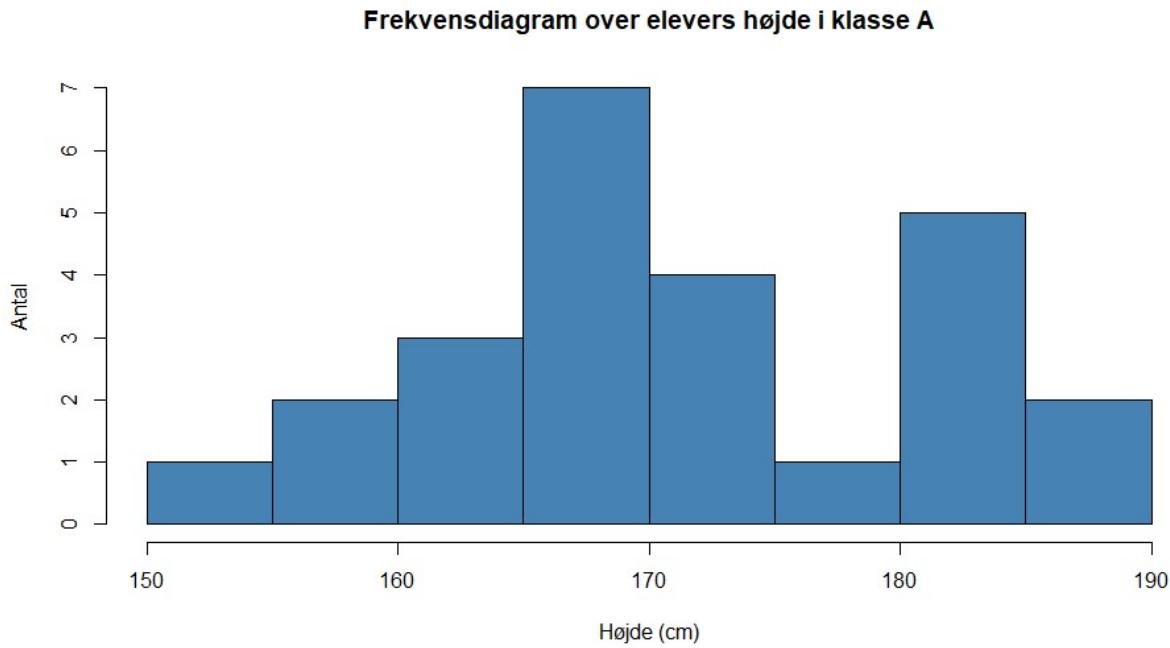
- Analogt til prikdiagrammet kan hyppighederne også vises i et stolpediagram:



- Her er lavet et histogram med R funktion: `hist()`.

Frekvensdiagram (histogram)

- Ved store datasæt kan det være en fordel at gruppere observationerne i en **frekvensfordeling**.
- Frekvensfordelingen kan præsenteres i et histogram
- Fordel: Større overblik. Ulempe: Detaljer mistet
- Histogrammet viser antal elever i hver højdekategori (lavet i R med **hist()**).

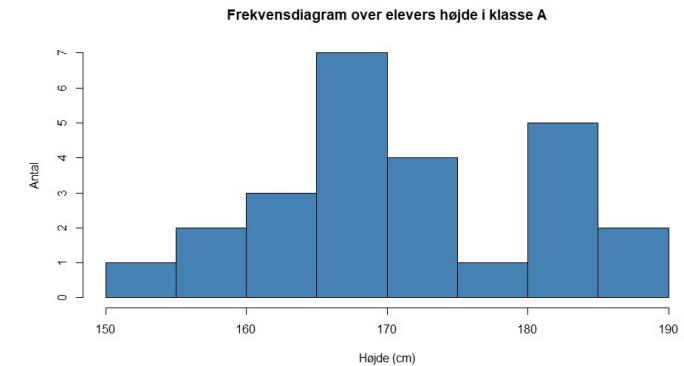
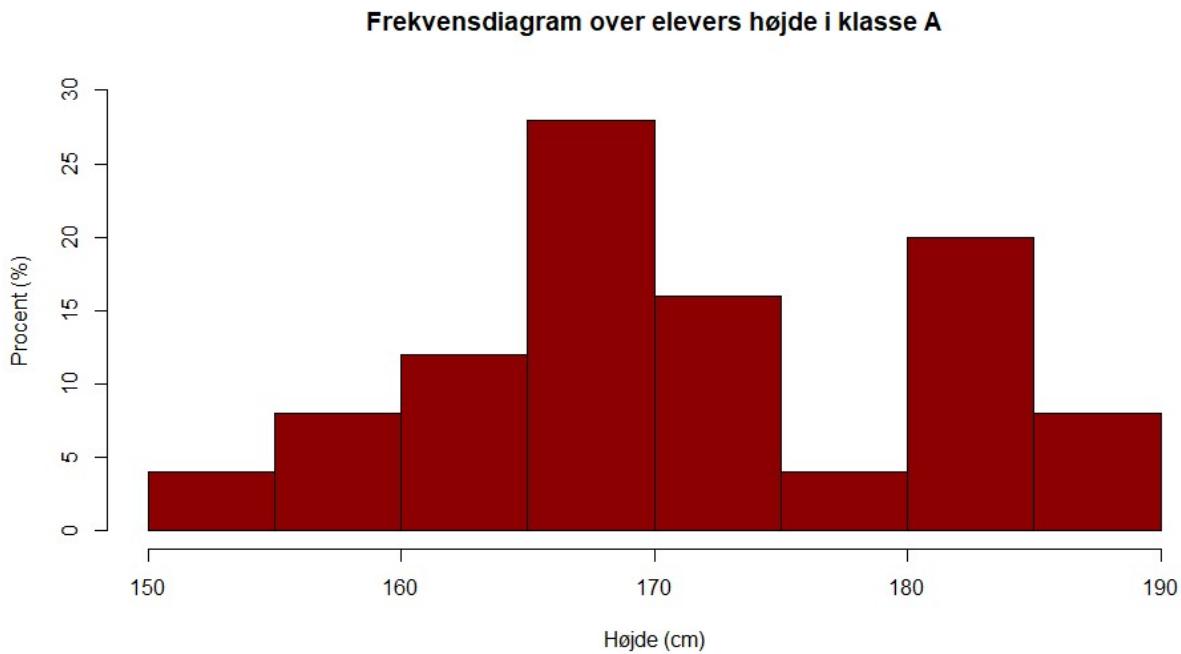


Frekvensfordeling:

Interval	Antal	Pct.
(150-155]	1	4%
(155-160]	2	8%
(160-165]	3	12%
(165-170]	7	28%
(170-175]	4	16%
(175-180]	1	4%
(180-185]	5	20%
(185-190]	2	8%
Total	25	100%

Frekvensdiagram (histogram)

- Vi kan vise det relative antal (procent) elever i hvert interval
- Fordel: Sammenligneligt med højdedata for andre grupper f.eks. med 100 personer
- Histogrammet viser pct. elever i hver højdekategori (lidt R-kodning nødvendig).

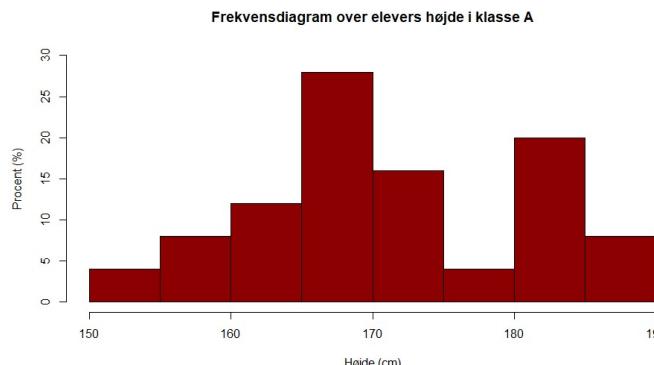
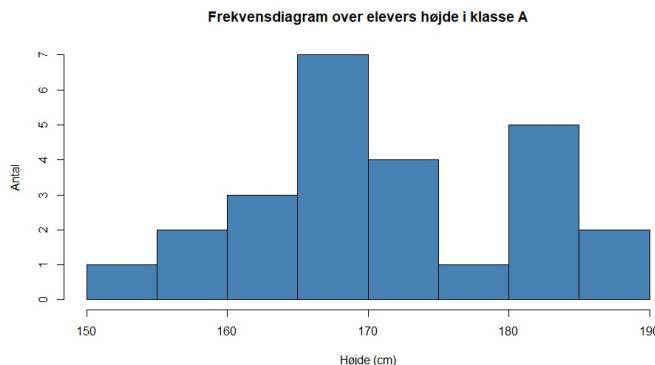
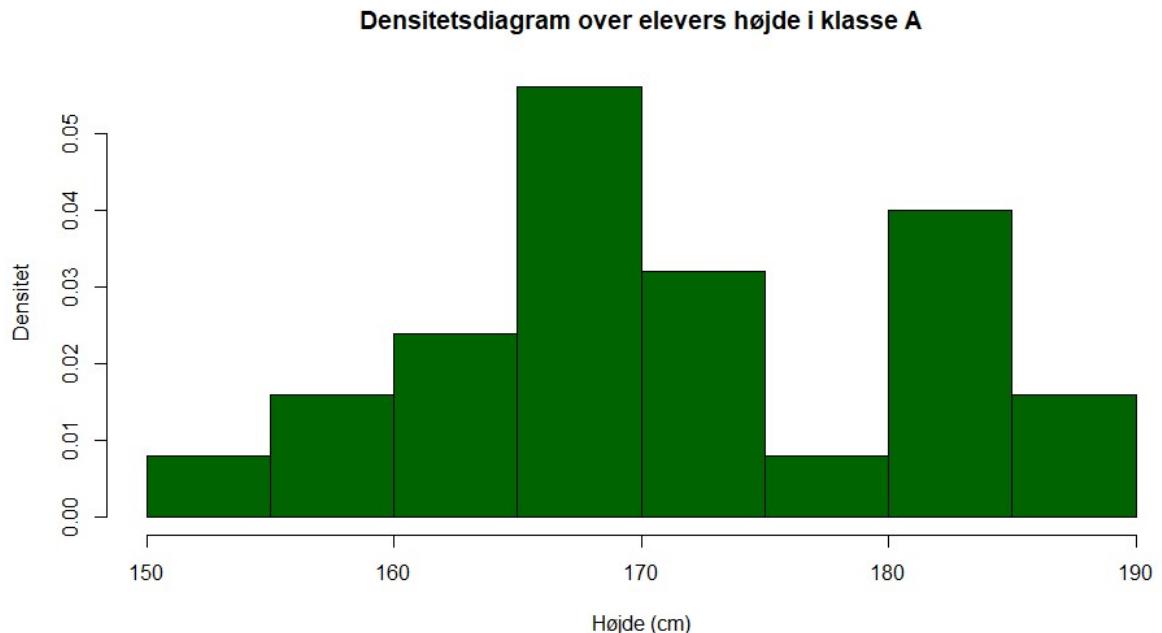


Frekvensfordeling:

Interval	Antal	Pct.
(150-155]	1	4%
(155-160]	2	8%
(160-165]	3	12%
(165-170]	7	28%
(170-175]	4	16%
(175-180]	1	4%
(180-185]	5	20%
(185-190]	2	8%
Total	25	100%

Densitetsdiagram (histogram)

- I densitetsdiagrammet er sjælernes samlede areal 1
- Laves også med `hist()`
- Bemærk at stolperne har samme form i de tre histogrammer, kun y-aksen er ændret
- Derfor bruger vi oftest kun almindeligt frekvensdiagram.



Der er forskel på () og []

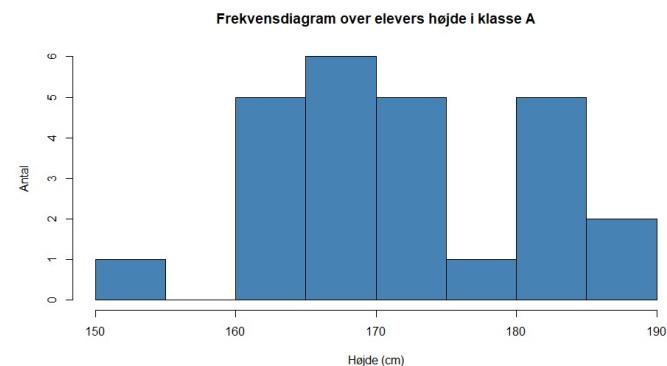
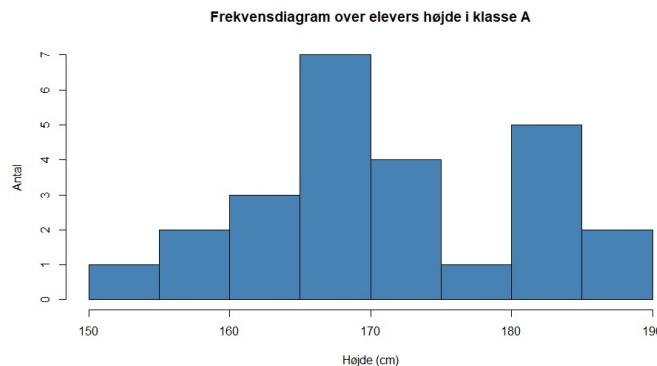
- Det kan give forskelligt udtryk, om det er nedre eller øvre intervalgrænse, der er inkluderet, især med små datasæt
- Bemærk f.eks., at de to elever på 160 cm ryger i forskellige kategorier

Øvre grænse inkluderet (standard):

Interval	Antal	Pct.
(150-155]	1	4%
(155-160]	2	8%
(160-165]	3	12%
(165-170]	7	28%
(170-175]	4	16%
(175-180]	1	4%
(180-185]	5	20%
(185-190]	2	8%
Total	25	100%

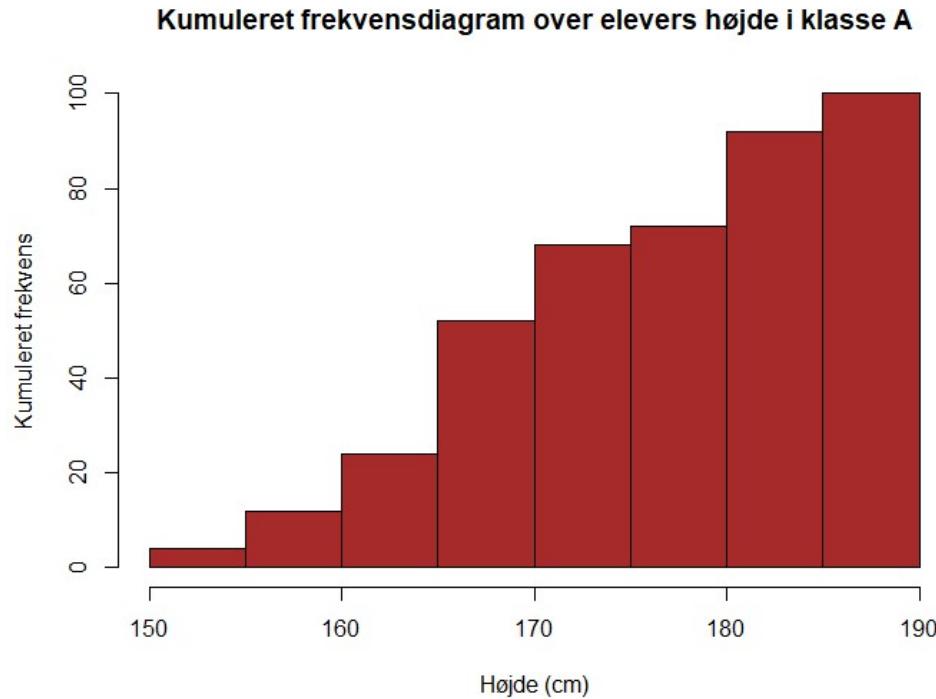
Nedre grænse inkluderet:

Interval	Antal	Pct.
[150-155)	1	4%
[155-160)	0	0%
[160-165)	5	20%
[165-170)	6	24%
[170-175)	5	20%
[175-180)	1	4%
[180-185)	5	20%
[185-190)	2	8%
Total	25	100%

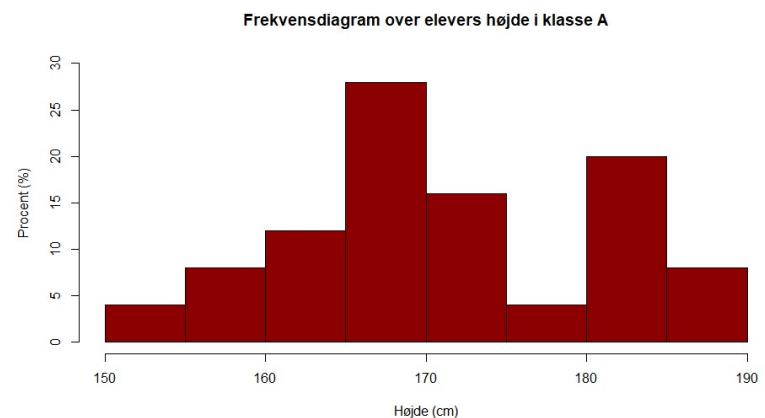


Kumuleret frekvensdiagram

- Intervallernes observationer akkumuleres, så hver søjle viser frekvensen med højde op til intervalgrænsen. F.eks. er 4 % 155 cm eller derunder og 72 % er 180 cm eller derunder
- Sidste søjle er 100 %, så alle elever er 190 cm eller derunder
- Laves i R med [hist\(\)](#), men der skal kodes lidt.



Interval	Antal	Pct.	Kumuleret	
			Antal	Pct.
(150-155]	1	4%	1	4%
(155-160]	2	8%	3	12%
(160-165]	3	12%	6	24%
(165-170]	7	28%	13	52%
(170-175]	4	16%	17	68%
(175-180]	1	4%	18	72%
(180-185]	5	20%	23	92%
(185-190]	2	8%	25	100%
Total	25	100%	25	100%



Stem-and-leaf display

- Dansk: Stængel-og-Blad plot?
- En metode til at få overblik over datasættets ”form” uden at miste detaljerne
- En slags histogram, der kan laves uden computer og med alle data bevaret
- Vi kommer til at bruge stem-and-leaf plot til at teste antagelser i vores statistiske modeller.

Stem	Leaves
0.19:	<u>3</u>
0.20:	1 <u>4</u>
0.21:	034578
0.22:	<u>3</u> 33468
0.23:	<u>1</u> 37

Stem-and-Leaf plot for højdedata

- Vi kan lave Stem-and-Leaf plot manuelt, f.eks. på vores højdedata

167	183	188	172	167
162	182	173	160	173
170	160	166	163	179
184	183	169	151	182
162	167	172	169	186

- Find min, max og bestem passende stammer
 - Skriv blade op
 - Man kan evt. sortere bladene
 - Diskutér diagrammet
- Resultat: min = 151, max = 188

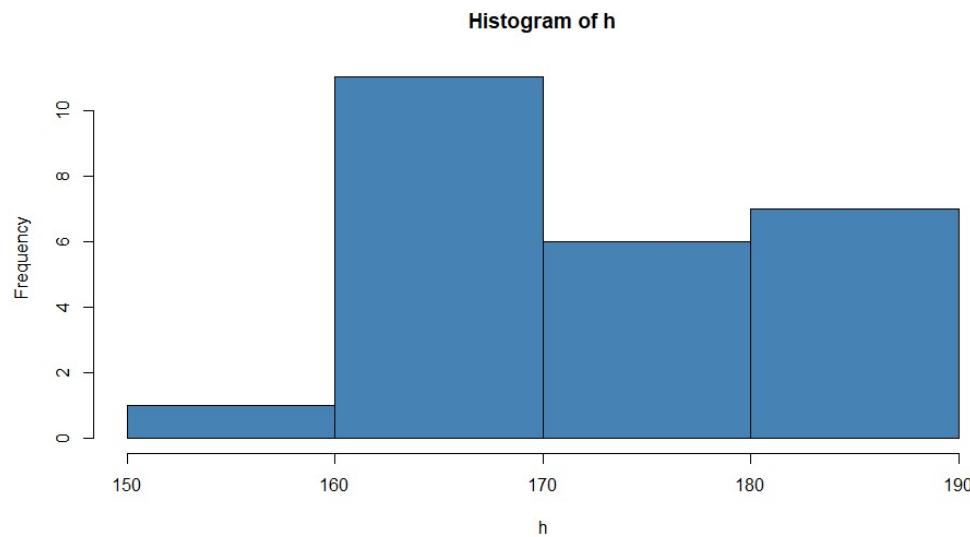
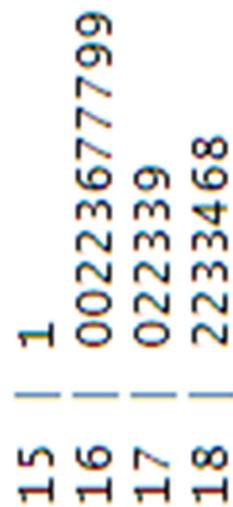
15 | 1
16 | 77200639279
17 | 233092
18 | 3824326

Output fra R funktionen `stem()`:

15 | 1
16 | 00223677799
17 | 022339
18 | 2233468

Vores anvendelse af Stem-and-Leaf

- Hvordan ser "formen" på data ud?
 - Symmetrisk?
 - Højre- eller venstrehalet?
 - Et eller flere toppunkter?
 - Outliers?
- Vi kan få nogenlunde tilsvarende information med et histogram.



Opsummering af R funktioner

Import af data

- `read.table()` til .txt og .csv, ikke Excel (.xls eller .xlsx)

Diagrammer

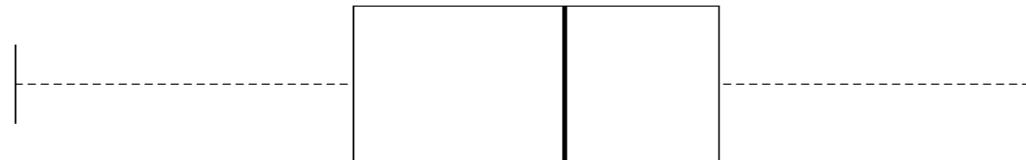
- `barplot()` Stolpediagram
- `stripchart()` Prikdiagram
- `hist()` Histogram
- `stem()` Stem-and-leaf plot

Deskriptorer

- `mean()` Middelværdi
- `var()` Varians (stikprøve)
- `sd()` Standardafvigelse (stikprøve)
- `min()` Minimum
- `max()` Maksimum
- `length()` Antal
- `median()` median

Boksplot = kassediagram

- Et kassediagram er en anden simpel måde at få overblik over data uden komplicerede beregninger. F.eks. er det enklere at bestemme median end middelværdi



- Et kassediagram deler data op i fire nogenlunde lige store grupper vha. såkaldte *kvartiler* Q_1 , Q_2 og Q_3 , hvor Q_2 er det samme som medianen.

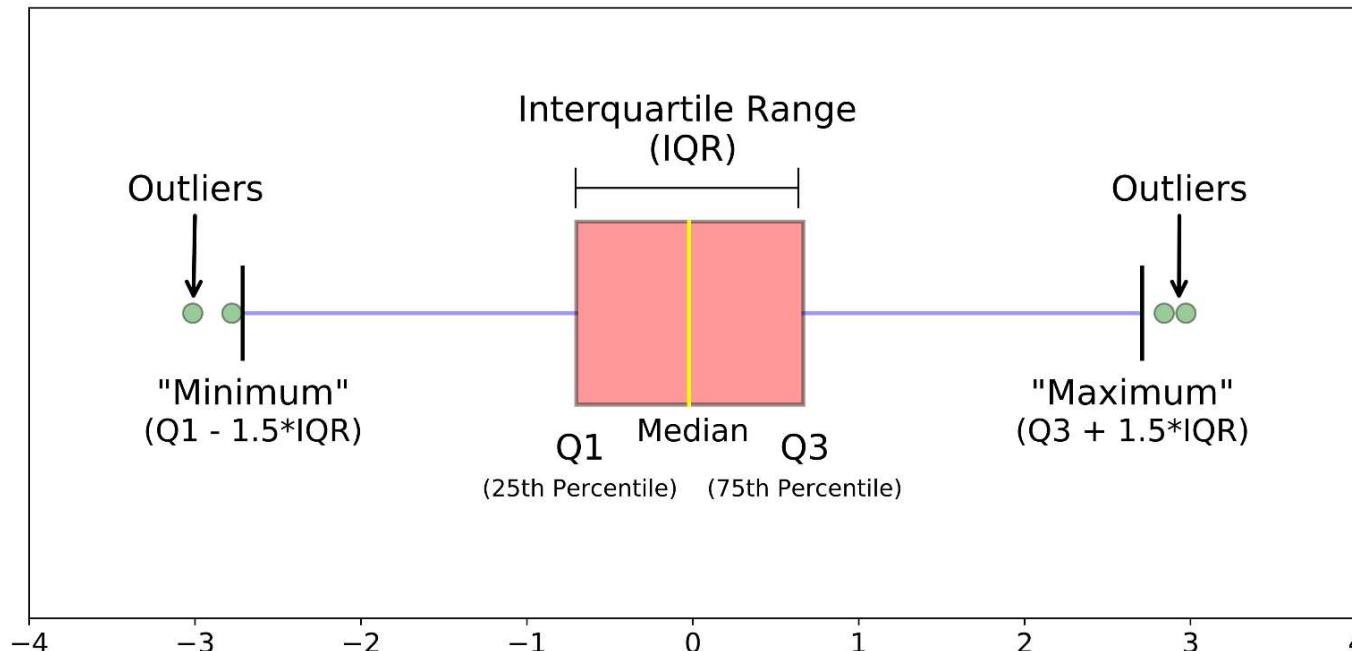
Fraktiler

Fraktiler (engelsk: **quantiles**) bruges til at dele et datasæt op i et antal lige store dele. Der er forskellige specialiserede fraktiler:

- **Median \tilde{y}** : Den midterste værdi i et sorteret datasæt y . Medianen deler datasættet op i to halvdele
- **Kvartiler**: Tre værdier, der deler det sorterede datasæt op i fire lige store dele.
 - Nedre kvartil, Q_1
 - Midterste kvartil, som er det samme som medianen, $Q_2 = \tilde{y}$
 - Øvre kvartil, Q_3
- **Percentiler**: 99 værdier, som deler datasættet op i 100 dele
 - Q_1 er den 25. percentil
 - \tilde{y} er den 50. percentil
 - Q_3 er den 75. percentil.

Boksplot = kassediagram

- Et kassediagram deler data op i fire nogenlunde lige store grupper vha. kvartilerne Q_1 , Q_2 og Q_3
- Den midterste halvdel er det interkvartile område (*interquartile range*, IQR)
- Det består af en kasse omkring medianen med koste (*whiskers*) på
- Kassens bredde er IQR, dvs. $Q_3 - Q_1$. IQR er et mål for datas spredning
- Kostenes længde kan defineres forskelligt, f.eks. så de går til min og max. Vi bruger figurens definition, så der kan identificeres outliers.

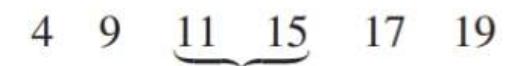


Algoritme til beregning af kvartilsæt

Vi har et datasæt med n observationer: y_1, y_2, \dots, y_n

Step 1. Beregning af medianen \tilde{y} :

1. Sortér datasættet: $y_{(1)}, y_{(2)}, \dots, y_{(n)}$
2. Bestem positionen af medianen: $\ell_m = \frac{n+1}{2}$
3. Bestem værdien af medianen:
 - a. Hvis n er **ulige**, så er ℓ_m et heltal, og så er medianen
$$\tilde{y} = y_{(\ell_m)}$$

 - b. Hvis n er **lige**, så er ℓ_m ikke et heltal, men indeholder $\frac{1}{2}$. Så beregnes medianen som gennemsnittet af de to observationer med position $\ell_m - \frac{1}{2}$ og $\ell_m + \frac{1}{2}$:
$$\tilde{y} = \frac{1}{2} (y_{(\ell_m - \frac{1}{2})} + y_{(\ell_m + \frac{1}{2})})$$


Algoritme til beregning af kvartilsæt

Step 2. Beregn kvartiler Q_1 og Q_3 :

1. Bestem positionen af kvartilerne ved at beregne ℓ_q :

a. Hvis n er ulige: $\ell_q = \frac{n+3}{4}$

b. Hvis n er lige: $\ell_q = \frac{n+2}{4}$

Nu er ℓ_q enten et heltal eller indeholder $\frac{1}{2}$

2. Bestem værdien af kvartilerne:

- a. Hvis ℓ_q er et heltal:

$$Q_1 = y_{(\ell_q)}$$

$$Q_3 = y_{(n+1-\ell_q)}$$

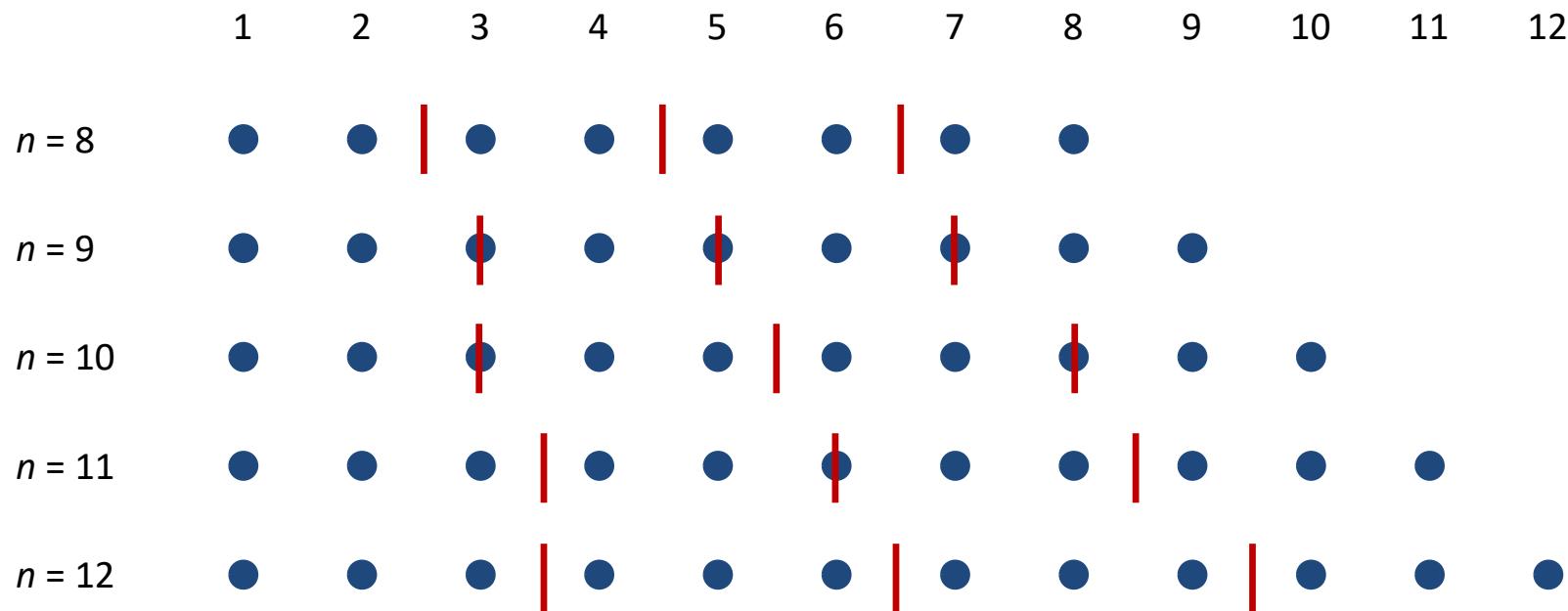
- b. Ellers gennemsnit af obs. omkring ℓ_q :

$$Q_1 = \frac{1}{2} (y_{(\ell_q - \frac{1}{2})} + y_{(\ell_q + \frac{1}{2})})$$

$$Q_3 = \frac{1}{2} (y_{(n+1-\ell_q - \frac{1}{2})} + y_{(n+1-\ell_q + \frac{1}{2})})$$

Beregning af kvartiler afhængigt af n

Antal	Position af Q_1	Position af \tilde{y}	Position af Q_3
n	$\ell_q = \frac{n+c}{4}; c = \begin{cases} 3, n \text{ ulige} \\ 2, n \text{ lige} \end{cases}$	$\ell_m = \frac{n+1}{2}$	$n+1 - \ell_q$
8	$(8+2)/4 = 2\frac{1}{2}$	$9/2 = 4\frac{1}{2}$	$8 + 1 - 2\frac{1}{2} = 6\frac{1}{2}$



Eksempel (flymotor)

Målt godstykke (inches) af aluminium køledel til flymotor

0.223	0.193	0.218	0.201	0.231	0.204
0.228	0.223	0.215	0.223	0.237	0.226
0.214	0.213	0.233	0.224	0.217	0.210

Step 1. Bestem medianen \tilde{y}

a. Sortér data y

$$y_{(1)} = 0.193, \quad y_{(2)} = 0.201, \quad \dots, \quad y_{(18)} = 0.237$$

b. Bestem position af medianen:

$$n = 18$$

$$\ell_m = \frac{n+1}{2} = 9\frac{1}{2}$$

c. Bestem værdien af medianen \tilde{y} :

$$\tilde{y} = \frac{1}{2} (y_{(\ell_m - \frac{1}{2})} + y_{(\ell_m + \frac{1}{2})})$$

$$\tilde{y} = \frac{1}{2} (y_{(9)} + y_{(10)})$$

$$\tilde{y} = \frac{1}{2} (0.2180 + 0.2230) = 0.2205.$$

0.1930
0.2010
0.2040
0.2100
0.2130
0.2140
0.2150
0.2170
0.2180
0.2230
0.2230
0.2230
0.2240
0.2260
0.2280
0.2310
0.2330
0.2370

Eksempel (flymotor)

Step 2. Beregn kvartiler Q_1 og Q_3

a. Positioner af kvartilerne. Da n er lige:

$$\ell_q = \frac{n+2}{4} = \frac{20}{4} = 5$$

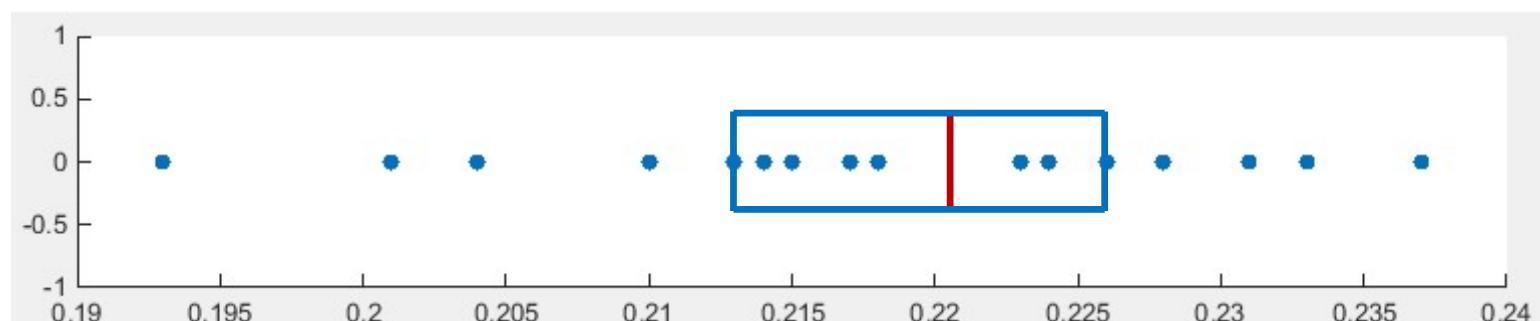
b. Værdier af kvartilerne:

Da ℓ_q er et heltal:

$$Q_1 = y_{(\ell_q)} = y_{(5)} = 0.2130$$

$$Q_3 = y_{(n+1-\ell_q)} = y_{(18+1-5)} = y_{(14)} = 0.2260$$

Step 3. Tegn kassen



Eksempel (flymotor)

Step 4. Tegn koste på:

- a. Beregn *step*:

$$step = 1.5 \cdot (Q_3 - Q_1) = 1.5 \cdot (0.226 - 0.213) = 0.0195$$

- b. Beregn mulige grænser for koste, Upper og Lower Inner Fence (*UIF* og *LIF*):

$$UIF = Q_3 + step = 0.226 + 0.0195 = 0.2455$$

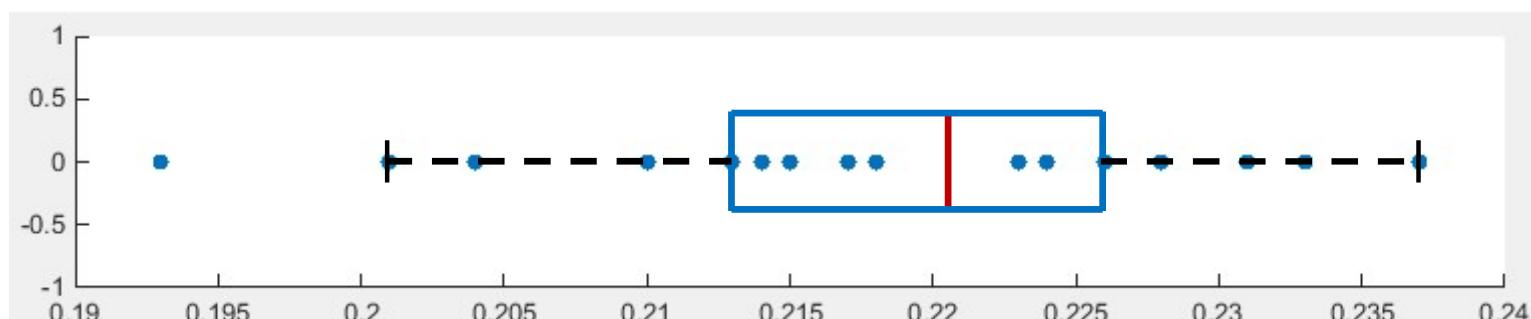
$$LIF = Q_1 - step = 0.213 - 0.0195 = 0.1935$$

- c. Find 'adjacents', nærmeste dataværdier inden for *UIF* og *LIF*:

Upper adjacent: 0.2370

Lower adjacent: 0.2010

- d. Tegn koste fra kassen til adjacents.

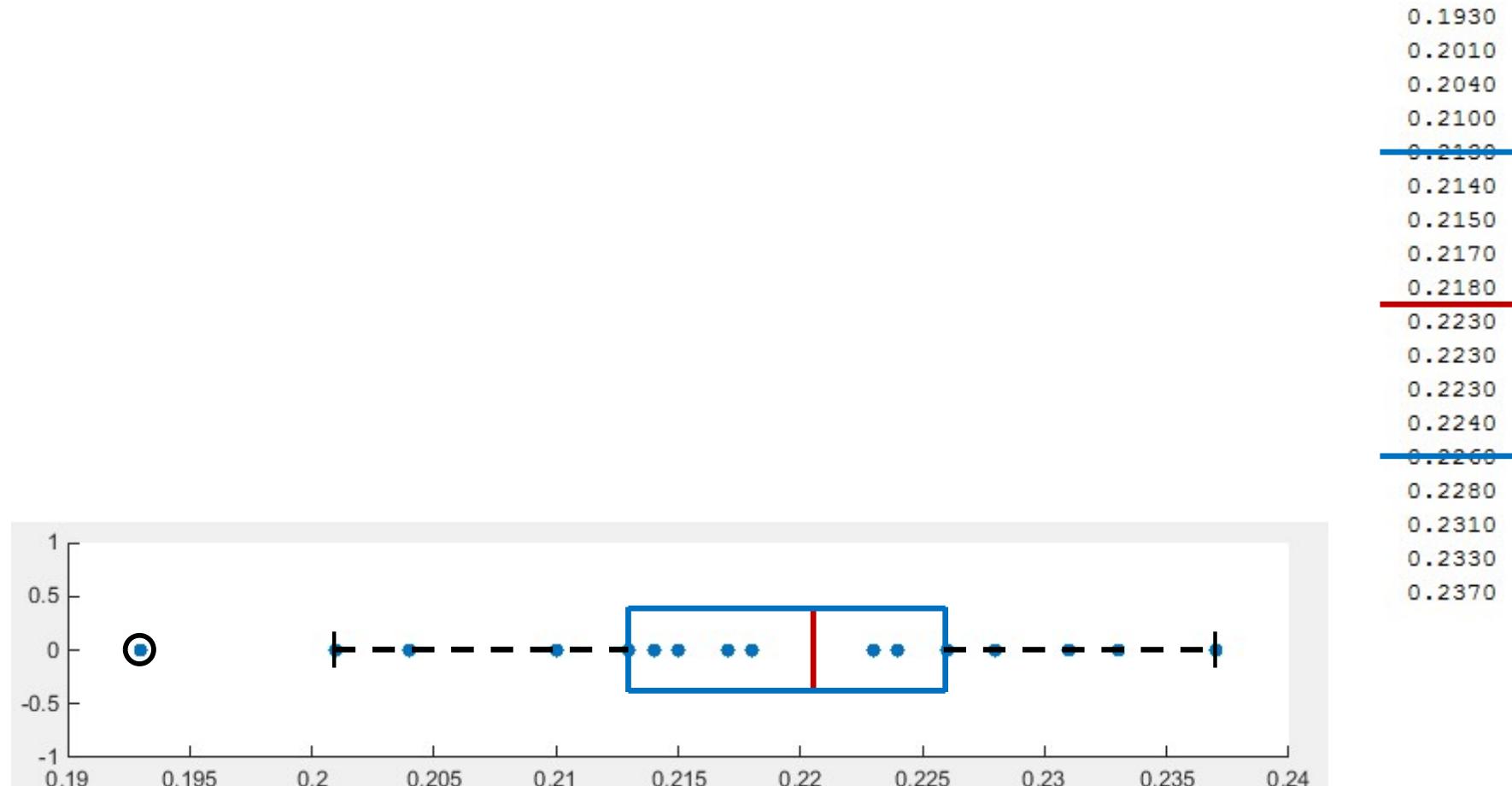


0.1930
0.2010
0.2040
0.2100
<u>0.2100</u>
0.2140
0.2150
0.2170
<u>0.2180</u>
0.2230
0.2230
0.2230
0.2240
<u>0.2260</u>
0.2280
0.2310
0.2330
0.2370

Eksempel (flymotor)

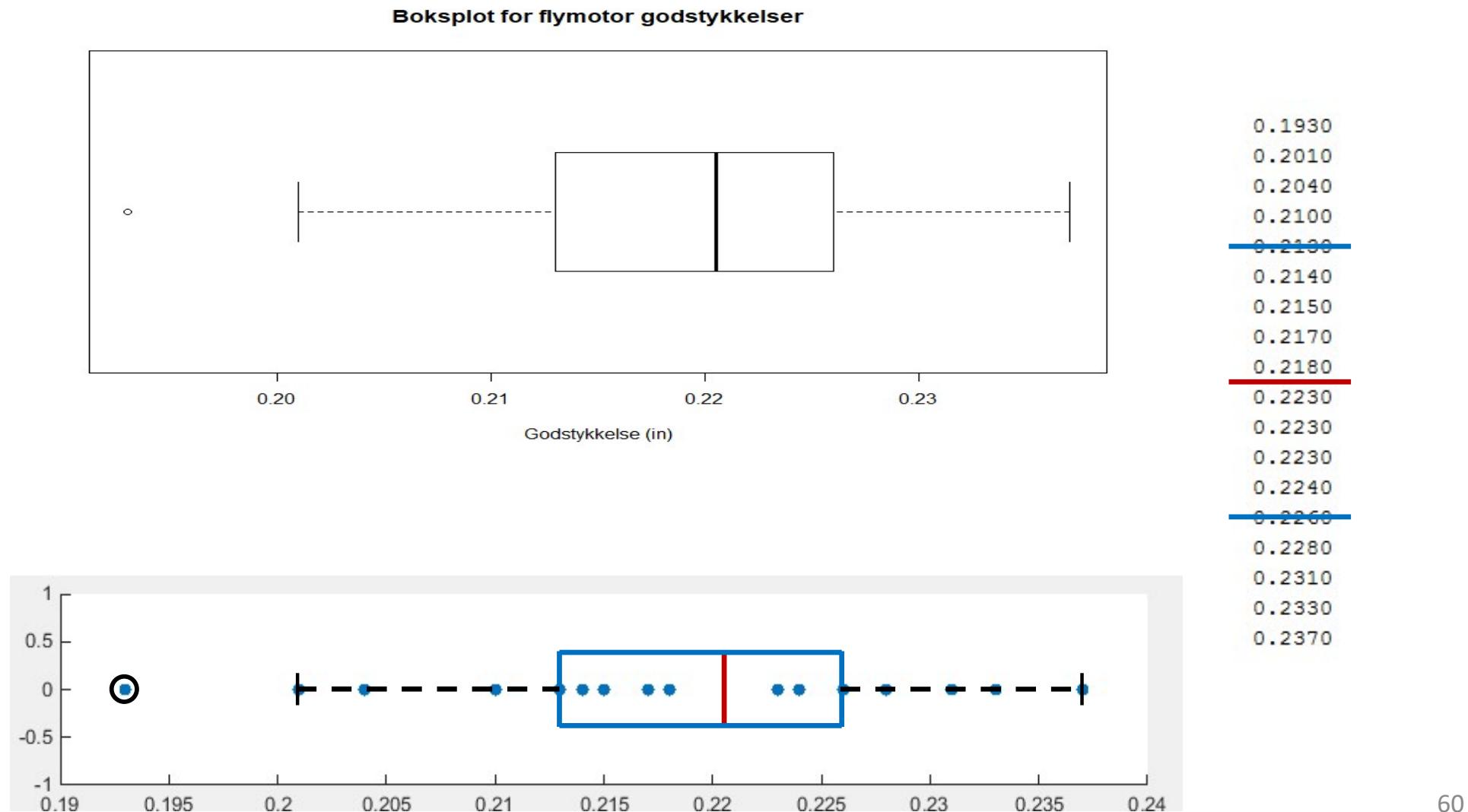
Step 5. Identificér outliers

- a. 0.193 er en outlier, for $0.193 < LIF = 0.1935$



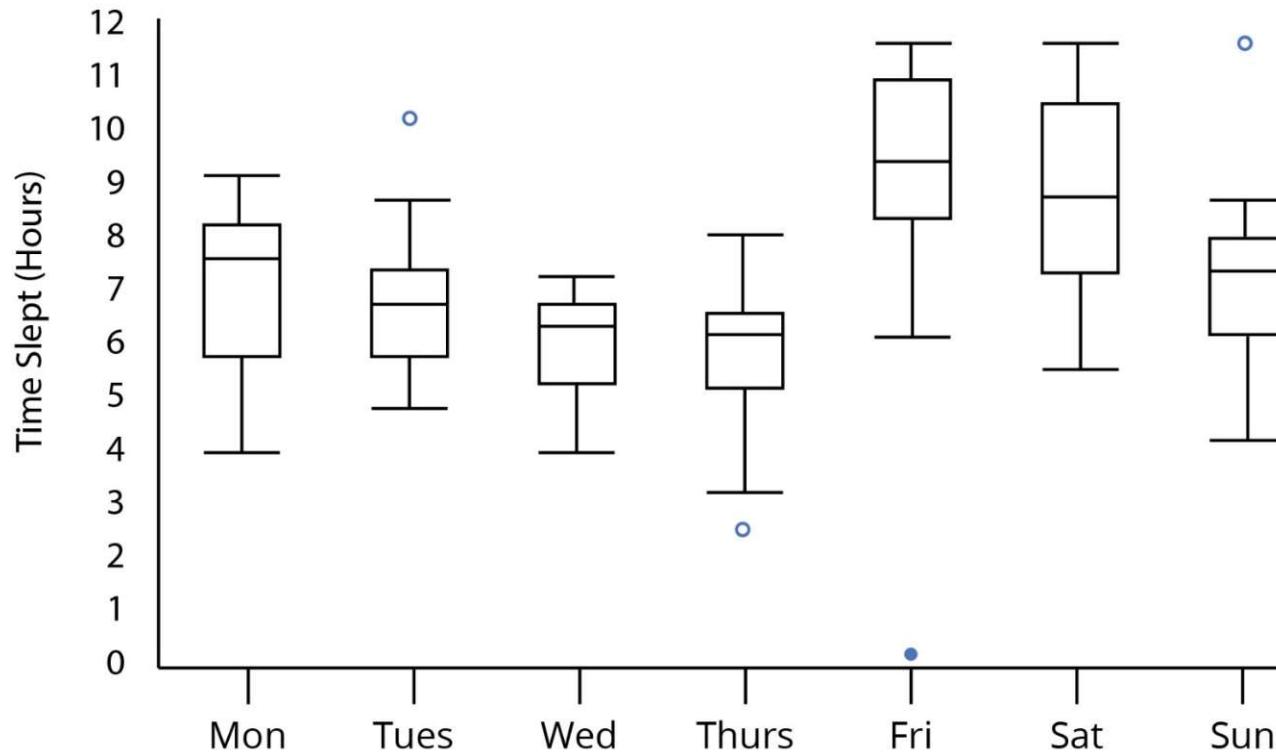
Eksempel (flymotor)

- Normalt vises data (prikkerne) ikke i et boksplot.
Med R ser boksplottet sådan ud:



Parallelle boksplots

- Vi bruger især parallelle boksplots til at sammenligne grupper af data
- Eksempel: Antal timers søvn fordelt på ugedage for universitetsstuderende



- Parallelt boksplot i R: `boxplot(Timer ~ Ugedag)`.

Opsummering af R funktioner

Import af data

- `read.table()` til .txt og .csv, ikke Excel (.xls eller .xlsx)

Deskriptorer

- `mean()` Middelværdi
- `var()` Varians (stikprøve)
- `sd()` Standardafvigelse (stikprøve)
- `min()` Minimum
- `max()` Maksimum
- `length()` Antal
- `median()` median
- `quantile()` Fraktiler (0, 25, 50, 75, 100 %)

Diagrammer

- `barplot()` Stolpediagram
- `stripchart()` Prikdiagram
- `hist()` Histogram
- `stem()` Stem-and-leaf plot
- `boxplot()` Boksplot.

Opgave om batterier til et eksoskelet

Et lille ingeniørfirma udvikler et eksoskelet, som de kalder for SwiftLift

Opgave - Batterier til eksoskelettet

En af de største udfordringer med udviklingen af SwiftLift er at gøre eksoskelettet fuldstændigt mobilt. Firmaet ønsker, at SwiftLift skal drives af genopladelige batterier, der skal bæres på ryggen af personen. Batterierne skal være enkle at udskifte i løbet af arbejdssagen.

Firmaet tester kvaliteten af tre typer 12 V batterier, der alle har den samme nominelle energikapacitet (50 Ah). For hver type batteri måles, hvor længe (antal sekunder) en motor kan drives, før batteriet løber tør for strøm. Hver batteritype måles med 5 gentagelser.

Resultatet af målingerne ses i følgende tabel:



Batteritype	Driftstid [s]				
1	11453	8850	10451	8317	10915
2	9027	9794	7493	9204	8319
3	10030	8968	10679	11446	10433

- a. Beregn middelværdi og standardafvigelse for alle målingerne af driftstid
- b. Beregn middelværdi og standardafvigelse for hver af de tre batterityper
- c. Lav et parallelt boksplot, der viser driftstiden for hver batteritype.

Sandsynlighedsteori og statistik

Kapitel 3. Sandsynlighed

(afsnit 3.1-3.7 (alt))

Allan Leck Jensen

alj@ece.au.dk

Det er vigtigt at kunne regne med sandsynligheder



Usikkerhed er en del af hverdagen

- Vi træffer beslutninger under usikkerhed hver dag:
 - Kan jeg nå over krydset inden lyset bliver rødt?
 - Skal jeg tage en paraply med?
 - Skal jeg tegne en ulykkesforsikring?
 - Behøver jeg regne opgaver til statistik-kurset?
 - Får mormor Covid-19, hvis jeg besøger hende?
 - Skal jeg sætte 100 kr på, at AGF vinder over Viborg?



- Man kvantificerer usikkerhed med sandsynligheder, så sandsynlighedsteori og statistik er det sprog, vi har brug for, for at træffe bedre beslutninger.

Tilfældige eksperimenter

At eksperimentet er **tilfældigt** betyder, at vi ikke kan forudsige udfaldet, og udfaldet af næste eksperiment er ikke afhængigt af de forrige (det ændrer ikke karakter)

Udfaldsrum (sample space): Mængden af mulige udfald af eksperimentet.

For eksempel:

- Plat og krone: $U = \{\text{Plat, Krone}\} = \{P, K\}$
- Kast en terning: $U = \{1, 2, 3, 4, 5, 6\}$
- Træk et kort: $U = \{\text{de 52 forskellige kort}\}$

Udfald (outcome): Resultatet af eksperimentet, f.eks. Plat, 3, Hjerter 6

Hændelse (event): En delmængde af udfaldsrummet, f.eks. "et lige antal øjne på terningen" eller "et es"

N.B. Alle udfald er også hændelser.

Tilfældige eksperimenter

Plat og krone

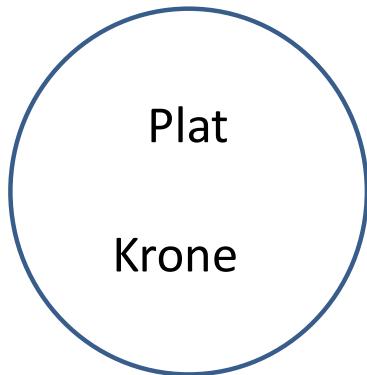
- Mulige udfald: $U = \{\text{Plat, Krone}\}$
- Udfald lige sandsynlige (tilfældigt kast med lige mønt)
- $P(\text{Plat}) = P(\text{Krone}) = \frac{1}{2}$



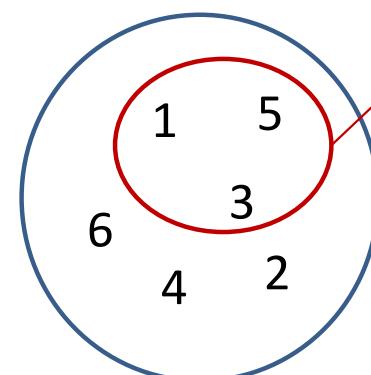
Kast med terning

- Mulige udfald: $U = \{1, 2, 3, 4, 5, 6\}$
- Udfald lige sandsynlige (tilfældigt kast)
- $P(1) = P(2) = \dots = P(6) = \frac{1}{6}$
- Hændelse: Terningen viser ulige øjne
- $P(\text{ulige}) = P(1) + P(3) + P(5) = \frac{1+1+1}{6} = \frac{3}{6} = \frac{1}{2}.$

Udfaldsrum



Udfaldsrum



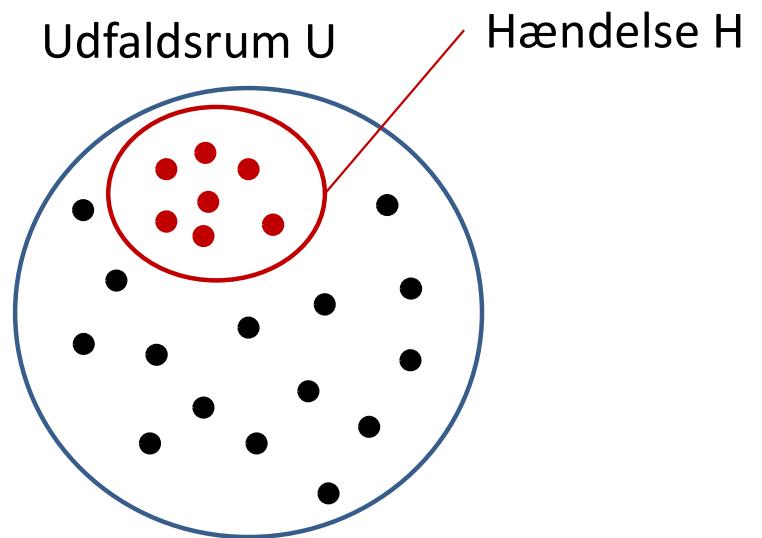
Hændelse
(ulige øjne)

Sandsynlighed

- **Definition (sandsynlighed):**

Hvis alle grundlæggende udfald i U er lige sandsynlige, er sandsynligheden for udfaldet H :

$$P(H) = \frac{\text{Antal udfald i } H}{\text{Antal udfald i } U} = \frac{N(H)}{N(U)}$$



- Hvad er sandsynligheden for at trække en rød konge i et spil kort?
 - Antal udfald i U : 52 lige sandsynlige kort
 - Antal udfald i H : 2 (hjerter og ruder konge)
 - $P(\text{rød konge}) = 2/52 = 0.038 \approx 4\%$.

Eksempel



I en polsk landsby er der kun født piger de seneste ti år

08. aug. 2019, 15:27



Ifølge New York Times er indbyggertallet i Miejsce Odrzanskie faldet fra 1200 til 272 siden Anden Verdenskrig. Foto: Screendump / Google Street View

- Hvor mange fødsler har der været på de 10 år ifølge artiklen?
- Hvad er sandsynligheden for det antal pigefødsler i træk, hvis vi antager $P(\text{pige}) = 0.5$
- Er det ekstremt?
- Artiklen oplyser antal drenge- og pigefødsler i Polen i 2007. Hvad er sandsynligheden $P(\text{polsk pige})$?
- Hvad er så sandsynligheden for det observerede antal pigefødsler i træk?

<https://nyheder.tv2.dk/udland/2019-08-08-i-en-polsk-landsby-er-der-kun-foedt-piger-de-seneste-ti-aar>

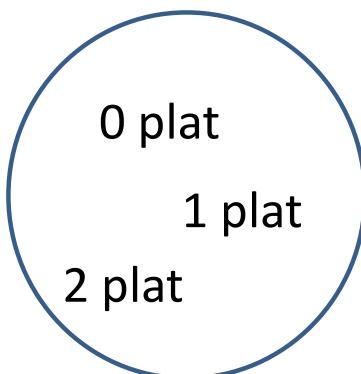
Tilfældige eksperimenter

Antal plat med to mønter

- $U = \{0, 1, 2\}$
- I 1700 tallet var der uenighed blandt matematikere, hvad sandsynligheden er for de tre udfald
- Jean le Rond d'Alembert:
 $P(0 \text{ plat}) = P(1 \text{ plat}) = P(2 \text{ plat}) = \frac{1}{3}.$

Forkert

Udfaldsrum

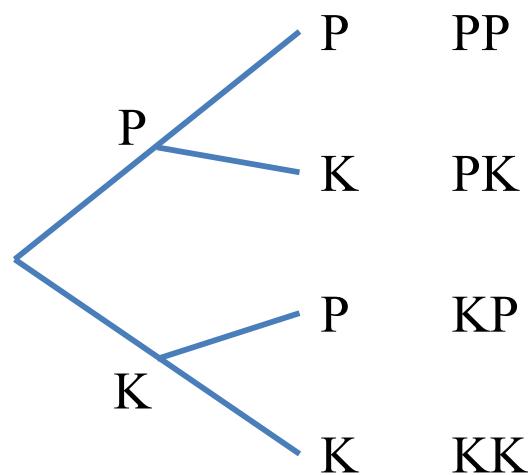
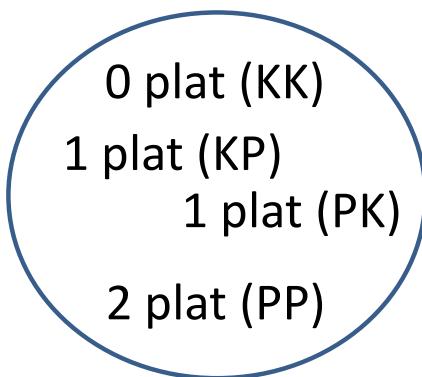


Tilfældige eksperimenter

Antal plat med to mønter

- $U = \{0, 1, 2\}$
- Pierre-Simon Laplace:
De grundlæggende udfald er PP, PK, KP, KK, og de er lige sandsynlige
- Derfor er sandsynligheden for hændelserne
 $P(0 \text{ plat}) = P(2 \text{ plat}) = \frac{1}{4}$ og
 $P(1 \text{ plat}) = \frac{2}{4} = \frac{1}{2}$.

Udfaldsrum



Tilfældige eksperimenter



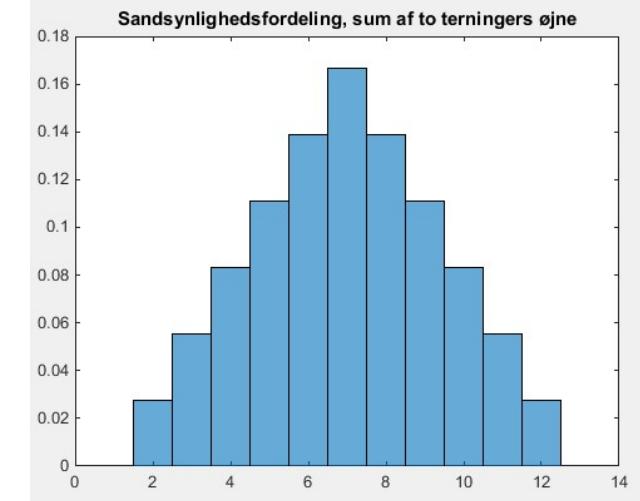
Antal øjne med to terninger

- $U = \{ 2, 3, 4, \dots, 12 \}$
- Hvad er sandsynligheden for 5?
- Vi antager at udfaldet af hver terning er tilfældigt og uafhængigt af hinanden
- Der er $6 \cdot 6 = 36$ grundlæggende udfald med lige sandsynlighed
- 4 af dem har summen 5, så
 $P(5) = \frac{4}{36} = \frac{1}{9}$.

		Terning 2					
		1	2	3	4	5	6
Terning 1	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Tilfældige eksperimenter

Sum af øjne	Kombinationer	Sandsynlighed
2		1 $1/36 = 0.027778$
3		2 $2/36 = 0.055555$
4		3 $3/36 = 0.083333$
5		4 $4/36 = 0.111111$
6		5 $5/36 = 0.138889$
7		6 $6/36 = 0.166667$
8		5 $5/36 = 0.138889$
9		4 $4/36 = 0.111111$
10		3 $3/36 = 0.083333$
11		2 $2/36 = 0.055555$
12		1 $1/36 = 0.027778$
Sum	36	$36/36 = 1$



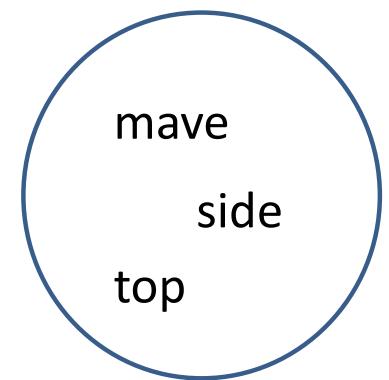
Ikke-symmetrisk sandsynlighedsmodel

Kast med tændstikæske

- Udfaldsrum:
 $U = \{\text{mave, side, top}\}$
- Her er sandsynlighederne for udfaldene ikke symmetriske (de er ikke lige sandsynlige)
- Hvordan kan vi fastslå sandsynlighederne for de mulige udfald?
- Jeg forsøgte at fastslå sandsynlighederne på tre måder:
 1. Subjektive **gæt**
 2. **Modellere** (hver sides andel af samlet overfladeareal.
Dimensioner $58 \times 35 \times 17 \text{ mm}$)
 3. Kaste tændstikæsken 100 gange og **tælle frekvens** af udfald



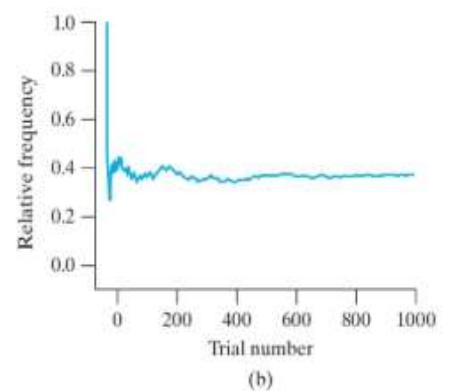
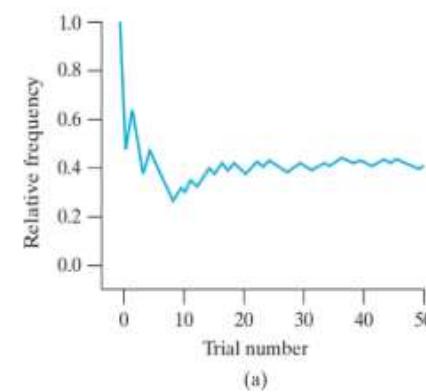
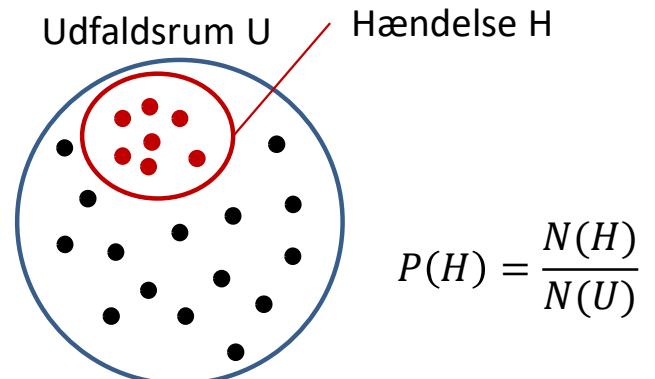
Udfaldsrum



Udfald	Gæt
Mave	70%
Side	20%
Top	10%

Fortolkninger af sandsynlighed

- Den klassiske definition af sandsynlighed kan ikke altid benyttes
- Somme tider kan man ikke tælle antal lige sandsynlige udfald:
Hvad er sandsynligheden for:
 - Det bliver regnvejr i morgen?
 - Mit fly styrter ned?
 - Jeg dumper til eksamen?
- “Der er 40 % sandsynlighed for regn i morgen.” Hvad betyder det?
Enten regner det, eller også gør det ikke
- Hvordan er sandsynligheden fastlagt? F.eks.:
 - Subjektiv vurdering
 - Modellering (40 ud af 100 scenarier)
 - Frekvensoptælling (40 % af lignende vejrførhold gav regn)
- Frekvensfortolkning: Sandsynligheden for en hændelse er andelen af gange, hændelsen sker i et stort antal forsøg.



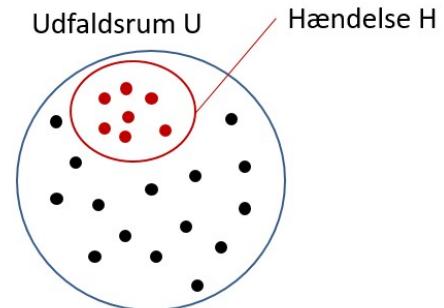
Tælleteknikker

- Som følge af definitionen af sandsynlighed skal vi ofte tælle antal udfald i H og i U
- Der findes forskellige tælle-teknikker fra kombinatorik og mængdelære:
 - Trædiagram
 - Permutationer og kombinationer
 - Venn diagrammer.

Definition (sandsynlighed):

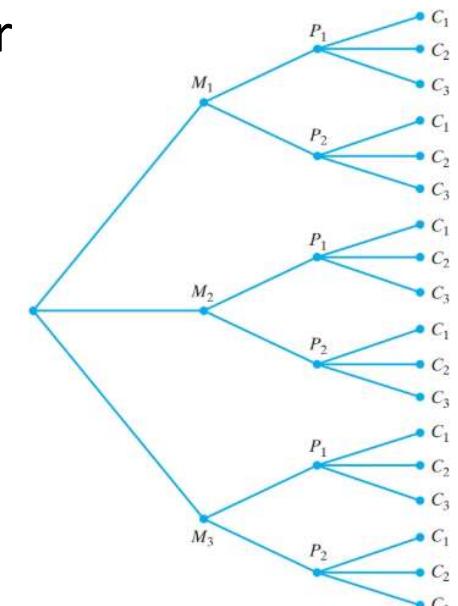
Hvis alle grundlæggende udfald i U er lige sandsynlige, er sandsynligheden for udfaldet H :

$$P(H) = \frac{\text{Antal udfald i } H}{\text{Antal udfald i } U}$$



Kombinatorik

- **Sætning 3.1 (Fundamental theorem of counting):**
Hvis mængderne A_1, A_2, \dots, A_k indeholder hhv. n_1, n_2, \dots, n_k elementer, så er der $n_1 n_2 \cdots n_k$ måder at vælge først et element fra A_1 , så et element fra A_2, \dots , og sidst et element fra A_k
- **Eksempel:** Klassificering af brugte biler efter 3 kriterier
 - Kilometertal (Mileage): M_1 : Lav, M_2 : Medium, M_3 : Høj
 - Pris (Price): P_1 : Moderat, P_2 : Høj
 - Driftsomkostn. (Cost): C_1 : Billig, C_2 : Medium, C_3 : Dyr
- Her er $n_1 = 3, n_2 = 2$ og $n_3 = 3$, så antal klasser af brugte biler er $n_1 \cdot n_2 \cdot n_3 = 3 \cdot 2 \cdot 3 = 18$
- **Eksempel:** På hvor mange måder kan man blande et spil kort med 52 kort?
- Svar: Vi bruger den fundamentale tællesætning med $k = 52$.
 - A_1 er hele kortspillet med $n_1 = 52$
 - A_2 er kortspillet, når første kort er trukket, så $n_2 = 51$, o.s.v.
 - A_{52} er kortspillet, når der kun er et kort tilbage, så $n_{52} = 1$
 - Dermed er antallet $n_1 n_2 \cdots n_k = 52 \cdot 51 \cdots 1 = 52! = 8.1 \cdot 10^{67}$
 - Det er mange – som antal atomer i solen, gange 80 milliarder!! .



Permutationer og kombinationer

- På hvor mange måder kan man trække 7 kort fra et normalt spil kort med 52 kort (rækkefølgen af de 7 kort tæller)?
- Svar: Første kort vælges tilfældigt blandt 52 kort, næste blandt 51 kort, osv.

$${}_{52}P_7 = 52 \cdot 51 \cdot 50 \cdot 49 \cdot 48 \cdot 47 \cdot 46 = 6.74 \cdot 10^{11}$$

- Generelt er antal **permutationer** af r elementer ud af n :

$${}_nPr = n(n - 1)(n - 2) \cdots (n - r + 1) = \frac{n!}{(n-r)!}$$

- I de fleste kortspil er rækkefølgen af en hånd ligegyldig. F.eks. poker:

$\text{K}\spadesuit, 10\spadesuit, 7\clubsuit, 5\heartsuit, A\diamond$ er samme hånd som $A\diamond, 5\heartsuit, 7\clubsuit, 10\spadesuit, K\spadesuit$

På hvor mange forskellige måder kan man blande denne hånd på 5 kort?

Svar: ${}_5P_5 = \frac{5!}{(5-5)!} = 5! = 120$

- Tilsvarende kan hver hånd på 7 kort blandes på $7! = 5040$ måder. Derfor er antal *forskellige* hænder på 7 kort: $\frac{6.74 \cdot 10^{11}}{5040} = 133,784,560$
- Generelt er antal **kombinationer** af r elementer ud af n :

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

Eksempel

- Dit lokale pizzeria har et tilbud på en vælg-selv-pizza for 50 kr
- Der er 30 forskellige slags fyld at vælge mellem, og man må vælge tre
- Du beslutter dig for systematisk at spise dig igennem alle kombinationer. Hver dag vil du vælge en pizza med en kombination af tre forskellige slags fyld, som du ikke har prøvet før
- Hvornår har du spist dig igennem alle pizza-kombinationerne?
- Vi kan lave et trædiagram, men løg-pepperoni-skinke er samme pizza som skinke-pepperoni-løg. Vi skal bruge antal kombinationer, ${}_{30}C_3$:
- $${}_{30}C_3 = \binom{30}{3} = \frac{30!}{3! \cdot 27!} = \frac{30 \cdot 29 \cdot 28}{3 \cdot 2 \cdot 1} = 5 \cdot 29 \cdot 28 = 4060$$

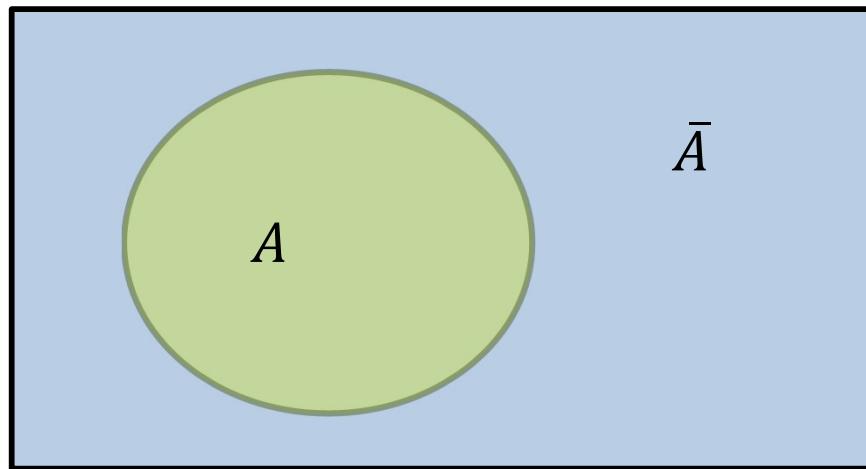
... altså efter 4060 dage, eller godt 11 år! Bon appétit!

- I R: `choose(30, 3) = 4060`
`factorial(5) = 5! = 120`
- **Hjemmeopgave:** Skinke er en af de 30 slags fyld. Hvad er sandsynligheden for, at du en given dag får pizza med skinke?

Venn diagram

- Lad hændelsen A betegne, at vi trækker en konge fra et spil kort. Vi kan illustrere det med et **Venn diagram**:

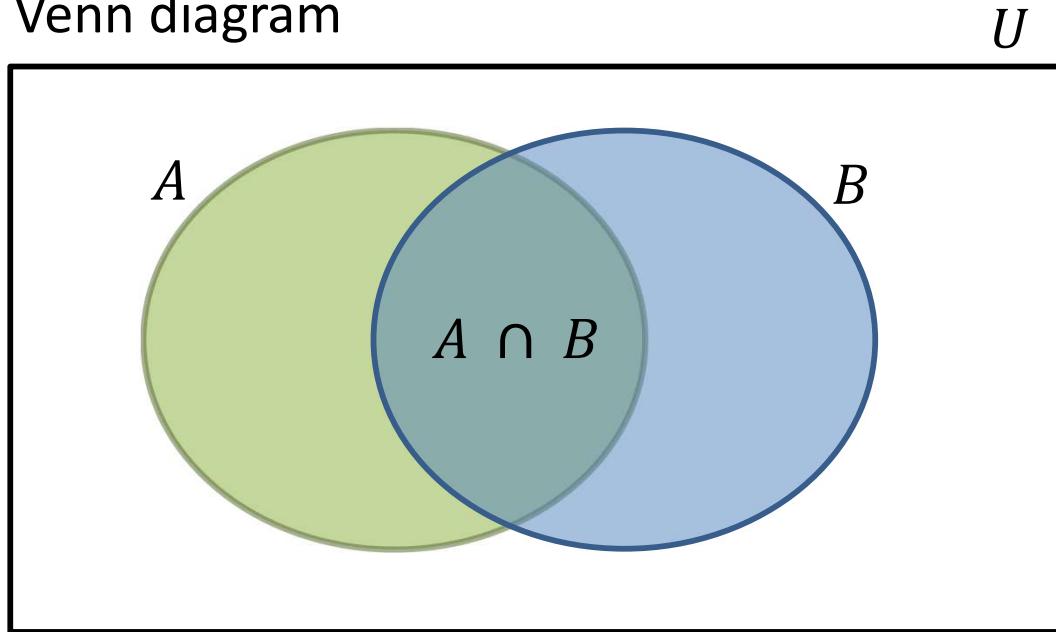
U



- Udfaldsrummet U er de 52 kort i spillet, vist som firkanten. A er de 4 konger, vist som den grønne oval
- Derfor er $P(A) = \frac{4}{52} = 0.077$
- Komplementærhændelsen til A er, at A ikke sker. I eksemplet er det, at det kort, vi trækker, ikke er en konge. Komplementærhændelsen til A skrives \bar{A} eller A^c
- Der er 48 kort, der ikke er konger, så $P(\bar{A}) = \frac{48}{52} = 0.923 = 1 - P(A)$.

Fælleshændelse og foreningshændelse

Venn diagram

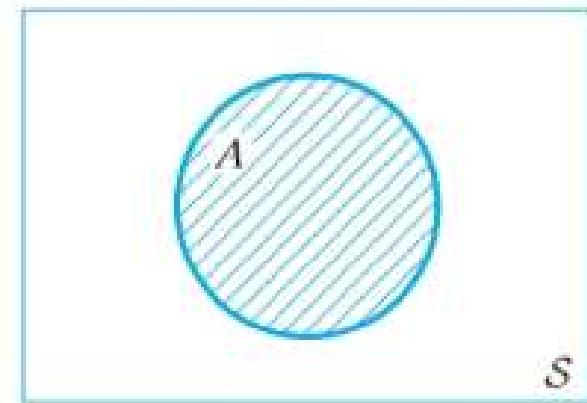


- Hændelsen A er stadig, at et kort i U er en konge. Hændelsen B er, at et kort er en hjerter
- Fælleshændelsen $A \cap B$ er, at kortet er både en konge og en hjerter.
Kun 1 kort (hjerter konge) opfylder $A \cap B$
- Foreningshændelsen $A \cup B$ er, at kortet er enten en konge eller en hjerter.
16 kort opfylder $A \cup B$: De 13 hjerter og de 4 konger, hvorved hjerter konge er talt med to gange.

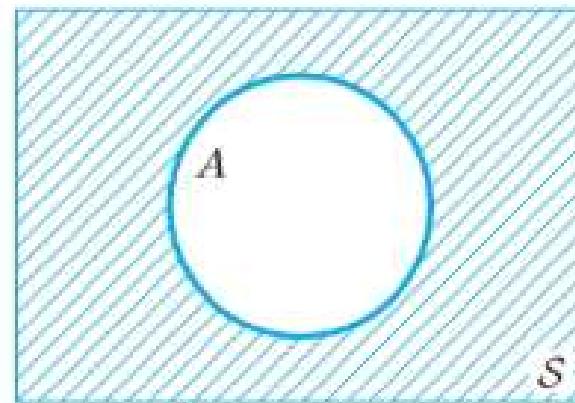
Brug af Venn diagrammer

- Venn diagrammer, hvor det skraverede areal svarer til:

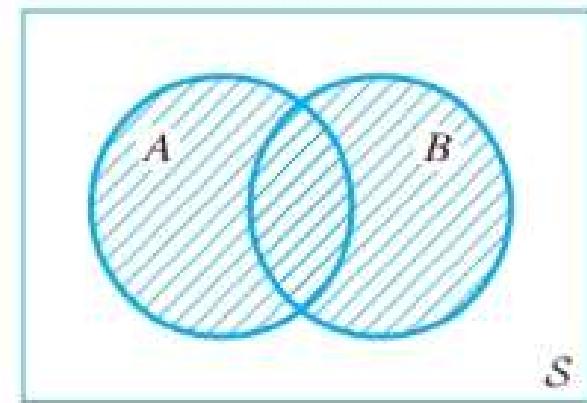
Hændelsen A



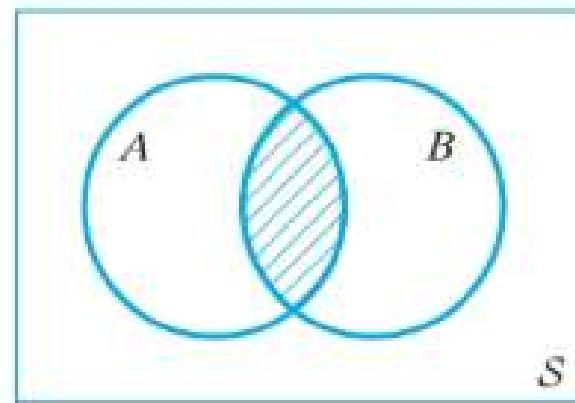
Komplementærhændelsen \bar{A}



Foreningshændelsen $A \cup B$



Fælleshændelsen $A \cap B$



Beregning af sandsynligheder

Fuldførte elever, lange videregående uddannelser efter tid, køn og uddannelse

	655920 Maskin- og skibsbygningsteknik	655940 Kemiteknik
2013		
Mænd	222	112
Kvinder	41	95

Kilde:
Danmarks Statistik,
Statistikbanken.dk

2013	Maskin	Kemi	Total
Mænd	222	112	334
Kvinder	41	95	136
Total	263	207	470

- Vi forestiller os, at vi har alle 470 ingeniører i et rum og trækker en tilfældigt.

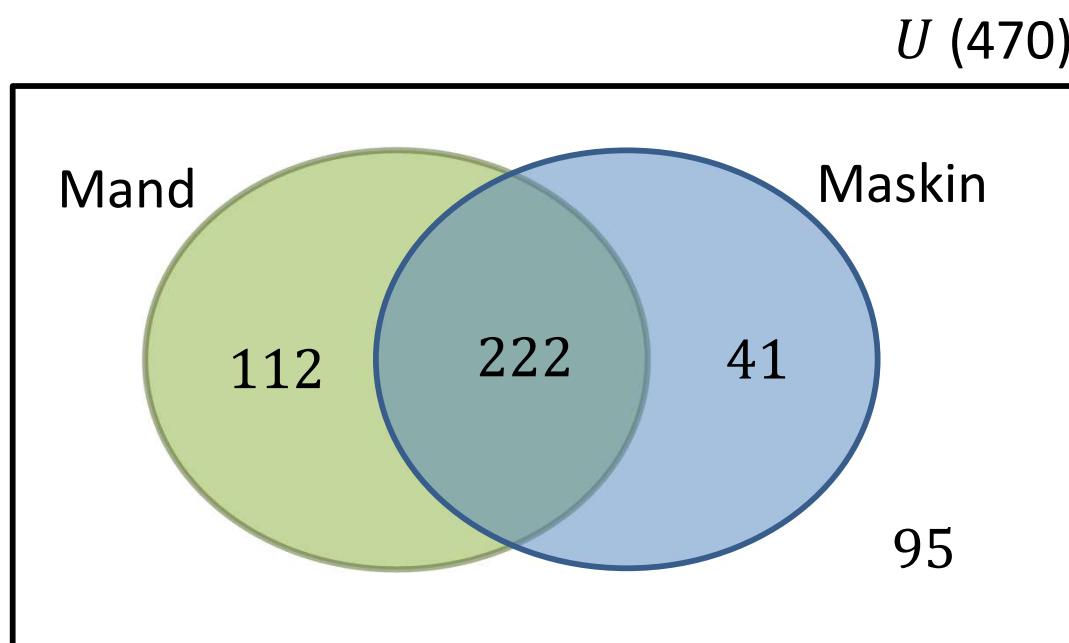
Beregning af sandsynligheder

2013	Maskin	Kemi	Total
Mænd	222	112	334
Kvinder	41	95	136
Total	263	207	470

Eksempler på hændelser:

Mand: Det er en mand

Maskin: Det er en maskiningeniør



Antal:

$$N(\text{Mand}) = 122 + 222 = 334$$

$$N(\text{Maskin}) = 222 + 41 = 263$$

$$N(\text{Mand} \cap \text{Maskin}) = 222$$

$$N(\text{Mand} \cap \overline{\text{Maskin}}) = 112$$

$$N(\overline{\text{Mand}} \cap \text{Maskin}) = 41$$

$$N(\overline{\text{Mand}} \cap \overline{\text{Maskin}}) = 95$$

Sandsynligheder:

$$P(\text{Mandlig kemiingeniør}) =$$

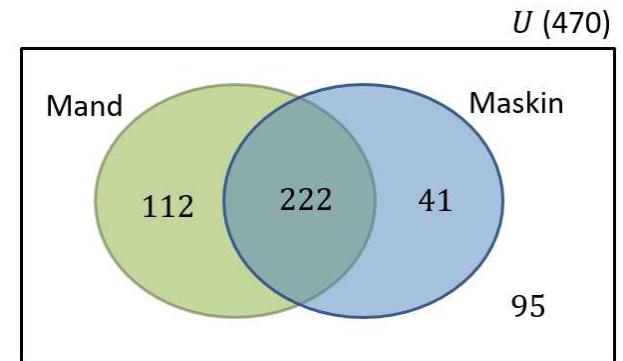
$$P(\text{Mand} \cap \overline{\text{Maskin}}) =$$

$$\frac{N(\text{Mand} \cap \overline{\text{Maskin}})}{N(U)} = \frac{112}{470} = 0.24$$

o.s.v.

Beregning af sandsynligheder

2013	Maskin	Kemi	Total
Mænd	222	112	334
Kvinder	41	95	136
Total	263	207	470



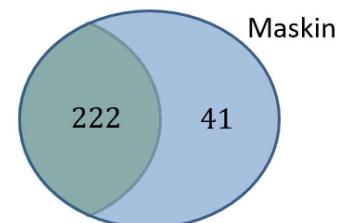
- Eksempler på beregning af sandsynligheder:
334 ud af 470 er mænd, så $P(\text{Mand}) = 334/470 = 0.71$
- $P(\overline{\text{Mand}}) = 1 - 0.71 = 0.29$.
Det samme som $P(\text{Kvinde}) = 136/470 = 0.29$
- 263 ud af 470 er maskiningeniører, så $P(\text{Maskin}) = 263/470 = 0.56$
- 222 ud af 470 er mandlige maskiningeniører, så
 $P(\text{Mand} \cap \text{Maskin}) = 222/470 = 0.47$.

Betingede sandsynligheder

2013	Maskin	Kemi	Total
Mænd	222	112	334
Kvinder	41	95	136
Total	263	207	470

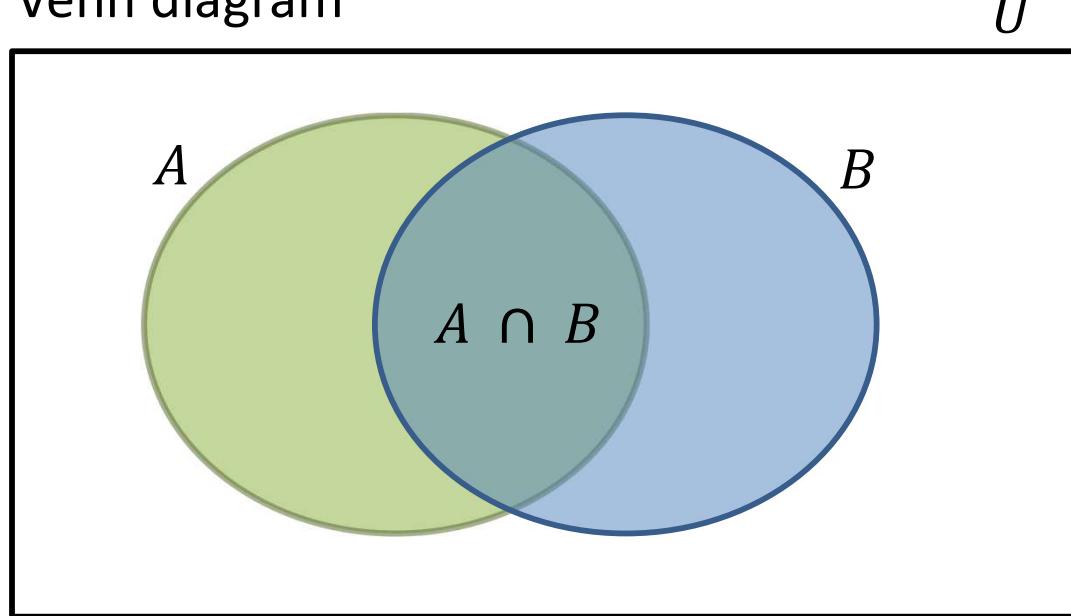
$$P(\text{Mand}) = 334/470 = 0.71$$
$$P(\text{Maskin}) = 263/470 = 0.56$$
$$P(\text{Mand} \cap \text{Maskin}) = 222/470 = 0.47$$

- Nu fokuserer vi på de 263 maskiningeniører. Hvad er sandsynligheden for at en tilfældigt udtrukket maskiningeniør er en mand?
- Vi siger: "Sandsynligheden for at det er en mand, givet (eller betinget af) at det er en maskiningeniør" og skriver det $P(\text{Mand} | \text{Maskin})$
- Ud af de 263 maskiningeniører er de 222 mænd, så
 $P(\text{Mand} | \text{Maskin}) = 222/263 = 0.84$
- $P(\text{Mand} | \text{Maskin}) = 222/263 = (222/470) / (263/470) = P(\text{Mand} \cap \text{Maskin}) / P(\text{Maskin})$
- Generelt:
$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$
- Bemærk forskellen mellem $P(\text{Mand} \cap \text{Maskin})$ og $P(\text{Mand} | \text{Maskin})$.



Betinget sandsynlighed

Venn diagram



$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

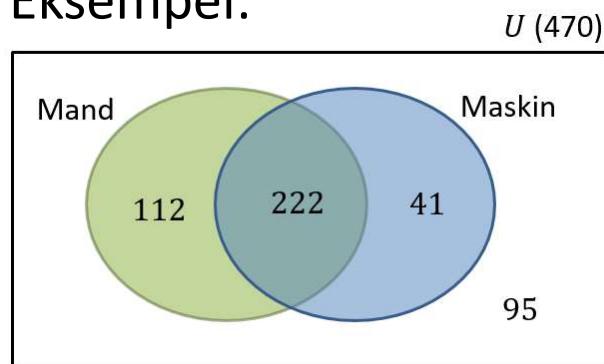
Uafhængighed

To hændelser A og B er uafhængige, hvis

$$P(A | B) = P(A)$$

Med andre ord: Information om B ændrer ikke vores opfattelse af A

Eksempel:



$$P(\text{Mand}) = \frac{(112+222)}{470} = 0.71$$

$$P(\text{Mand} | \text{Maskin}) = \frac{222}{(222+ \quad)} = 0.84$$

Da $P(\text{Mand} | \text{Maskin}) \neq P(\text{Mand})$ er hændelserne Mand og Maskin ikke uafhængige (der er ikke samme kønsfordeling for de to ingenørretninger).

Uafhængighed

Antag at de to hændelser A og B er uafhængige. Så ved vi, at

$$P(A | B) = P(A)$$

Det følger af ligningen for betinget sandsynlighed, at

$$\begin{aligned} P(A | B) &= \frac{P(A \cap B)}{P(B)} \\ \Rightarrow P(A \cap B) &= P(A | B) \cdot P(B) \\ \Rightarrow P(A \cap B) &= P(A) \cdot P(B) \end{aligned}$$

Med andre ord kan vi beregne sandsynligheden for fælleshændelsen af A og B som produktet af sandsynligheden for enkelthændelserne. For eksempel: A er kast med mønt, B er kast med terning.

$$P(\text{Plat} \cap 6) = P(\text{Plat}) \cdot P(6) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$$

I eksemplet med ingeniørerne var hændelserne ikke uafhængige, så

$$P(\text{Mand}) \cdot P(\text{Maskin}) = 0.71 \cdot 0.56 = 0.40 \neq 0.47 = P(\text{Mand} \cap \text{Maskin}).$$

Opgave om design af et eksoskelet

Opgaven handler om udvikling af ekso-skelettet SwiftLift. Der skal vælges design



Firmaet har planer om at kunne sælge SwiftLift både i byggesektoren og i sundhedssektoren. I byggesektoren er de fleste ansatte mænd, mens der er flest kvinder i sundhedssektoren. Firmaet forestiller sig, at der kan være kønsforskelle på præferencerne. Derfor udvikler de to prototyper af SwiftLift med forskellige designs, som de kalder for SL1 og SL2. De samler et testhold bestående af 19 kvinder og 18 mænd. Hver testperson afprøver begge prototyper og bliver bedt om at vurdere, om de foretrækker designet af SL1 eller SL2. Det viser sig, at 6 kvinder og 11 mænd foretrækker SL1, mens 13 kvinder og 7 mænd foretrækker SL2. Resultatet af brugerundersøgelsen vises i følgende tabel:

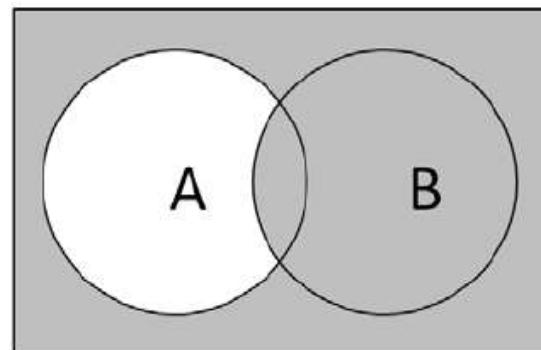
Foretrukket design	Køn		
	Kvinder	Mænd	I alt
SL1	6	11	17
SL2	13	7	20
I alt	19	18	37

Opgave om design af et eksoskelet

Foretrukket design	Køn		
	Kvinder	Mænd	I alt
SL1	6	11	17
SL2	13	7	20
I alt	19	18	37

Lad hændelsen A angive at en tilfældigt udvalgt testperson foretrækker prototypedesignet SL1, og lad hændelsen B angive at en tilfældigt udvalgt testperson er kvinde.

- Beregn sandsynlighederne $P(A)$, $P(A^c)$, $P(B)$ og $P(B^c)$ på baggrund af data fra brugerundersøgelsen.
- Beregn sandsynlighederne $P(A \cap B)$ og $P(A | B)$.
- Lader A og B til at være uafhængige hændelser?
- Nedenfor vises et Venn diagram med hvide og grå områder:
 - Beskriv med ord de testpersoner, der svarer til det hvide område i Venn diagrammet.
 - Hvilken eller hvilke af følgende hændelser svarer til det grå område?:
 - $(A \cap B) \cup B$
 - $(A \cap B) \cup A^c$
 - $B \cup A^c$
 - $A^c \cap B$



De 7 basale regler for sandsynlighed

1. $0 \leq P(A) \leq 1$

2. $P(\emptyset) = 0; P(U) = 1$

\emptyset er den tomme eller umulige hændelse.

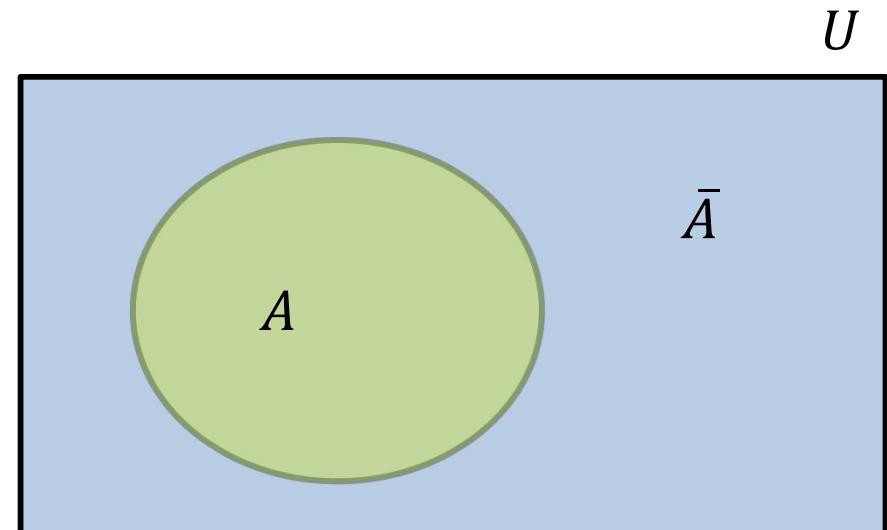
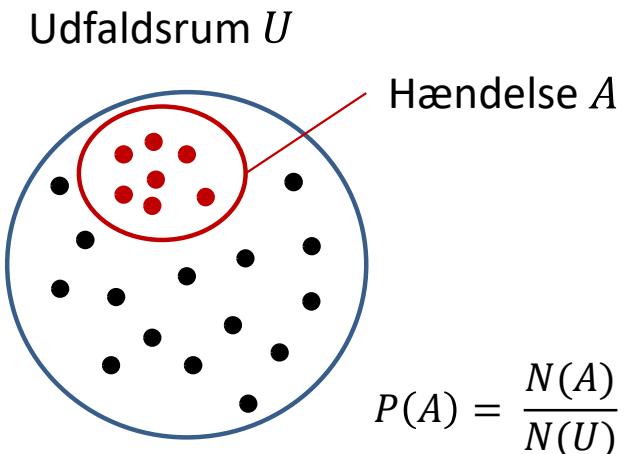
F.eks. at slå en 7'er med en almindelig terning

3. **Komplementær sandsynlighed:**

Den komplementære hændelse til A er mængden af udfald i U , som ikke tilhører A .

$$P(\bar{A}) = 1 - P(A)$$

F.eks.: Vi slår ikke en 6'er, kortet vi trækker er ikke en konge, o.s.v.



Eksempel for de basale regler for sandsynlighed

2013	Maskin	Kemi	Total
Mænd	222	112	334
Kvinder	41	95	136
Total	263	207	470

1. $0 \leq P(A) \leq 1$

Følger af definitionen af sandsynlighed $P(H) = \frac{N(H)}{N(U)}$

2. $P(\emptyset) = 0; P(U) = 1$

$$P(\text{Hund}) = \frac{0}{470} = 0; P(\text{Menneske}) = \frac{470}{470} = 1$$

3. Komplementær lov:

$$P(\bar{A}) = 1 - P(A)$$

$$P(\text{Ikke Maskin}) = 1 - P(\text{Maskin}) = 1 - \frac{263}{470} = 1 - 0.56 = 0.44 .$$

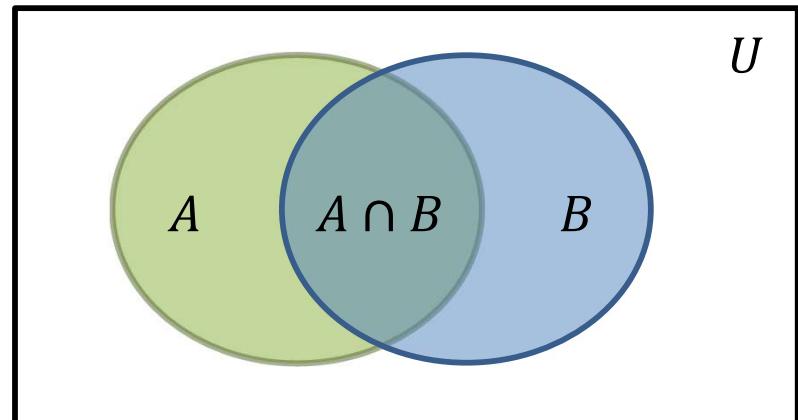
De 7 basale regler for sandsynlighed

4. Additiv lov for sandsynligheder (foreningshændelse):

Foreningshændelsen mellem A og B er at et udfald fra A eller B sker

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

F.eks.: Den person vi trækker tilfældigt er enten maskiningeniør eller mand
(eller begge dele)

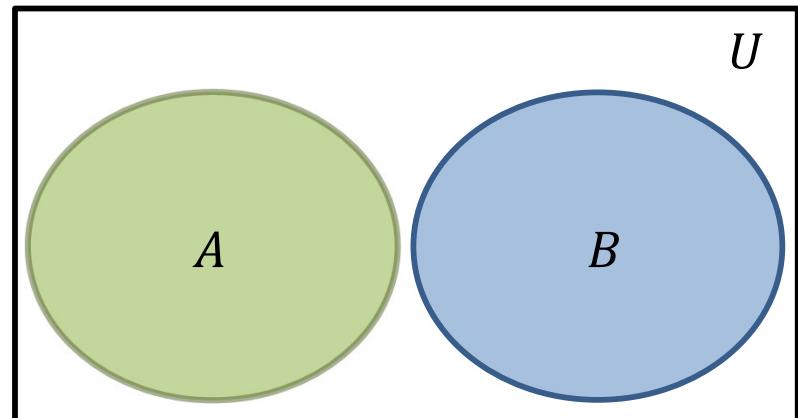


Hvis A og B ikke kan ske samtidig, kaldes de "gensidigt udelukkende"
(mutually exclusive).

Så er fælleshændelsen tom, så

$$P(A \cup B) = P(A) + P(B)$$

F.eks.: Vi slår en 6'er eller en 3'er.



Eksempel for de basale regler for sandsynlighed

2013	Maskin	Kemi	Total
Mænd	222	112	334
Kvinder	41	95	136
Total	263	207	470

4. Additiv lov: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$P(\text{Maskin eller Mand}) = P(\text{Maskin} \cup \text{Mand}) = \frac{222+41+112}{470} = \frac{375}{470} = 0.80$$

$$P(\text{Maskin}) + P(\text{Mand}) - P(\text{Maskin} \cap \text{Mand}) = \frac{263}{470} + \frac{334}{470} - \frac{222}{470} = \frac{375}{470} = 0.80 .$$

Eksempel

En produktion er præget af to typer af fejl.

- 15 % af de producerede emner har fejl 1 (og måske også fejl 2)
 - 10 % har fejl 2 (og måske også fejl 1)
 - 2.5 % af emnerne har begge fejl
-
- a) Hvor stor en andel af emnerne har mindst én fejl
 - b) Hvor stor en andel af emnerne er fejlfri?
 - c) Er forekomsten af de to typer af fejl uafhængige? .

De 7 basale regler for sandsynlighed

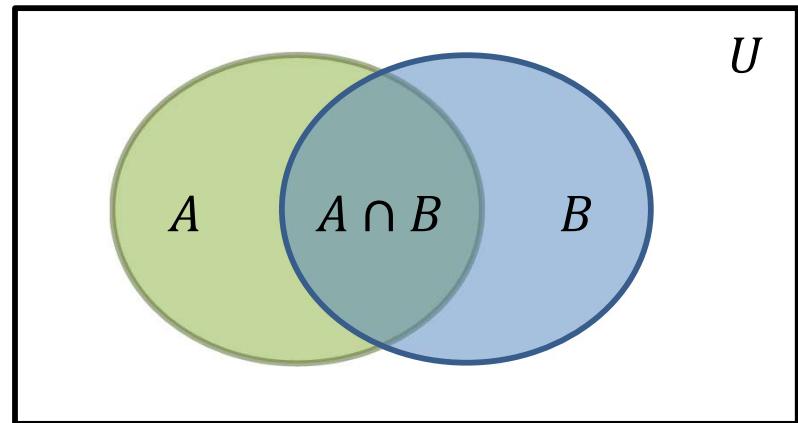
5. Multiplikative lov (fælleshændelse):

Fælleshændelsen mellem A og B er at udfaldet ligger i både A og B

$$P(A \cap B) = P(A | B) \cdot P(B)$$

$$P(A \cap B) = P(B | A) \cdot P(A)$$

F.eks.: Den person vi trækker er
både mand og maskiningeniør



Hvis A og B er uafhængige er $P(A | B) = P(A)$, så:

$$P(A \cap B) = P(A) \cdot P(B)$$

F.eks.: Vi slår plat med mønten og 6 med terningen.

Eksempel for de basale regler for sandsynlighed

2013	Maskin	Kemi	Total
Mænd	222	112	334
Kvinder	41	95	136
Total	263	207	470

5. Multiplikativ lov: $P(A \cap B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$

$$P(\text{Maskin og Mand}) = P(\text{Maskin} \cap \text{Mand}) = \frac{222}{470} = 0.47$$

$$P(\text{Maskin}) \cdot P(\text{Mand} | \text{Maskin}) = \frac{263}{470} \cdot \frac{222}{263} = \frac{222}{470} = 0.47$$

$$P(\text{Mand}) \cdot P(\text{Maskin} | \text{Mand}) = \frac{334}{470} \cdot \frac{222}{334} = \frac{222}{470} = 0.47.$$

De 7 basale regler for sandsynlighed

6. Loven om den totale sandsynlighed

Del B op i den del, der er fælles med A og den del, der ikke er.

Benyt først den additive lov (regel 4), så den multiplikative lov (regel 5):

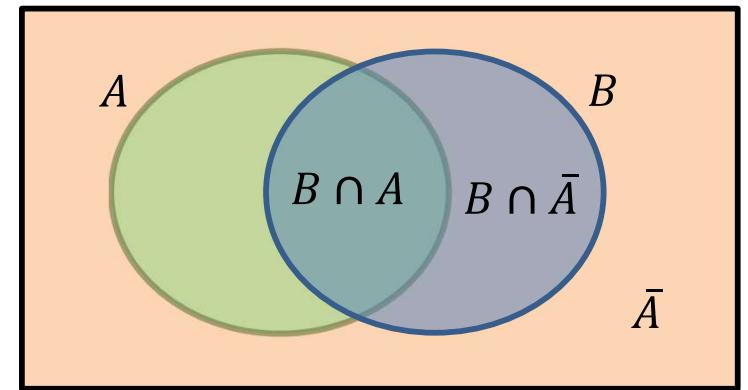
U

$$B = B \cap (A \cup \bar{A}) = (B \cap A) \cup (B \cap \bar{A}) \Rightarrow$$

$$P(B) = P((B \cap A) \cup (B \cap \bar{A})) \Rightarrow$$

$$P(B) = P(B \cap A) + P(B \cap \bar{A}) \Rightarrow$$

$$P(B) = P(B | A) \cdot P(A) + P(B | \bar{A}) \cdot P(\bar{A})$$



F.eks.:

$$P(\text{Mand}) = P(\text{Mand} | \text{Maskin}) \cdot P(\text{Maskin}) + P(\text{Mand} | \text{Kemi}) \cdot P(\text{Kemi})$$

Generelt: Ikke kun maskin- og kemiingeniører, men mange typer ingenior.

Så er A opdelt i n gensidigt udelukkende hændelser $A = A_1 \cup A_2 \cup \dots \cup A_n$

$$P(B) = \sum_{i=1}^n P(B | A_i) \cdot P(A_i)$$

Eksempel for de basale regler for sandsynlighed

2013	Maskin	Kemi	Total
Mænd	222	112	334
Kvinder	41	95	136
Total	263	207	470

6. Total sandsynlighed: $P(B) = \sum_{i=1}^n P(B|A_i) \cdot P(A_i)$

$$P(\text{Mand}) = \frac{334}{470} = 0.71$$

$$P(\text{Mand} | \text{Maskin}) \cdot P(\text{Maskin}) + P(\text{Mand} | \text{Kemi}) \cdot P(\text{Kemi}) = \\ \frac{222}{263} \cdot \frac{263}{470} + \frac{112}{207} \cdot \frac{207}{470} = \frac{222}{470} + \frac{112}{470} = \frac{334}{470} = 0.71.$$

De 7 basale regler for sandsynlighed

7. Bayes' regel

Bayes' regel følger af sætningen om betinget sandsynlighed:

$$P(A \cap B) = P(A | B) \cdot P(B)$$

$$P(A \cap B) = P(B | A) \cdot P(A)$$

Dermed er

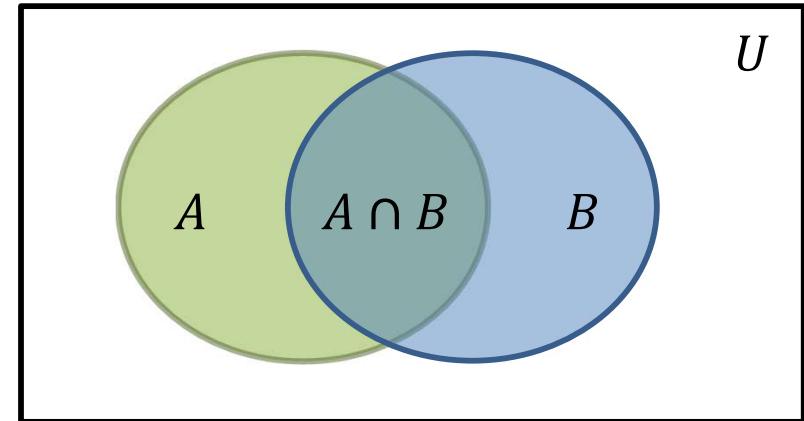
$$P(A | B) \cdot P(B) = P(B | A) \cdot P(A) \Rightarrow$$

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$

Dette kaldes Bayes' regel

Vi bruger regel 6 om den totale sandsynlighed i nævneren:

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{\sum_{i=1}^n P(B | A_i) \cdot P(A_i)}$$



Eksempel for de basale regler for sandsynlighed

2013	Maskin	Kemi	Total
Mænd	222	112	334
Kvinder	41	95	136
Total	263	207	470

7. Bayes' regel: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \Rightarrow P(A | B) \cdot P(B) = P(B | A) \cdot P(A)$

$$P(\text{Mand} | \text{Maskin}) \cdot P(\text{Maskin}) = \frac{222}{263} \cdot \frac{263}{470} = \frac{222}{470} = 0.47$$

$$P(\text{Maskin} | \text{Mand}) \cdot P(\text{Mand}) = \frac{222}{334} \cdot \frac{334}{470} = \frac{222}{470} = 0.47 .$$

Eksempel (medicinsk)

En mand går til lægen med et særligt symptom. Symptomet kan være tegn på, at han har en bestemt sygdom.

- Man ved, at manden tilhører en genetisk subpopulation, hvor 10 % lider af sygdommen
 - Af dem, der lider af sygdommen, har 75 % symptomet
 - 5 % af dem, der ikke lider af sygdommen har alligevel symptomet
-
- a) Hvad er sandsynligheden for, at manden har sygdommen?
 - b) Ændrer denne sandsynlighed sig, hvis forekomsten af sygdommen i subpopulationen er 1 % snarere end 10 %?.

Sandsynlighedsteori og statistik

Kapitel 4. Sandsynlighedsfordelinger (afsnit 4.1-4.2, 4.4, 4.6-4.7)

Allan Leck Jensen

alj@ece.au.dk

Allans terningspil

- Det koster 50 kr at spille
- Et spil består i et kast med en terning
- Hvis du slår 6, vinder du 200 kr.
Hvis du slår 1, vinder du 50 kr.
Hvis du slår 2, 3, 4 eller 5 vinder du ingenting
- Kan du forvente at tabe eller vinde penge i det lange løb?
- Det kan vi beregne med en **stokastisk variabel**.



Stokastisk variabel (tilfældighedsvariabel)

- En **stokastisk variabel** (*random variable*) er en variabel (funktion), der bruges til at beskrive tilfældige eksperimenter, hvor udfaldet ikke kendes på forhånd
- En stokastisk variabel X er en funktion fra udfaldsrummet U til de reelle tal
- F.eks. kan vi beskrive eksperimentet at slå plat og krone med en stokastisk variabel X , hvor vi tilknytter værdierne 1 og 2 til hhv. plat og krone
- Vi ved at X giver enten 1 eller 2 hver gang, men vi ved ikke hvilken
- I eksperimentet med to terninger er X en stokastisk variabel, der beskriver samlet antal øjne, f.eks. tilknyttes værdien $x = 5$ til udfaldet, at der slås en to'er og en tre'er eller en en'er og en fir'er
- Det er normalt at bruge store bogstaver (f.eks. X og Y) til at betegne en stokastisk variabel og små bogstaver (f.eks. x og y) til at betegne variablens værdier.

Eksempler på stokastiske variable

- Vi kan ikke forudsige udfaldet af en stokastisk variabel, men vi kan udtale os om sandsynligheden for ethvert udfald. F.eks.:
- Plat og krone:

$$P(X = 1) = \frac{1}{2}$$

$P(X = 1)$ angiver "Sandsynligheden for at den stokastiske variabel X antager værdien 1", m.a.o. sandsynligheden for at slå plat

- Antal plat med to mønter:

$$P(X = 1) = \frac{1}{2}; \quad P(X = 0) = P(X = 2) = \frac{1}{4}$$

- Terning:

$$P(X = x) = \frac{1}{6}, \quad x = 1, 2, \dots, 6$$

- Antal øjne med to terninger:

$$P(X = 5) = \frac{4}{36} = \frac{1}{9}.$$

		Terning 2					
		1	2	3	4	5	6
Terning 1	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Stokastisk vs deterministisk variabel

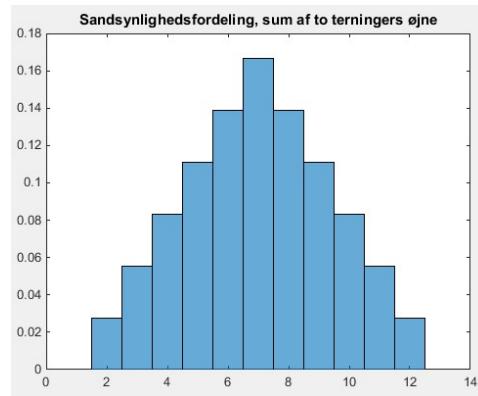
- En **deterministisk variabel** er en variabel (funktion), der bruges til at beskrive hændelser, hvor udfaldet kan forudsiges
- En deterministisk variabel giver samme (forudsigelige) output (resultat) med samme input, f.eks. faltdiden af en kugle fra samme højde
- En **stokastisk variabel** giver forskelligt output (resultat) med samme input, f.eks. faltdiden af en fjer fra samme højde
- Ingeniører og fysikere ignorerer ofte variabilitet og bruger deterministiske variable i modeller, fordi det er enklere end at bruge stokastiske variable.

Diskrete vs kontinuerte variable

- En **diskret** variabel kan antage et (i princippet) tælleligt antal værdier
- En **kontinuert** variabel kan (i princippet) antage alle værdier i de reelle tal eller i et interval heraf, f.eks. $[-10; 10]$
- Dette gælder uanset om variablen er **stokastisk** eller **deterministisk**
- I dag beskæftiger vi os med **diskrete**, stokastiske variable og beskriver deres ‘opførsel’ med **sandsynlighedsfordelinger**.

Sandsynlighedsfordeling for en diskret stokastisk variabel

- (Sandsynligheds)fordeling (engelsk: *(probability) distribution*)
- For hvert muligt udfald x_i tilknyttes sandsynligheden for udfaldet: $P(X = x_i)$
- For kast med to terninger har vi allerede beregnet sandsynlighedsfordelingen for den tilhørende stokastiske variabel

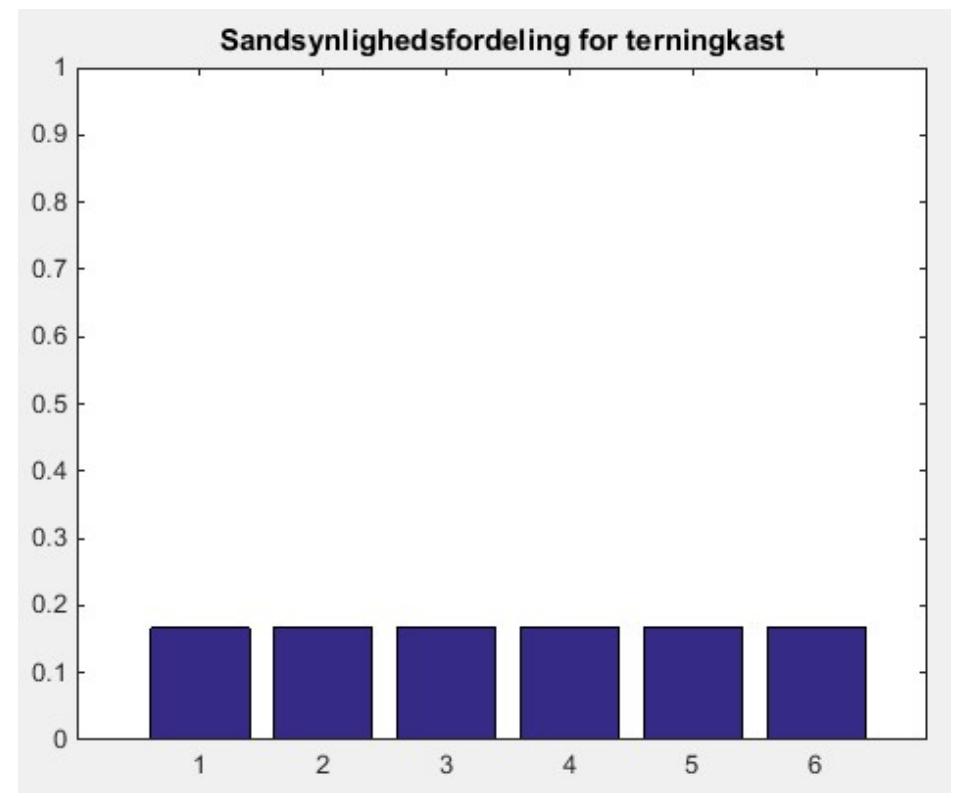
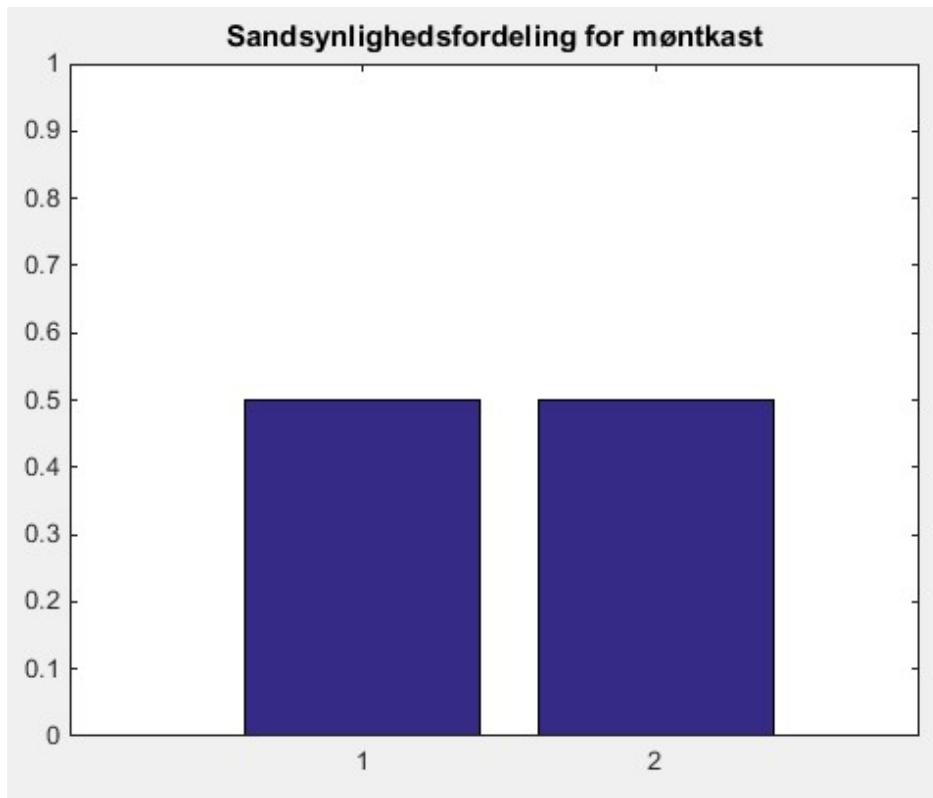


Sum af øjne	Kombinationer	Sandsynlighed
2	1	1/36 = 0.027778
3	2	2/36 = 0.055555
4	3	3/36 = 0.083333
5	4	4/36 = 0.111111
6	5	5/36 = 0.138889
7	6	6/36 = 0.166667
8	5	5/36 = 0.138889
9	4	4/36 = 0.111111
10	3	3/36 = 0.083333
11	2	2/36 = 0.055555
12	1	1/36 = 0.027778
Sum	36	36/36 = 1

- Nogle sandsynlighedsfordelinger er så karakteristiske, at de har et navn.

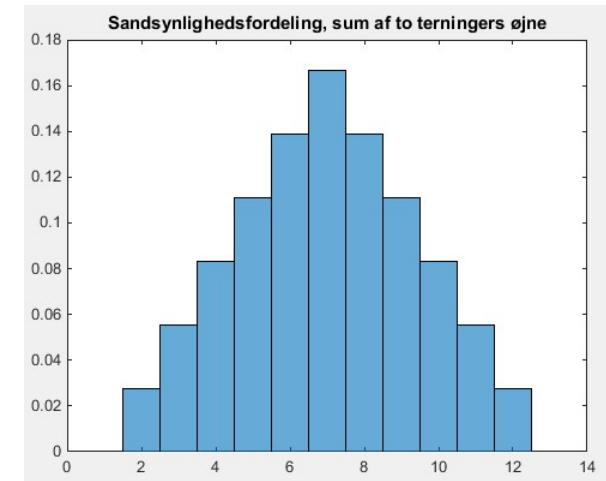
Uniform fordeling

- Sandsynlighedsfordelingen for en stokastisk variabel, hvor alle udfald er lige sandsynlige kaldes **uniform**
- Den uniforme fordeling kaldes også for **kassefordelingen**.



Sandsynlighedsfunktion

- Lad f være funktionen så
$$f(x) = P(X = x) \quad \text{for } x \in U$$
- Funktionen f kaldes **sandsynlighedsfunktionen** (*probability function*) for X
- Det er klart, at der gælder:
$$0 \leq f(x) \leq 1$$
$$\sum_{(x \in U)} f(x) = 1$$
- Statistikere kalder også **sandsynlighedsfunktionen** $f(x)$ for **fordelingsfunktionen** og **tæthedsfunktionen**
- Vi kalder den desuden pdf (*probability density function*).



Kumuleret fordelingsfunktion

- *Cumulative Distribution Function (cdf)*

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

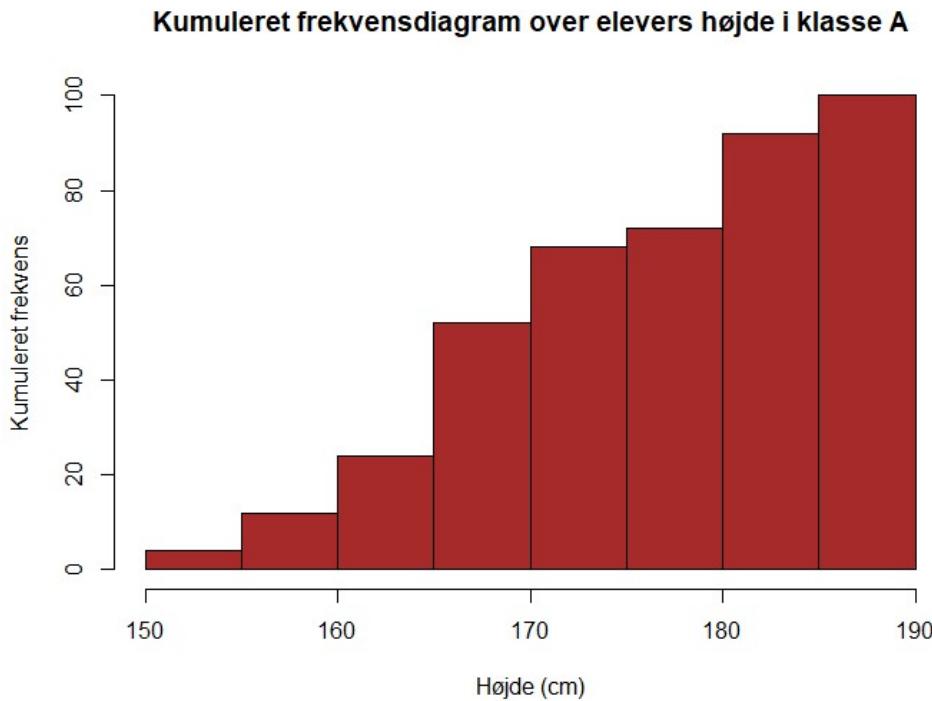
- Eksempel: PDF ($f(x)$) og CDF ($F(x)$) for antal øjne ved kast med to terninger:

x	f(x)	F(x)
2	$1/36 = 0.02778$	$1/36 = 0.02778$
3	$2/36 = 0.05556$	$3/36 = 0.08333$
4	$3/36 = 0.08333$	$6/36 = 0.16667$
5	$4/36 = 0.11111$	$10/36 = 0.27778$
6	$5/36 = 0.13889$	$15/36 = 0.41667$
7	$6/36 = 0.16667$	$21/36 = 0.58333$
8	$5/36 = 0.13889$	$26/36 = 0.72222$
9	$4/36 = 0.11111$	$30/36 = 0.83333$
10	$3/36 = 0.08333$	$33/36 = 0.91667$
11	$2/36 = 0.05556$	$35/36 = 0.97222$
12	$1/36 = 0.02778$	$36/36 = 1.00000$

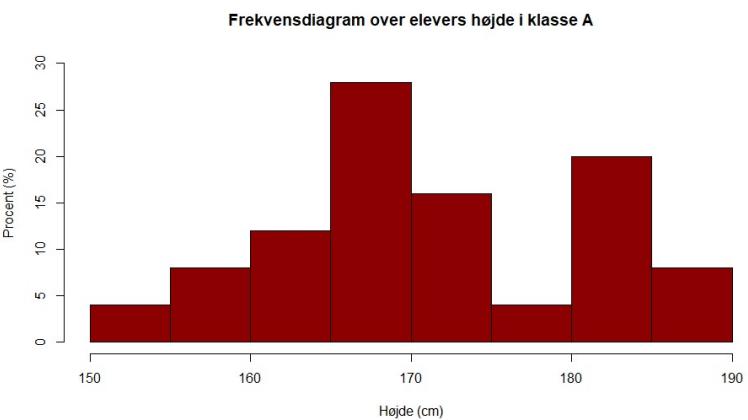
- Vi ser, at sandsynligheden for at få 5 øjne eller derunder er 27.8 %:
 $P(X \leq 5) = F(5) = 0.27778.$

Fra kap. 2: Kumuleret frekvensdiagram

- Intervallernes observationer akkumuleres, så hver søjle viser frekvensen med højde op til intervalgrænsen. F.eks. er 4 % 155 cm eller derunder og 72 % er 180 cm eller derunder
- Sidste søjle er 100 %, så alle elever er 190 cm eller derunder
- Laves i R med `hist()`, men der skal kodes lidt.



Interval	Antal	Pct.	Kumuleret	
			Antal	Pct.
(150-155]	1	4%	1	4%
(155-160]	2	8%	3	12%
(160-165]	3	12%	6	24%
(165-170]	7	28%	13	52%
(170-175]	4	16%	17	68%
(175-180]	1	4%	18	72%
(180-185]	5	20%	23	92%
(185-190]	2	8%	25	100%
Total	25	100%	25	100%



Binomialfordelingen

- Eksempler fra bogen på statistiske problemer, hvor vi bruger binomialfordelingen. Hvad er sandsynligheden for:
 - at 1 ud af 5 nitter knækker i en trækstyrketest
 - 45 ud af 300 stoppede bilister kørte uden sikkerhedssele
 - 66 ud af 200 TV-seere kan huske den reklame de så
 - Generelt: Sandsynligheden for x succeser i n forsøg (og dermed $n - x$ fejl)
- Vi bruger binomialfordelingen til serier af forsøg, der kaldes **Bernoulli-forsøg**:
 - Hvert forsøg har to mulige udfald, typisk kaldet Succes (S) og Fejl (F)
 - Sandsynligheden for Succes er ens for hvert forsøg, $P(S) = p$. Dermed er $P(F) = 1 - p$
 - Resultatet af de n Bernoulli-forsøg er uafhængige af hinanden
- Kan en politirazzia, hvor biler stoppes og testes for om chaufføren har sikkerhedssele på opfattes som et Bernoulli-forsøg?

Eksempel 4.3: Reparation af 3 mobilmaster

Et firma påstår at de reparerer en mobilmast indenfor en time i 90 % af tilfældene. De næste tre nedbrud af mobilmaster undersøges

- List alle mulige udfald for 3 nedbrud, hvor S betyder, at en mast blev repareret indenfor en time, og F betyder, at det gjorde den ikke
- Find sandsynlighedsfordelingen for den stokastiske variabel X , der angiver antal S for de tre reparationer

Løsning:

- Der er 2 mulige udfald for hver reparation, så disse $2^3 = 8$ udfald:

FFF	FFS	FSS	SSS
FSF	SFS		
SFF	SSF		

$$X = 0 \quad X = 1 \quad X = 2 \quad X = 3$$

- Vi antager, at reparationerne er uafhængige, og $P(S) = p = 0.9$, $P(F) = 1 - p = 0.1$

$$P(X = 3) = P(SSS) = p \cdot p \cdot p = p^3 = (0.9)^3 = \mathbf{0.729}$$

Tilsvarende:

$$P(X = 0) = P(FFF) = (1 - p) \cdot (1 - p) \cdot (1 - p) = (0.1)^3 = \mathbf{0.001}$$

Eksempel 4.3: Reparation af 3 mobilmaster

b. (fortsat)

$$\begin{aligned} P(X = 2) &= P(FSS \cup SFS \cup SSF) \\ &= P(FSS) + P(SFS) + P(SSF) \end{aligned}$$

$$P(FSS) = (1 - p) \cdot p \cdot p$$

$$P(SFS) = p \cdot (1 - p) \cdot p$$

$$P(SSF) = p \cdot p \cdot (1 - p)$$

$$\begin{aligned} P(X = 2) &= 3 \cdot p^2 \cdot (1 - p) \\ &= 3 \cdot (0.9)^2 \cdot (1 - 0.9) = \mathbf{0.243} \end{aligned}$$

Og endelig:

$$P(X = 1) = P(FFS) + P(FSF) + P(SFF)$$

$$\begin{aligned} P(X = 1) &= 3 \cdot p \cdot (1 - p)^2 \\ &= 3 \cdot 0.9 \cdot (1 - 0.9)^2 = \mathbf{0.027} \end{aligned}$$

Generelt for x gange S ud af 3 forsøg:

$$\begin{aligned} P(X = x) &= (\text{Antal måder at få } x \text{ gange } S \text{ ud af 3}) \cdot p^x \cdot (1 - p)^{3-x} \\ &= \binom{3}{x} \cdot p^x \cdot (1 - p)^{3-x} \text{ for } x = 0, 1, 2, 3 \end{aligned}$$

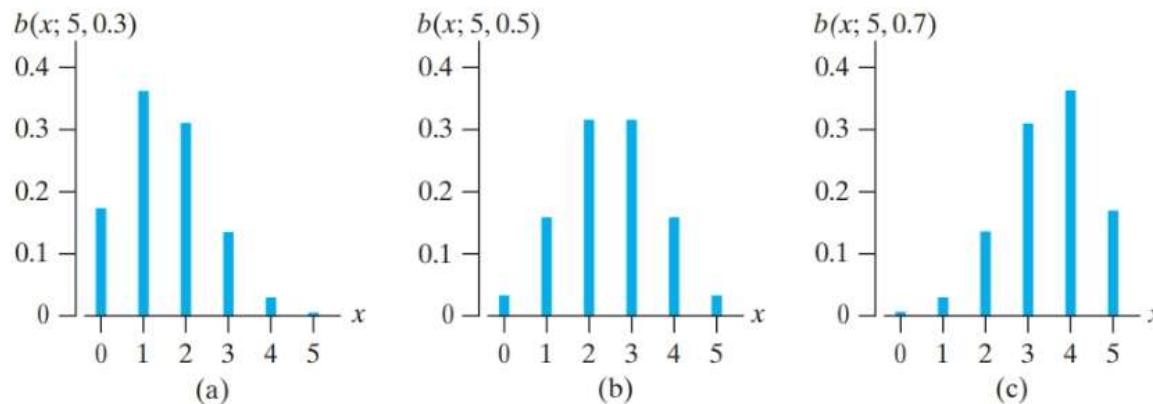
hvor $\binom{3}{x}$ er binomialkoefficienten (kapitel 3), så $\binom{3}{x} = \frac{3!}{x!(3-x)!}$.

FFF	FFS	FSS	SSS
FSF	SFS	SFF	SSF
$X = 0$	$X = 1$	$X = 2$	$X = 3$

x	f(x)	F(x)
0	0.001	0.001
1	0.027	0.028
2	0.243	0.271
3	0.729	1.000

Binomialfordelingen

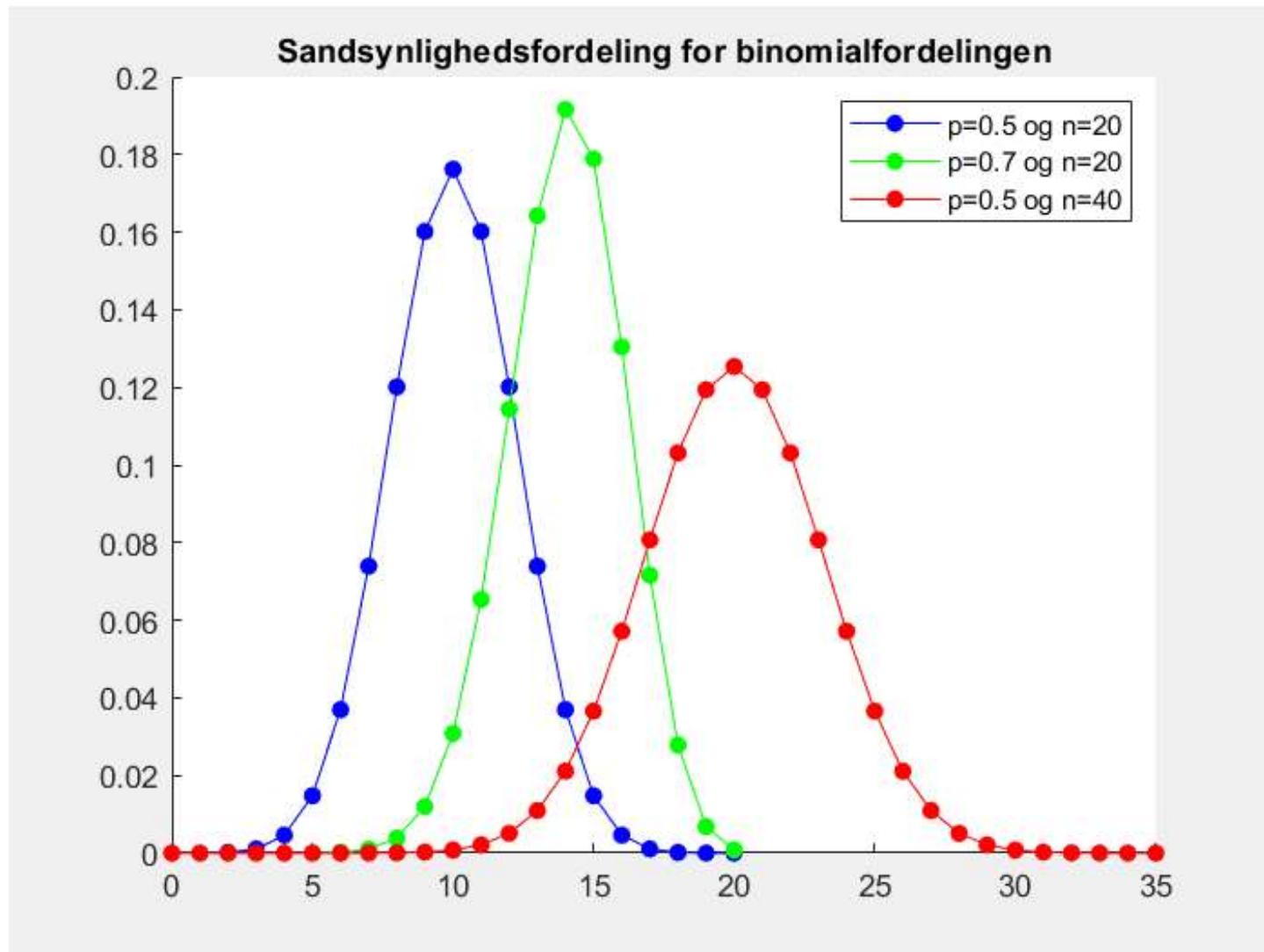
- Sandsynlighedsfunktionen for binomialfordelingen kaldes $b(x; n, p)$:
$$b(x; n, p) = \binom{n}{x} \cdot p^x \cdot (1 - p)^{n-x} \text{ for } x = 0, 1, 2, \dots, n$$
- Binomialfordelingens form afhænger af parametrene n og p :



- Når $p = 0.5$ (b): $b(x; n, p)$ er symmetrisk omkring $n/2$
$$b(x; n, p) = \binom{n}{x} \cdot (0.5)^n$$
- Når $p < 0.5$ (a): $b(x; n, p)$ har en hale opadtil (højre-hale), fordi lave værdier af x er mere sandsynlige
- Når $p > 0.5$ (c): $b(x; n, p)$ har en hale nedadtil (venstre-hale) fordi høje værdier af x er mere sandsynlige.

Binomialfordeling

Afhængighed af p og n



PDF OG CDF for binomialfordelingen

- Sandsynlighedsfunktion PDF: $P(X = x)$

$$b(x; n, p) = \binom{n}{x} \cdot p^x \cdot (1 - p)^{n-x} \text{ for } x = 0, 1, 2, \dots, n$$

- I R: `dbinom(x, n, p)`

- Kumuleret fordelingsfunktion CDF: $P(X \leq x)$

$$B(x; n, p) = \sum_{k=0}^x b(k; n, p)$$

- I R: `pbinom(x, n, p)`

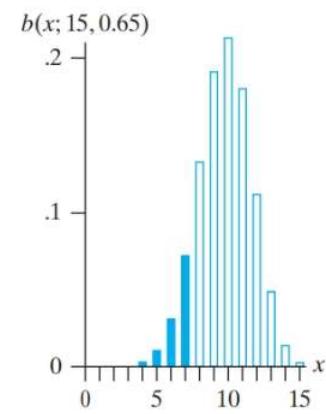
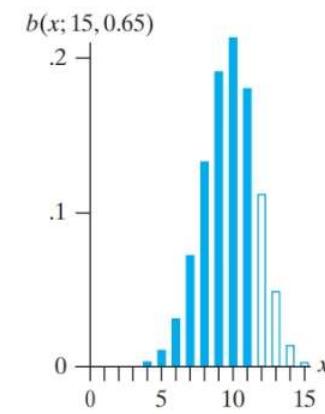
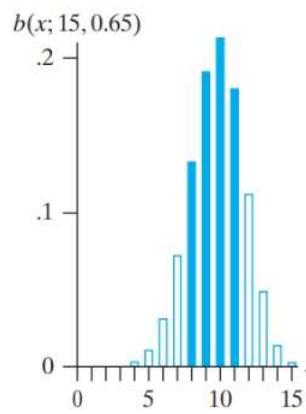
Eksempel 4.6. Kunstig sportsjournalist

På et sportssite bliver 65 % af artiklerne lavet af en AI-robot. Ud af de næste 15 artikler, hvad er sandsynligheden for:

- a. Præcis 11 er skrevet af robotten
- b. Mindst 10 er skrevet af robotten
- c. Mellem 8 og 11 (begge inclusive) er skrevet af robotten

Løsning i R:

- a. $\text{dbinom}(11, 15, 0.65)$
= 0.1792
- b. $1 - \text{pbinom}(9, 15, 0.65)$
 $= 1 - 0.4357 = \underline{0.5643}$
- c. $\text{pbinom}(11, 15, 0.65)$
 $- \text{pbinom}(7, 15, 0.65)$
 $= 0.8273 - 0.1132 = \underline{0.7141}$



$$P(8 \leq X \leq 11) = B(11; 15, 0.65) - B(7; 15, 0.65)$$

Deskriptorer

- Vi kan tale om **deskriptorer** for en stokastisk variabel, ligesom for et datasæt:
 - Middelværdi, Varians, Standardafvigelse
- Vi skelner mellem om det er for en ‘stikprøve’ eller for hele ‘populationen’
- Lad f.eks. X være en stokastisk variabel, der modellerer terningkast:
$$f(x) = P(X = x) = 1/6 \quad \text{for } x \in \{1, 2, \dots, 6\}$$
- En stikprøve giver 2, 6, 1, 2.
Stikprøve-middelværdien er gennemsnit af stikprøven:
$$\bar{x} = \frac{1}{4}(2 + 6 + 1 + 2) = \frac{11}{4} = 2.25$$
- **Populations-middelværdien** er det langsigtede gennemsnit:
$$\mu = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5 .$$

Populations-middelværdi μ

- Forventet værdi (*expected value*) $E(X)$ er det langsigtede gennemsnit
- $E(X)$ og μ er to alternative måder at betegne middelværdien på:

$$\mu = E(X) = \sum_x x \cdot f(x).$$

Populations-variens σ^2

- Variansen (*variance*) er et mål for gennemsnitlig afvigelse fra middelværdien
- Afvigelse (*deviation*) af et datapunkt x_i :

$$x_i - \mu$$

- Gennemsnitlig afvigelse:

$$\mathbb{E}(X - \mu) = 0$$

Derfor er afvigelsen et dårligt mål for varians, så vi bruger kvadratet af afvigelsen:

$$\mathbb{E}[(X - \mu)^2]$$

- Populations varians:

$$\sigma^2 = \text{Var}(X) = \mathbb{E}[(X - \mu)^2] = \sum_x (x - \mu)^2 \cdot f(x)$$

- Variansen kan lettere beregnes som:

$$\sigma^2 = \sum_x (x^2 \cdot f(x)) - \mu^2.$$

Populations-standardafvigelse σ

- Standardafvigelse (eller *spredning*) (*standard deviation*) defineres som:

$$\sigma = \sqrt{\sigma^2}$$

- Standardafvigelsen har samme enhed som dataværdierne
- Den empiriske regel:

Normalt ligger næsten alle data i intervallet

$$\mu \pm 3\sigma$$

(empirisk betyder ‘tommelfingerregel baseret på erfaring’).

Deskriptorer for terningkast

- Lad igen X være en stokastisk variabel, der modellerer terningkast:

$$f(x) = P(X = x) = 1/6 \quad \text{for } x \in \{1, 2, \dots, 6\}$$

- Populations-middelværdien:

$$\mu = \sum_{i=1}^6 x_i \cdot f(x_i) = \frac{1}{6} \cdot \sum_{i=1}^6 x_i = \frac{1}{6} \cdot 21 = 3.5$$

- Populations-variанс:

$$\begin{aligned}\sigma^2 &= \sum_{i=1}^6 x_i^2 \cdot f(x_i) - \mu^2 = \frac{1}{6} \cdot \sum_{i=1}^6 x_i^2 - \mu^2 \\ &= \frac{1}{6} \cdot (1 + 4 + 9 + 16 + 25 + 36) - (3.5)^2 \\ &= 15.16667 - 12.25 = 2.9167\end{aligned}$$

- Populations-standardafvigelse:

$$\sigma = \sqrt{2.9167} = 1.7078$$

- Empirisk interval:

$$\mu \pm 3\sigma = 3.5 \pm 3 \cdot 1.7078 = [-1.6; 8.6] = [1; 6].$$

Deskriptorer for binomialfordelingen

- Man kan vise, at deskriptorerne for binomialfordelingen med parametre n og p er:

$$\mu = E(X) = np$$

$$\sigma^2 = \text{Var}(X) = np(1 - p)$$

$$\sigma = \sqrt{np(1 - p)}.$$

* Opgave: Metaltræthed i stålbjælker

En producent af stålbjælker har et problem i produktionen, som bevirker, at 10 % af stålbjælkerne udviser metaltræthed efter cirka 10 år. Normale stålbjælker udviser ikke metaltræthed, selv efter 50 år. Man kender ikke årsagen til problemet, og den eneste måde man med sikkerhed kan påvise, om en nyproduceret bjælke har svagheden, er ved at udsætte den for en destruktiv styrketest. Man er selvfølgelig ikke interesseret i at ødelægge bjælkerne for at undersøge, om deres kvalitet er i orden. Heldigvis har man opdaget, at en ultralydsscanning giver et særligt mønster for de bjælker, der er svage. Desværre giver metoden ikke en sikker indikation: 87 % af de svage bjælker udviser det særlige mønster ved scanningen, men det gør 7 % af de stærke bjælker også.

Lad S betegne hændelsen at en bjælke er svag, og lad R betegne hændelsen at en bjælke udviser det særlige mønster ved ultralydsscanning. Lad S^c og R^c betegne komplementærhændelsen til henholdsvis S og R . Dermed følger det for eksempel af opgaveteksten, at $P(R|S) = 87\% = 0.87$ og $P(R^c|S) = 1 - P(R|S) = 0.13$.

- Angiv værdien af følgende sandsynligheder:
 $P(S)$, $P(S^c)$, $P(R|S^c)$, $P(R^c|S^c)$.
- Hvad er sandsynligheden for, at en tilfældigt udvalgt stålbjælke vil udvise det særlige mønster, når den bliver scannet?
- En tilfældigt udvalgt stålbjælke bliver scannet, og det viser sig, at den udviser det særlige mønster. Hvad er sandsynligheden for, at den er svag?
- En anden tilfældigt udvalgt stålbjælke bliver også scannet, og her viser det sig, at den ikke udviser det særlige mønster. Hvad er sandsynligheden for, at den alligevel er svag?

Poisson-fordelingen

- En fabrik producerer i gennemsnit 4.2 defekte produkter om dagen. Hvad er sandsynligheden for at den producerer præcis 7 defekte i morgen?
- Poisson-fordelingen bruges, når man tæller antal ‘succes’er’ (x) og kender det forventede antal pr. enhed eller tidsrum (λ)
- Sandsynlighedsfunktionen for Poisson-fordelingen er:

$$f(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \quad \text{for } x = 0, 1, 2, \dots \text{ og } \lambda > 0$$

- Bemærk, at x ikke har en øvre grænse, i modsætning til binomialfordelingen, hvor $0 \leq x \leq n$
- Eksemplet: $f(7; 4.2) = \frac{(4.2)^7}{7!} e^{-(4.2)} = 0.0686$
- Middelværdi, varians og standardafvigelse for Poisson-fordelingen:

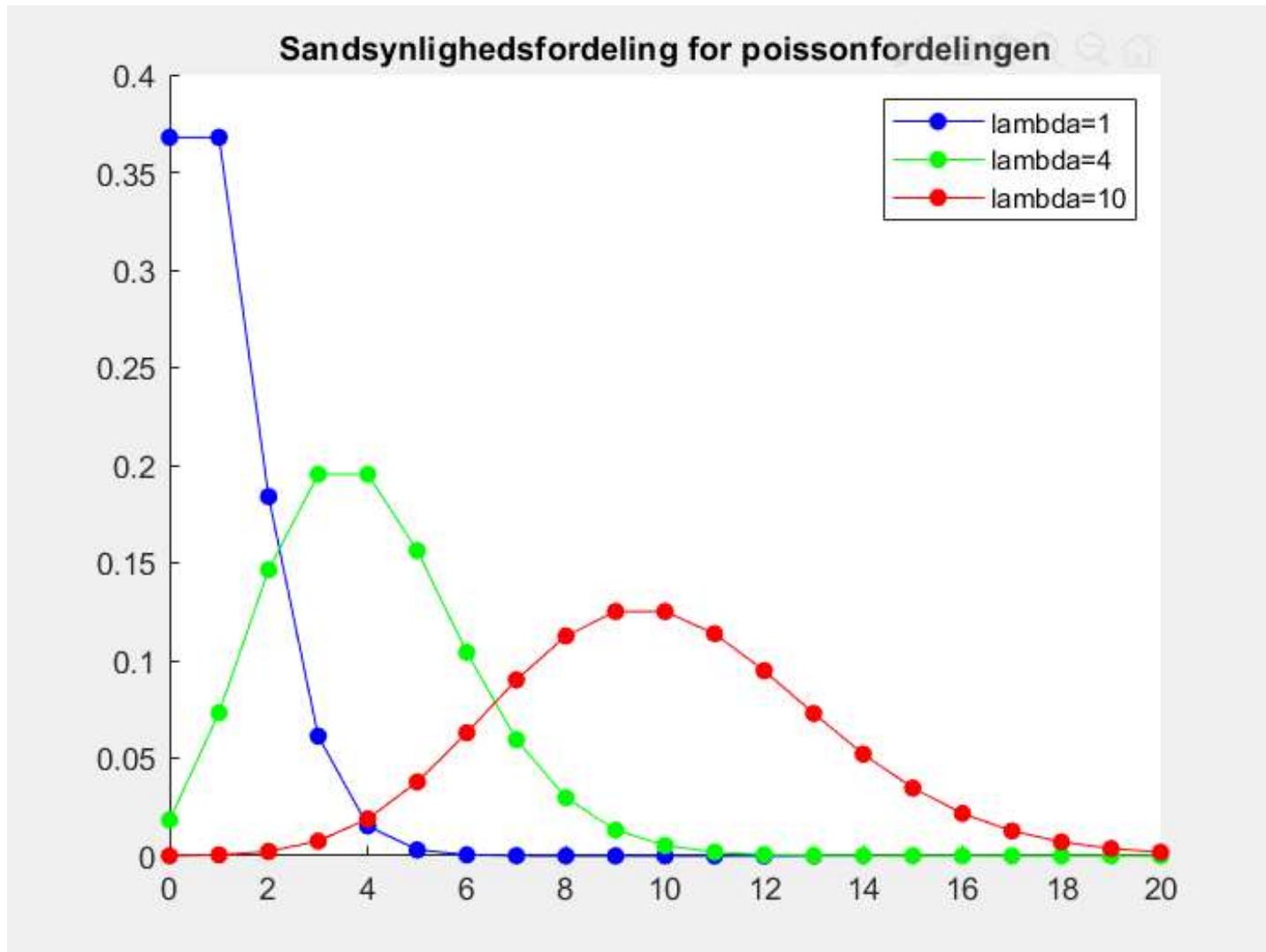
$$\mu = E(X) = \lambda$$

$$\sigma^2 = \text{Var}(X) = \lambda$$

$$\sigma = \sqrt{\lambda}.$$

Poisson-fordelingen

Afhængighed af λ :



PDF OG CDF for Poisson-fordelingen

- Sandsynlighedsfunktion PDF: $P(X = x)$

$$f(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \quad \text{for } x = 0, 1, 2, \dots \text{ og } \lambda > 0$$

- I R: `dpois(x, lambda)`

- Kumuleret fordelingsfunktion CDF: $P(X \leq x)$

$$F(x; \lambda) = \sum_{k=0}^x f(k; \lambda)$$

- I R: `ppois(x, lambda)`.

Eksempel 4.25: Nedbrud af Internet-service

En internetudbyder påstår, at en bestemt service har nedbrud i gennemsnit 0.2 gange per uge. Hvad er sandsynligheden for

- a. Præcis 1 nedbrud i de næste 3 uger
- b. Mindst 2 nedbrud i de næste 5 uger
- c. Højst 1 nedbrud i de næste 15 uger

Løsning:

- a. Med 0.2 nedbrud per uge vil vi forvente $\lambda = 3 \cdot 0.2 = 0.6$ nedbrud på tre uger.

$$P(X = 1) = \text{dpois}(1, 0.6) = 0.329$$

- b. Nu ser vi på 5 uger, så vi forventer $\lambda = 5 \cdot 0.2 = 1.0$ nedbrud.

$$P(X \geq 2) = 1 - \text{ppois}(1, 1.0) = 1 - 0.736 = 0.264$$

- c. På 15 uger forventer vi $\lambda = 15 \cdot 0.2 = 3.0$ nedbrud.

$$P(X \leq 1) = \text{ppois}(1, 3.0) = 0.199.$$

Generelle R funktioner til fordelinger

- **Tæthedsfunktion** (Probability Density Function, **PDF**)

Den funktion, der beskriver sandsynlighedsfordelingen. Tæthedsfunktionen giver $P(X = x)$ for hvert x , som den stokastiske variabel X kan antage

R: `dxxx()`, hvor 'xxx' erstattes af en forkortelse for fordelingen

- **Kumuleret fordelingsfunktion** (Cumulated Distribution Function, **CDF**)

En funktion, der beregner $p = P(X \leq x)$ for hvert x , som den stokastiske variabel X kan antage

R: `pxxx()`

- **Invers kumuleret fordelingsfunktion**

Den inverse funktion til CDF, så for en given sandsynlighed p beregner funktionen den værdi af x , så der gælder at $P(X \leq x) = p$

R: `qxxx()`

- **Tilfældighedsgenerator**

En funktion, der kan generere tilfældige tal af fordelingen

R: `rxxx()`.

R funktioner til fordelinger, eksempel

- En fabrik producerer 200 enheder dagligt. 1 % enheder har fejl. Hvad er sandsynligheden for 0 enheder med fejl i produktionen i morgen?
 $\text{dbinom}(0, 200, 0.01) = 0.134$
- Hvad er sandsynligheden for 5 fejl eller færre?
 $\text{pbinom}(5, 200, 0.01) = 0.984$
- Hvor mange fejl kan fabrikken гаранtere at være under i 99 % af dagene?
 $\text{qbinom}(0.99, 200, 0.01) = 6$
(Altså kun i 1 % af dagene er der mere end 6 fejl (7 eller flere))
- Lav en simulering af antal fejl de næste 5 dage:
 $\text{rbinom}(5, 200, 0.01) = [2, 3, 1, 2, 0].$

R funktioner til diskrete fordelinger

	PDF (dxxx)	CDF (pxxx)	Invers (qxxx)	Random (rxxx)
Binomial (xbinom)	<code>dbinom()</code>	<code>pbinom()</code>	<code>qbinom()</code>	<code>rbinom()</code>
Poisson (xpois)	<code>dpois()</code>	<code>ppois()</code>	<code>qpois()</code>	<code>rpois()</code>

Binomial- eller Poisson-fordeling?

- Eksempel:

En bilfabrik producerer 25 biler i timen. 5 % af dem må kasseres pga. graverende produktionsfejl (d.v.s. i gennemsnit kasseres 1.25 biler/time). Hvad er sandsynligheden for at der kasseres hhv. 2 og 26 biler den næste time?

 - $P(X=2) = \binom{25}{2} \cdot 0.05^2 \cdot (1 - 0.05)^{25-2} = \text{dbinom}(2, 25, 0.05) = \underline{0.2305}$
 - $P(X=26) = \text{dbinom}(26, 25, 0.05) = \underline{0}$ (*umuligt*)
- Alternativ opgaveformulering:

En bilfabrik må kassere 1.25 producerede biler i timen pga. graverende produktionsfejl. Hvad er sandsynligheden for at der kasseres hhv. 2 og 26 biler den næste time?

 - $P(X=2) = \frac{1.25^2}{2!} e^{-1.25} = \text{dpois}(2, 1.25) = \underline{0.2238}$
 - $P(X=26) = \text{dpois}(26, 1.25) = \underline{2.35e-25}$ (*usandsynligt, ikke umuligt*).

Opgave om robot til sprøjtemaling

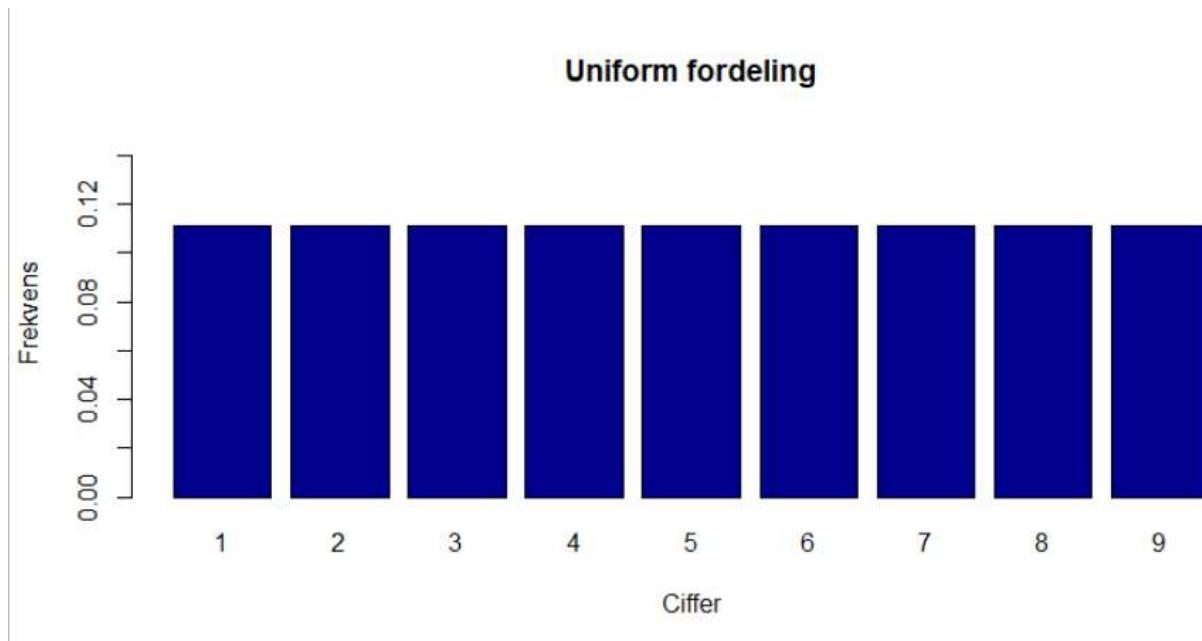
En robot til sprøjtemaling laver utilsigtede 'helligdage', d.v.s. små pletter, der ikke er blevet dækket af maling. Robotten laver i gennemsnit 0.8 helligdage per malet kvadratmeter. Robotten skal male 70 cirkelformede skiver på forsiden. Hver skive har en diameter på 1.2 m



- Hvor mange helligdage må der forventes at være på en tilfældig skive?
- Hvilken sandsynlighedsfordeling vil du bruge til at beskrive antal helligdage på en skive, og hvad er fordelingens middelværdi, varians og standardafvigelse?
- Hvad er sandsynligheden for, at ingen af de 70 skiver har helligdage?
- Beregn det forventede antal skiver med henholdsvis 0, 1, 2, 3 og 4 eller flere helligdage.

Benfords lov (ikke pensum)

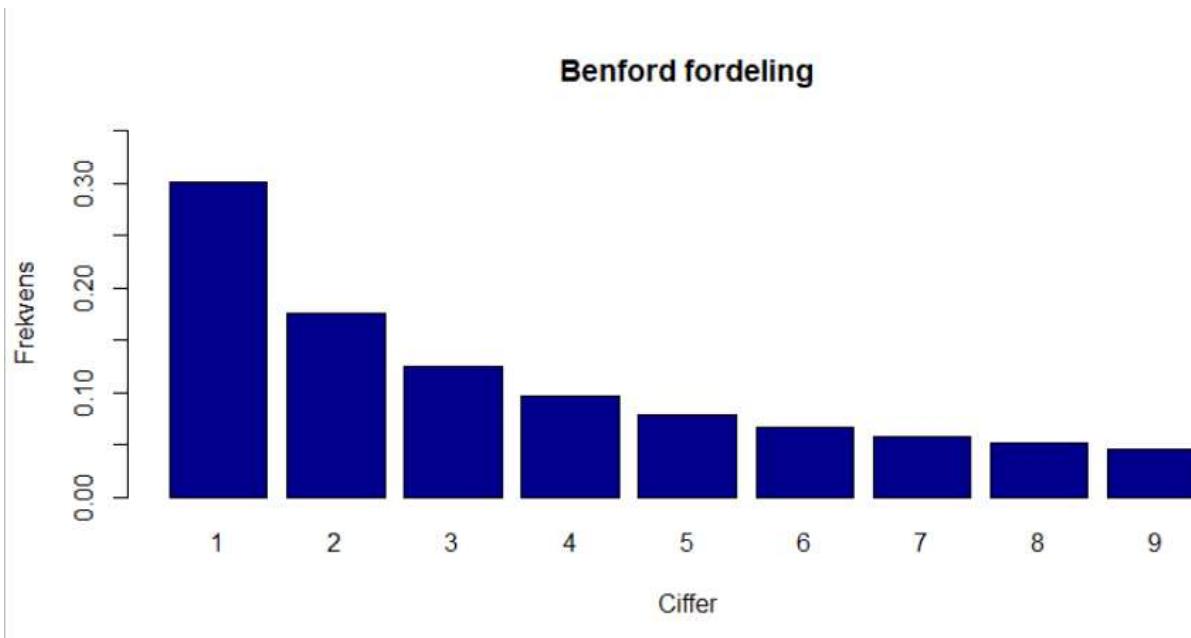
- Frank Benford (1883-1948, fysiker og el-ingeniør) undersøgte naturligt forekommende data, f.eks. 104 fysiske konstanter, 1800 molekylvægte, størrelse af 3259 befolkninger, overfladeareal af 335 floder
- For hvert datasæt tog han første ciffer af alle tallene og så på fordelingen af 1, 2, 3, ..., 9
- Umiddelbart ville man forvente, at de følger en uniform fordeling:



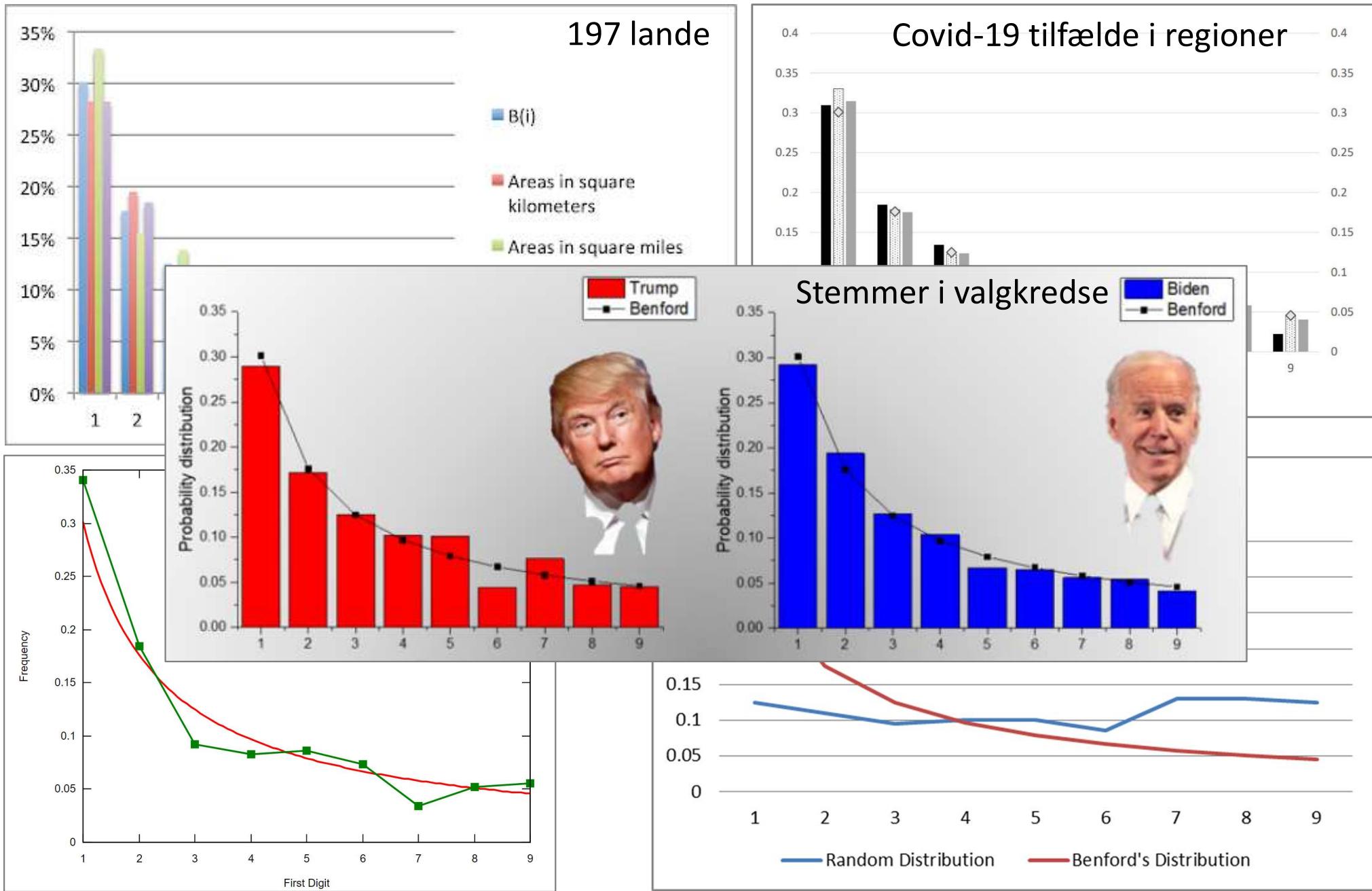
Benfords lov (ikke pensum)

- Men det viser sig, at der forekommer flest 1-taller og færrest 9-taller
- Hvorfor??
- Sandsynlighedsfunktionen:

$$f(x) = \log_{10}(x + 1) - \log_{10}(x) = \log_{10}\left(1 + \frac{1}{x}\right) \quad x \in \{1, \dots, 9\}$$



Eksempler



Sandsynlighedsteori og statistik

Kapitel 5. Kontinuerte stokastiske variable og deres sandsynlighedsfordelinger (afsnit 5.1-5.2, 5.4-5.5, 5.7, 5.10, 5.12-5.13)

Allan Leck Jensen
alj@ece.au.dk

Kontinuert stokastisk variabel

- En stokastisk variabel X kaldes **kontinuert**, hvis den kan antage (i princippet) alle værdier i et interval $x_1 \leq x \leq x_2$
- F.eks.: X er højden af en tilfældigt udtrukket person, eller tiden der går til næste ulykke, eller trækstyrken af en metalstang
- Der findes ingen mennesker på *præcis* 180 cm (hvis vi måler uendeligt præcist), så $P(X = 180) = 0$
- I stedet kan vi tale om f.eks. $P(179.5 < X \leq 180.5)$
- Generelt: Da der er uendeligt antal værdier af X er $P(X = x) = 0$ for en kontinuert stokastisk variabel X .

Kontinuert stokastisk variabel

Kontinuerte stokastiske variable karakteriseres af deres **tæthedsfunktion** (*Probability Density Function (pdf)*). Bogen kalder det ‘density function’. Tæthedsfunktionen $f(x)$ er en funktion, hvor der gælder:

1. $f(x) \geq 0$ for alle $x \in X$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

F.eks. **Uniform fordeling** (aka. **kassefordeling**, engelsk: *Uniform distribution*):

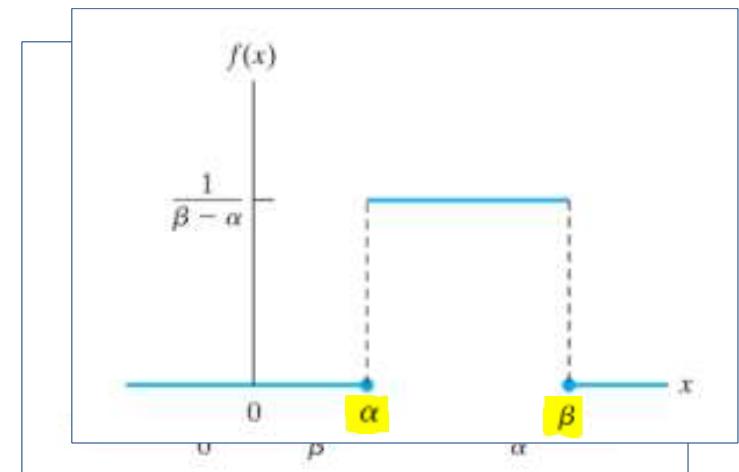
$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{for } \alpha < x < \beta \\ 0 & \text{ellers} \end{cases}$$

Bemærk, at $f(x)$ er en tæthedsfunktion:

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

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

Vigtigt: Hvis $\alpha < x_0 < \beta$, så er $f(x_0) = \frac{1}{\beta - \alpha}$ men $P(X = x_0) = 0$.



Kontinuert stokastisk variabel

Tæthedsfunktionen $f(x)$ bruges til at beskrive sandsynlighed, dog ikke for et punkt, men for et interval, som vi integrerer over

Se på den kumulerede fordelingsfunktion (engelsk: *Cumulated Distribution Function*, cdf.), $F(x)$:

$$F(x_0) = P(X \leq x_0) = \int_{-\infty}^{x_0} f(x)dx$$

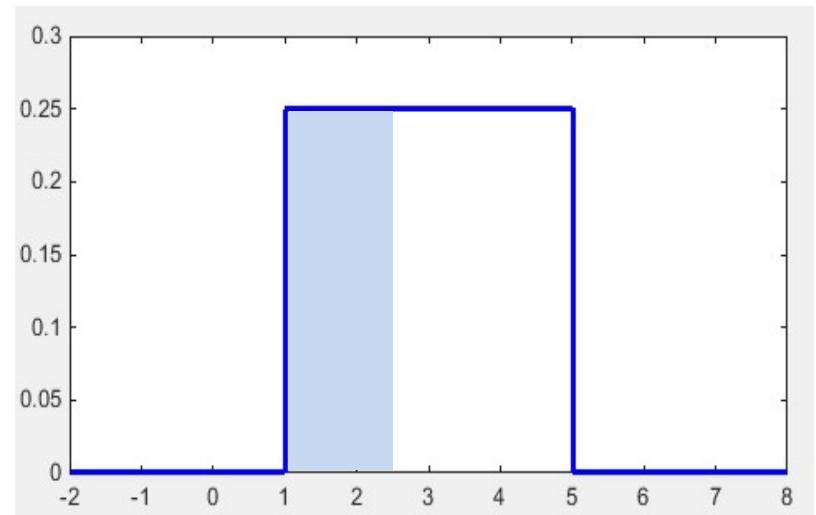
Bogen kalder $F(x)$ for ‘distribution function’

F.eks. i Uniform fordeling mellem 1 og 5 er

$$f(x) = \frac{1}{(5-1)} = \frac{1}{4} \text{ når } 1 < x < 5$$

$$F(2.5) = P(X \leq 2.5) = \int_{-\infty}^{2.5} \frac{1}{4} dx$$

$$\begin{aligned} &= \frac{1}{4} \int_1^{2.5} dx = \frac{1}{4} (2.5 - 1) = \frac{3}{8} \\ &= 0.375. \end{aligned}$$

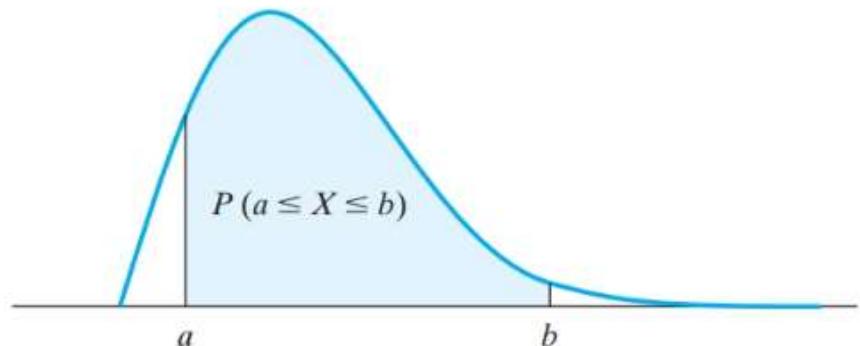


Kontinuert stokastisk variabel

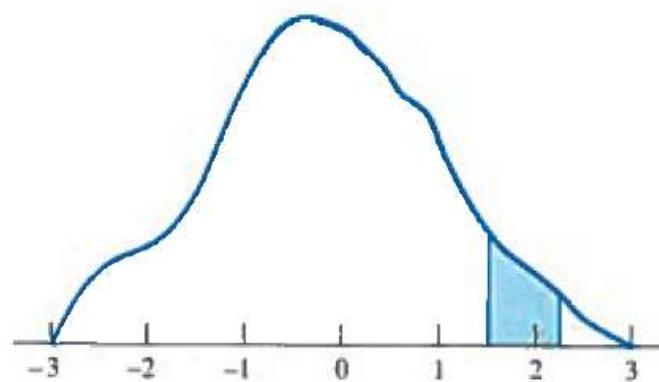
Sandsynlighed bliver målt som arealet under kurven f

Hvis a og b er to værdier, vi er interesserede i:

$$\begin{aligned} P(a \leq X \leq b) &= \int_a^b f(x)dx \\ &= F(b) - F(a) \end{aligned}$$



For en stokastisk variabel, X , med tæthedsfunktion som vist i figuren, er $P(1.50 \leq X \leq 2.25)$ det farvede areal under kurven



N.B. For kontinuerte stokastiske variable bruger vi kun **pdf** til at tegne kurven. Vi bruger **cdf** til at beregne sandsynligheder for et interval som areal under kurven.

Kontinuert stokastisk variabel

Da $P(X = x_0) = 0$ har vi:

$$\begin{aligned} P(X \leq x_0) &= P(X < x_0) + P(X = x_0) \\ &= P(X < x_0) + 0 \\ &= P(X < x_0) \end{aligned}$$

Tilsvarende:

$$\begin{aligned} P(x_1 \leq X \leq x_2) &= P(x_1 < X \leq x_2) \\ &= P(x_1 \leq X < x_2) \\ &= P(x_1 < X < x_2). \end{aligned}$$

Eksempel: Måling af en stangs trækstyrke

- En stang på 20 cm udsættes for træk, indtil den bryder.
Antag at brudstedet er lige sandsynligt i hele stangens længde
- Diskutér med hinanden:
 - Hvad er sandsynligheden for at brudet sker præcis 12 cm fra det øverste fæstningspunkt?
 - Hvad er sandsynligheden for at brudet sker mellem 5 og 7 cm fra det øverste fæstningspunkt?
 - Hvad er sandsynligheden for at brudet sker mellem 11.95 og 12.05 cm fra det øverste fæstningspunkt?



Overblik for en **diskret stokastisk variabel** X

- **Sandsynlighedsfunktion / Tæthedsfunktion (pdf):**

$$f(x) = P(X = x) \text{ for } x \in U$$

- **Kumuleret fordelingsfunktion (cdf):**

$$F(x_0) = P(X \leq x_0) = \sum_{x \leq x_0} f(x)$$

- **Middelværdi:**

$$\mu = \sum_x x \cdot f(x)$$

- **Varians:**

$$\sigma^2 = \sum_x (x^2 \cdot f(x)) - \mu^2$$

- **Standardafvigelse:**

$$\sigma = \sqrt{\sigma^2}$$

- **Den empiriske regel:**

Normalt ligger næsten alle data i intervallet

$$\mu \pm 3\sigma$$

Overblik for en **kontinuert** stokastisk variabel X

- **Tæthedsfunktion (pdf):**

$$f(x) \quad \text{N.B. } f(x) \text{ er ikke en sandsynlighed}$$

- **Kumuleret fordelingsfunktion (cdf):**

$$F(x_0) = P(X \leq x_0) = \int_{-\infty}^{x_0} f(x) dx$$

- **Middelværdi:**

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

- **Varians:**

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

- **Standardafvigelse:**

$$\sigma = \sqrt{\sigma^2}$$

- **Den empirisk regel:**

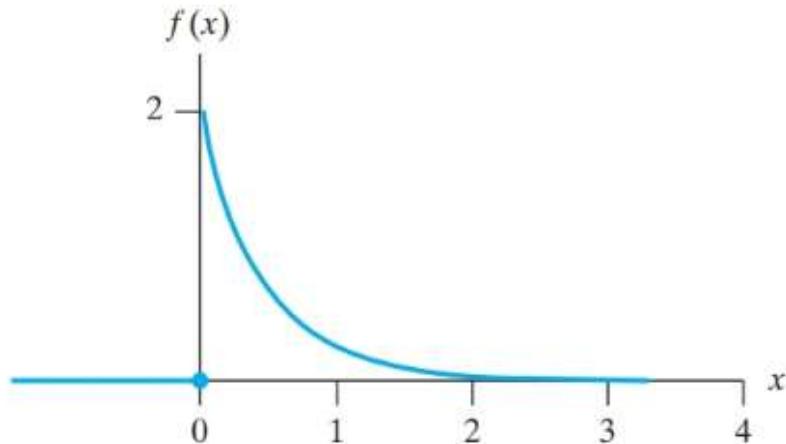
Normalt ligger næsten alle data i intervallet

$$\mu \pm 3\sigma$$

Eksempel 5.1, s. 137

En stokastisk variabel X har følgende tæthedsfunktion:

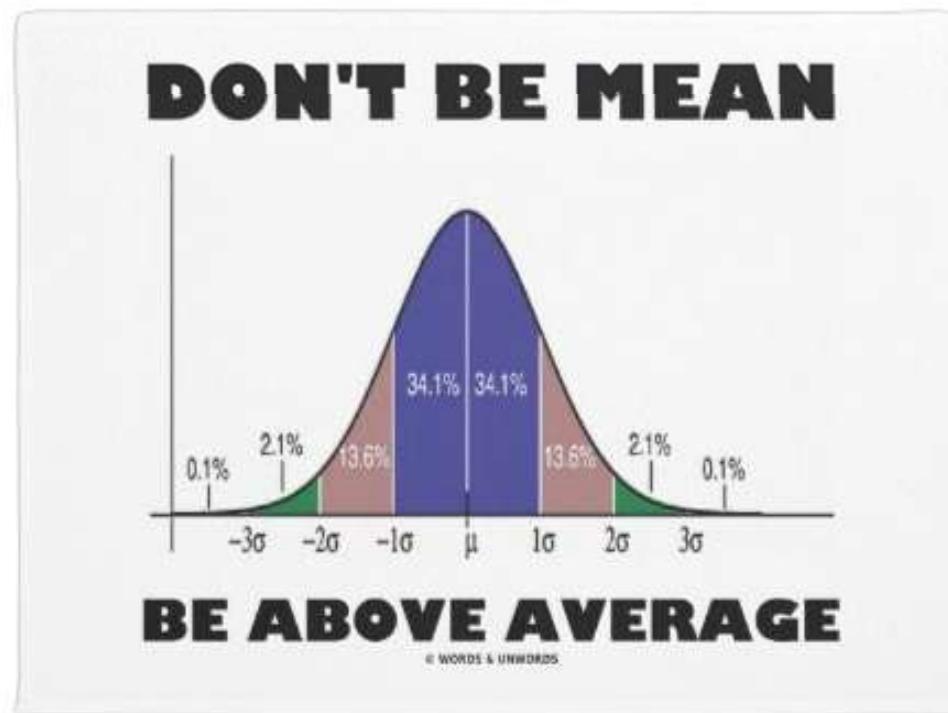
$$(f(x) = \begin{cases} 2e^{-2x} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases})$$



- a) Hvad er $P(1 < X < 3)$?
- b) Hvad er $P(X > 0.5)$?
- c) Er $(f(x))$ overhovedet en tæthedsfunktion (evt. hjemmearbejde)?

Normalfordelingen

Verdens vigtigste fordeling

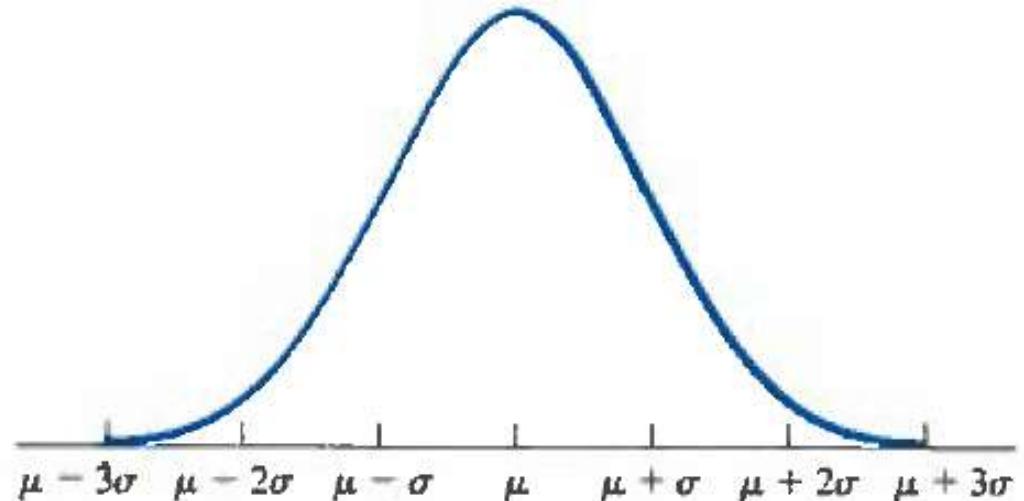


Normalfordelingen

- Egenskaber: Klokkeformet, symmetrisk omkring middelværdien
- Tæthedsfunktion pdf:

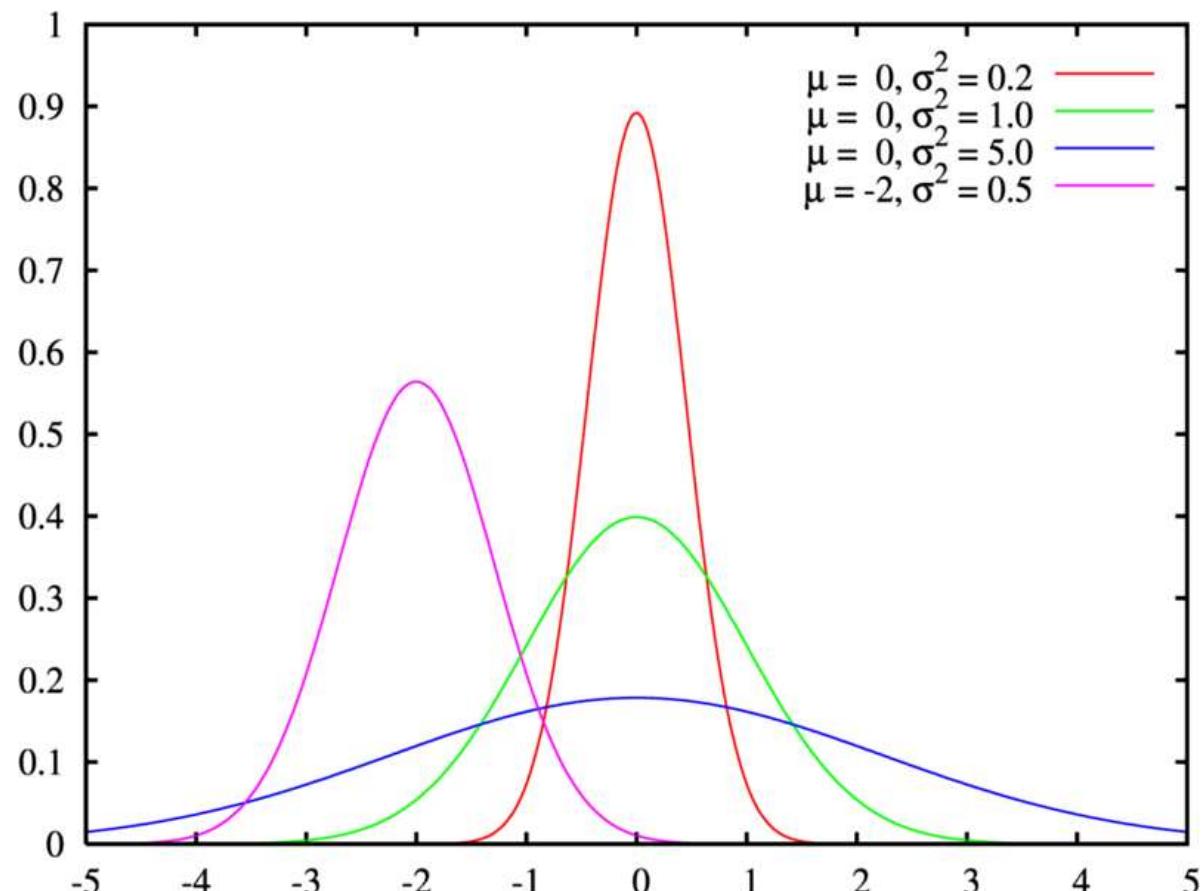
$$f(x) = f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right)$$

- Bemærk at μ og σ er parametre af tæthedsfunktionen, så de skal ikke beregnes.
Derimod afhænger klokkeformen af μ og σ .

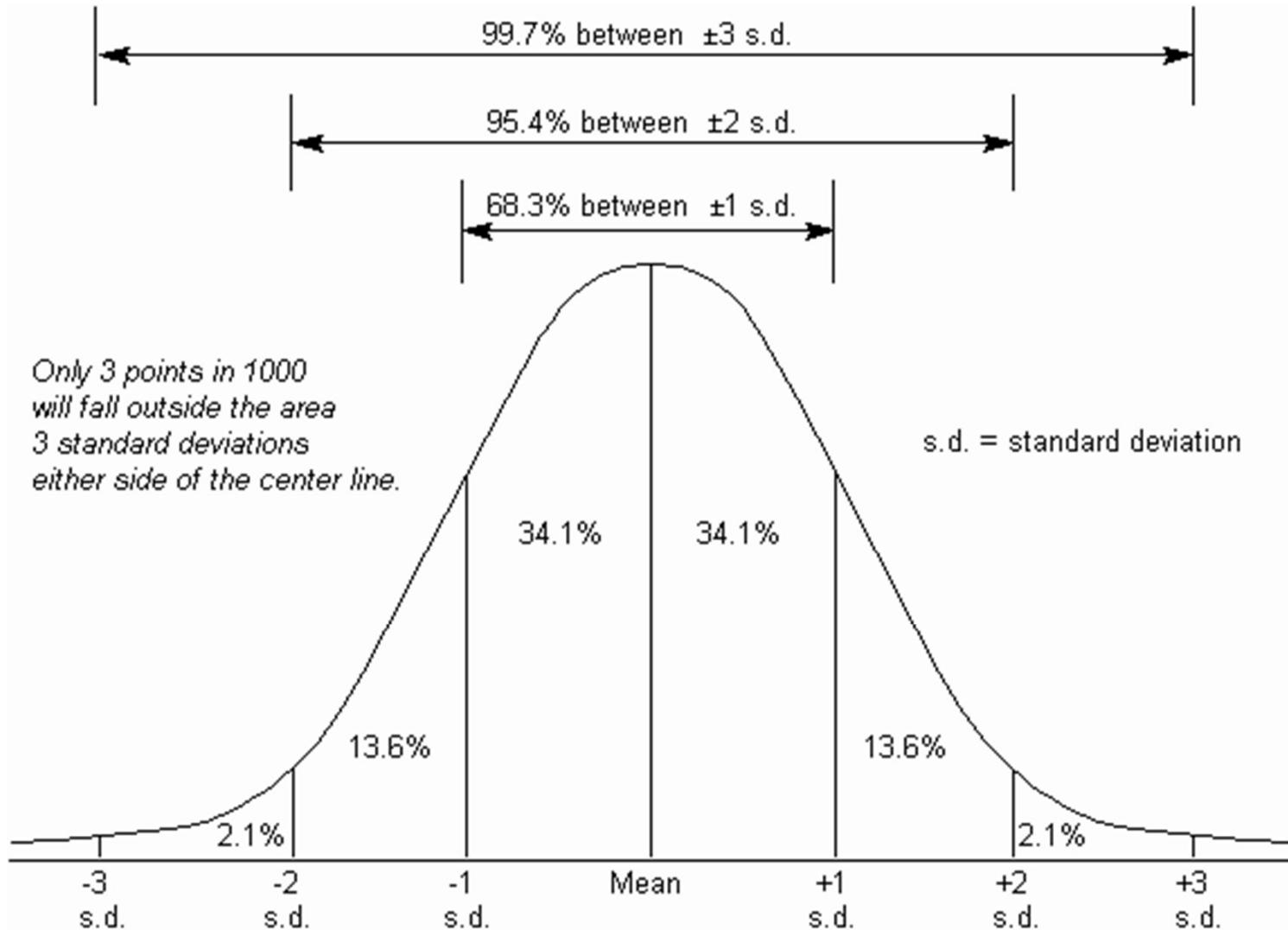


Normalfordeling med forskellige μ og σ

- Bemærk: Altid symmetrisk klokkeform omkring μ
Ændring i μ parallelforskyder kurven langs x-aksen (samme form)
Lille værdi af σ gør kurven spids og stor værdi gør den flad
- Hvis X er normalfordelt siger vi, at $X \sim N(\mu, \sigma)$.

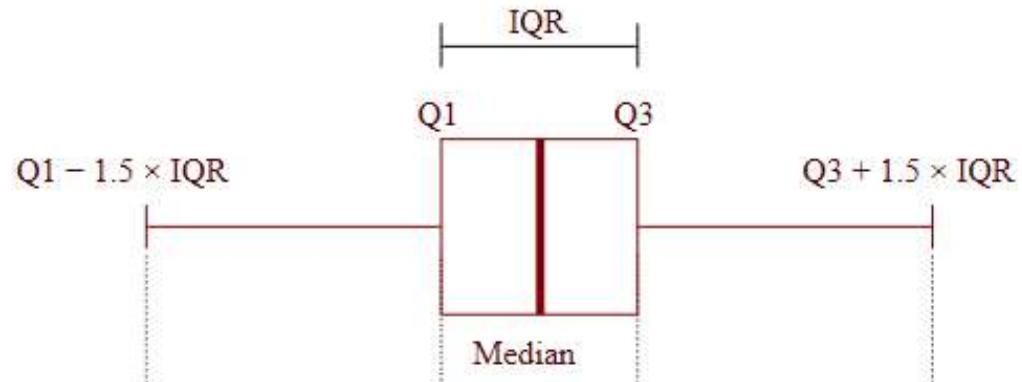


Fordeling af sandsynlighedsmassen



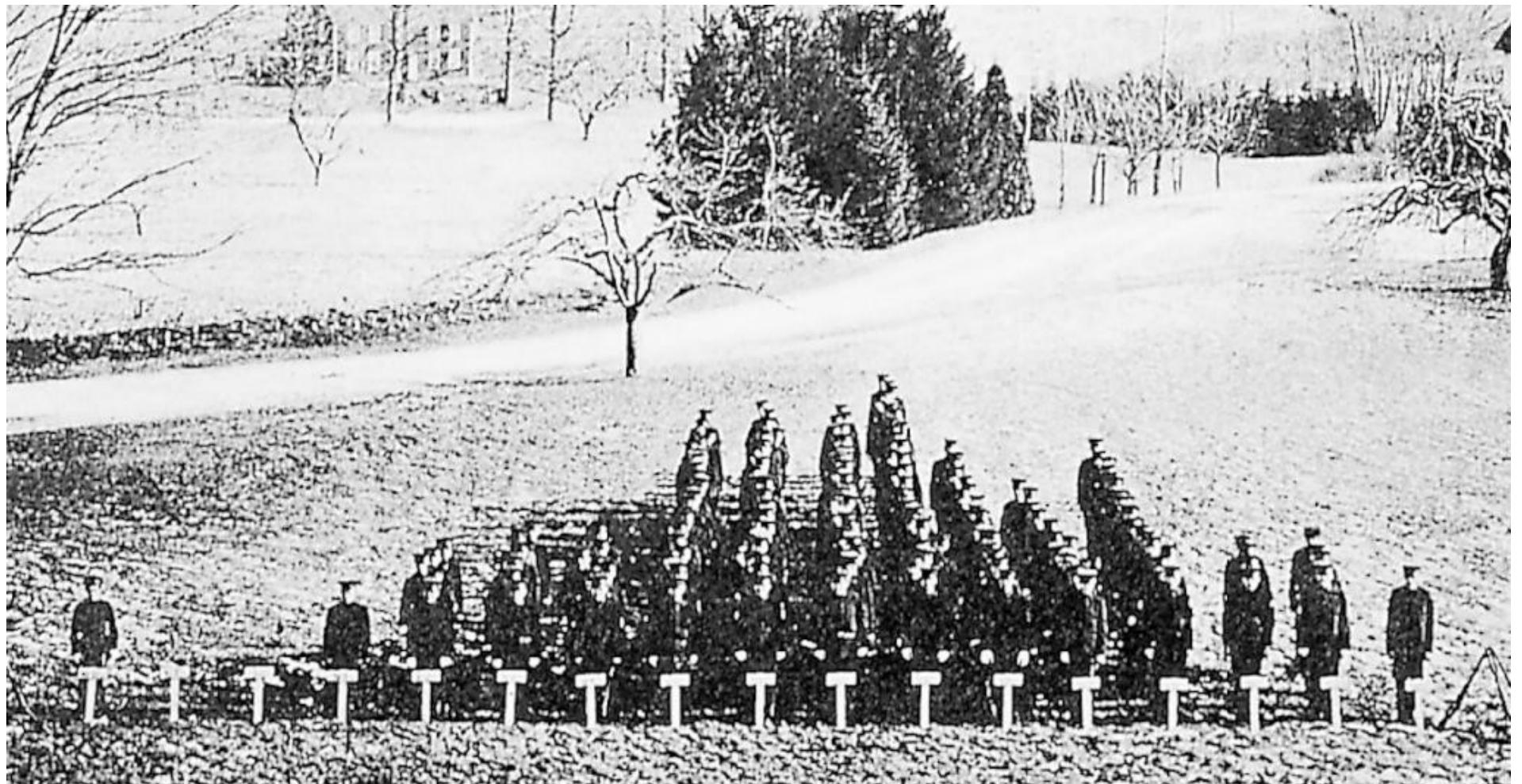
- Den empiriske regel: 99.73% af observationer ligger indenfor $\mu \pm 3\sigma$.

Normalfordeling og boksplot



Anvendelse af normalfordelingen

Normalfordelingen kan bruges til at beskrive variationen i mange fysiske og biologiske egenskaber, f.eks. menneskers højde



Anvendelse af normalfordelingen

- Gengivne målinger har måleusikkerhed. Måleusikkerheden afhænger bl.a. af måleudstyrets præcision
- Vi har målte data: x_1, x_2, \dots, x_n
- Vi kan modellere dette som
$$x_i = x_{\text{sand}} + \epsilon_i \quad \text{for } i = 1, 2, \dots, n$$
- Nu antages måleafvigelsen ϵ_i at være normalfordelt med middelværdi $\mu = 0$ og standardafvigelse σ : $\epsilon_i \sim N(0, \sigma)$. Variansens størrelse afhænger af måleudstyrets præcision
- Fun fact: Normalfordelingen blev oprindeligt kaldt ‘the normal curve of errors’, da det blev opdaget i 1700 tallet, at målefejl følger det samme mønster (en klokkeformet fordeling).

Den standardiserede normalfordeling

- En særlig variant af normalfordelingen er den med $\mu = 0$ og $\sigma = 1$. Den kaldes **standard normalfordelingen** eller **den standardiserede normalfordeling**
- Den stokastiske variabel for den standardiserede normalfordeling kaldes traditionelt Z . Vi siger: Z er $N(0,1)$
- Tæthedsfunktionen for den **generelle normalfordeling**:

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

- Tæthedsfunktionen for **standard normalfordelingen**:

$$\begin{aligned} f(z) &= \frac{1}{\sqrt{2\pi \cdot 1^2}} \exp\left(-\frac{1}{2}\left(\frac{z-0}{1}\right)^2\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \end{aligned}$$

Den kumulerede fordelingsfunktion, cdf

- Den kumulerede fordelingsfunktion (cdf) for den generelle normalfordeling er:

$$F(x_0) = \int_{-\infty}^{x_0} f(x)dx = \int_{-\infty}^{x_0} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx$$

- Der findes ikke en analytisk løsning til dette integral (heller ikke for det simplere udtryk for standard normalfordelingen), så man må bruge tabeller eller software for at løse f.eks. $P(Z \leq 1.96)$
- Tabel over den standardiserede normalfordeling Z findes bagerst i bogen.

Tabel over standard normalfordelingen Z

Table 3 Standard Normal Distribution Function

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-5.0	0.0000003									
-4.0	0.00003									
-3.5	0.0002									
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002	
-3.3	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0006	0.0003	
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005	
-3.1	0.0010	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007	
-3.0	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010	
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016					
-2.8	0.0026	0.0025	0.0024	0.0023	0.0022					
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031					
-2.6	0.0047	0.0045	0.0044	0.0043	0.0042					
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055					
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073					
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096					
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125					
-2.1	0.0179	0.0174	0.0170	0.0166	0.0161					
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207					
-1.9	0.0287	0.0281	0.0274	0.0268	0.0261					
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329					
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Tabel 3, de sidste to sider i M&F:
 Man kan slå $P(Z \leq z)$ op for
 $-3.40 \leq z \leq 3.49$
 dvs. mere end det empiriske interval

Table 3

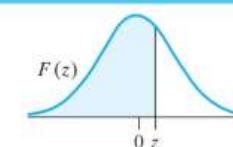
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5973	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8150	0.8186	0.8217	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621				
1.1	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830				
1.2	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015				
1.3	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177				
1.4	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319				
1.5	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441				
1.6	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545				
1.7	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633				
1.8	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706				
1.9	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767				
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998									
4.0	0.99997									
5.0	0.999997									

Tabel over standard normalfordelingen Z

- F.eks. $P(Z \leq 1.96)$
- $P(Z \leq 1.96)$
 $= P(Z \leq (1.9 + 0.06))$
 $= 0.975.$

Table 3

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-t^2/2} dt$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5973	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9705
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9705
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9705
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9705
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9705
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9705
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9705
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9705
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.991	

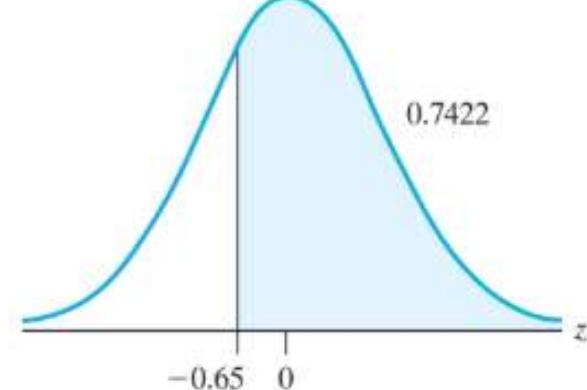
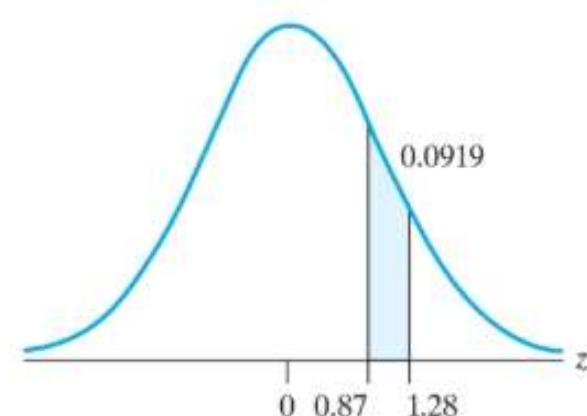
Eksempel 5.5, s. 141

Bestem sandsynligheden for, at en stokastisk variabel, der følger standard normalfordelingen vil antage en værdi i disse intervaller:

- a) Mellem 0.87 og 1.28
- b) Mellem -0.34 og 0.62
- c) Over 0.85
- d) Over -0.65

$$\begin{aligned} \text{a) } P(0.87 < Z < 1.28) \\ &= F(1.28) - F(0.87) \\ &= 0.8997 - 0.8078 \\ &= 0.0919 \end{aligned}$$

$$\begin{aligned} \text{d) } P(Z > -0.65) &= 1 - P(Z < -0.65) \\ &= 1 - F(-0.65) \\ &= 1 - 0.2578 = 0.7422. \end{aligned}$$

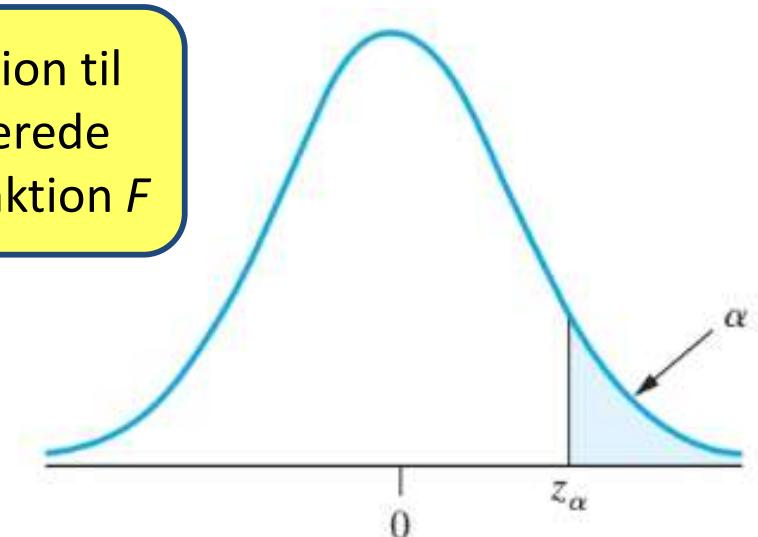


Beregning af sandsynligheder med $N(0,1)$

- Vi er ofte interesserede i at finde den værdi af Z , som giver et bestemt areal ude i halerne
- Lad os sige, at vi gerne vil finde den værdi af Z , som gør, at arealet af højre hale er α . Lad os kalde denne værdi af Z for z_α

$$\begin{aligned} P(Z > z_\alpha) &= \alpha \Rightarrow \\ 1 - P(Z \leq z_\alpha) &= \alpha \Rightarrow \\ P(Z \leq z_\alpha) &= 1 - \alpha \\ F(z_\alpha) &= 1 - \alpha \\ z_\alpha &= F^{-1}(1 - \alpha) \end{aligned}$$

Invers funktion til
den kumulerede
fordelingsfunktion F



- F.eks. for $\alpha = 0.025$:

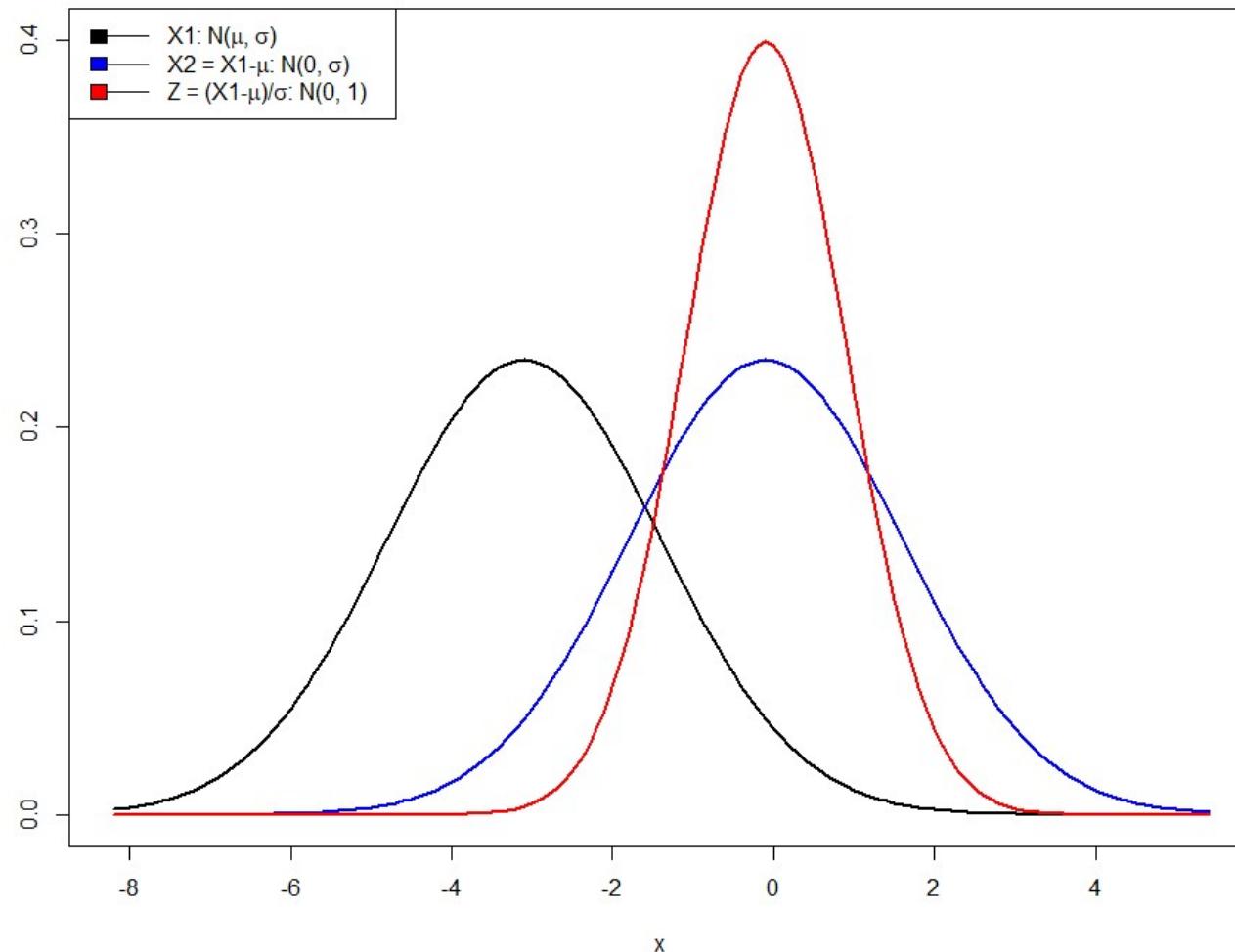
$$\begin{aligned} P(Z > z_{0.025}) &= 0.025 \Rightarrow \\ P(Z \leq z_{0.025}) &= 1 - 0.025 = 0.975 \end{aligned}$$

- Vi har tidligere vist, at $F(1.96) = P(Z \leq 1.96) = 0.975$, så $z_{0.025} = F^{-1}(1 - 0.025) = F^{-1}(0.975) = 1.96$.

Hvad så hvis X følger *generel* normalford.?

- Hvis $X \sim N(\mu, \sigma)$, så er $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

Transformering fra X til Z , hvor X er $N(-3.1, 1.7)$



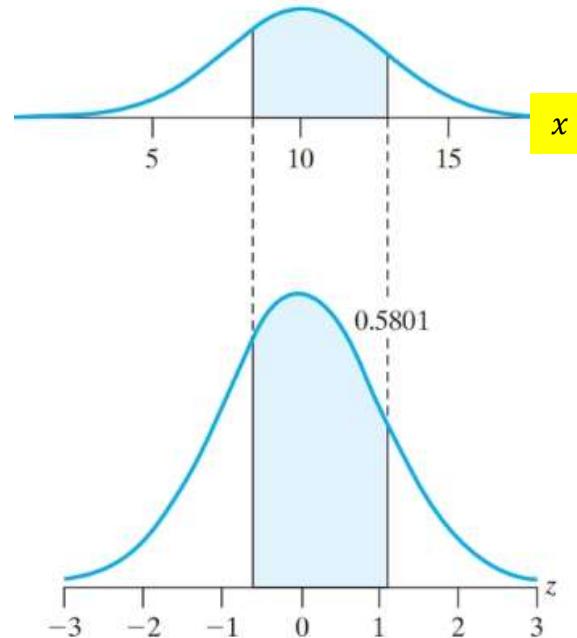
Eksempel 5.7, s. 144

For en stregkodeskanner mÅler man, hvor meget man kan dæmpe signalet, sÅ den lige netop kan aflæse en stregkode. Dæmpningen afhænger af stregkoden for forskellige fødevarer. Det viser sig, at den følger en normalfordeling med $\mu = 10.1$ dB og $\sigma = 2.7$ dB.

- a) Hvad er sandsynligheden for at dæmpningen for den næste stregkode er mellem 8.5 og 13.0 dB

Svar: $X \sim N(10.1, 2.7)$,
sÅ $Z = \frac{X-10.1}{2.7} \sim N(0,1)$

$$\begin{aligned}P(8.5 < X < 13.0) &= F\left(\frac{13.0-10.1}{2.7}\right) - F\left(\frac{8.5-10.1}{2.7}\right) \\&= F(1.07) - F(-0.59) \\&= 0.8577 - 0.2776 \\&= 0.5801\end{aligned}$$



N.B.: De to blå arealer under kurverne er ens, nemlig 0.5801.

Beregning med R

Standard normalfordeling $N(0, 1)$

- PDF: $\text{dnorm}(z)$ f.eks. $f(1.96) = \text{dnorm}(1.96) = 0.0584$
- CDF: $\text{pnorm}(z)$ f.eks. $F(1.96) = \text{pnorm}(1.96) = 0.975$
- Invers: $\text{qnorm}(p)$ f.eks. $F^{-1}(0.975) = \text{qnorm}(0.975) = 1.96$

Generel normalfordeling $N(\mu, \sigma^2)$

- PDF: $\text{dnorm}(x, \text{mu}, \text{sigma})$
- CDF: $\text{pnorm}(x, \text{mu}, \text{sigma})$
- Invers: $\text{qnorm}(p, \text{mu}, \text{sigma})$

Tilfældighedsgenerator (funktion der giver n tilfældige, normalfordede tal)

- Random: $\text{rnorm}(n, \text{mu}, \text{sigma}).$

Opgave

Opgave (normalfordeling)

En dansk virksomhed er specialiseret i produktion af fjedre af mange forskellige typer og specifikationer. For en bestemt type galvaniseret trækfjeder oplyser virksomheden, at fjedre af denne type har en forventet fjederkonstant på 4.0 N/mm . Mere præcist oplyser virksomheden, at fjederkonstanten for fjedrene er normalfordelt med middelværdi 4.0 N/mm og standardafvigelse 0.1 N/mm .



Beregn følgende under forudsætning af, at virksomhedens oplysninger er korrekte:

- a. Beregn sandsynligheden for at en tilfældig fjeder af den omtalte type har en fjederkonstant på præcis 4.0 N/mm .
- b. Beregn sandsynligheden for at fjederens fjederkonstant er under 3.7 N/mm .
- c. Beregn sandsynligheden for at fjederens fjederkonstant er større end 4.1 N/mm .
- d. Beregn et interval omkring middelværdien, som 99% af fjedrene vil tilhøre.

Gammafordelingen (eksponentiafford.)

- En generisk fordeling, der har flere fordelinger som specialtilfælde
- Tæthedsfunktion:

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

hvor gammafunktionen $\Gamma(\alpha)$ er defineret som

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x/\beta} dx = (\alpha - 1)! \quad (\text{kan man vise})$$

- Gammafordelingen har:
 - Middelværdi $\mu = \alpha \beta$
 - Varians $\sigma^2 = \alpha \beta^2$
- **Eksponentiaffordelingen** er Gammaford. med $\alpha = 1$ (og $\lambda = 1/\beta$):
$$f(x) = \begin{cases} \frac{1}{\beta} e^{-x/\beta} & \text{for } x > 0, \beta > 0 \\ 0 & \text{ellers} \end{cases}$$
- Eksempel 5.1 fra tidligere er eksponentiaffordelt med $\lambda = 2$.

Eksponentiafordelingen

- Eksponentiafordelingen (*exponential distribution*) beskriver f.eks. tiden imellem begivenheder, der indtræffer tilfældigt, så som ulykker, orkaner, epidemier, maskiners levetid eller radioaktivt henfald
- Eksponentiafordelingen kan opfattes som en kontinuert version af Poisson-fordelingen. Poisson: Sands. for 7 fejl på 1 dag (diskret). Eksponential: Sands. for mere end 10 timer før næste fejl (kontinuert)
- Tæthedsfunktion pdf:

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

- λ er gennemsnitlig antal hændelser per tidsenhed (f.eks. antal orkaner per år, antal maskinnedbrud per dag, antal pandemier pr 100 år)
- $\mu = 1/\lambda$ er gennemsnitlig tid imellem hændelserne.

Eksponentialfordelingen

- Den kumulerede fordelingsfunktion cdf:

$$\begin{aligned}F(x_0) &= P(X \leq x_0) \\&= \int_{-\infty}^{x_0} f(x) dx \\&= \int_0^{x_0} \lambda e^{-\lambda x} dx \\&= [-e^{-\lambda x}]_0^{x_0} \\&= -e^{-\lambda x_0} - (-e^{-\lambda \cdot 0}) \\&= 1 - e^{-\lambda x_0}\end{aligned}$$

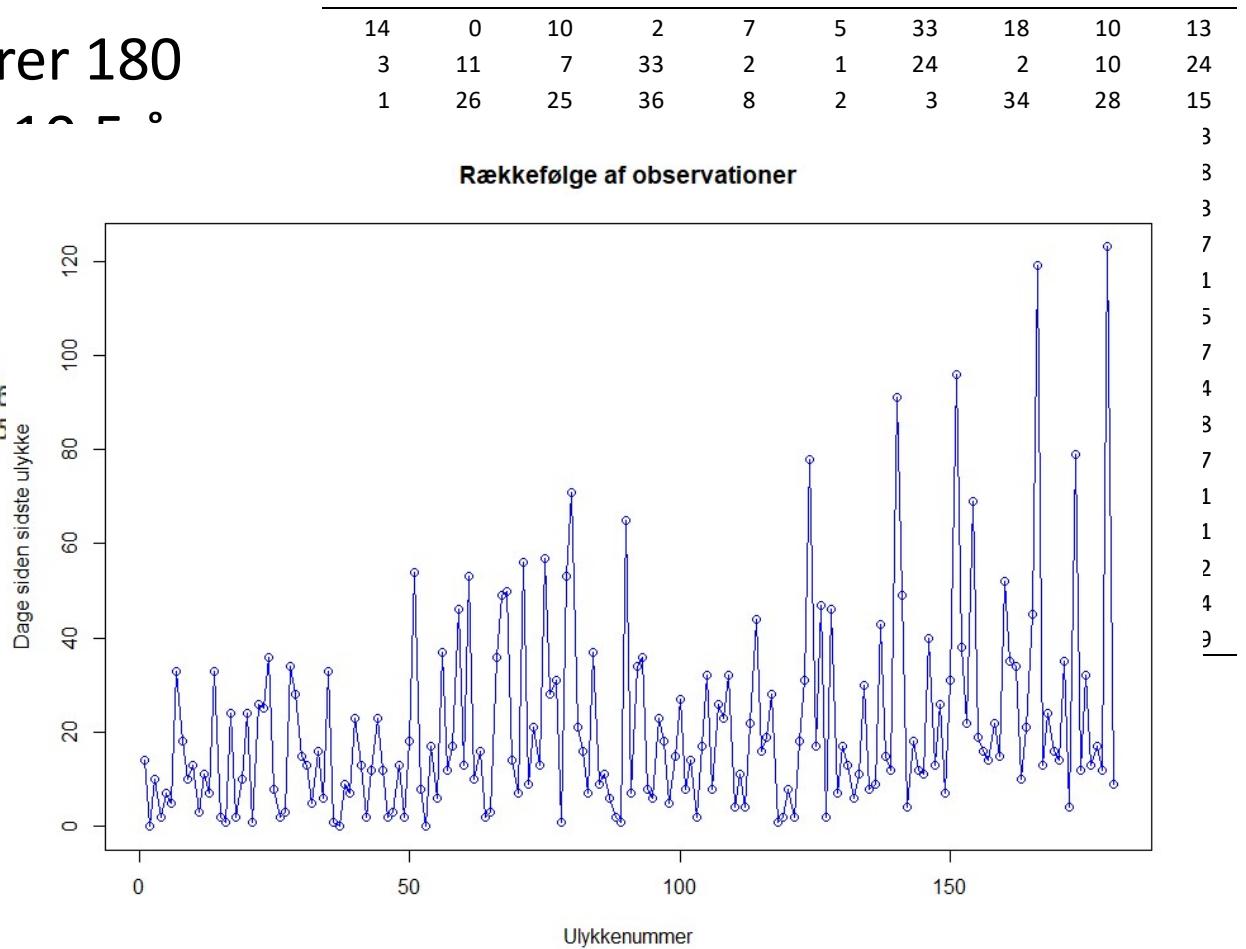
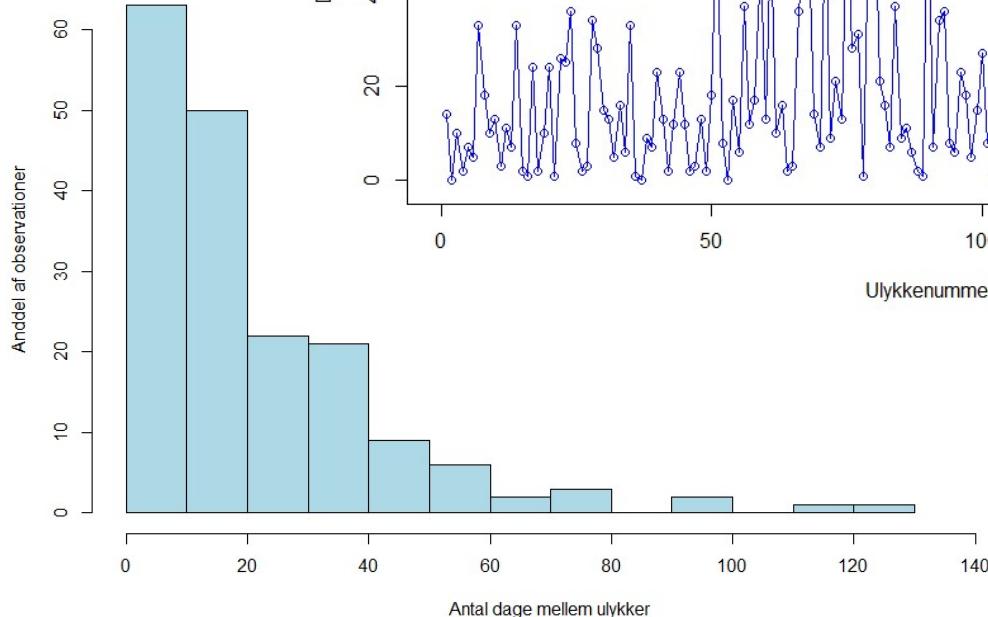
- Middelværdi og standardafvigelse:

- $\mu = \frac{1}{\lambda}$
- $\sigma = \frac{1}{\lambda}$.

Eksempel – tid mellem arbejdsulykker

- En virksomhed registrerer 180 arbejdsulykker over ca. 100 dage
- I snit går der 21.3 dage mellem ulykker

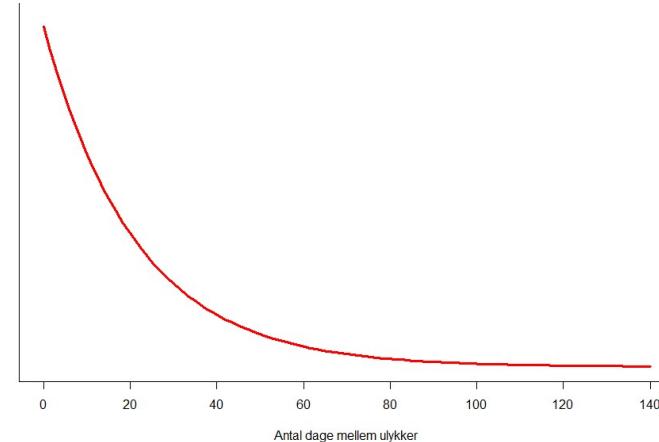
0 | 0001111112222222223333444455566
1 | 00000111112222223333333344444555
2 | 111222333344456667888
3 | 0111222333444556667778
4 | 034566799
5 | 0233467
6 | 59
7 | 189
8 |
9 | 16
10 |
11 | 9
12 | 3



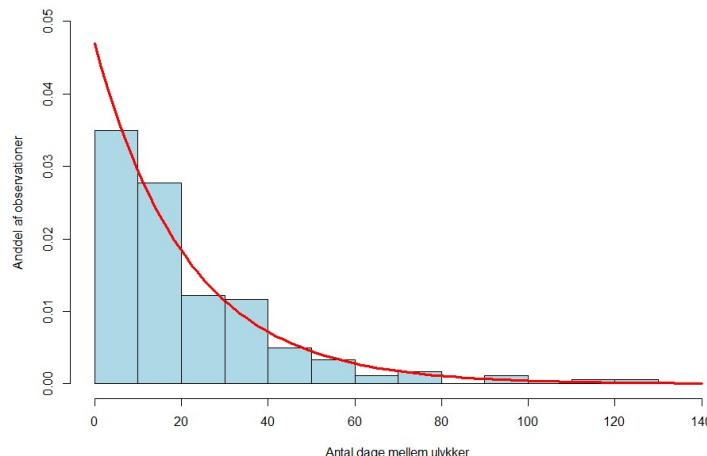
Eksempel – tid mellem arbejdsulykker

- Dette kan modelleres med eksponentialfordelingen.
 $\mu = 21.3$ dage pr. ulykke, så
 $\lambda = 1/\mu = 1/21.3 = 0.047$ ulykker pr. dag. Derfor:

$$F(x) = 1 - e^{-\lambda x} = 1 - e^{-0.047x}$$



- Tæthedsfunktionen for eksponentialfordelingen med $\lambda = 0.047$
- Sammenligning mellem de målte data og den teoretiske fordeling.



Eksponentialfordelingen i R

- **dexp(x, lambda)**
Tæthedsfunktionen (kan bruges til at tegne fordelingen)
- **pexp(x, lambda)**
Kumuleret fordelingsfunktion (bruges til at beregne sandsynligheder)
F.eks. $pexp(20, 0.047) = 0.609$ giver sandsynligheden for, at der går mindre end 20 dage til næste arbejdsulykke
- **qexp(p, lambda)**
Den inverse funktion til fordelingsfunktionen (bruges til at beregne den værdi, der svarer til en given sandsynlighed)
F.eks. $qexp(0.95, 0.047) = 63.7$ giver at 95 % af ulykker sker med mellemrum under 63.7 dage
- **rexp(n, lambda)**
Funktion til generering af n tilfældige tal fra eksponentialfordelingen (bruges til at simulere en tilfældig stikprøvetagning, f.eks. ulykker).

Simultane fordelinger (*joint distributions*)

- Ofte ser man på problemstillinger, hvor der er flere stokastiske variable samtidig
- Vi ser først på *diskrete* stokastiske variable
- For eksempel fra kap. 3:

- Terning 1: $X_1 \sim U(1,6)$
- Terning 2: $X_2 \sim U(1,6)$
- $P((X_1 = x_1) \cap (X_2 = x_2))$
 $= P(X_1 = x_1, X_2 = x_2)$
 $= f(x_1, x_2)$
 $= \frac{1}{36}$

for $x_1, x_2 = 1, \dots, 6$

- Vi kalder $f(x_1, x_2)$ for den simultane fordeling for X_1 og X_2
- Ud fra den simultane fordeling kan vi beregne f.eks.

$$P(X_1 + X_2 = 5) = \frac{4}{36} = \frac{1}{9}.$$

Tilfældige eksperimenter

Antal øjne med to terninger

- $U = \{ 2, 3, 4, \dots, 12 \}$
- Hvad er sandsynligheden for 5?
- Vi antager at udfaldet af hver terning er tilfældigt og uafhængigt af hinanden
- Der er $6 \cdot 6 = 36$ grundlæggende udfald med lige sandsynlighed
- 4 af dem har summen 5, så
 $P(5) = \frac{4}{36} = \frac{1}{9}$.



		Terning 2					
		1	2	3	4	5	6
Terning 1	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Simultane fordelinger (*joint distributions*)

- Nyt spil med 2 terninger:
 - Terning 1: $X_1 \sim U(1,4)$
 - Terning 2: $X_2 \sim U(1,6)$
 - $P((X_1 = x_1) \cap (X_2 = x_2))$
 $= P(X_1 = x_1, X_2 = x_2)$
 $= f(x_1, x_2) = \frac{1}{24}$
for $x_1 = 1, \dots, 4$, $x_2 = 1, \dots, 6$
 - Vi ser stadig på summen af de to terningers øjne, men i dette spil tæller terning 1's øjne dobbelt:
 $2X_1 + X_2$
 - Hvad er sandsynligheden for at score over 10 i spillet?
 - $P(2X_1 + X_2 > 10) = \frac{6}{24} = \frac{1}{4}$.



		X ₁ : Terning 1			
		1	2	3	4
X ₂ : Terning 2	1	3	5	7	9
	2	4	6	8	10
	3	5	7	9	11
	4	6	8	10	12
	5	7	9	11	13
	6	8	10	12	14

Simultan fordeling for diskrete variable

- To diskrete stokastiske variable, X_1 og X_2
- Vi kalder funktionen $f(x_1, x_2) = P((X_1 = x_1) \cap (X_2 = x_2)) = P(X_1 = x_1, X_2 = x_2)$ for den *simultane fordeling* for X_1 og X_2

- Eksempel 5.20. s. 162:**

- Find $P(X_1 + X_2 > 1)$
- Find tæthedsfunktionen f_1 for X_1 :
 $f_1(x_1) = P(X_1 = x_1)$ for $x_1 = 0, 1, 2$

- Svar:**

- $X_1 + X_2 > 1$ gælder for kombinationerne (1,1), (2,0) og (2,1). Ifølge den additive lov:

$$P(X_1 + X_2 > 1) = f(1,1) + f(2,0) + f(2,1) = 0.2 + 0.1 + 0 = 0.3$$

- Den *marginale* fordeling for X_1 findes ved at addere sandsynlighederne for X_2 ud:

$$f_1(x_1) = \sum_{x_2} f(x_1, x_2)$$

		Joint Probability Distribution $f(x_1, x_2)$ of X_1 and X_2		
		0	1	2
x_2	0	0.1	0.4	0.1
	1	0.2	0.2	0

		0	1	2	Total $f_2(x_2)$
x_2	0	0.1	0.4	0.1	0.6
	1	0.2	0.2	0	0.4
Total	$f_1(x_1)$	0.3	0.6	0.1	1.0

Simultan fordeling for diskrete variable

- **Betinget sandsynlighed:** $P(X_1 = x_1 | X_2 = x_2) = f_1(x_1 | x_2)$:

$$f_1(x_1 | x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}, \text{ hvis } f_2(x_2) \neq 0$$

- F.eks. $P(X_1 = 1 | X_2 = 0)$

$$\begin{aligned} &= f_1(1|0) = \frac{f(1,0)}{f_2(0)} \\ &= \frac{0.4}{0.6} = \frac{2}{3} \end{aligned}$$

		0	1	2	Total $f_2(x_2)$
x_2	0	0.1	0.4	0.1	0.6
	1	0.2	0.2	0	0.4
Total	$f_1(x_1)$	0.3	0.6	0.1	1.0

- **Uafhængighed:**

X_1 og X_2 er uafhængige

$$\Leftrightarrow f_1(x_1 | x_2) = f_1(x_1) \text{ for alle } x_1, x_2$$

$$\Leftrightarrow f(x_1, x_2) = f_1(x_1)f_2(x_2) \text{ for alle } x_1, x_2$$

- Eksempel 5.20: X_1 og X_2 er ikke uafhængige, for (f.eks.):

$$f_1(1 | 0) = \frac{2}{3} \neq 0.6 = f_1(1)$$

Tilsvarende: $f(1,0) = 0.4 \neq 0.36 = (0.6) \cdot (0.6) = f_1(1)f_2(0)$.

Simultan fordeling for kontinuerte variable

- To kontinuerte stokastiske variable, X_1 og X_2
Den simultane tæthedsfunktion for X_1 og X_2 , $f(x_1, x_2)$, bruges til at beregne sandsynligheder:

$$P(a_1 < X_1 < b_1, a_2 < X_2 < b_2) = \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2) dx_1 dx_2$$

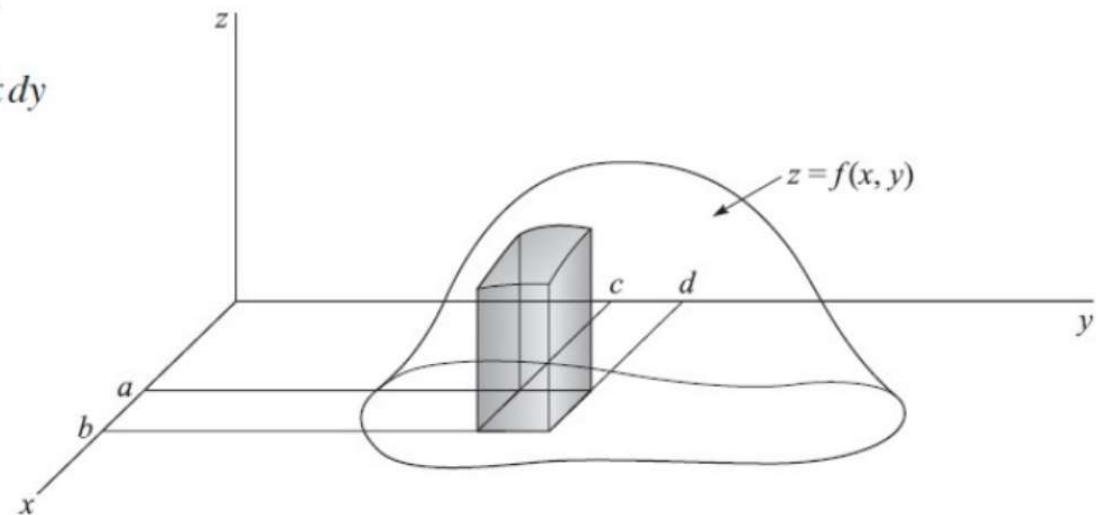
- Illustration fra SlideShare (med lidt anderledes notation):

Probability Surface

$$P(a < X < b, c < Y < d) = \int_{x=a}^b \int_{y=c}^d f(x, y) dx dy$$

Probability Function

$$P(A) = \iint_{\mathcal{R}_A} f(x, y) dx dy$$



Simultan fordeling for kontinuerte variable

- To kontinuerte stokastiske variable, X_1 og X_2
Den simultane tæthedsfunktion for X_1 og X_2 , $f(x_1, x_2)$, bruges til at beregne sandsynligheder:

$$P(a_1 < X_1 < b_1, a_2 < X_2 < b_2) = \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2) dx_1 dx_2$$

- For at være en simultan tæthedsfunktion skal $f(x_1, x_2)$ opfylde:
 - $f(x_1, x_2) \geq 0$
 - $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 dx_2 = 1$
- Den **simultane, kumulative fordelingsfunktion**:
$$F(x_1, x_2) = P(X_1 < x_1, X_2 < x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(x_1, x_2) dx_1 dx_2$$
- De **marginale tæthedsfunktioner** for X_1 og X_2 :

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2$$

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 .$$

Simultan fordeling for kontinuerte variable

- **Uafhængighed:**

X_1 og X_2 er uafhængige

$$\Leftrightarrow F(x_1, x_2) = F_1(x_1)F_2(x_2) \text{ for alle } x_1, x_2$$

$$\Leftrightarrow f(x_1, x_2) = f_1(x_1)f_2(x_2) \text{ for alle } x_1, x_2 \text{ (kan man vise)}$$

- **Betinget sandsynlighed:**

$$f_1(x_1 | x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}, \text{ hvis } f_2(x_2) \neq 0$$

Linearkombination af stokastiske variable

- Lad X_1 og X_2 være uafhængige stokastiske variable med middelværdi og varians hhv.
 - Middelværdi: $E(X_1) = \mu_{X_1}$ og $E(X_2) = \mu_{X_2}$
 - Varians: $Var(X_1) = \sigma_{X_1}^2$ og $Var(X_2) = \sigma_{X_2}^2$
- Vi ser på den stokastiske variabel Y , der er en linearkombination af X_1 og X_2 :
$$Y = a_1 X_1 + a_2 X_2 \quad \text{for konstante værdier } a_1 \text{ og } a_2$$
- Y har middelværdi og varians:
 - Middelværdi: $E(Y) = \mu_Y = a_1 E(X_1) + a_2 E(X_2)$
 - Varians: $Var(Y) = (a_1)^2 Var(X_1) + (a_2)^2 Var(X_2)$
- Hvis vi har
$$Y = a_1 X_1 + a_2 X_2 + a_3 \quad \text{for konstante værdier } a_1, a_2 \text{ og } a_3$$
så gælder:
 - Middelværdi: $E(Y) = \mu_Y = a_1 E(X_1) + a_2 E(X_2) + a_3$
 - Varians: $Var(Y) = (a_1)^2 Var(X_1) + (a_2)^2 Var(X_2).$

Eksempel 5.31 s. 171

Vi har to stokastiske variable:

- X_1 har middelværdi 4 og varians 9
- X_2 har middelværdi -2 og varians 6
- Bestem middelværdi og varians for $Y = 2X_1 + X_2 - 5$

Løsning:

- $$\begin{aligned} E(Y) &= E(2X_1 + X_2 - 5) = 2E(X_1) + E(X_2) - 5 \\ &= 2 \cdot 4 - 2 - 5 = 1 \end{aligned}$$
- $$\begin{aligned} Var(Y) &= Var(2X_1 + X_2 - 5) = 2^2 \cdot Var(X_1) + 1^2 \cdot Var(X_2) \\ &= 2^2 \cdot 9 + 6 = 42. \end{aligned}$$

Eksempel 5.28 s. 168

Vi har en stokastisk variabel X med middelværdi μ og standardafvigelse σ

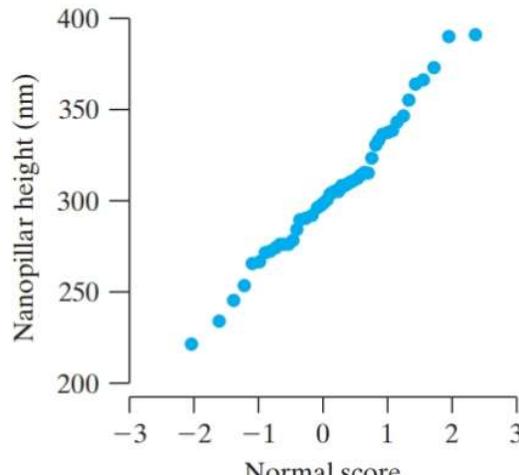
- Bestem middelværdi og standardafvigelse for $Z = \frac{X-\mu}{\sigma}$

Løsning:

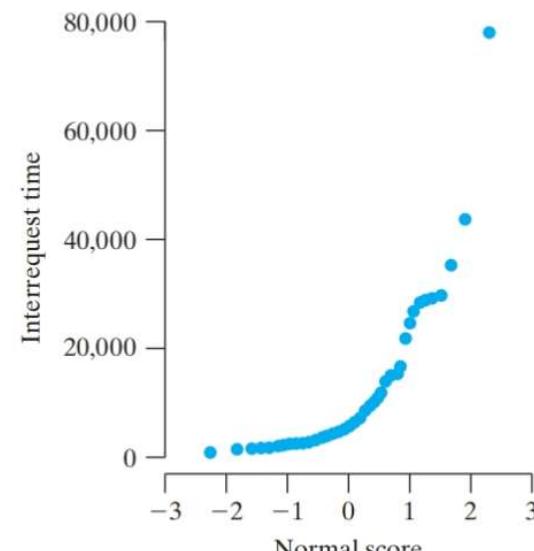
- $Z = \frac{X-\mu}{\sigma} = \frac{1}{\sigma}X - \frac{\mu}{\sigma}$, så
 $E(Z) = E\left(\frac{1}{\sigma}X - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma}E(X) - \frac{\mu}{\sigma} = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0$
- $Var(Z) = Var\left(\frac{1}{\sigma}X - \frac{\mu}{\sigma}\right) = \left(\frac{1}{\sigma}\right)^2 \cdot Var(X) = \frac{\sigma^2}{\sigma^2} = 1$
- Dette gælder uanset fordelingen af X , men
hvis X er **normalfordelt**, $X \sim N(\mu, \sigma)$, så er $Z = \frac{X-\mu}{\sigma}$ **standard normalfordelt**, $Z \sim N(0,1)$.

Er data normalfordelte?

- Senere i kurset kommer vi til at antage, at data er normalfordelte, eller kommer fra en fordeling, der ligner normalfordelingen (en ‘pæn’ fordeling)
- Vi kan teste antagelsen ved at lave et stem-and-leaf plot eller et histogram over data og se, om det ligner normalfordelingen
- Vi kan desuden teste antagelsen med et **normalfordelingsplot**. Hvis data er normalfordelte vil de ligge *nogenlunde* på en ret linje i normalfordelingsplottet



Normalfordelte data



Ikke normalfordelte data.

Er data normalfordelte?

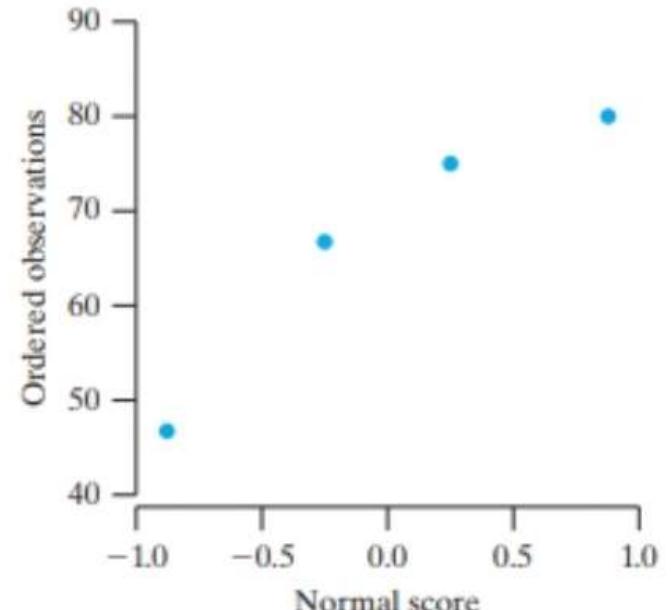
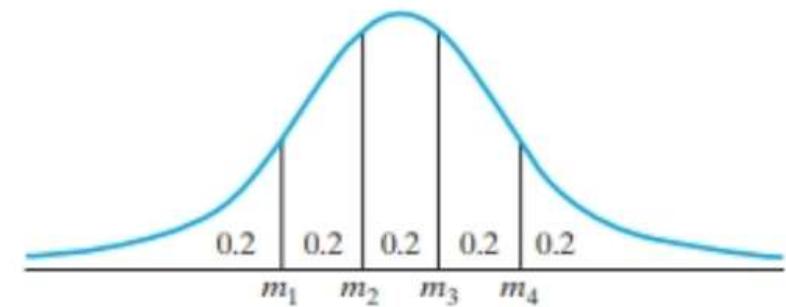
- Et normalfordelingsplot er principielt sammenligneligt med logaritmepapir
- På enkeltlogaritmisk papir er skalaen på y -aksen lavet, så den eksponentielle funktion $y = a \cdot b^x$ bliver vist lineært

No
Image

- I et normalfordelingsplot er skalaen lavet, så normalfordelte data bliver vist nogenlunde lineært (kun nogenlunde pga. tilfældig støj).

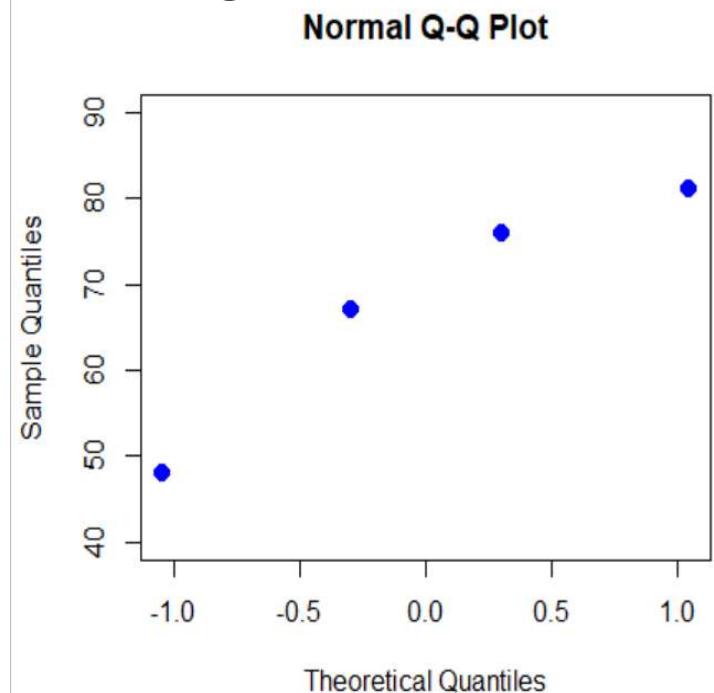
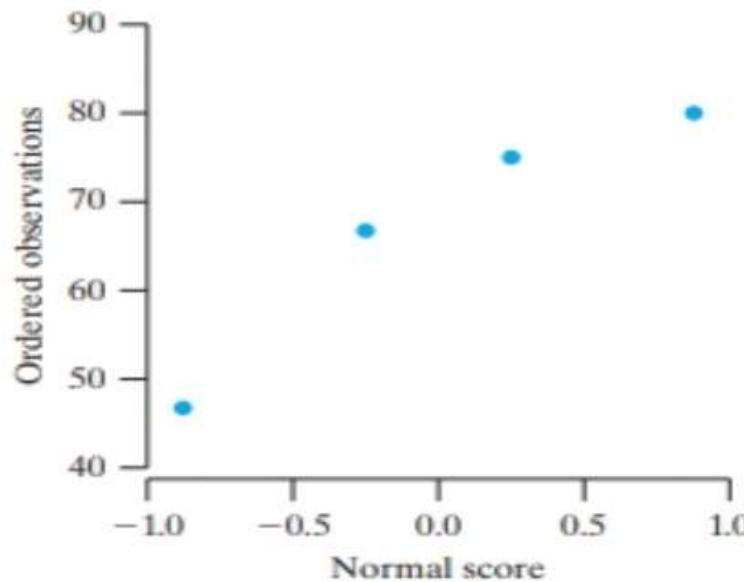
Er data normalfordelte?

- Eksempel: $n = 4$ observationer: 67, 48, 76, 81. Kommer de fra en normalfordeling? (Obs. det er et *forsimpleret* eksempel, normalt har vi brug for større antal observationer, mindst 15-20 observationer)
- Bogens metode:
 1. Sortér data efter størrelse: 48, 67, 76, 81
 2. Beregn *normalscorer* for $n = 4$:
$$m_1 = -z_{0.20} = -0.84$$
$$m_2 = -z_{0.40} = -0.25$$
$$m_3 = z_{0.40} = 0.25$$
$$m_4 = z_{0.20} = 0.84$$
 3. Plot den i-te observation mod den i-te normalscore for $i = 1, \dots, n$
- Her ser data ikke ud til at være normalfordelte, for de ligger ikke på en ret linje.



Normalfordelingsplot i R

- De fleste statistikprogrammer har funktioner til at lave normalfordelingsplots, men de benytter ofte lidt forskellige metoder
- I R findes funktionen `qqnorm()`.
(Normalfordelingsplot kaldes også ‘normal quantile-quantile plot’ eller ‘normal Q-Q plot’ eller bare ‘QQ plot’. NB: quantiles = fraktiler fra K2)
- Bemærk at `qqnorm` ikke benytter normalscorer som bogen, så y-akserne er ens, men x-akserne er lidt forskellige.



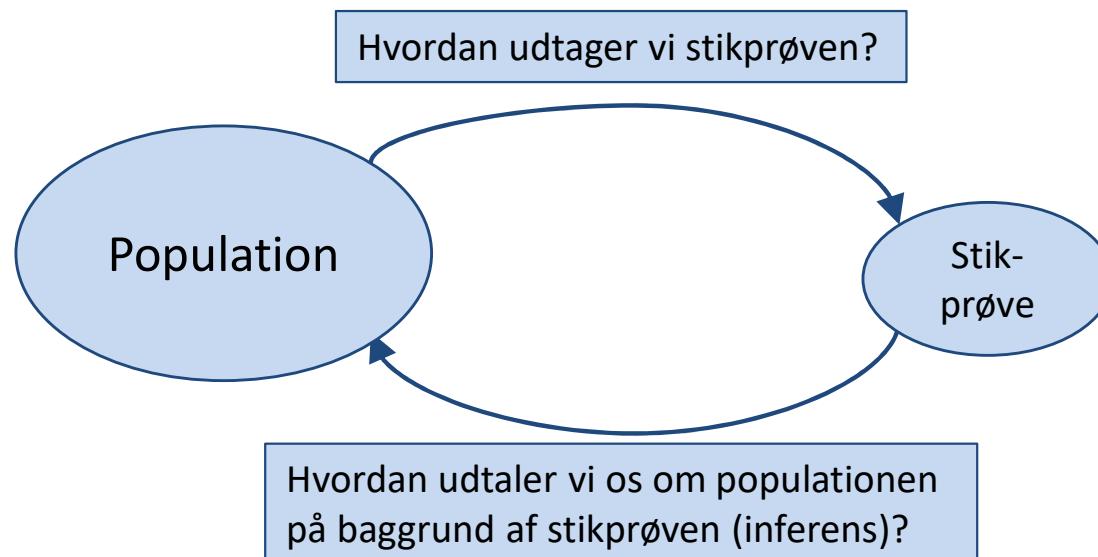
Sandsynlighedsteori og statistik

Kapitel 6. Stikprøver og deres fordelinger (afsnit 6.1-6.4)

Allan Leck Jensen
alj@ece.au.dk

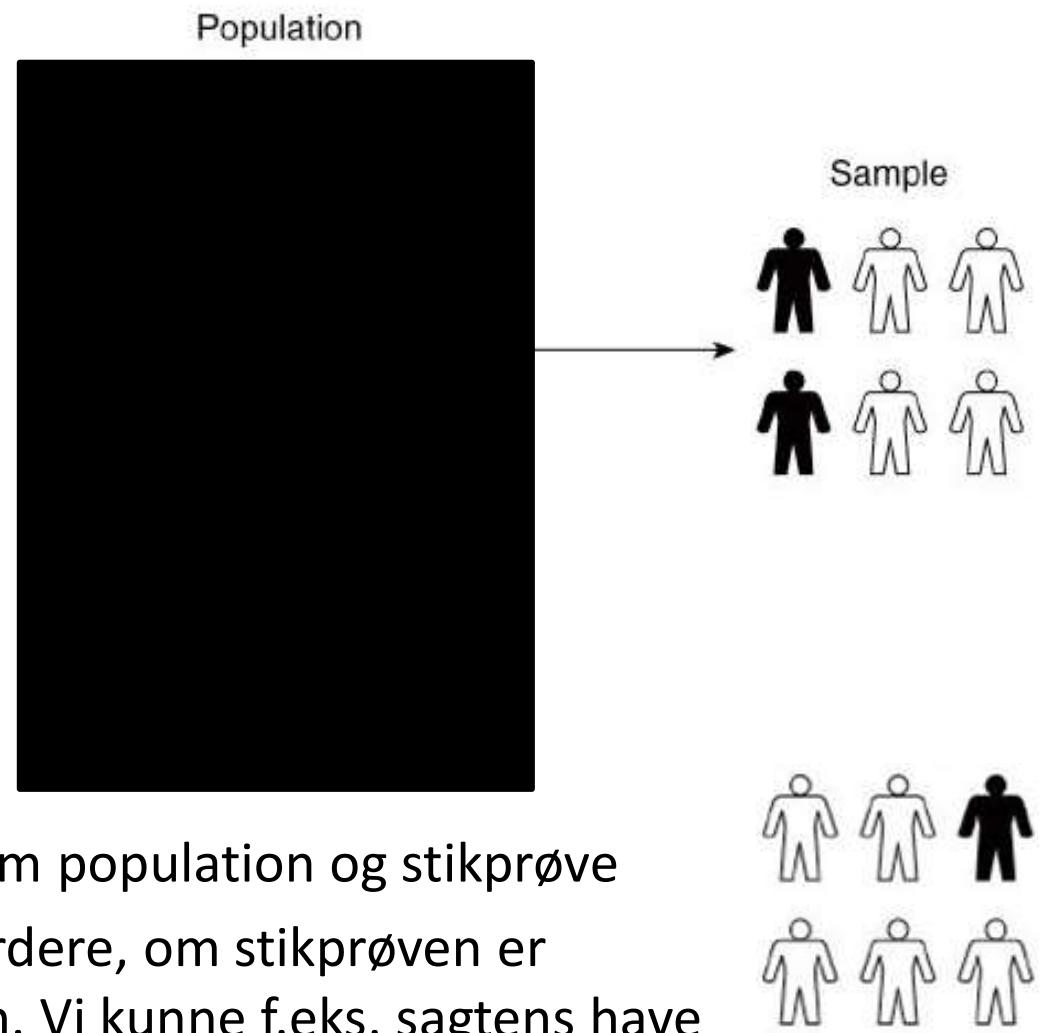
Population og stikprøve

- Hvad er gennemsnitshøjden af danske ingenørstuderende?
- Hvordan ville mandatfordelingen i folketinget se ud, hvis der var folketingsvalg i morgen?
- Hvad kan jeg forvente, at brudstyrken af min stålbjælke er?
- Er torsken ved at uddø i Vesterhavet?
- En ‘population’ behøver ikke være en samling mennesker. Den kan være uendelig, f.eks. sandsynligheden for plat med en given mønt
- Hvordan kan vi udtales os om ‘populationen’?



Population og stikprøve

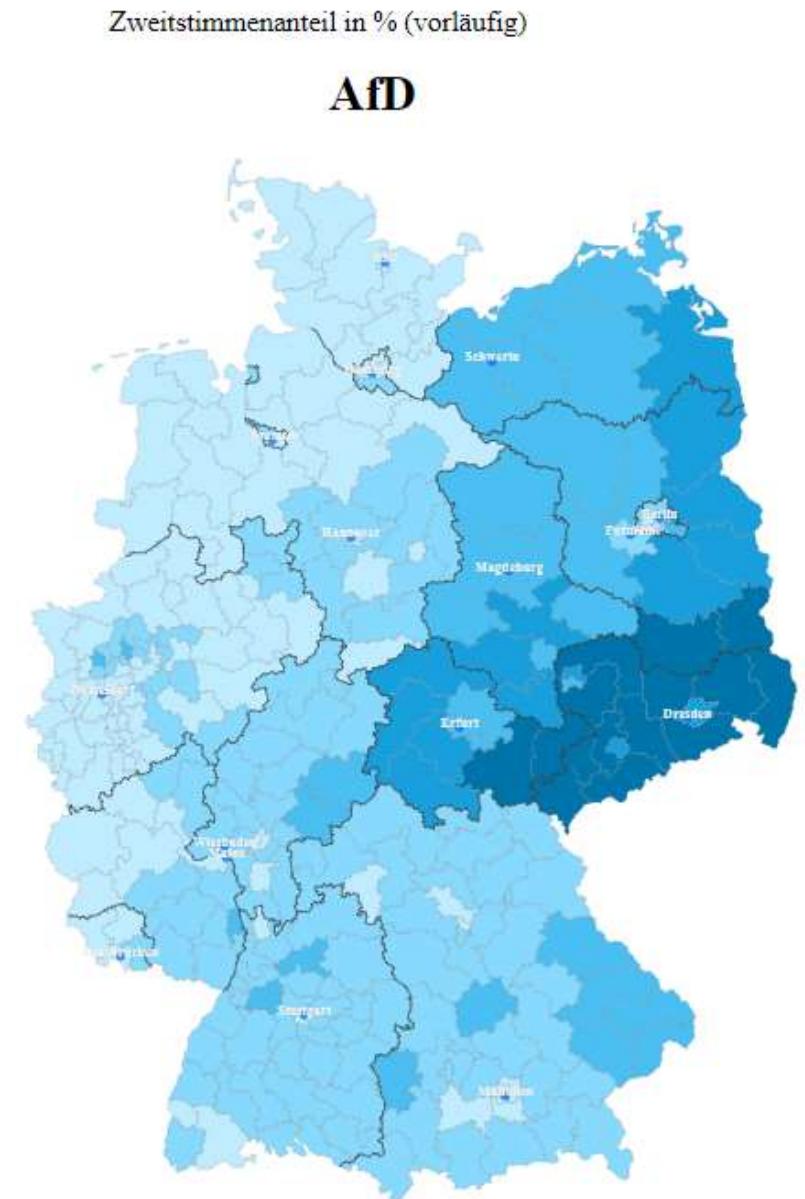
- Hvad er andelen af hvide i populationen?
- Vi udtager en stikprøve, hvor andelen af hvide er $\frac{4}{6} = 0.67$
- Den sande værdi er $\frac{15}{25} = 0.60$, men den værdi kender vi ikke
- Da vi ikke ved bedre, estimerer vi på baggrund af vores stikprøve, at andelen af hvide i populationen også er 0.67
- Når vi udtaler os om statistiske parametre som middelværdi, varians og standardafvigelse er det nødvendigt at skelne mellem population og stikprøve
- Det er nødvendigt at kunne vurdere, om stikprøven er **repræsentativ** for populationen. Vi kunne f.eks. sagtens have trukket en stikprøve med en anden andel hvide



Population og stikprøve

Bundestagswahl 2017 in Deutschland

- Det er ikke ligetil at udtage en repræsentativ stikprøve
- Hvis jeg f.eks. skal udtales mig om holdningen til flygtninge i Tyskland skal jeg lave en tilfældig stikprøve, der er repræsentativ på tværs af geografi, alder, køn, indtægt, politisk holdning, etnicitet, uddannelse, ...



Eksempel: Præsidentvalg i USA, 1936



- To meningsmålinger om udfaldet
 1. Tidsskriftet The Literary Digest:
2.400.000 personer: Landon får 55% af stemmerne
 2. Marketing-ekspert George Gallup:
30.000 personer: Roosevelt får 60% af stemmerne
- Gallup fik ret, Roosevelt vandt præsidentvalget klart (60.8%)
- The Literary Digest's fejl: de havde fundet respondenterne i telefonbøger og registre over bilnummerplader (dvs. blandt de rige)
- Derfor var stikprøven ikke repræsentativ for de amerikanske vælgere.

Notation

Population

- En *parameter* er en kvantitativ størrelse, der beskriver en egenskab ved populationen
- F.eks.
Populations-middelværdi: μ
Populations-standardafvigelse: σ
Generel parameter ('theta'): θ
- Vi bruger de beregnede *statistikker* som *estimatorer* for populationens *parametre*, f.eks. \bar{x} som estimator for μ og s som estimator for σ .
Generelt $\hat{\theta}$ som estimator for θ .

Stikprøve (*sample*)

- En *statistik* er en kvantitativ størrelse, beregnet fra en stikprøve, der beskriver en egenskab ved stikprøven
- F.eks.
Stikprøve-middelværdi: \bar{x}
Stikprøve-standardafvigelse: s
Generel statistik ('theta hat'): $\hat{\theta}$

Stikprøver til estimering af populationer

- De vigtigste parametre for en population, som vi gerne vil estimere ud fra en stikprøve er populationens middelværdi μ og varians σ^2 (og dermed standardafvigelse σ)
- Vi kan bruge hhv. stikprøvens middelværdi \bar{x} og varians s^2
- Stikprøve-middelværdi (*Sample mean*):

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

- Stikprøve-variens (*Sample variance*):

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

- Hvorfor dividerer vi med $n - 1$ og ikke n ?

Fordi så er den gennemsnitlige varians af mange stikprøver σ^2 , altså:

$$E(s^2) = \sigma^2 \text{ (kan man vise)}$$

- Stikprøve-standardafvigelse (*Sample standard deviation*):

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

Stikprøvefordeling

- Normalt uttaler man sig om en population på baggrund af en (1) stikprøve, f.eks. om populationens middelværdi μ ud fra stikprøvens middelværdi \bar{x}
- Stikprøven består af et antal observationer. Antallet kaldes **stikprøvestørrelsen**
- Hvis vi lavede endnu en stikprøve ville vi nok få en anden stikprøvemiddelværdi pga. tilfældigheder
- For at lære noget generelt om fordelingen af stikprøvers middelværdi vil vi i det næste se på **serier** af stikprøver
- Ved at se på mange stikprøver kan vi erfare, hvordan stikprøvemiddelværdien fordeler sig.

Serier af stikprøver



- Terningkast

X : stokastisk variabel med diskret uniform fordeling $U(1,6)$

- Der gælder om $U(a,b)$:

$$\mu = \frac{a+b}{2} = \frac{1+6}{2} = 3.5$$

$$\sigma = \sqrt{\frac{(b-a+1)^2 - 1}{12}} = \sqrt{\frac{(6-1+1)^2 - 1}{12}} = \sqrt{\frac{35}{12}} = 1.708$$

Mean	$\frac{a+b}{2}$
Variance	$\frac{(b-a+1)^2 - 1}{12}$

- Vi laver 100 tilfældige stikprøver, hver med en stikprøvestørrelse på 4
- Hver stikprøve består af 4 terningkast, og vi beregner stikprøve-middelværdi \bar{x} for hver stikprøve (gennemsnitligt antal øjne i de 4 kast)
- Derved får vi 100 værdier af \bar{x} : $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{100}$
- Vi opfatter værdierne \bar{x} som kommende fra en stokastiske variabel, vi kalder \bar{X} . Det viser sig, at middelværdi og standardafvigelse for \bar{X} er:

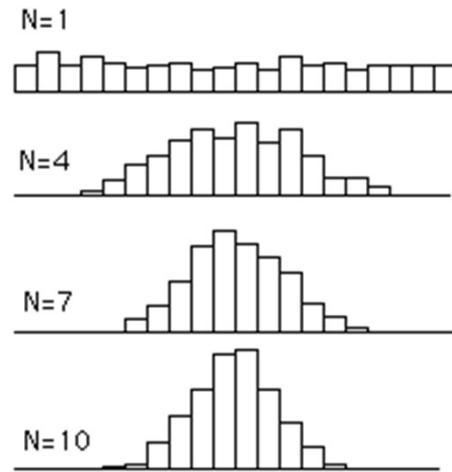
$$\mu_{\bar{X}} = E(\bar{X}) = E(X) = \mu = 3.5$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1.708}{\sqrt{4}} = 0.854$$

- Diskutér med sidemanden: Er det intuitivt at $\mu_{\bar{X}} = \mu$ men $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \leq \sigma$?? 197

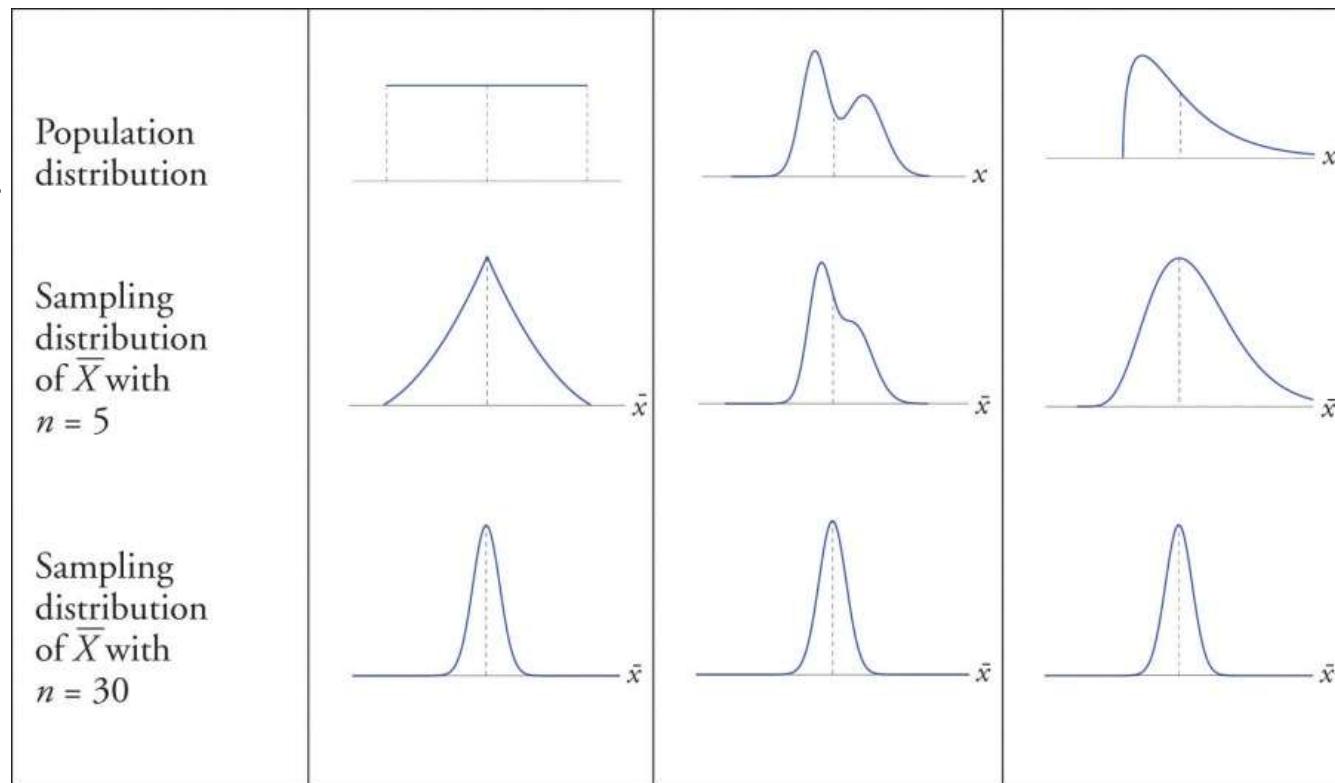
Stikprøvefordeling for middelværdi af stikprøver

- Fordelingen af stikprøvemiddelværdien afhænger af stikprøvestørrelsen (her N):



- Uanset populationsfordelingen kommer stikprøvefordelingen til at ligne **normalfordelingen**, når stikprøvestørrelsen (her n) er tilpas stor:

Dette fænomen kaldes
Den Centrale
Grænseværdi-sætning.



Den centrale grænseværdi-sætning

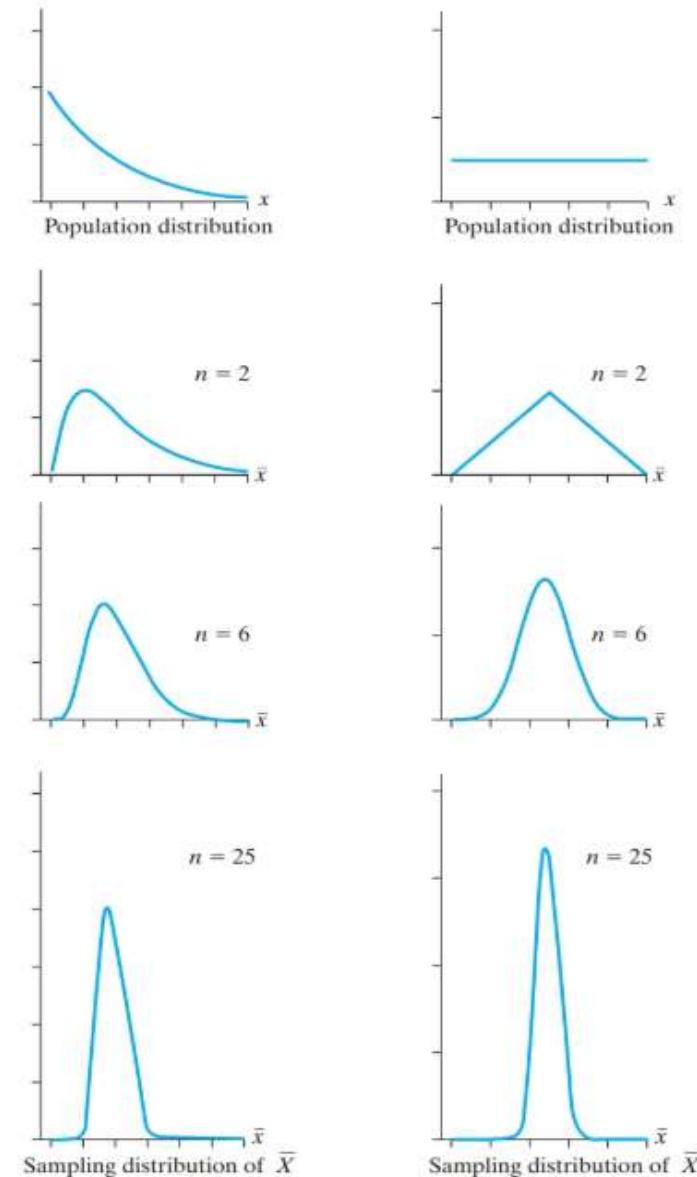
Den centrale grænseværdi-sætning (Central Limit Theorem)

- Lad X være en stokastisk variabel med middelværdi μ og standardafvigelse σ (Vi kender ikke fordelingen af X)
- Lad \bar{X} være den stokastiske variabel, der beskriver gennemsnittet af stikprøver med stikprøvestørrelse n , der er trukket af X
- Hvis n er ‘tilstrækkelig stor’, så er \bar{X} normalfordelt med middelværdi μ og standardafvigelse σ/\sqrt{n} $\bar{X}: N(\mu, \sigma/\sqrt{n})$
- Dermed er $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ standard normalfordelt $Z: N(0,1)$
- Standardafvigelsen for stikprøvens gennemsnit, $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ kaldes ‘standardfejlen af middelværdien’ (standard error of the mean)
- Hvad er en ‘tilstrækkelig stor’ værdi af n ?:
 - Hvis X er normalfordelt er $n = 1$ tilstrækkelig
 - Hvis X er symmetrisk med ét toppunkt er $n = 3$ til 5 tilstrækkelig
 - Hvis X er uniformt fordelt er $n = 6$ til 12 tilstrækkelig
 - For de fleste fordelinger er $n \geq 30$ tilstrækkelig.

Den centrale grænseværdi-sætning

Figur 6.3

To forskellige populationsfordelinger.
Med stigende stikprøvestørrelse n
kommer fordelingen af stikprøve-
middelværdi til at ligne normal-
fordelingen mere og mere



Eksempel: Hydraulikpumper til eksoskelet

- Små batteridrevne 12V hydraulikpumper.
Leverandøren lover maks. pumpetryk på $\mu = 105$ bar med standardafvigelse på $\sigma = 5.0$ bar
- En stikprøve på $n = 8$ pumper købes, og der måles maks. pumpetryk:

Maksimalt tryk for 8 hydraulikpumper [bar]			
106	97	102	96
102	110	108	104

- Vi får $\bar{x} = 103.1$. Betyder det, at leverandørens specifikationer er forkerte?
 - Vi antager den centrale grænseværdidisætning (CGS): $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$. Så er $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ standard normalfordelt $N(0,1)$
- $$P(\bar{X} \leq 103.1) = P\left(Z \leq \frac{103.1 - 105}{5.0/\sqrt{8}}\right) = P(Z \leq -1.0607) = 0.144$$
- Med andre ord, hvis leverandørens specifikationer er korrekte, vil 14.4 % af stikprøver have en middelværdi på 103.1 eller derunder
 - Det er ikke usandsynligt, så vi tror på specifikationerne.



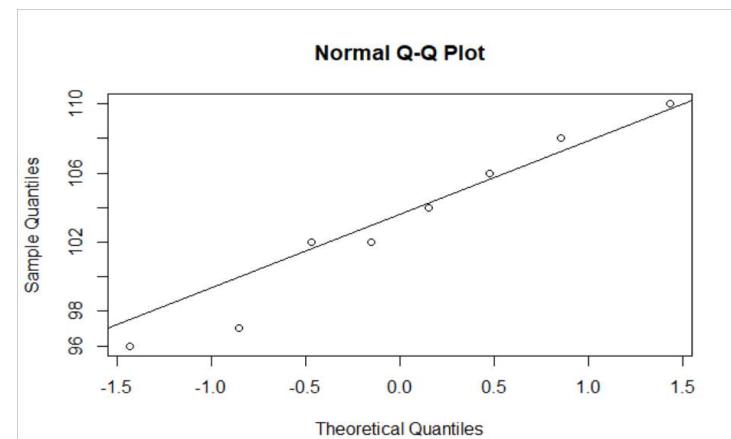
Eksempel: Hydraulikpumper til eksoskelet

- Forudsætningen for at beregne sandsynligheden for at trække en stikprøve med gennemsnit 103.1, hvis populationen har gennemsnit 105 bar er, at vores antagelse om den centrale grænseværdisætning holder.
Er $n = 8$ ‘tilstrækkeligt stort’?
- Stem-and-leaf plot viser en rimeligt symmetrisk fordeling med et enkelt toppunkt, så $n = 8$ burde være ‘tilstrækkeligt stort’
- Når data plottes i et **normalfordelingsplot** fås en nogenlunde lineær sammenhæng, hvilket bekræfter antagelsen: Fordelingen for X ‘ligner’ en normalfordeling, så $n = 8$ burde være ‘tilstrækkeligt stort’.

Hvad er en ‘tilstrækkelig stor’ værdi af n ?:

- Hvis X er normalfordelt er $n = 1$ tilstrækkelig
- Hvis X er symmetrisk med ét toppunkt er $n = 3$ til 5 tilstrækkelig
- Hvis X er uniformt fordelt er $n = 6$ til 12 tilstrækkelig
- For de fleste fordelinger er $n \geq 30$ tilstrækkelig.

9		67
10		224
10		68
11		0



Stikprøver, hvor variansen ikke kendes

- I eksemplet vidste vi, at $\mu = 105$ bar og $\sigma = 5.0$ bar (populationen)
- Vi kunne bruge den centrale grænseværdisætning:

$z_0 = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ er standard normalfordelt, hvis n er tilstrækkelig stor

- Hvis vi *ikke* kender σ , kan vi estimere den ved stikprøve-standardafvigelsen s , men

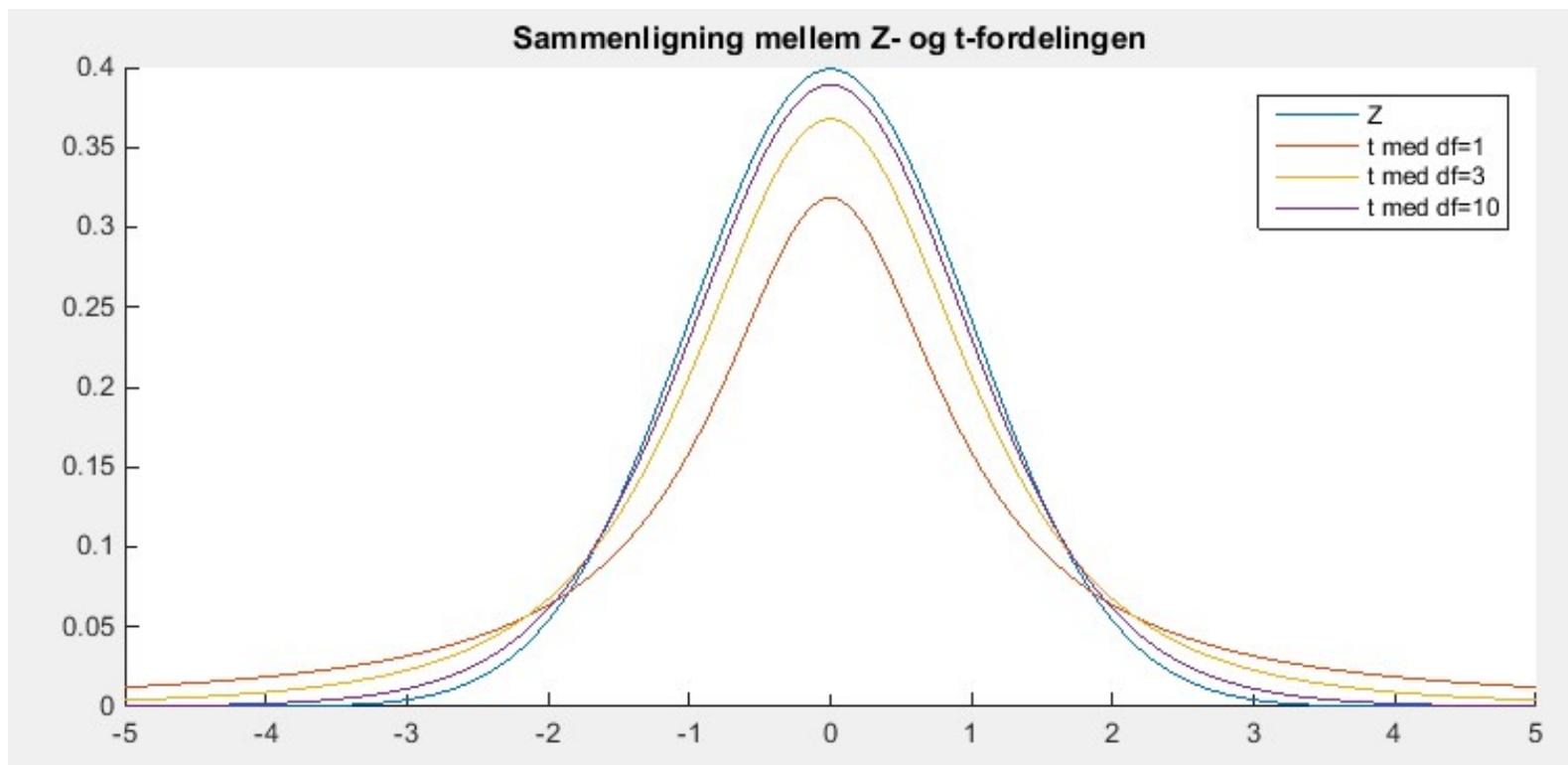
$t_0 = \frac{\bar{x} - \mu}{s / \sqrt{n}}$ er *ikke* standard normalfordelt

- I udtrykket for z_0 er det kun \bar{x} der er stokastisk, resten (μ, σ og n) antages kendt
- I udtrykket for t_0 er det både \bar{x} og s , der bidrager med variabilitet
- Man kan vise, at

$t_0 = \frac{\bar{x} - \mu}{s / \sqrt{n}}$ følger en ***t-fordeling*** med $v = n - 1$ frihedsgrader
(degrees of freedom (*df*)).

t-fordelingen

- Hedder egentlig *Student's t-fordeling* efter William Gosset, som arbejdede med små stikprøver på Guinness bryggeriet i Dublin. Han publicerede sine resultater videnskabeligt under pseudonymet 'Student' i 1908
- *t-fordelingen* ligner standard normalfordelingen Z , d.v.s. symmetrisk omkring middelværdien 0, men med mere spredning, d.v.s. 'tungere' haler
- Jo flere frihedsgrader (df), desto mindre spredning har t , og dermed desto tættere på Z , standard normalfordelingen.



Frihedsgrader

- Frihedsgrader (degrees of freedom, df) er et mål for hvor meget variabilitet, der er i vores estimering
- Bogen bruger det græske bogstav ν ('ny') til at betegne frihedsgrader
- Frihedsgrader er defineret som:
Antal frihedsgrader =
$$\text{Antal observationer} - \text{Antal parametre der estimeres}$$
- Her estimeres der 1 parameter (stikprøve-standardafvigelsen s), så
Antal frihedsgrader = $\nu = n - 1$.

t-fordelingen i R

- `dt(t, df)`
Tæthedsfunktionen (kan bruges til at tegne fordelingen)
- `pt(t, df)`
Kumuleret fordelingsfunktion (bruges til at beregne sandsynligheder)
F.eks. `pt(1.5, 7)` giver sandsynligheden for t er mindre end 1.5 med 7 frihedsgrader
- `qt(p, df)`
Den inverse funktion til fordelingsfunktionen (bruges til at beregne den værdi af t , der svarer til en given sandsynlighed)
F.eks. `qt(0.95, 7)` giver den værdi t_0 , så $pt(t_0, 7) = 0.95$
- `rt(n, df)`
Funktion til generering af n tilfældige tal fra *t*-fordelingen (bruges til at simulere en tilfældig stikprøvetagning).

Eksempel 6.52, s. 206

En kemisk fabrik udleder spildevand til en flod. Fabrikken påstår, at koncentrationen af et udledt stof er under 40 mg/l i gennemsnit

En stikprøve på $n = 20$ vandprøver har $\bar{x} = 46$ og $s = 9.4 \text{ mg/l}$

Kan vi på baggrund af stikprøven afvise fabrikkens påstand?

Løsning:

- $t_0 = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{46 - 40}{9.4/\sqrt{20}} = 2.855$
 t_0 følger en t -fordeling med $n - 1 = 19$ frihedsgrader
- $P(t > t_0) = 1 - P(t < t_0) = 1 - pt(2.855, 19) = 0.00507$
- Vi tror ikke på fabrikkens påstand, for så vil det være ekstremt usandsynligt at få en stikprøve, som den vi har fået
- Vi bør dog tjekke, at forudsætningerne for CGS holder ('pæn' fordeling), men bogen har ikke oplyst stikprøvens målinger.

Fordelingen af en stikprøves varians

- Lad X være en stokastisk variabel for en population med middelværdi μ og varians σ^2
- Vi tager en stikprøve af X med stikprøvestørrelse n og beregner stikprøvemiddelværdi \bar{x} og stikprøvevariанс s^2
- En ny stikprøve vil have en anden værdi af \bar{x} og s^2 pga. tilfældigheder
- Derfor har vi opfattet \bar{x} som en værdi af en stokastisk variabel, vi har kaldt \bar{X} . CGS sagde, at \bar{X} er normalfordelt $N(\mu, \sigma/\sqrt{n})$, hvis n er tilstrækkelig stor
- Tilsvarende kan vi opfatte stikprøvevariansen s^2 som en værdi af en stokastisk variabel, som vi kan kalde S^2 . Vi er interesserede i, hvordan fordelingen er for S^2 , så vi kan regne med sandsynligheder.

Fordelingen af en stikprøves varians

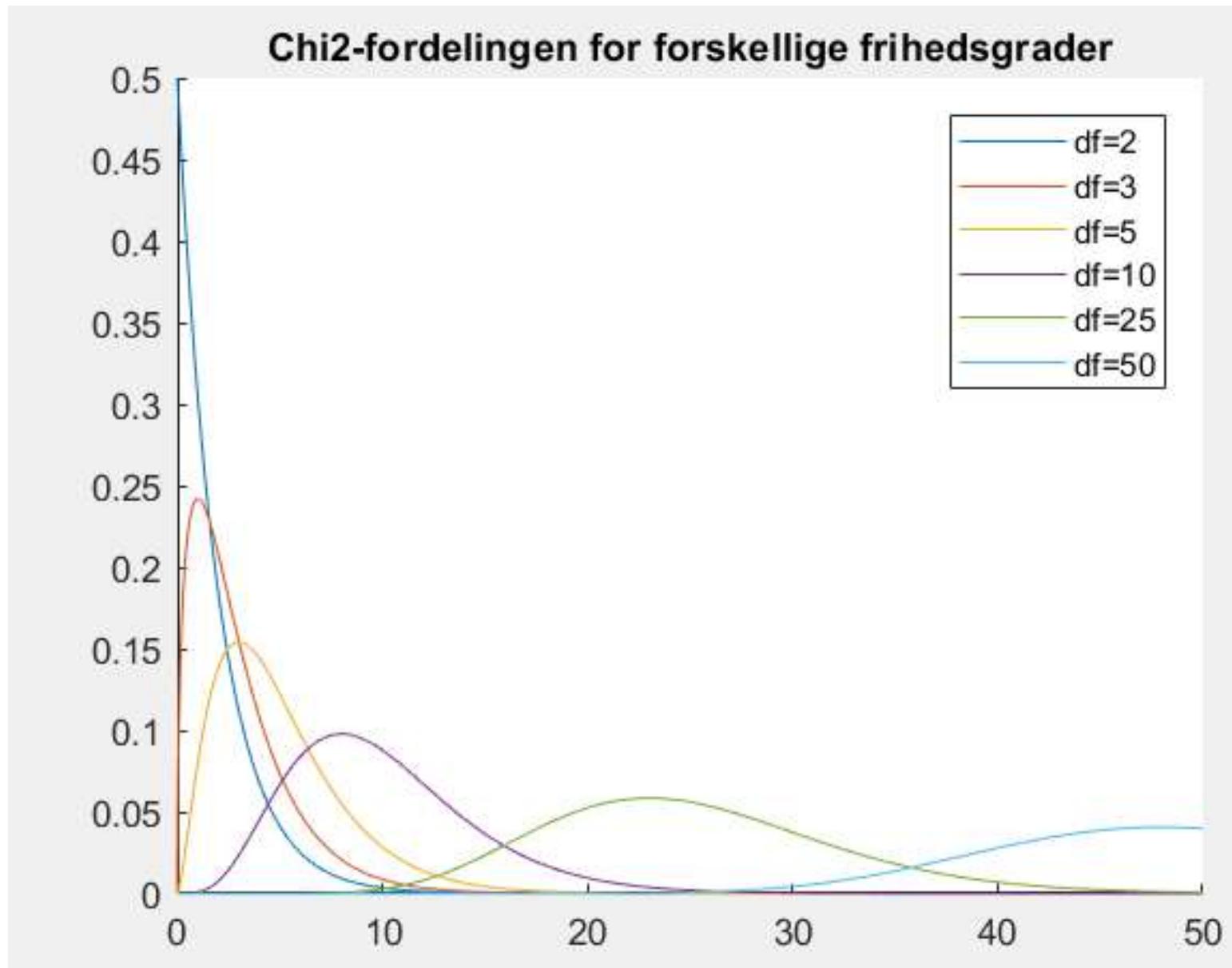
- Lad s^2 være variansen for en stikprøve med størrelse n , taget fra en normalfordelt population med middelværdi μ og varians σ^2
- Så er

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma^2}$$

χ^2 (chi-i-anden) fordelt med $\nu = n - 1$ frihedsgrader

- Ny kontinuert sandsynlighedsfordeling: χ^2 fordelingen.

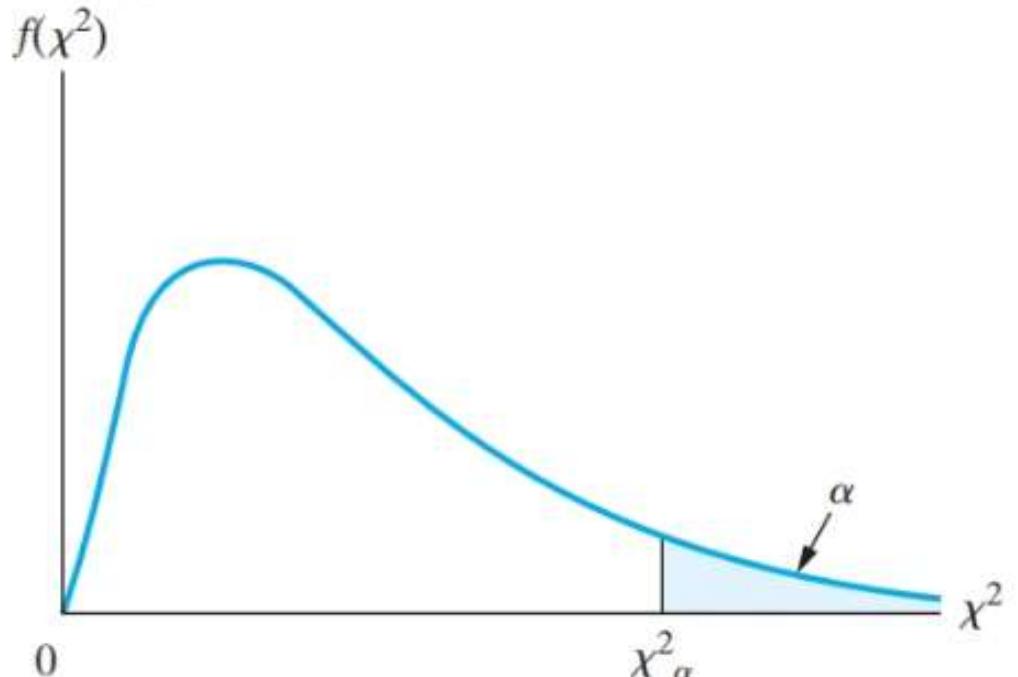
χ^2 fordelingen



χ^2 fordelingen

- Usymmetrisk, højreskæv fordeling, især for få frihedsgrader
- $\chi^2 \geq 0$

χ^2 distribution with v degrees of freedom



- Funktioner i R:
 - `dchisq(x, df):` pdf for x med df frihedsgrader
 - `pchisq(x, df):` cdf for x, d.v.s. $P(\chi^2 \leq x)$
 - `qchisq(p, df):` Invers cdf, finder det x , så $P(\chi^2 \leq x) = p$
 - `rchisq(n, df):` Generering af n tilfældige, χ^2 -fordelte tal

Eksempel 6.53, s. 208

En maskine producerer plastikfolie. Når maskinen fungerer korrekt, er tykkelsen af folien normalfordelt med standardafvigelse $\sigma = 1.35$ mm. Ind imellem bliver variationen på tykkelsen større. For at fange dette, tages der jævnligt en stikprøve med størrelse $n = 20$, og hvis stikprøvens standardafvigelse overstiger $s = 1.40$ undersøger man maskinen

- Hvad er sandsynligheden for, at en stikprøve giver anledning til undersøgelse af maskinen?

Løsning:

- Vi kan beregne χ^2 -statistikken (NB: Der er fejl i bogens beregning):
$$\chi_0^2 = \frac{(n - 1)s^2}{\sigma^2} = \frac{(20 - 1) \cdot (1.40)^2}{(1.35)^2} = 20.4$$
- $P(\chi^2 \geq 20.4) = 1 - P(\chi^2 < 20.4) = 1 - \text{pchisq}(20.4, 19) = \mathbf{0.369}$
- 37 % af stikprøverne giver altså en alarm, selv om $\sigma = 1.35$ mm.

Fordeling af varians for **to** stikprøver

- Vi har to stikprøver med hhv. stikprøvestørrelse n_1 og n_2
- Stikprøverne kommer fra to **normalfordelte** populationer med samme varians, evt. fra den samme normalfordelte population
- Vi har beregnet de to stikprøvevarianser til hhv. s_1^2 og s_2^2
- På grund af tilfældigheder i stikprøverne forventer vi ikke, at de to stikprøvevarianser er helt ens, men dog at de er tæt på hinanden. Derfor forventer vi, at forholdet mellem dem er tæt på 1
- Forholdet mellem stikprøvevarianserne

$$F_0 = \frac{s_1^2}{s_2^2}$$

er en stokastisk variabel, der er F-fordelt med $\nu_1 = n_1 - 1$ frihedsgrader i tælleren og $\nu_2 = n_2 - 1$ frihedsgrader i nævneren

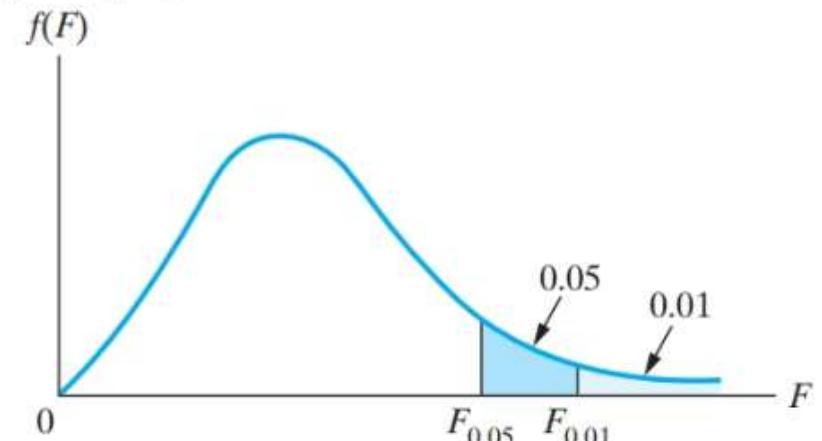
- Ny kontinuert sandsynlighedsfordeling: **F fordelingen**.

F-fordelingen

- F -fordelingen er højreskæv, ligesom χ^2 fordelingen

- F -fordelingen har to parametre for frihedsgrader, fordi der er frihedsgrader for både tælleren og nævneren i udtrykket (fordi der kan være forskellige stikprøvestørrelser)

F distribution with v_1 and v_2 degrees of freedom



- F -fordelingen konvergerer mod χ^2 fordelingen, når nævnerens frihedsgrader bliver stor

- Funktioner i R:

- $df(x, df1, df2)$: pdf for x med $df1$ og $df2$ frihedsgrader for hhv. tæller og nævner
- $pf(x, df1, df2)$: cdf for x , d.v.s. $P(F \leq x)$
- $qf(p, df1, df2)$: Invers cdf, finder det x , så $P(F \leq x) = p$
- $rf(n, df1, df2)$: Generering af n tilfældige, F -fordelte tal

Eksempel 6.54, s. 209

To stikprøver tages fra en normalfordelt population. Den første har størrelse $n_1 = 7$ og den anden $n_2 = 13$. Hvad er sandsynligheden for, at den første stikprøves varians er mindst tre gange større end den andens?

Løsning:

- $s_1^2 \geq 3s_2^2 \Leftrightarrow \frac{s_1^2}{s_2^2} \geq 3 \Leftrightarrow F_0 \geq 3$
- $P(s_1^2 \geq 3s_2^2) = P(F_0 \geq 3) = 1 - P(F_0 < 3)$
- $v_1 = n_1 - 1 = 7 - 1 = 6$
 $v_2 = n_2 - 1 = 13 - 1 = 12$
- $P(s_1^2 \geq 3s_2^2) = 1 - P(F_0 < 3) = 1 - \text{pf}(3, 6, 12) = \mathbf{0.051}$
- Tilsvarende kan vi beregne sandsynligheden for, at det er den anden stikprøves varians, der er mindst 3 gange større end den førstes:
 $P(s_2^2 \geq 3s_1^2) = P(\frac{s_2^2}{s_1^2} \leq \frac{1}{3}) = P(F_0 \leq \frac{1}{3}) = \text{pf}(\frac{1}{3}, 6, 12) = \mathbf{0.094}.$

R funktioner til kontinuerte fordelinger

	PDF (dxxx)	CDF (pxxx)	Invers (qxxx)	Random (rxxx)
Normal (xnorm)	dnorm()	pnorm()	qnorm()	rnorm()
Eksponentiel (xexp)	dexp()	pexp()	qexp()	rexp()
t (xt)	dt()	pt()	qt()	rt()
Chi-i-anden (xchisq)	dchisq()	pchisq()	qchisq()	rchisq()
F (xf)	df()	pf()	qf()	rf()

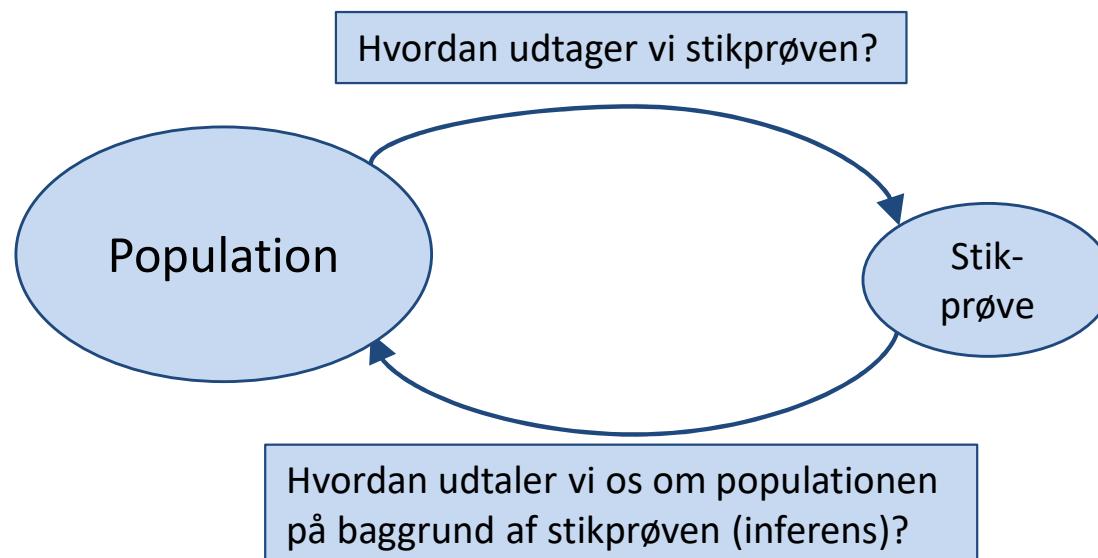
Sandsynlighedsteori og statistik

Kapitel 7. Inferens om middelværdi (afsnit 7.1-7.3, 7.4-7.8)

Allan Leck Jensen
alj@ece.au.dk

Fra kap. 6: Population og stikprøve

- Hvad er gennemsnitshøjden af danske ingenørstuderende?
- Hvordan ville mandatfordelingen i folketinget se ud, hvis der var folketingsvalg i morgen?
- Hvad kan jeg forvente, at brudstyrken af min stålbjælke er?
- Er torsken ved at uddø i Vesterhavet?
- En ‘population’ behøver ikke være en samling mennesker. Den kan være uendelig, f.eks. sandsynligheden for plat med en given mønt
- Hvordan kan vi udtales os om ‘populationen’?



Fra kap. 6: Notation

Population

- En *parameter* er en kvantitativ størrelse, der beskriver en egenskab ved populationen
- F.eks.
Populations-middelværdi: μ
Populations-standardafvigelse: σ
Generel parameter ('theta'): θ
- Vi bruger de beregnede *statistikker* som *estimatorer* for populationens *parametre*, f.eks. \bar{x} som estimator for μ og s som estimator for σ .
Generelt $\hat{\theta}$ som estimator for θ .

Stikprøve (*sample*)

- En *statistik* er en kvantitativ størrelse, beregnet fra en stikprøve, der beskriver en egenskab ved stikprøven
- F.eks.
Stikprøve-middelværdi: \bar{x}
Stikprøve-standardafvigelse: s
Generel statistik ('theta hat'): $\hat{\theta}$

Tre typer af inferens

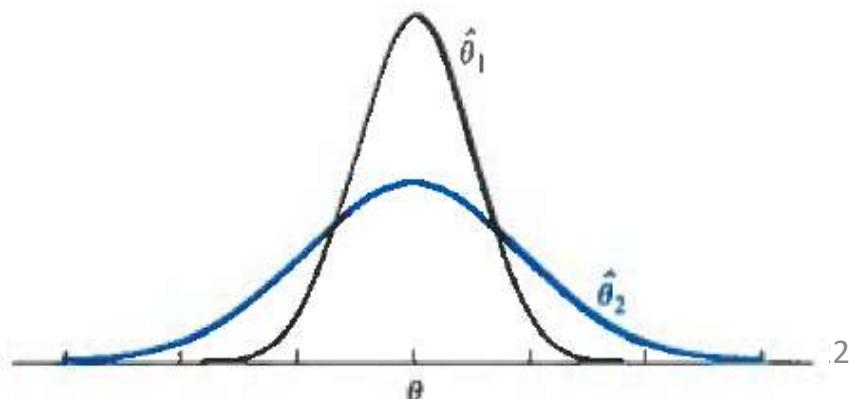
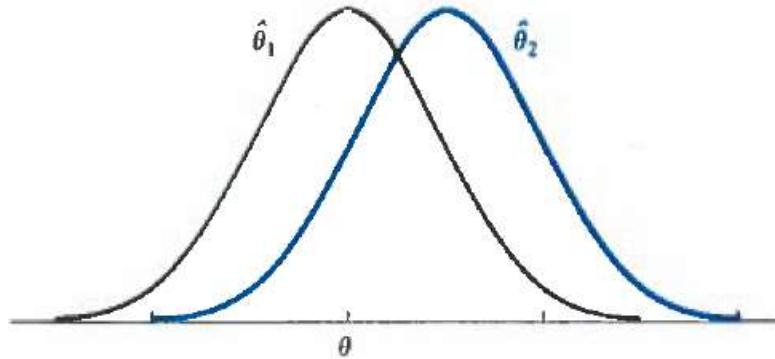
- Vi ønsker at udtale os om populationen vha. en stikprøve. Det kaldes *inferens* eller *generalisering*
- I dette kapitel er den parameter, vi primært vil lave inferens for, populationens middelværdi, μ
- Der er tre typer af inferens vi kan vælge imellem:
 1. Punkt-estimering: Vi estimerer værdien af μ ud fra stikprøven
 2. Interval-estimering: Vi estimerer et sandsynligt interval for værdien af μ
 3. Hypotesetest: Vi bruger en metode til at afgøre, om μ har en bestemt værdi eller ej.

Valg af estimator

- Som estimator af populations-middelværdien kunne man vælge:
 - Stikprøve-middelværdien
 - Stikprøve-medianen
 - Gennemsnit af stikprøvens maksimums- og minimumsværdier ([midrange](#))
 - Stikprøvens [typeværdi](#) (den værdi, der er observeret hyppigst i stikprøven)
- Som estimator af populations-standardafvigelsen kunne man vælge:
 - Stikprøve-standardafvigelsen
 - En konstant gange interkvartile-bredden (interquartile range):
$$c(Q_3 - Q_1)$$
 - Forskellen på største og mindste værdi (range) i stikprøven
- Ikke alle estimatorer er lige gode, så hvad er kvalitetskriteriet?

Hvad er en god estimator?

- Analogi: Lad θ (theta) være den præcise tid lige nu. Vi kender den ikke eksakt, men vi har ure til at estimere θ .
Med andre ord angiver vores ur $\hat{\theta}$
- Hvad er et godt ur?
- Accuracy: Uret er sat så tæt som muligt til den præcise tid, θ
- Precision: Uret kan vise tiden præcist (det har sekundviser), og det taber eller vinder ikke tid
- Akkuratesse for en estimator: Unbiased: $E(\hat{\theta}) = \theta$
(Bias betyder skæv, partisk, forudindtaget. Unbiased oversættes til middelret)
- Præcision for en estimator: Mindre variation
 $\hat{\theta}_1$ er unbiased, $\hat{\theta}_2$ er biased $\hat{\theta}_1$ er mere præcis end $\hat{\theta}_2$.

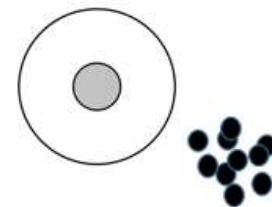


Akkuratesse og præcision

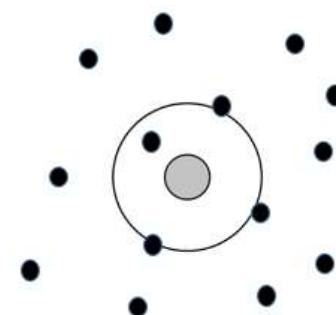
God akkuratesse
God præcision



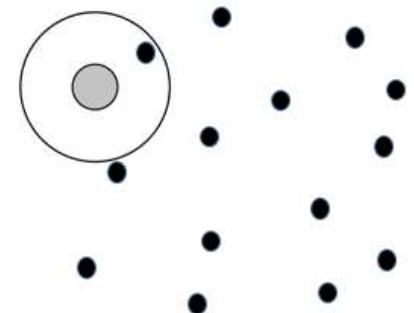
Dårlig akkuratesse
God præcision



God akkuratesse
Dårlig præcision

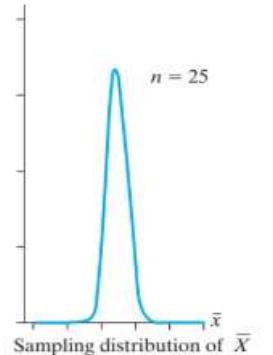
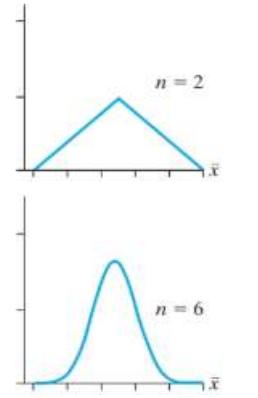
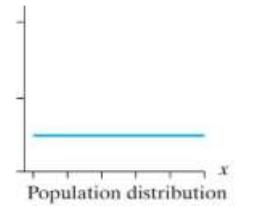


Dårlig akkuratesse
Dårlig præcision



\bar{x} som estimator for μ

- Vi ved fra den centrale grænseværdidisætning, at stikprøve-middelværdien \bar{x} er normalfordelt med middelværdi μ og standardafvigelse σ/\sqrt{n}
- Det vil sige: $E(\bar{x}) = \mu$
- Derfor er \bar{x} en **unbiased estimator** for μ
- Størrelsen σ/\sqrt{n} kaldes **standardfejlen (standard error)**
- Jo større stikprøvestørrelse n , desto mindre standardfejl, og derfor bliver \bar{x} et både mere akkurat og præcist estimat for μ jo større n
- Når vi ikke kender populationens standardafvigelse σ , kan vi estimere den med stikprøvens standardafvigelse s . Så kaldes s/\sqrt{n} for den **estimerede standardfejl**.



Eksempel 7.1-2: Punktestimering af μ

- Gammel beton genbruges til ny vejbelægning. I en stikprøve af vejbelægninger på $n = 18$ måles ‘modulus of resilience’ (~elasticiteten), som er et mål for, hvor meget man skal påvirke et materiale, for at det ikke længere kan vender tilbage til sin oprindelige form
- Resultater (MPa):

136	143	147	151	158	160
161	163	165	167	173	174
181	181	185	188	190	205

The descriptive summary for the sample is

sample mean $\bar{x} = 168.2$

sample median 166

sample standard deviation $s = 18.10$

first quartile 158 third quartile 181

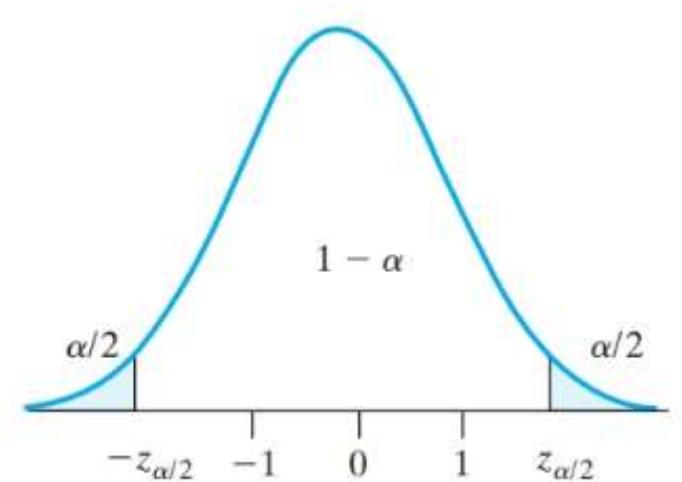
- Punktestimat af μ : $\bar{x} = \mathbf{168.2}$
- Estimeret standardfejl: $\frac{s}{\sqrt{n}} = \frac{18.10}{\sqrt{18}} = \mathbf{4.27}$
- Vores bedste estimat for værdien af μ er $\bar{x} = 168.2$, men da værdien er kontinuert, ved vi, at $P(\mu = \bar{x}) = 0!$ Øv.

Fejlen på vores punktestimat

- Fejlen E på punktestimatet kan udtrykkes som forskellen på estimat og korrekt værdi, dvs. $E = |\bar{x} - \mu|$
- Ifølge CGS: Hvis n er tilstrækkelig stor, er

$$z_0 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$
 standard normalfordelt

- Vi kan vælge en værdi af α , f.eks. $\alpha = 0.05$. Når vi laver en stikprøve, beregner \bar{x} og dernæst z_0 , så vil værdien af z_0 være i det hvide område under kurven med sandsynligheden $1 - \alpha$. z_0 vil være i hvert af de to blå områder med sandsynligheden $\alpha/2$. Grænsen mellem det hvide og de blå områder kaldes $-z_{\alpha/2}$ og $z_{\alpha/2}$



- $P\left(-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = P\left(\frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$
- $P(|\bar{x} - \mu| \leq z_{\alpha/2} \cdot \sigma/\sqrt{n}) = 1 - \alpha$
- Den **maksimale fejl på estimatet** er med sandsynlighed $1 - \alpha$:
$$E = z_{\alpha/2} \cdot \sigma/\sqrt{n}$$

Eks. 7.3, s. 226: Maksimal estimeringsfejl

En ingeniør vil punktestimere middelværdien for en population med standardafvigelse $\sigma = 6.2$. Han vil bruge en stikprøve med størrelse $n = 150$ og beregne \bar{x} . Hvad er den maksimale estimeringsfejl med 99 % sandsynlighed?

Løsning:

- $n = 150, \sigma = 6.2, \alpha = 1 - 0.99 = 0.01$.
Så er $z_{\alpha/2} = z_{0.005} = \text{qnorm}(1 - 0.005) = 2.576$.
- $E = z_{\alpha/2} \cdot \sigma / \sqrt{n} = 2.576 \cdot 6.2 / \sqrt{150} = 1.30$
Ingeniøren er altså 99 % sikker på, at estimeringsfejlen vil være på højst 1.30.

Estimeringsfejl, når σ er ukendt

- Når σ er ukendt kan vi estimere den med s og udnytte, at
$$t_0 = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$
er t -fordelt med $n - 1$ frihedsgrader, når n er stor
- Analogt til situationen med kendt σ får vi:
Den **maksimale fejl på estimatet** er med sandsynlighed $1 - \alpha$:
$$E = t_{\alpha/2} \cdot s / \sqrt{n}.$$

Eksempel 7.1-2: Punktestimering af μ

- Punktestimat af μ : $\bar{x} = \mathbf{168.2}$
- Estimeret standardfejl: $\frac{s}{\sqrt{n}} = \frac{18.10}{\sqrt{18}} = \mathbf{4.27}$
- Maksimal fejl af punktestimatet:
Vi vælger $\alpha = 0.05$. Så er $t_{\alpha/2} = t_{0.025} = qt(1 - 0.025, 17) = 2.11$
 $E = t_{\alpha/2} \cdot s/\sqrt{n} = 2.11 \cdot 4.27 = \mathbf{9.0}.$

Interval-estimater

- Vi er interesserede i at kunne angive et interval, hvor μ ligger indenfor med en vis sandsynlighed
- F.eks. ‘ μ ligger imellem 120 og 140 med 95 % sikkerhed’
D.v.s.: $P(120 < \mu < 140) = 0.95$
- Sådan et interval kaldes et **interval-estimat**
- I eksemplet kaldes intervallet mellem 120 og 140 for **95 % konfidensintervallet for μ**
(confidence \sim tillid)
- Generelt kan vi vælge α mellem 0 og 1 og tale om **$(1 - \alpha) \cdot 100\%$ konfidensintervallet**. For $\alpha = 0.05$ får vi 95 % konfidensintervallet.

Beregning af konfidensinterval, kendt σ

$$\begin{aligned}1 - \alpha &= P\left(-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) \\&= P\left(-z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{x} - \mu \leq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) \\&= P\left(-\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) \\&= P\left(\bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) \\&= P\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)\end{aligned}$$

- Med andre ord: $(1 - \alpha) \cdot 100\%$ konfidensintervallet for μ er

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

- Bemærk: $(1 - \alpha) \cdot 100\%$ konfidensintervallet for μ er

$$\bar{x} \pm E$$

hvor E er den maksimale estimeringsfejl.

Beregning af stikprøvestørrelse

- $(1 - \alpha) \cdot 100\%$ konfidensinterval:

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = \bar{x} \pm E$$

- Vi kan reducere E og gøre konfidensintervallet smallere ved at øge stikprøvestørrelsen n
- For at reducere E med en faktor 10 skal vi øge n med en faktor 100 (pga. kvadratroden)
- Vi kan beregne hvor stor n skal være for at få en bestemt værdi af E
- Hvad skal stikprøvestørrelsen være, for at 95 % konfidensintervallet er $\bar{x} \pm B$ for en vilkårligt valgt værdi af B ?

$$\begin{aligned} z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} &\leq B \Rightarrow \\ z_{\alpha/2} \cdot \sigma &\leq B \cdot \sqrt{n} \Rightarrow \\ \sqrt{n} &\geq \frac{z_{\alpha/2} \cdot \sigma}{B} \Rightarrow \\ n &\geq \left(\frac{z_{\alpha/2} \cdot \sigma}{B} \right)^2 \end{aligned}$$

Eks. 7.3, s. 226: Stikprøvestørrelse

Tidligere: En ingeniør vil punktestimere middelværdien for en population med standardafvigelse $\sigma = 6.2$. Han vil bruge en stikprøve med størrelse $n = 150$ og beregne \bar{x} . Hvad er den maksimale estimeringsfejl med sandsynlighed 0.99?

$$E = z_{\alpha/2} \cdot \sigma / \sqrt{n} = 2.576 \cdot 6.2 / \sqrt{150} = 1.30$$

Nu: Ingeniøren vil beregne stikprøvestørrelsen, der er nødvendig for at opnå en maksimal estimeringsfejl på 1.0, d.v.s. så 99 % konfidensintervallet er $\bar{x} \pm B$ med $B = 1.0$

Løsning:

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sigma}{B} \right)^2 \Rightarrow$$

$$n \geq \left(\frac{2.576 \cdot 6.2}{1.0} \right)^2 \Rightarrow$$

$$n \geq 255.05 \Rightarrow$$

$$n \geq 256$$

- Husk at runde op til nærmeste heltal.

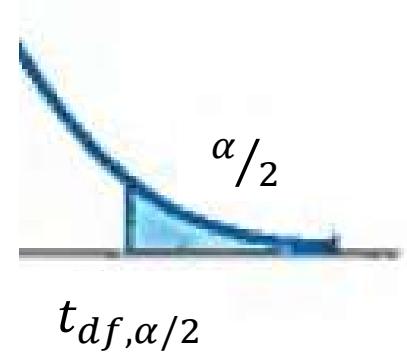
Beregning af konfidensinterval, ukendt σ

- Hvis vi ikke kender populations-standardafvigelsen σ , så følger statistikken

$$t_0 = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad t\text{-fordelingen med } df = n - 1 \text{ frihedsgrader}$$

- Helt analogt er

$$\begin{aligned}1 - \alpha &= \\&= P\left(-t_{df,\alpha/2} \leq \frac{\bar{x} - \mu}{s/\sqrt{n}} \leq t_{df,\alpha/2}\right) \\&= P\left(\bar{x} - t_{df,\alpha/2} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{df,\alpha/2} \cdot \frac{s}{\sqrt{n}}\right)\end{aligned}$$



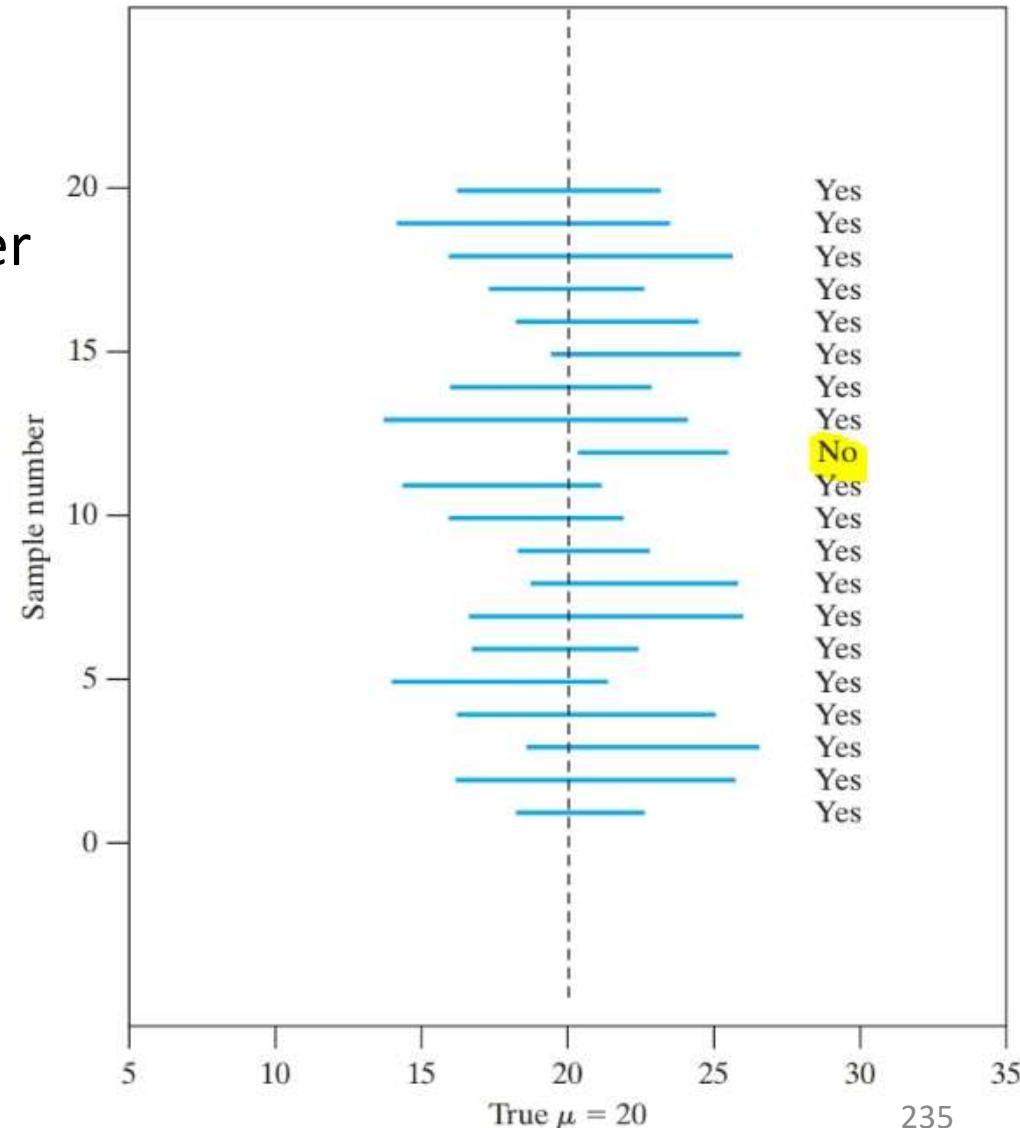
hvor $t_{df,\alpha/2}$ er den værdi af t , hvor arealet i halen er $\alpha/2$ i t -fordelingen med df frihedsgrader

- $(1 - \alpha) \cdot 100\%$ konfidensintervallet for μ er:

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

Konfidensintervallet er usikkert

- Et konfidensintervallet afhænger af stikprøven. Vi kender ikke den sande værdi for μ , så måske indeholder konfidensintervallet ikke μ
- Bogen har simuleret 20 stikprøver hver med $n = 10$ fra $N(20,5)$
- For hver stikprøve blev der beregnet 95 % konfidensinterval
- Som det ses, var der 19 konfidensintervaller, der indeholdt $\mu = 20$, og en enkelt, der ikke gjorde
- 19 ud af 20 svarer til 95 %.



Prædiktionsinterval

- 95 % konfidensintervallet giver et interval, hvor populations-middelværdien μ ligger indenfor med 95 % sikkerhed:

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

- 95 % prædiktionsintervallet giver et interval, hvor de enkelte observationer (stikprøve-middelværdier) ligger indenfor med 95 % sikkerhed
- Definition på $(1 - \alpha) \cdot 100$ % prædiktionsinterval:

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

Eksempel 7.1-2: Prædiktionsinterval

- Punktestimat af μ : $\bar{x} = \mathbf{168.2}$
- Estimeret standardfejl: $\frac{s}{\sqrt{n}} = \frac{18.10}{\sqrt{18}} = \mathbf{4.27}$
- Maksimal fejl af punktestimatet: $E = t_{\alpha/2} \cdot s / \sqrt{n} = \mathbf{9.0}$
- 95 % konfidensinterval:
 $168.2 \pm 9.0 = [159.2; 177.2]$)
- 95 % prædiktionsinterval:
$$\bar{x} \pm t_{df,\alpha/2} \cdot s \cdot \sqrt{1 + \frac{1}{n}} = 168.2 \pm 39.2 = [129.0; 207.4].$$

Hypotesetest – indledende eksempel

- En forskergruppe har udviklet en ny type lithium-batteri til elbiler. De vil gerne kunne påstå, at batterierne kan klare opladning mere end 1600 gange. De vil tage en stikprøve på 36 nye batterier og genoplade dem, indtil de fejler
- Stikprøvens middelværdi \bar{x} skal bruges som bevismateriale for påstanden (hypotesen) at $\mu > 1600$
- Hvordan skal de fortolke resultatet af stikprøven?
 - Hvis f.eks. $\bar{x} = 1610$, så tyder det på, at påstanden holder, men det kunne sagtens være tilfældigt p.g.a. stikprøven
 - Hvis f.eks. $\bar{x} = 1590$, så tyder det på, at påstanden ikke holder, men det kunne sagtens være tilfældigt p.g.a. stikprøven
 - Hvis f.eks. $\bar{x} = 1720$, så tyder det på, at påstanden holder. Vi er mere overbeviste, for det virker usandsynligt, at en tilfældig stikprøve af en population med $\mu \leq 1600$ kommer ud med et snit så langt over populationens middelværdi.

Hypotesetest – indledende eksempel

- Forskergruppen vælger, at de vil forkaste påstanden, med mindre $\bar{x} > 1660$. De vurderer (lidt arbitrært), at en stikprøve med $\bar{x} > 1660$ vil være et tilstrækkeligt stærkt bevis for påstanden om at $\mu > 1600$
- $\bar{x} > 1660$: Vi tror på påstanden $\mu > 1600$
 $\bar{x} \leq 1660$: Vi tror ikke på påstanden $\mu > 1600$
- Selv med disse krystalklare regler kan vi begå fejl:
 - En population med $\mu \leq 1600$ kan give en stikprøve med $\bar{x} > 1660$, og så godkender vi påstanden, selv om den er forkert
 - En population med $\mu > 1600$ kan give en stikprøve med $\bar{x} < 1660$, og så forkaster vi påstanden, selv om den er korrekt.

Nul- og alternativhypoteser

- For at formulere problemstillingen som en hypotesetest skal forskergruppen opstille to hypoteser:
 - Nulhypotesen (H_0)
 - Alternativhypotesen (H_1 (eller H_a))
- Som regel formuleres **alternativhypotesen** som det, man tror er sandt, eller det, man gerne vil bekræfte. Her:
$$H_1: \mu > 1600$$
- Alternativhypotesen er alternativet til **nulhypotesen**, så umiddelbart formuleres H_0 som det modsatte af H_1 :
$$H_0: \mu \leq 1600$$
- Hypotesetesten foregår ved at antage, at nulhypotesen H_0 er sand. Hvis de statistiske undersøgelser viser, at det har usandsynlige konsekvenser, så må H_0 forkastes, og dermed kan vi acceptere H_1
- For at kunne beregne konsekvenser af H_0 formuleres den som regel med '=':
 - $H_0: \mu = 1600$
 - $H_1: \mu > 1600$.

Hypotesetest i retssalen

- En mand er anklaget for mord
- Hypotesetest:
 - Nulhypotesen H_0 : Den anklagede er uskyldig
 - Den alternative hypotese H_1 : Den anklagede er skyldig
- Den anklagede er uskyldig indtil det modsatte er bevist (uskyldsformodningen), og det er anklageren, der skal løfte bevisbyrden
- På engelsk/amerikansk: A person is innocent unless and until proven guilty with evidence that is beyond **reasonable doubt**
- Anklageren fremlægger beviser, der skal forkaste H_0 . Hvis det lykkes, accepteres H_1 og den anklagede dømmes skyldig
- Hvis det ikke lykkes anklageren at forkaste H_0 , så bliver den anklagede frikendt
- Hvis den anklagede bliver frikendt, betyder det ikke, at han er uskyldig. Han begik måske mordet i virkeligheden – det kunne bare ikke bevises ‘beyond reasonable doubt’, at det var ham, der gjorde det
- Vi har ikke bevist H_0 – der var bare ikke bevis nok til at kunne forkaste den

Tvivlen skal komme den anklagede til gode

[NYHEDER](#)[SPORT](#)[MERE ▾](#)[BT SHOP](#)[TIP OS](#)**KRIMI**

Polak frifundet trods klokkeklares DNA-beviser: 'Det er den HELT rigtige dom'

SANNE FAHNØE — ⌂ 25. JAN. 2017 - 8:07

EDEL

En mistænkt blev frifundet for indbrud selvom hans DNA forbandt ham til forbrydelsen.

Hans DNA blev nemlig ikke fundet på selve gerningsstedet, kun på det brækjern, der blev brugt til indbruddet.

Man kunne ikke bevise, at han havde været på gerningsstedet.

To typer af fejl

H_0 : Anklagede er uskyldig

H_1 : Anklagede er skyldig

Anklagedes
sande tilstand

		Rettens afgørelse	
		Dømt	Frikendt
Anklagedes sande tilstand	Uskyldig (H_0)	Type I fejl	OK
	Skyldig (H_a)	OK	Type II fejl

Type I fejl:

- Vi dømmer en uskyldig
- Vi forkaster H_0 , selvom den er sand
- Vi vurderer, at der er noget galt, selvom der ikke er
- Sandsynligheden $\alpha = P(\text{forkaster } H_0 \mid H_0 \text{ er sand})$
- $\alpha \sim$ hvor stærke beviser kræves (beyond reasonable doubt).

Type II fejl:

- Vi frikender en skyldig
- Vi forkaster ikke H_0 , selvom den er falsk
- Vi opdager ikke, at der er noget galt
- Sandsynligheden $\beta = P(\text{accepterer } H_0 \mid H_1 \text{ er sand})$

Test for sygdom

H_0 : Patienten er rask

H_a : Patienten er syg

		Resultat af testen	
		Syg (positiv)	Rask (negativ)
Patientens sande tilstand	Rask (H_0)	Falsk positiv	Sand negativ
	Syg (H_a)	Sand positiv	Falsk negativ

- Type I fejl: Testen siger, at en rask person er syg (testen er falsk positiv)
- Type II fejl: Testen siger, at en syg person er rask (testen er falsk negativ)
- Hvis man gerne vil reducere sandsynligheden for type I fejl (reducere α) kan testens følsomhed reduceres, så færre testes positiv
- Dette vil typisk øge sandsynligheden for type II fejl, da den ændrede følsomhed får testen til at reducere i antal positive testresultater, både for raske og syge
- Tilsvarende problem med f.eks. røgalarmer – en følsom alarm bipper ved madlavning, en mindre følsom reagerer måske for sent på en brand
- Hvilken fejltype er mest alvorlig?.

Indledende eksempel om bilbatterier

$$H_0: \mu = 1600$$

$$H_1: \mu > 1600$$

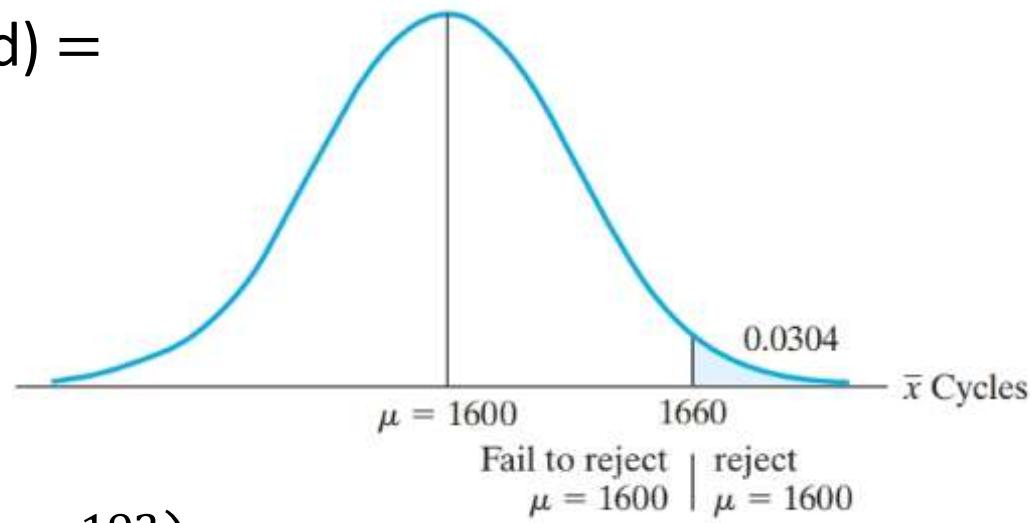
**Den sande
tilsand**

		Resultat	
		H_0 forkastes	H_0 accepteres
H_0 er sand.	H_1 er sand.	Type I fejl	OK
	H_1 ikke sand.	OK	Type II fejl

- Type I fejl: Vi forkaster H_0 , fordi vi får $\bar{x} > 1660$, men det skyldes en tilfældighed. Vi tror, at batterierne er bedre end de reelt er. Vi risikerer klager, sagsanlæg og dårlig publicity
- Type II fejl: Vi kan ikke bevise, at batteriet har så høj kvalitet, som det reelt har, og derfor kan vi ikke sætte prisen så højt.

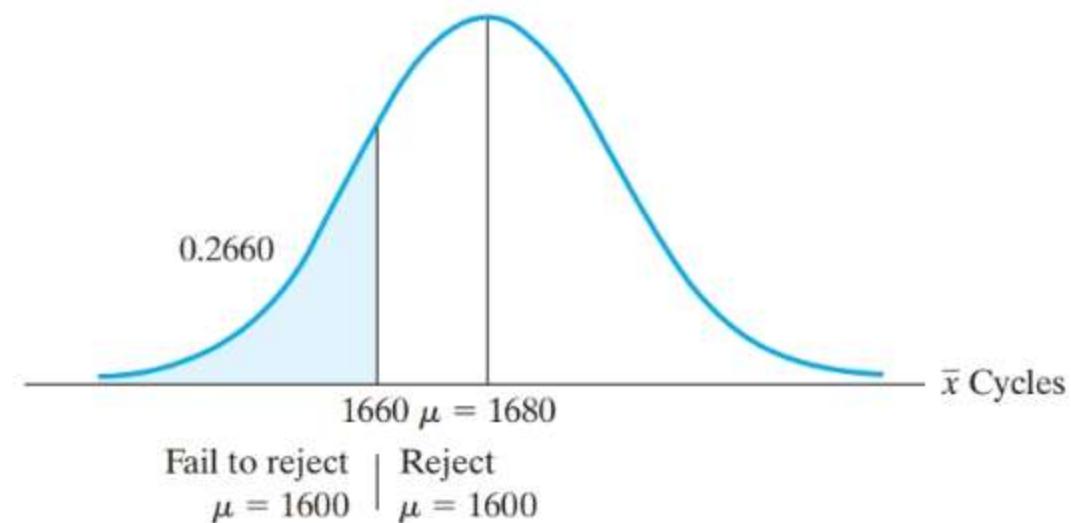
Hypotesetest – indledende eksempel

- Hvad er sandsynligheden α for at begå en type I fejl?
- Hvordan kan vi regne sandsynligheder? Vi ved fra CGS, at \bar{x} er $N(\mu, \frac{\sigma}{\sqrt{n}})$ for tilstrækkeligt stor n . Lad os antage, at vi kender $\sigma = 192$
- $\alpha = P(\text{forkaster } H_0 \mid H_0 \text{ er sand}) = P(\bar{x} > 1660 \mid \mu = 1600)$
- Sandsynligheden α for type I fejl er det blå område i den øvre hale i figuren
- $\alpha = P(\bar{x} > 1660 \mid \mu = 1600) = 1 - \text{pnorm}\left(1660, 1600, \frac{192}{\sqrt{36}}\right) = 0.0304$
- Hvis $\mu < 1600$ svarer det til at rykke kurven mod venstre, og så vil det blå areal blive mindre.



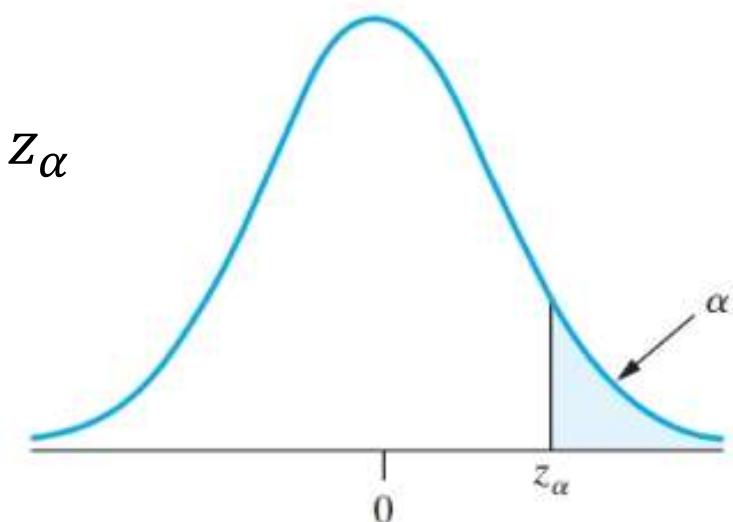
Hypotesetest – indledende eksempel

- Hvad er sandsynligheden β for at begå en type II fejl?
- $\beta = P(\text{accepterer } H_0 \mid H_1 \text{ er sand}) = P(\bar{x} \leq 1660 \mid \mu > 1600)$
- Når H_1 er sand, ved vi bare at værdien af μ er over 1600, vi kender ikke værdien. Det er svært at regne sandsynligheden, når det, vi betinger ikke er en fast værdi
- Grafen viser fordelingen af stikprøvemiddelværdier, hvis $\mu = 1680$ (eksempelvis). Det blå areal viser sandsynligheden for at få en stikprøve, der forkastes
- $\beta' = P(\bar{x} < 1660 \mid \mu = 1680) = \text{pnorm}\left(1660, 1680, \frac{192}{\sqrt{36}}\right) = 0.2660$
- Vi kommer ikke til at beskæftige os mere med sandsynligheden β .



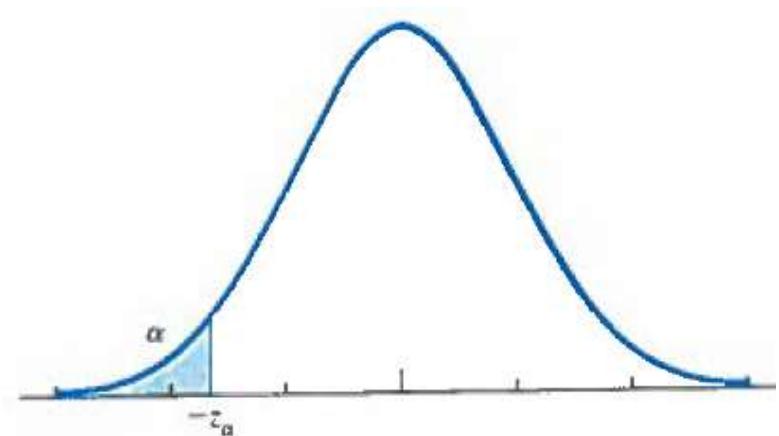
Hypotesetest – indledende eksempel

- Hvordan afgør vi om H_0 kan forkastes?
- I det indledende eksempel havde forskergruppen valgt at forkaste nulhypotesen, hvis $\bar{x} > 1660$. Vi så, at det bevirkede, at
$$\alpha = P(\bar{x} > 1660 \mid \mu = 1600) = 0.0304$$
- Alternativt kan man vælge en værdi af α og beregne den tilhørende **kritiske grænse**, z_α
- Vi forkaster nulhypotesen, hvis
$$z_0 = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} > z_\alpha$$
- α kaldes også for **signifikansniveauet**. Typiske værdier for α er 0.1, 0.05 og 0.01.
Hertil svarer værdier af z_α på hhv. 1.282, 1.645 og 2.326. Omregnet til antal genopladninger svarer det til hhv. 1641, 1653 og 1674
- Hvis vi vælger signifikansniveau $\alpha = 0.05$, skal vi forkaste nulhypotesen, hvis $\bar{x} > 1653$. Der er så 5 % risiko for at vi har begået en fejl type I, dvs. at H_0 er sand, selv om vi har forkastet den.



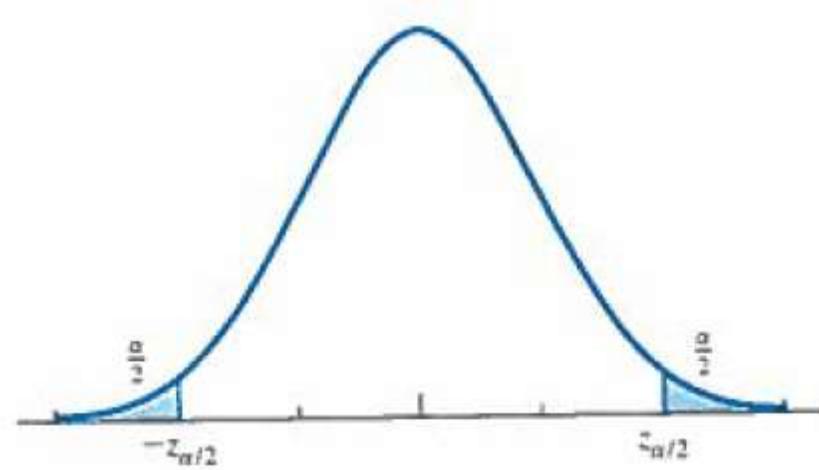
Ensidet hypotesetest

- I eksemplet med bilbatterier var alternativhypotesen formuleret med ‘>’, så den kritiske region var den øvre hale. Det kaldes en **ensidet hypotesetest**
- Man kan også have en ensidet hypotesetest, hvor alternativhypotesen er formuleret med ‘<’. Hvis man f.eks. gerne vil vise, at spildevand indeholder mindre end grænseværdien på 10 ppm af et stof, kan hypoteserne være
 - $H_0: \mu = 10$
 - $H_1: \mu < 10$Her vil nulhypotesen blive forkastet, hvis stikprøvens middelværdi er tilstrækkeligt langt under 10. Den kritiske grænse er den blå, venstre hale i figuren, afgrænset af $-z_\alpha$
- Man kalder også de to ensidede hypotesetest for hhv. **højre-** og **venstre-halede** hypotestest.



Tosidet hypotesetest

- En hypotesetest kan også være **tosidet**. Det er den, hvis alternativhypotesen er formuleret med ' \neq '
- Hvis man f.eks. vil teste om en kaffeautomat doserer mere eller mindre end de angivne 20 cl (for lidt doseret kaffe giver klager, for meget giver tab).
Så kan hypoteserne være:
 - $H_0: \mu = 20$
 - $H_1: \mu \neq 20$
- Figuren viser standard normalfordelingen for stikprøver, givet H_0 er sand. Hvis teststørrelsen z_0 er tilstrækkeligt langt fra middelværdien (0), forkaster vi H_0
- Vi vælger α , beregner $z_{\alpha/2}$, og hvis stikprøvens teststørrelse er mere ekstrem, dvs. hvis $|z_0| > z_{\alpha/2}$, så forkaster vi H_0
- Sandsynligheden for at vi forkaster H_0 , givet den er sand, er det samlede blå areal, $\alpha/2 + \alpha/2 = \alpha$. $P(H_0 \text{ forkastes} \mid H_0 \text{ er sand}) = \alpha$.



Hypotesetestens 5 skridt

1. Formulér hypoteser

- $H_0: \mu = \mu_0$
- $H_1:$ Enten $\mu < \mu_0$, $\mu > \mu_0$ eller $\mu \neq \mu_0$
(er det en ensidet test på venstre eller højre Hale, eller en tosidet test?)

2. Vælg signifikansniveau

Typisk vælges $\alpha = 0.1, 0.05$ eller 0.01 . α er sandsynligheden for type I fejl

3. Opstil kriterier for test af nulhypotesen mod alternativet

- Der opstilles en formel for teststørrelsen, som er den værdi, der skal vurderes mod det kritiske interval
- Det nævnes hvilken fordeling, teststørrelsen følger
- På baggrund af fordelingen, H_1 og signifikansniveauet α beregnes det kritiske interval, der angiver grænser for, om H_0 kan forkastes eller ej

4. Beregn værdien af teststørrelsen

På baggrund af stikprøvens data kan værdien af teststørrelsen beregnes

5. Drag konklusioner og test antagelser

Konkludér om nulhypotesen kan forkastes eller ej. Vurdér om antagelser for data holder.

Hypotesetest for batterieksemplet

1. Formulér hypoteser

- $H_0: \mu = \mu_0 = 1600$
- $H_1: \mu > 1600$
(det en højre-halet, ensidet test)

2. Vælg signifikansniveau

Vi vælger $\alpha = 0.05$

3. Opstil kriterier for test af nulhypotesen mod alternativet

- Formlen for teststørrelsen følger af den centrale grænseværdidisætning (CGS), nemlig at \bar{x} er $N(\mu, \frac{\sigma}{\sqrt{n}})$:

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- Teststørrelsen z_0 er standard normalfordelt
- Nulhypotesen forkastes, hvis $z_0 > z_\alpha$, hvor

$$z_\alpha = qnorm(1 - \alpha) = qnorm(0.95) = 1.645.$$

Hypotesetest for batterieksemplet

4. Beregn værdien af teststørrelsen

Bogen antager, at populations-standardafvigelsen er kendt, $\sigma = 192$.

Lad os sige, at stikprøven har $n = 36$ og $\bar{x} = 1670$

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{1670 - 1600}{192/\sqrt{36}} = 2.1875$$

5. Drag konklusioner og test antagelser

Da $z_0 = 2.1875$ og $z_\alpha = 1.645$ er $z_0 > z_\alpha$. Teststørrelsen er altså i det kritiske interval, så vi forkaster nulhypotesen.

Vi har således vist på 5 % signifikansniveau, at batterierne kan oplades mere end 1600 gange.

Vi har antaget CGS, men da stikprøvestørrelsen er stor ($n > 30$), så holder CGS. Ellers kunne vi teste antagelsen med et normalfordelingsplot af data

Hypotesesten viser, at $\mu > 1600$, men ikke *hvor* meget større. Måske er $\mu = 1601$. Et konfidensinterval kan give plausible værdier af μ , her med $\alpha = 0.05$:

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 1670 \pm 1.96 \cdot \frac{192}{\sqrt{36}} = 1670 \pm 62.7$$

Det giver et 95 % konfidensinterval på [1607; 1733].

Eksempel: Tosidet hypotesetest (s. 249)

I en produktion af cementblokke skal blokkene have termisk konduktivitet på 0.340. Vi vil undersøge, om det opnås med en stikprøve på 35 blokke, på 5 % signifikansniveau. Fra lignende studier vides det, at $\sigma = 0.01$

1. Formulér hypoteser

- $H_0: \mu = 0.340$
- $H_1: \mu \neq 0.340$
(tosidet test)

2. Vælg signifikansniveau

$$\alpha = 0.05$$

3. Opstil kriterier for test af nulhypotesen mod alternativet

Teststørrelsen $z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ er standard normalfordelt

Da vi har en tosidet test, forkastes H_0 , hvis $z_0 > z_{\alpha/2}$ eller hvis $z_0 < -z_{\alpha/2}$

$$z_{\alpha/2} = qnorm(1 - \alpha/2) = qnorm(0.975) = 1.96$$

$$-z_{\alpha/2} = qnorm(\alpha/2) = qnorm(0.025) = -1.96$$

Vi forkaster H_0 , hvis $|z_0| > z_{\alpha/2}$.

Eksempel: Tosidet hypotesetest

4. Beregn værdien af teststørrelsen

Det viser sig, at stikprøvens middelværdi er $\bar{x} = 0.343$.

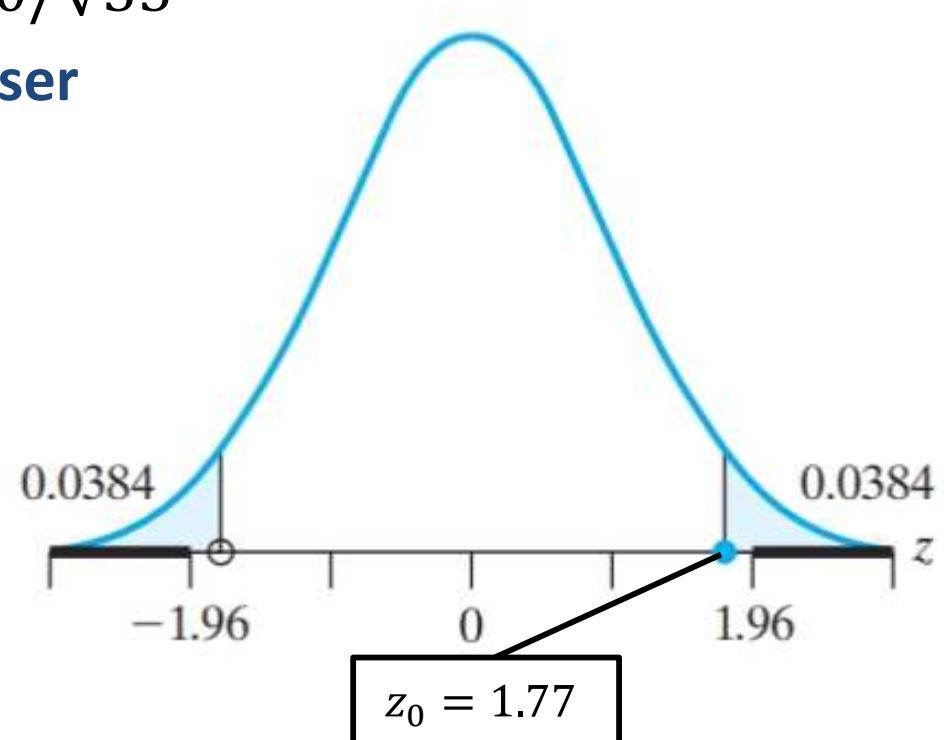
$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{0.343 - 0.340}{0.010/\sqrt{35}} = 1.77$$

5. Drag konklusioner og test antagelser

Da $|z_0| = 1.77$ og $z_{\alpha/2} = 1.96$

er teststørrelsen udenfor det kritiske interval, så vi kan *ikke* forkaste nulhypotesen.

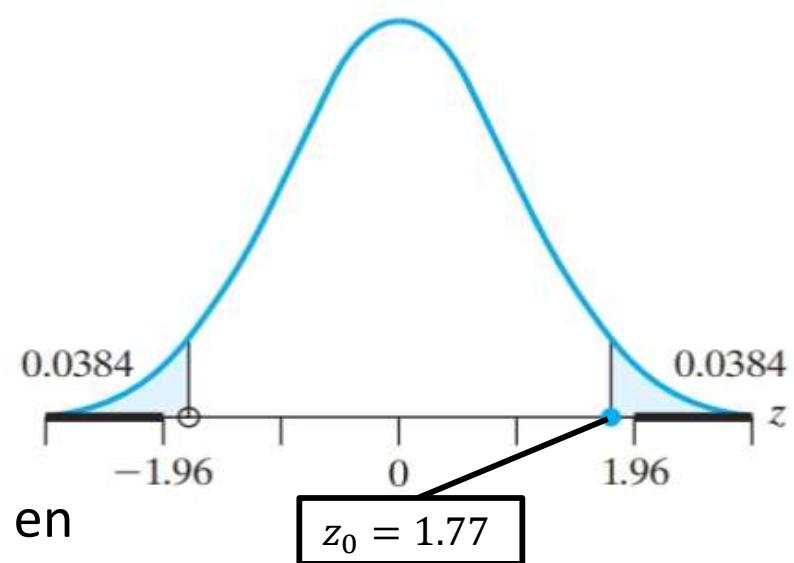
Stikprøven med $\bar{x} = 0.343$ er ikke et stærkt nok bevis til, at vi kan afvise, at $\mu = 0.340$



Vi har antaget CGS, men da stikprøvestørrelsen er stor ($n > 30$), så holder den.

Valg af signifikansniveau α

- Kritisk område for en tosidet test med $H_a: \mu \neq \mu_0$:
 - $\alpha = 0.10: z_{\alpha/2} = z_{0.05} = \text{qnorm}(1 - 0.05) = 1.64$
 - $\alpha = 0.05: z_{\alpha/2} = z_{0.025} = \text{qnorm}(1 - 0.025) = 1.96$
 - $\alpha = 0.01: z_{\alpha/2} = z_{0.005} = \text{qnorm}(1 - 0.005) = 2.58$
- Her var $z_0 = 1.77$, så H_0 vil bliver forkastet på 10 % med signifikansniveau, men ikke på 5 % eller 1 %
- I stedet for at vælge signifikansniveauet α og kritisk interval, så kan man beregne den værdi af α der svarer til testværdien:
$$\begin{aligned}P(Z > z_0) &= 1 - P(Z \leq z_0) \\&= 1 - \text{pnorm}(1.77) = 0.0384\end{aligned}$$
- Tilsvarende er $P(Z < -z_0) = 0.0384$
- Hvis H_0 er sand, er sandsynligheden for at få en stikprøve med \bar{x} eller endnu længere fra μ_0 lig med $0.0384 + 0.0384 = 0.0768$
- Vi forkaster H_0 med signifikansniveau $\alpha = 0.0768$ eller derover
- Denne værdi af α kaldes **p-værdien**.



p-værdi

- *p*-værdien er den mindste sandsynlighed for type I fejl, som tillader os at forkaste H_0
- *p*-værdien kaldes også det opnåede signifikansniveau
- *p*-værdien afhænger af valg af hypoteser:
 - For $H_1: \mu > \mu_0$: $p = P(Z > z_0)$
 - For $H_1: \mu < \mu_0$: $p = P(Z < z_0)$
 - For $H_1: \mu \neq \mu_0$: $p = 2 \cdot P(Z > |z_0|)$.

Hypotesetest for μ med ukendt σ

- Som regel kender man hverken μ eller σ for populationen.
Så kan vi udnytte, at teststørrelsen

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

følger t -fordelingen med $df = n - 1$ frihedsgrader

Eksempel 7.20 s. 254

- Test om blyindhold i vand er under grænseværdien på 2.25 ($\mu\text{g/L}$).
Stikprøve på 12 vandprøver. σ kendes ikke. Der ønskes 2.5 % signifikansniveau (!)
- Data: 2.4 2.9 2.7 2.6 2.9 2.0 2.8 2.2 2.4 2.4 2.0 2.5
- Vi regner eksemplet i R.

Sandsynlighedsteori og statistik

Kapitel 8. Inferens med to stikprøver (afsnit 8.1-8.5 (alt))

Allan Leck Jensen

alj@ece.au.dk

Eksperimentelt design

- Det giver kun mening at sammenligne to stikprøver, hvis de har næsten alt til fælles, og kun adskiller sig på ganske få punkter
- Vi sammenligner to populationer ved hjælp af to stikprøver. Vi er interesserede i, om der er forskel på de to populationers middelværdi
- **Eksempel:** Et nyt produkt til at gøre sko vandtætte skal testes. Forskellige test-personer skal gå med sko, der enten er imprægneret med det nye eller med det nuværende produkt, i en måned, hvorefter skoenes grad af vandtæthed måles
- **Design 1:** 10 personer får nye sko imprægneret med det nye produkt, 10 personer får gamle sko imprægneret med det gamle produkt
- *Er det et godt design?*
- Nej! Skoene skal have samme alder.



Eksperimentelt design

- **Design 2:** 10 personer får nye sko imprægneret med det nye produkt, 10 personer får nye sko imprægneret med det gamle produkt
- Skoen er den eksperimentelle enhed (*experimental unit*). Hver sko bliver utsat for en af to mulige behandlinger (*treatments*), og det vi mäter er skoens grad af vandtæthed efter 1 måneds brug, d.v.s. responsen på behandlingen (*response*)
- Population 1 er sko med ny behandling, population 2 med gammel. Begge stikprøver har en stikprøvestørrelse på 20 (2 sko pr. person)
- **Randomisering:** Vi bør lade det være tilfældigt, hvilke personer der modtager sko med den ene og den anden imprægnering for at jævne forskelle mellem de to grupper ud. F.eks. kan det give skævheder, hvis de yngste testpersoner får ny imprægnering og de ældste får gammel
- *Er det et godt design? Hvad giver variation i data?*
- *Er de to stikprøver uafhængige?*
- Ja, men det behøver de ikke at være – se design 3.

Eksperimentelt design

- **Design 3:** 20 personer får et nyt par sko, hvor den venstre er imprægneret med det nye produkt og den højre med det gamle
- Vi har udnyttet, at hver testperson formodentlig kommer til at bruge venstre og højre sko lige meget. Vi har fjernet variation fra, at nogle testpersoner bruger deres sko mere end andre i løbet af testperioden
- *Er det et godt design? Kan vi fjerne mere uønsket variation?*
- **Design 4:** 20 personer får et nyt par sko, hvor den ene er imprægneret med det nye produkt og den anden med det gamle. Det er tilfældigt (randomiseret) for hver person, hvilken sko (højre/venstre), der er imprægneret med hvilket produkt
- Måske slider højrehåndede personer mere på højre sko end på venstre, fordi de sparker til ting med højre fod?

Eksperimentelt design

- Der er basalt set to forskellige designs, vi skal se på:
 1. To uafhængige stikprøver
 2. To stikprøver, hvor observationerne er parvist afhængige
- Desuden skelner vi mellem, om begge stikprøver har stor stikprøvestørrelse (ca. 30 eller derover), eller ej.

To store, uafhængige stikprøver

Antagelser:

1. X_1, X_2, \dots, X_{n_1} er en tilfældig stikprøve med størrelse n_1 af population 1, som har middelværdi μ_1 og varians σ_1^2
 2. Y_1, Y_2, \dots, Y_{n_2} er en tilfældig stikprøve med størrelse n_2 af population 2, som har middelværdi μ_2 og varians σ_2^2
 3. De to stikprøver er uafhængige
- Vi er interesseret i forskellen på de to populationers middelværdi:
$$\delta = \mu_1 - \mu_2$$
 - Fra den Centrale Grænseværdidisætning følger:
 - $\bar{X} \sim N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right)$, så $E(\bar{X}) = \mu_1$ og $Var(\bar{X}) = \frac{\sigma_1^2}{n_1}$
 - $\bar{Y} \sim N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$, så $E(\bar{Y}) = \mu_2$ og $Var(\bar{Y}) = \frac{\sigma_2^2}{n_2}$
 - Desuden: $E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$ og $Var(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

To store, uafhængige stikprøver

- Dermed:

$$z_0 = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

er standard normalfordelt, når n_1 og n_2 er store

- Som regel kendes σ_1^2 og σ_2^2 ikke, men så længe n_1 og n_2 er store kan de estimeres fra stikprøven:

$$z_0 = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

Konfidensinterval for store stikprøver

$$\begin{aligned}
1 - \alpha &= P(-z_{\alpha/2} < z_0 < z_{\alpha/2}) = P\left(-z_{\alpha/2} < \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{(s_1^2)/n_1 + (s_2^2)/n_2}} < z_{\alpha/2}\right) \\
&= P\left(-z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < (\bar{x} - \bar{y}) - (\mu_1 - \mu_2) < z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right) \\
&= P\left(-(\bar{x} - \bar{y}) - z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < -(\mu_1 - \mu_2) < -(\bar{x} - \bar{y}) + z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right) \\
&= P\left(\bar{x} - \bar{y} + z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} > \mu_1 - \mu_2 > \bar{x} - \bar{y} - z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right) \\
&= P\left(\bar{x} - \bar{y} - z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < \bar{x} - \bar{y} + z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)
\end{aligned}$$

100 · (1 - α)% konfidensinterval for $\mu_1 - \mu_2$:

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

Hypotesetest for store stikprøver

- Vi er interesseret i forskellen på de to populationers middelværdi: $\delta = \mu_1 - \mu_2$. Nulhypotesen er, at forskellen har en given værdi, δ_0 , typisk $\delta_0 = 0$ (dvs. der er ingen forskel).
- Hypoteser:

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

$$H_1: \mu_1 - \mu_2 < \delta_0 \text{ eller } \mu_1 - \mu_2 > \delta_0 \text{ eller } \mu_1 - \mu_2 \neq \delta_0$$

- Teststørrelse:

Når $n_1, n_2 \geq 30$ er

$$z_0 = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

standard normalfordelt.

Eksempel 8.5: Modstand i kabler

Et kabel lavet med en ny legering påstås at kunne reducere modstanden med mere end 0.050Ω i forhold til et standard kabel.

En stikprøve på $n_1 = 32$ standard kabler havde $\bar{x} = 0.136 \Omega$ og $s_1 = 0.004 \Omega$, mens en stikprøve på $n_2 = 32$ nye kabler havde $\bar{y} = 0.083 \Omega$ og $s_2 = 0.005 \Omega$.

Kan de to stikprøver understøtte påstanden på 5 % signifikansniveau?

Løsning

1. Nulhypotese: $H_0: \mu_1 - \mu_2 = \delta_0 = 0.050$
Alternativ hypotese: $H_1: \mu_1 - \mu_2 > 0.050$
2. Signifikansniveau: $\alpha = 0.05$

3. Kriterier: Teststørrelsen $z_0 = \frac{(\bar{x} - \bar{y}) - \delta_0}{\sqrt{(s_1^2)/n_1 + (s_2^2)/n_2}}$

er standard normalfordelt, da n_1 og n_2 er store. Vi har en ensidet, højrehalet test, så vi forkaster nulhypotesen, hvis $z_0 > z_\alpha$, hvor $z_\alpha = qnorm(1 - \alpha) = 1.645$.

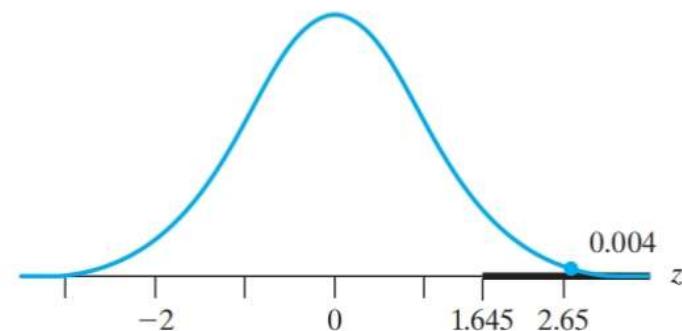
Eksempel 8.5: Modstand i kabler

4. Beregninger:

$$z_0 = \frac{(\bar{x} - \bar{y}) - \delta_0}{\sqrt{(s_1^2)/n_1 + (s_2^2)/n_2}} = \frac{0.136 - 0.083 - 0.050}{\sqrt{\frac{(0.004)^2}{32} + \frac{(0.005)^2}{32}}} = 2.65$$

5. Beslutning: Da $z_0 = 2.65 > 1.645 = z_\alpha$ forkaster vi nulhypotesen. Modstanden i de nye kabler er reduceret med mere end 0.050Ω .

P-værdien er $p = 1 - \text{pnorm}(2.65) = 0.004$, så hvis H_0 er sand vil vi se en stikprøve som denne, eller mere ekstrem i kun 4 ud af 1000 tilfælde



95 % konfidensinterval:

$$\begin{aligned}\bar{x} - \bar{y} &\pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 0.136 - 0.083 \pm 1.96 \sqrt{\frac{(0.004)^2}{32} + \frac{(0.005)^2}{32}} \\ &= 0.053 \pm 0.002 = [0.051; 0.054]\end{aligned}$$

Vi ser, at 0.050 ikke ligger i 95 % konfidensintervallet.

To små, uafhængige stikprøver

Antagelser (som før med store stikprøver):

1. X_1, X_2, \dots, X_{n_1} er en tilfældig stikprøve med størrelse n_1 af population 1, som har middelværdi μ_1 og varians σ_1^2
2. Y_1, Y_2, \dots, Y_{n_2} er en tilfældig stikprøve med størrelse n_2 af population 2, som har middelværdi μ_2 og varians σ_2^2
3. De to stikprøver er uafhængige

Yderligere antagelser:

4. Begge populationer er normalfordelte
 5. Populationerne har samme standardafvigelse, $\sigma_1 = \sigma_2 = \sigma$
-
- Heldigvis er metoden ikke så følsom overfor disse ekstra antagelser. Tommelfingerregel: Hvis det er ‘pæne’ fordelinger og den ene standardafvigelse ikke er mere end 4 gange den anden, så kan metoden bruges.

To små, uafhængige stikprøver

- Normalt er $\sigma_1 = \sigma_2 = \sigma$ ukendt, så σ eller σ^2 skal estimeres
- Både s_1^2 og s_2^2 er estimater for σ^2 . For at bruge alle observationer fra begge stikprøver beregner vi s_p^2 , som kaldes det puljede (pooled) estimat for σ^2 , ved at vægte stikprøve-varianserne med stikprøvestørrelserne:

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

- Teststørrelse:

$$t_0 = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

er t -fordelt med $\nu = n_1 + n_2 - 2$ frihedsgrader.

Konfidensinterval for små stikprøver

- $100 \cdot (1 - \alpha)\%$ konfidensinterval for $\mu_1 - \mu_2$:

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

hvor $t_{\alpha/2}$ er baseret på $\nu = n_1 + n_2 - 2$ frihedsgrader.

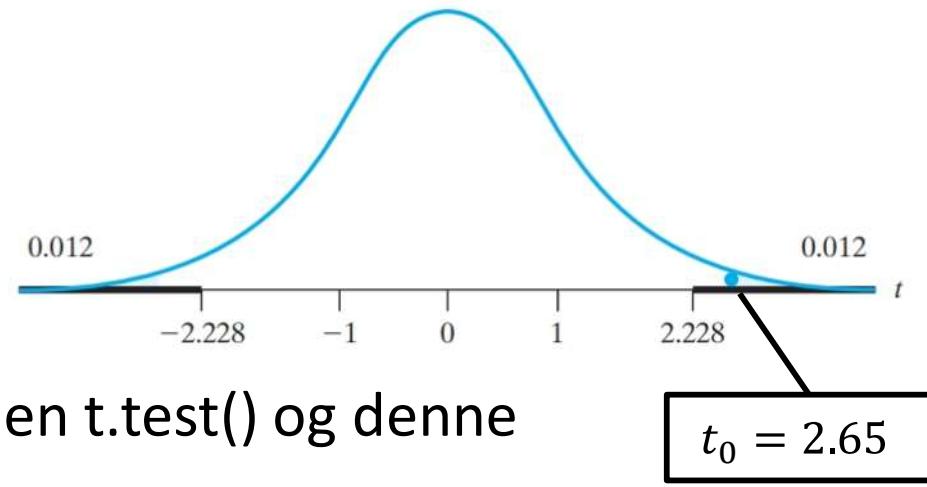
Eksempel 8.7 s. 274

- Beton kan knuses og genbruges i belægningsmateriale. Som i eksempel 7.1 måles styrken af materialet med elasticiteten, ‘modulus of resilience’. Elasticiteten af beton fra to lokaliteter sammenlignes i to små stikprøver på $n_1 = n_2 = 6$ observationer:

<i>Location 1 :</i>	707	632	604	652	669	674
<i>Location 2 :</i>	552	554	484	630	648	610

Er der forskel på elasticiteten på 5 % signifikansniveau?

- Vi regner eksemplet i R, men resultatet er, at der er forskel, og p-værdien er 2.4 %



- Opgaven kan løses med R-funktionen `t.test()` og denne ‘trylleformular’:

```
t.test(x, y, var.equal=T, mu=0, conf.level=0.95).
```

To parvist afhængige stikprøver

- Med visse eksperimentelle designs kan man (istedet for at sammenligne to uafhængige stikprøver) sammenligne to stikprøver, hvor observationerne er parvist afhængige (f.eks. højre og venstre sko med forskellig imprægnering)
- Metoden kaldes '**parret t-test**' eller '**matchede par t-test**'
- Vi har to stikprøver $X = x_1, x_2, \dots, x_n$ og $Y = y_1, y_2, \dots, y_n$, hvor hvert par af observationer (x_i, y_i) er afhængige. Derfor ser vi på forskellene: $d_i = x_i - y_i$
- Eksempler (før-og-efter-tests):
 - Personers vægt før og efter en slankekur
 - Bilens brændstofferbrug før og efter en motorrens
 - Produktionens kvalitet før og efter en justering
- Andre eksempler:
 - Måling af ansigtsgenkendelse med to forskellige algoritmer
 - BMI af mor og datter eller far og søn eller samboende partnere.

Eksempel 8.12, s. 281

- På 10 fabrikker er sikkerheden målt som det ugentlige antal mistede arbejdstimer som følge af ulykker. Der måles før og efter indførelsen af et sikkerhedsprogram. Har sikkerhedsprogrammet bevirket en forbedring af sikkerheden på 5 % signifikansniveau?
- Data:

<i>Before:</i>	45	73	46	124	33	57	83	34	26	17
<i>After:</i>	36	60	44	119	35	51	77	29	24	11

- Der er forskel på niveauet af antal mistede timer på de ti fabrikker (se f.eks. nr. 4 og nr. 10), sikkert fordi der er forskel på produktionen og på antal ansatte
- Vi ser på forskelle i antal mistede arbejdstimer før og efter:
 $d_i = x_i - y_i$: 9 13 2 5 -2 6 6 5 2 6
- Vi kan nu se bort fra de to oprindelige stikprøver og arbejde videre med $D = X - Y$.

Eksempel 8.12, s. 281

- Vi behandler $D = d_1, d_2, \dots, d_n$ som en enkelt stikprøve af en population med middelværdi $\delta = \delta_0$

9 13 2 5 -2 6 6 5 2 6

- Stikprøvens middelværdi, $\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$ er vores estimator for δ_0
- Dette er analogt til metoden fra kap. 7, hvor vi antog, at stikprøven med middelværdi \bar{x} kom fra en population med middelværdi $\mu = \mu_0$
- Stikprøvens varians s_d^2

$$s_d^2 = \frac{n \sum_{i=1}^n d_i^2 - (\sum_{i=1}^n d_i)^2}{n(n-1)}$$

- Stikprøvens standardafvigelse s_d : $s_d = \sqrt{s_d^2}$
- Teststørrelsen: $t_0 = \frac{\bar{d} - \delta_0}{s_d / \sqrt{n}}$ er t -fordelt med $n - 1$ frihedsgrader
- Hvis $n \geq 30$ kunne vi bruge at $z_0 = \frac{\bar{d} - \delta_0}{s_d / \sqrt{n}}$ er standard normalfordelt.276

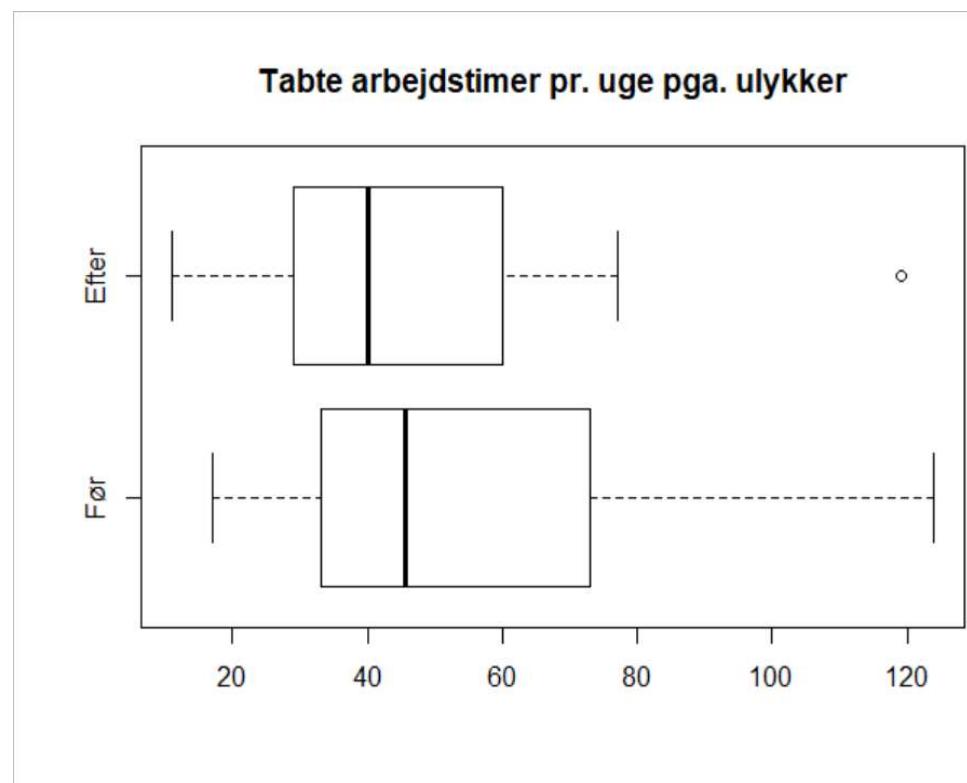
Eksempel 8.12, s. 281

Løsning:

1. Nulhypotese: $H_0: \delta = \delta_0 = 0$
Alternativ hypotese: $H_1: \delta > 0$
2. Signifikansniveau: $\alpha = 0.05$
3. Kriterier: Teststørrelsen $t_0 = \frac{\bar{d} - \delta_0}{s_d / \sqrt{n}}$ er t-fordelt med $n - 1$ frihedsgrader. Vi har ensidet test med højrehale, så vi forkaster H_0 , hvis $t_0 > t_\alpha$, hvor $t_\alpha = qt(1 - \alpha, n - 1) = 1.833$
4. Beregninger: $\bar{d} = 5.2$, $s_d = 4.08$, $n = 10$, så
$$t_0 = \frac{5.2 - 0}{4.08 / \sqrt{10}} = 4.03$$
5. Konklusion: Vi forkaster nulhypotesen, da $t_0 = 4.03 > 1.83 = t_\alpha$
P-værdien er 0.0015.

Eksempel 8.12, s. 281

- Parallelt boxplot af de to stikprøver viser stor variation mellem fabrikker. Medianerne er ikke langt fra hinanden
- Hvis vi ikke udnyttede, at observationerne er parvist afhængige, men i stedet brugte metoden til sammenligning af to stikprøver, så ville vi ikke kunne forkaste H_0 . Vi ville ikke kunne vise en effekt af sikkerhedspakken
- P-værdi for denne test er 0.34 (mod 0.0015 for parret t-test).



Sandsynlighedsteori og statistik

Kapitel 9. Inferens med varianser (afsnit 9.1-9.3 (alt))

Allan Leck Jensen

alj@ece.au.dk

Estimering af varianser

- Vi har en stikprøve $X = x_1, x_2, \dots, x_n$ taget fra en population med middelværdi μ og varians σ^2
- Vi kender ikke værdierne af μ og σ^2 , men vi vil gerne estimere dem ud fra stikprøven
- I kap. 7 og 8 har vi estimeret μ fra en og to stikprøver.
Vi kan tilsvarende estimere σ^2 og σ ud fra stikprøver på tre måder:
 - Punkt-estimering
 - Interval-estimering
 - Hypotesetest.

Punkt-estimering af varians

- I kap. 7 så vi, at stikprøvemiddelværdien

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

er en unbiased estimator for μ

- Man kan tilsvarende vise, at stikprøvevariansen

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

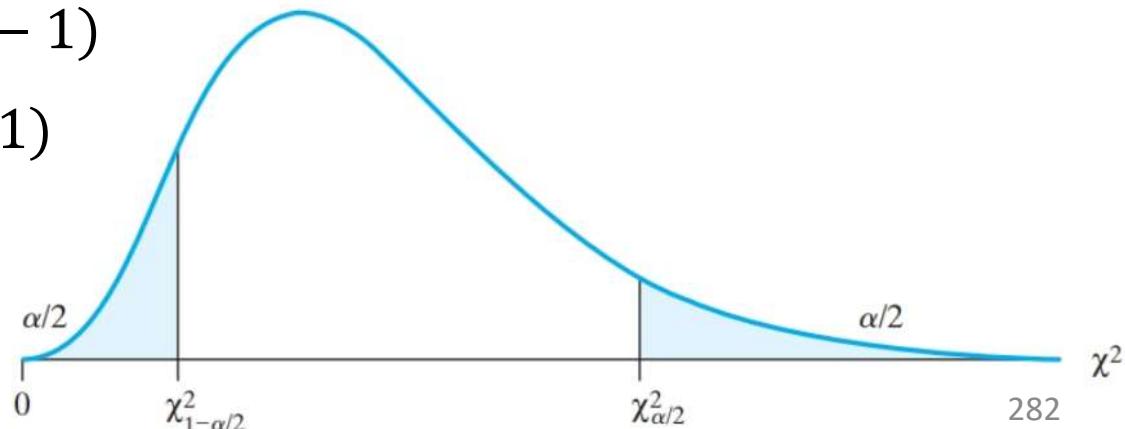
er en unbiased estimator for σ^2

- Desværre ved vi jo, at vi ikke estimerer korrekt, for $P(\sigma^2 = s^2) = 0$.
Derfor kan en interval-estimering være mere relevant.

Interval-estimering af varians

I kapitel 6 så vi sætning 6.5 (s. 207):

- Hvis s^2 er variansen for en stikprøve med størrelse n , taget fra en normalfordelt population med middelværdi μ og varians σ^2 , så er
$$\chi_0^2 = \frac{(n-1)s^2}{\sigma^2}$$
 χ^2 (chi-i-anden) fordelt med $\nu = n - 1$ frihedsgrader
- Det kan vi bruge til at beregne et $100(1 - \alpha)$ % konfidensinterval for σ^2
- Figuren viser χ^2 fordelingen med $n - 1$ frihedsgrader. De to værdier $\chi_{1-\alpha/2}^2$ og $\chi_{\alpha/2}^2$ er beregnet, så de blå arealer under kurvens venstre og højre haler begge er $\alpha/2$:
 - $\chi_{\alpha/2}^2 = \text{qchisq}(1 - \alpha/2, n - 1)$
 - $\chi_{1-\alpha/2}^2 = \text{qchisq}(\alpha/2, n - 1)$
- Dermed er det hvide areal imellem værdierne lig med $1 - \alpha$.



Beregning af konfidensinterval

- Sandsynligheden for at teststørrelsen χ_0^2 ligger i det hvide område:

$$P\left(\chi_{1-\alpha/2}^2 < \frac{(n-1)s^2}{\sigma^2} < \chi_{\alpha/2}^2\right) = 1 - \alpha$$

- Når vi tager det reciprokke og vender ulighedstegnene gælder det stadig:

$$P\left(\frac{1}{\chi_{1-\alpha/2}^2} > \frac{\sigma^2}{(n-1)s^2} > \frac{1}{\chi_{\alpha/2}^2}\right) = 1 - \alpha$$

- Vi vender 'læseretningen':

$$P\left(\frac{1}{\chi_{\alpha/2}^2} < \frac{\sigma^2}{(n-1)s^2} < \frac{1}{\chi_{1-\alpha/2}^2}\right) = 1 - \alpha$$

Beregning af konfidensinterval

- (overført)

$$P\left(\frac{1}{\chi^2_{\alpha/2}} < \frac{\sigma^2}{(n-1)s^2} < \frac{1}{\chi^2_{1-\alpha/2}}\right) = 1 - \alpha$$

- Vi ganger alle led med $(n-1)s^2$ (som er positivt):

$$P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right) = 1 - \alpha$$

- Dermed har vi beregnet $(1 - \alpha) \cdot 100\%$ konfidensintervallet for σ^2 :

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

(bemærk at den høje værdi for kritisk område bruges til den lave værdi for konfidensintervallet (fordi der divideres med den) – og omvendt).

Eksempel 9.2 s. 292 (data s. 145)

I en osteproduktion tages en stikprøve med vægten af $n = 80$ oste. Man ønsker at bruge stikprøven med $\bar{x} = 68.45$ og $s = 9.583$ pund til at lave et 95% konfidensinterval for produktionens standardafvigelse σ . Vægten af de producerede oste kan antages at være normalfordelte

Løsning:

- $\chi^2_{\alpha/2} = \text{qchisq}(1 - \alpha/2, n - 1) = \text{qchisq}(0.975, 79) = 105.473$
- $\chi^2_{1-\alpha/2} = \text{qchisq}(\alpha/2, n - 1) = \text{qchisq}(0.025, 79) = 56.309$
- $\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$
- $\frac{(79)(9.583)^2}{105.473} < \sigma^2 < \frac{(79)(9.583)^2}{56.309}$
- $68.8 < \sigma^2 < 128.8$
- Vi tager kvadratroden, for at få 95 % konfidensinterval for σ
 $8.29 < \sigma < 11.35$.

Hypotesetest om varians med 1 stikprøve

- Vi skal bruge stikprøven til at undersøge påstanden om, at populationens varians har en given værdi, σ_0^2 . Dermed er nulhypotesen $H_0: \sigma^2 = \sigma_0^2$
- Alternativhypotesen H_1 er, afhængig af situationen, en af mulighederne:
 - $\sigma^2 < \sigma_0^2$
 - $\sigma^2 > \sigma_0^2$
 - $\sigma^2 \neq \sigma_0^2$
- Teststørrelsen følger af sætning 6.5:

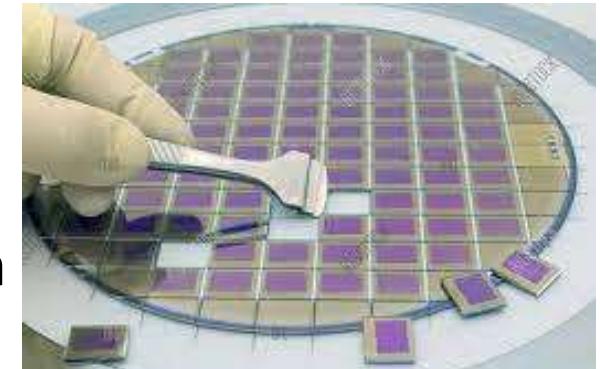
$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

er χ^2 fordelt med $n - 1$ frihedsgrader

- Vi må antage, at data er normalfordelte.

Eksempel 9.3, s. 294

- Siliciumskiver (silicon wafers) er ekstremt tynde skiver af halvledermateriale, der bl.a. bruges til integrerede kredsløb og solceller
- I en bestemt produktion skal standardafvigelsen være højst 0.50 mil
(1 mil = 1/1000 inch = 0.00254 cm)
- En stikprøve på $n = 15$ skiver har standardafvigelse $s = 0.64$ mil. Beslut på 5 % signifikansniveau, om produktionen skal kasseres?



Løsning:

1. Hypoteser:

$$H_0: \sigma = \sigma_0 = 0.50$$

$$H_1: \sigma > 0.50$$

2. Signifikansniveau: $\alpha = 0.05$

Eksempel 9.3, s. 294

3. **Kriterier:** Teststørrelsen $\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2}$ er χ^2 fordelt med $n - 1$ frihedsgrader. Det er en ensidet test med højrehale, så vi forkaster H_0 , hvis $\chi_0^2 > \chi_{0.05}^2 = \text{qchisq}(0.95, 15 - 1) = 23.685$

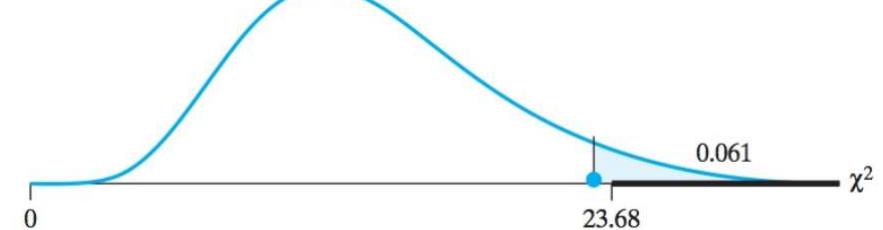
4. **Beregninger:**

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(15-1)(0.64)^2}{(0.50)^2} = 22.94$$

5. **Konklusion:**

Da $\chi_0^2 = 22.94 < 23.685 = \chi_{0.05}^2$ kan vi *ikke* forkaste nulhypotesen. Selv om stikprøvens standardafvigelse var større end 0.50 mil, så er der ikke stærkt nok bevis på 5 % signifikansniveau til at konkludere, at produktionens tolerance er overskredet.

$$\begin{aligned} \text{P-værdi: } & 1 - \text{pchisq}(22.94, 14) \\ & = 0.061 \end{aligned}$$



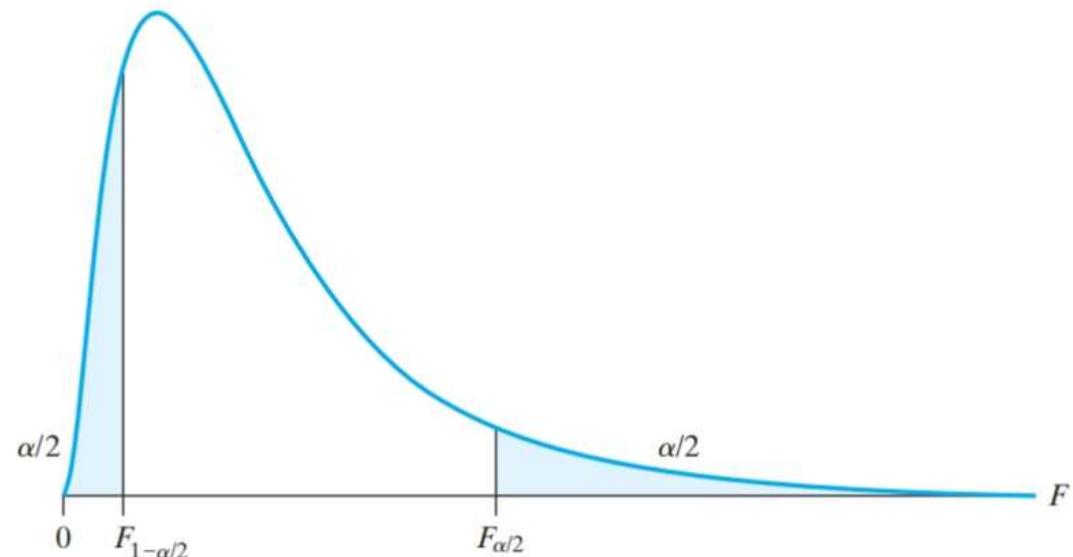
NB. Vi burde teste antagelsen om at data er normalfordelte.

Estimering af varians for 2 stikprøver

- Vi har to uafhængige stikprøver med størrelse hhv. n_1 og n_2 og med stikprøvevariанс hhv. s_1^2 og s_2^2 . Vi antager, at stikprøverne kommer fra to normalfordelte populationer med varians hhv. σ_1^2 og σ_2^2
- $\frac{(n_1-1)s_1^2}{\sigma_1^2}$ er χ^2 fordelt med $n_1 - 1$ frihedsgrader
- $\frac{(n_2-1)s_2^2}{\sigma_2^2}$ er χ^2 fordelt med $n_2 - 1$ frihedsgrader
- Sætning 6.6 (s. 209):

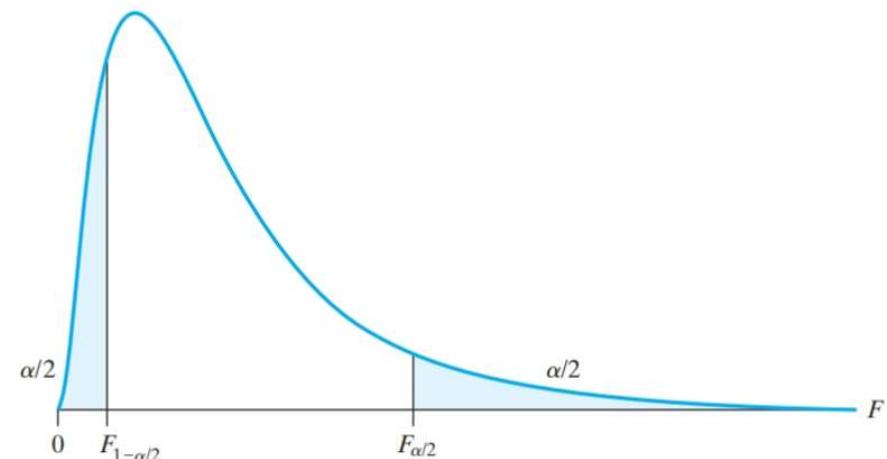
$$F_0 = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

er F fordelt med $n_1 - 1$ frihedsgrader i tælleren og $n_2 - 1$ frihedsgrader i nævneren.



Konf.interval for forhold mel. 2 varianser

- Figuren viser $F(n_1 - 1, n_2 - 1)$, d.v.s. F fordelingen med $n_1 - 1$ frihedsgader i tælleren og $n_2 - 1$ frihedsgrader i nævneren
- De to værdier $F_{1-\alpha/2}$ og $F_{\alpha/2}$ er beregnet, så de blå arealer under kurvens venstre og højre haler begge er $\alpha/2$:
 - $F_{\alpha/2} = \text{qf}(1 - \alpha/2, n_1 - 1, n_2 - 1)$
 - $F_{1-\alpha/2} = \text{qf}(\alpha/2, n_1 - 1, n_2 - 1)$
- Dermed er det hvide areal imellem værdierne lig med $1 - \alpha$.



Konf.intervall for forhold mel. 2 varianser

- $1 - \alpha = P\left(F_{1-\alpha/2} < \frac{(s_1^2)/(\sigma_1^2)}{(s_2^2)/(\sigma_2^2)} < F_{\alpha/2}\right)$
 $= P\left(F_{1-\alpha/2} \cdot \frac{s_2^2}{s_1^2} < \frac{\sigma_2^2}{\sigma_1^2} < F_{\alpha/2} \cdot \frac{s_2^2}{s_1^2}\right)$
- $100(1 - \alpha)\%$ konfidensinteval for $(\sigma_2^2)/(\sigma_1^2)$ for 2 normalfordelte populationer:

$$F_{1-\alpha/2}(n_1 - 1, n_2 - 1) \frac{s_2^2}{s_1^2} < \frac{\sigma_2^2}{\sigma_1^2} < F_{\alpha/2}(n_1 - 1, n_2 - 1) \frac{s_2^2}{s_1^2}$$

Eksempel 9.6, s. 298 (data s. 277)

Produktion af biobrændstof af sukker med 2 katalysatorer

Catalyst 1	0.63	2.64	1.85	1.68	1.09	1.67	0.73	1.04	0.68
Catalyst 2	3.71	4.09	4.11	3.75	3.49	3.27	3.72	3.49	4.26

Vi har $n_1 = n_2 = 9$, $s_1^2 = 0.4548$, $s_2^2 = 0.1089$.

Beregn et 98 % konfidensinterval for $\frac{\sigma_2^2}{\sigma_1^2}$

Løsning:

- Der er 8 frihedsgrader i både tæller og nævner. $\alpha/2 = 0.01$
- $F_{\alpha/2} = F_{0.01} = \text{qf}(0.99, 8, 8) = 6.029$
- $F_{1-\alpha/2} = F_{0.99} = \text{qf}(0.01, 8, 8) = 0.166$
- 98 % konfidensinterval for $\frac{\sigma_2^2}{\sigma_1^2}$ er:
$$\left[F_{1-\alpha/2} \frac{s_2^2}{s_1^2}; F_{\alpha/2} \frac{s_2^2}{s_1^2} \right] = \left[0.166 \frac{0.1089}{0.4548}; 6.029 \frac{0.1089}{0.4548} \right] = [0.04; 1.44]$$
- Det kan ikke udelukkes, at forholdet mellem de to varianser er 1.

Hypotesetest om varians med 2 stikprøver

- Denne type hypotesetest undersøger, om to uafhængige stikprøver kommer fra populationer med samme varians
- Vi skal antage at de to populationer er normalfordelte
- Så følger det af sætning 6.6 s. 209, at teststørrelsen

$$F_0 = \frac{s_1^2}{s_2^2}$$

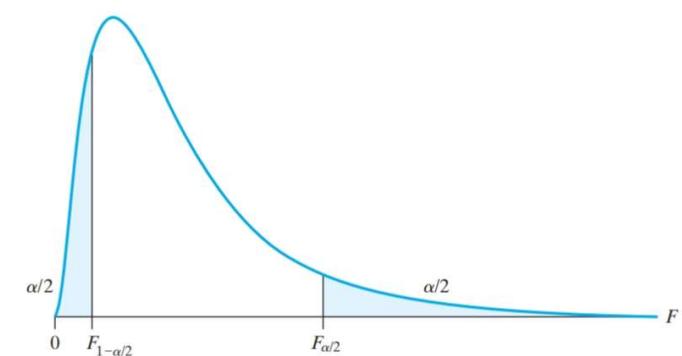
er F -fordelt med $n_1 - 1$ frihedsgrader i tælleren og $n_2 - 1$ frihedsgrader i nævneren

- Nulhypotesen er $\sigma_1^2 = \sigma_2^2$. Så forventer vi at $F_0 \approx 1$

- Alternativhypotesen:

Enten $\sigma_1^2 > \sigma_2^2$, $\sigma_1^2 < \sigma_2^2$ eller $\sigma_1^2 \neq \sigma_2^2$.

- Har vi f.eks. $H_1: \sigma_1^2 > \sigma_2^2$, så kan det bekræftes, hvis $s_1^2 \gg s_2^2$, så $F_0 \gg 1$
- Har vi $H_1: \sigma_1^2 < \sigma_2^2$, så kan det bekræftes, hvis $s_1^2 \ll s_2^2$, så $F_0 \ll 1$.



Hypotesetest om varians med 2 stikprøver

- **Fiktivt eksempel:** Vi har to stikprøver: Nr. 1 har $n_1 = 10$ og $s_1^2 = 4$, nr. 2 har $n_2 = 12$ og $s_2^2 = 8$. Så er $F_0 = \frac{4}{8} = 0.5$ F -fordelt med d.f. 9 i tæller og 11 i nævner: $F(9,11)$
- Hvis vi havde nummereret stikprøverne omvendt, så ville vi få, at $F_0 = \frac{8}{4} = 2.0$ er F -fordelt med d.f. 11 i tæller og 9 i nævner: $F(11,9)$
- Generelt for givet α , n_1 og n_2 :

$$F_{1-\alpha}(n_1 - 1, n_2 - 1) = \frac{1}{F_\alpha(n_2 - 1, n_1 - 1)}$$

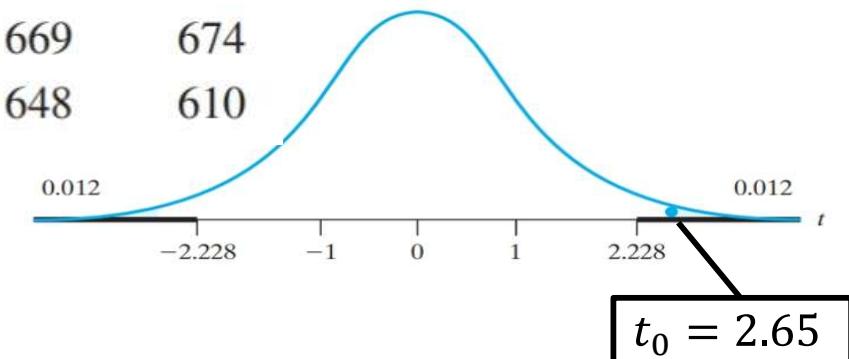
- Vi kan vælge den stikprøve med størst stikprøvevariанс som nr. 1, så vi altid opnår, at $F_0 = \frac{s_1^2}{s_2^2} \geq 1$. Kriterier for at forkaste nulhypotesen:
 - $H_1: \sigma_1^2 > \sigma_2^2$: H_0 forkastes hvis $F_0 > F_\alpha(n_1 - 1, n_2 - 1)$
 - $H_1: \sigma_1^2 < \sigma_2^2$: H_0 forkastes altid, for $F_0 > 1$
 - $H_1: \sigma_1^2 \neq \sigma_2^2$: H_0 forkastes hvis $F_0 > F_{\alpha/2}(n_1 - 1, n_2 - 1)$.

Eksempel 9.5, s. 296

- I eksempel 8.7 s. 274 så vi på to stikprøver med elasticiteten af genbrugt beton til belægning. Vi forkastede nulhypotesen om, at de kom fra populationer med ens middelværdi

Location 1 : 707 632 604 652 669 674

Location 2 : 552 554 484 630 648 610



- Brug 2 % signifikansniveau til at undersøge, om varianserne er ens

Løsning:

Stikprøvevariansenene for lokation 1 og 2 beregnes til hhv. 1277.87 og 3739.87. Lokation 2 har altså den største stikprøvevarians, så den får nr. 1

1. Hypoteser:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

2. Signifikansniveau: $\alpha = 0.02$.

Eksempel 9.5, s. 296

3. Kriterier: Teststørrelsen $F_0 = \frac{s_1^2}{s_2^2}$ er F -fordelt med $n_1 - 1$ frihedsgrader i tælleren og $n_2 - 1$ frihedsgrader i nævneren. Vi forkaster H_0 , hvis

$$\begin{aligned} F_0 &> F_{\alpha/2}(n_1 - 1, n_2 - 1) = F_{0.01}(5, 5) \\ &= \text{qf}(0.99, 5, 5) = 10.97 \end{aligned}$$

4. Beregninger:

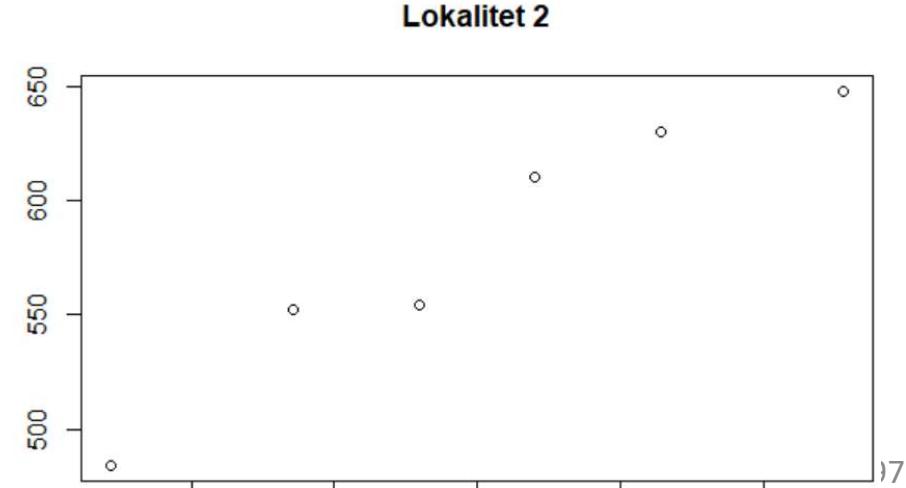
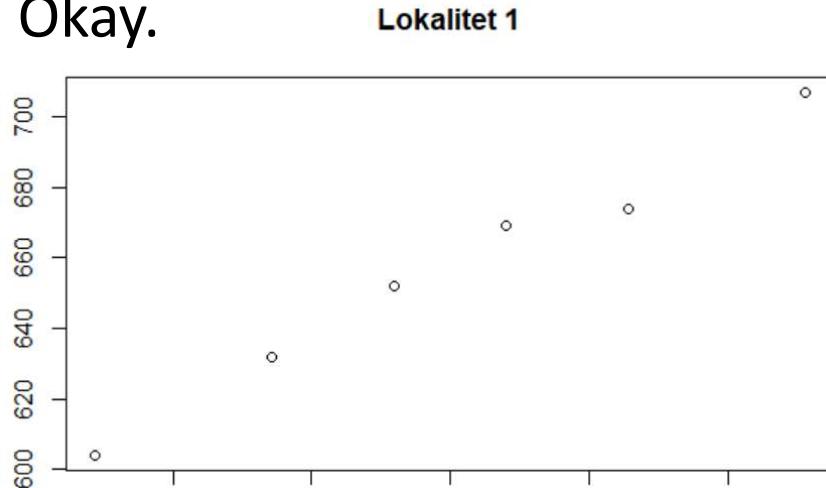
$$F_0 = \frac{s_1^2}{s_2^2} = \frac{3739.87}{1277.87} = 2.93$$

5. Konklusion: Da $F_0 = 2.93 < 10.97 = F_{\alpha/2}$ kan vi ikke forkaste nulhypotesen. De to lokaliteter kan have samme varians.

Eksempel 9.5, s. 296

Bemærk:

- I metoden fra kap. 8 til hypotesetest om forskel på to populationers **middelværdi** måtte vi antage, at populationerne var normalfordelte med ens varians, når stikprøvestørrelsen var lille
- Heldigvis er metoden ikke særlig følsom for antagelserne
- I metoden her til hypotesetest om forskel på to populationers **variанс** måtte vi også antage, at populationerne var normalfordelte, også for store stikprøvestørrelser
- Metoden er følsom for antagelserne, så vi tester med normalford.plot:
Okay.



Sandsynlighedsteori og statistik

Kapitel 10. Chi-i-anden tests (afsnit 10.4-10.5)

Allan Leck Jensen
alj@ece.au.dk

2 metoder til hypotesetest med χ^2

- Kontingenstabeller
Hypotesetest om to kategoriseringer af observationer er uafhængige
- Goodness of Fit
Hypotesetest om observationer kommer fra en bestemt fordeling.

Kontingenstabeller

- Et eksempel fra statistikbanken.dk (Danmarks Statistik):

25 - 45 åriges videregående uddannelse efter tid, uddannelsesstatus og køn

	Mænd	Kvinder
2019		
FULDFØRT VIDEREGÅENDE UDDANNELSE	247 919	347 337
IGANGVÆRENDE VIDEREGÅENDE UDDANNELSE	29 888	35 149
INGEN VIDEREGÅENDE UDDANNELSE	496 830	372 572

- Her vises uddannelsesstatus for 25-45 årige, fordelt på køn
- Data er præsenteret i en tabel med tre rækker og to søjler. Det kaldes en 3×2 tabel
- Generelt præsenterer en kontingenstabell to egenskaber i en tabel med r rækker og c søjler: en $r \times c$ tabel
- Er de to egenskaber uafhængige af hinanden? Er uddannelsesstatus f.eks. uafhængigt af køn?

Kontingenstabeller

- Et andet eksempel på kontingenstabell:
Er venstre-håndthed ligeså udbredt blandt kvinder som mænd?
M.a.o.: Er venstre-håndthed uafhængigt af køn?

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Totals	87	13	100

Fra Kap. 3: Uafhængighed

To hændelser A og B er uafhængige, hvis

$$P(A | B) = P(A)$$

Med andre ord: Information om B ændrer ikke vores forventning af A

Det følger af ligningen for betinget sandsynlighed, at

$$P(A | B) = \frac{P(A \cap B)}{P(B)} \Rightarrow$$

$$P(A \cap B) = P(A | B) \cdot P(B) = P(A) \cdot P(B)$$

når A og B er uafhængige.

Med andre ord kan vi beregne sandsynligheden for fælleshændelsen af A og B som produktet af sandsynligheden for enkelthændelserne.

For eksempel: A er kast med mønt, B er kast med terning.

$$P(\text{Plat og 6}) = P(\text{Plat}) \cdot P(6) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}.$$

Kontingenstabeller

- Hændelse A: Køn = ‘mand’. $P(A) = 52/100 = 0.52$
Hændelse B: Hånd = ‘højre’ $P(B) = 87/100 = 0.87$
- Hvis foretrukken hånd er uafhængig af køn, er
 $P(A \cap B) = P(A) \cdot P(B) = 0.52 \cdot 0.87 = 0.4524$
- Vi ville altså forvente, at 45 ud af 100 personer var højrehåndede mænd, hvis de to egenskaber er uafhængige. Vi så 43 i stikprøven
- Vi ville forvente, at $100 \cdot 0.48 \cdot 0.13 \cong 6$ ud af 100 personer var venstrehåndede kvinder, men vi så 4
- Tilfældigt?

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Totals	87	13	100

Kontingenstabeller generelt

- En stikprøve på n observationer af en population beskrives med to egenskaber
- Den første egenskab har r kategorier (rækker) og den anden egenskab har c kategorier (kolonner)
- o_{ij} angiver antal observationer i den i 'te kategori af egenskab 1 og den j 'te kategori af egenskab 2
- Bemærk: $n = \sum_{i=1}^r \sum_{j=1}^c o_{ij}$.

		Columns				
		1	2	...	c	
		O_{11}	O_{12}	...	O_{1c}	
Rows	1	O_{11}	O_{12}	...	O_{1c}	
	2	O_{21}	O_{22}	...	O_{2c}	
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
	r	O_{r1}	O_{r2}	...	O_{rc}	

Kontingenstabeller generelt

- Vi vil beregne det forventede (expected) antal observationer i hver tabelcelle, e_{ij} , og sammenligne det med det observerede, o_{ij}
- Hvis de to egenskaber (A og B) er uafhængige er

$$P(A \cap B) = P(A)P(B) \Leftrightarrow$$

$$P((A = i) \cap (B = j)) = P(A = i) \cdot P(B = j) \text{ for alle } i, j$$

- Det kan vi også skrive som

$$p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$$

- Vi estimerer $p_{i\cdot} = P(A = i)$:

$$p_{i\cdot} = \frac{1}{n} \sum_{j=1}^c o_{ij}$$

- Tilsvarende estimerer vi $p_{\cdot j} = P(B = j)$:

$$p_{\cdot j} = \frac{1}{n} \sum_{i=1}^r o_{ij}$$

- Dermed er det forventede antal i hver celle:

$$e_{ij} = n \cdot p_{ij} = n \cdot p_{i\cdot} \cdot p_{\cdot j} = \frac{1}{n} \cdot \sum_{j=1}^c o_{ij} \cdot \sum_{i=1}^r o_{ij} .$$

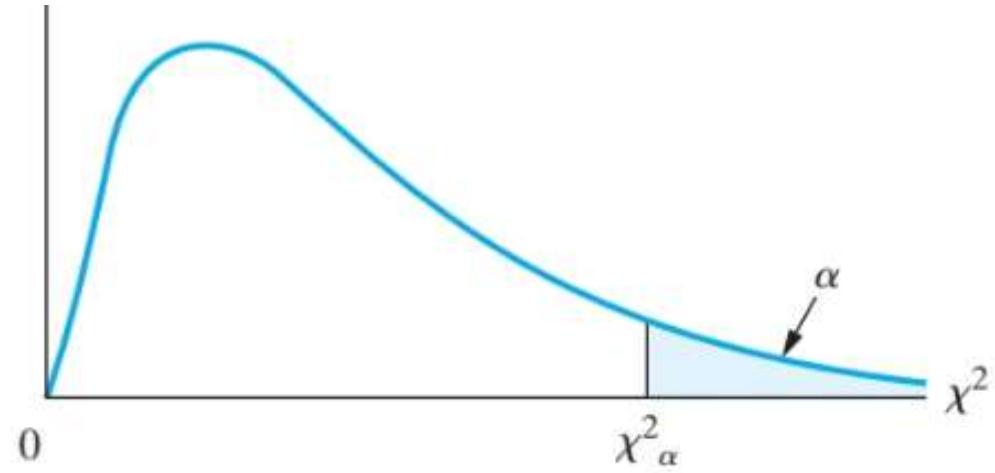
		Columns				
		1	2	...	c	
Rows	1	O_{11}	O_{12}	...	O_{1c}	
	2	O_{21}	O_{22}	...	O_{2c}	
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
		r	O_{r1}	O_{r2}	...	O_{rc}

Chi-i-anden test for kontingenstabeller

- Nulhypotesen er, at de to egenskaber er uafhængige, d.v.s.
$$H_0: p_{ij} = p_{i \cdot} \cdot p_{\cdot j} \text{ for alle } i, j$$
- Alternativhypotesen er, at de to egenkaber ikke er uafhængige
- Teststørrelsen for hypotesetesten sammenligner de observerede o_{ij} med de forventede e_{ij} :

$$\chi^2_0 = \sum_{i=1}^r \sum_{j=1}^c \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- Teststørrelsen er χ^2 fordelt med $(r - 1)(c - 1)$ frihedsgrader
- Bemærk $\chi^2_0 \geq 0$, og jo tættere på 0, teststørrelsen er, desto mere tror vi på nulhypotesen
- Vi forkaster H_0 , hvis $\chi^2_0 > \chi^2_\alpha$ så altid ensidet, højrehalet test.



Eksempel 10.13 og 10.14, s. 320

Et firma sætter nyansatte i et træningsprogram ved ansættelsen. De vil gerne undersøge, om der er sammenhæng mellem, hvor godt de ansatte klarer sig i træningsprogrammet og hvor godt, de senere klarer sig i jobbet

Lav en chi-i-anden test med signifikans-niveau 0.01, om der er sammenhæng

	<i>Success in job (employer's rating)</i>	<i>Performance in training program</i>			<i>Total</i>
		<i>Below average</i>	<i>Average</i>	<i>Above average</i>	
<i>Poor</i>	23	60	29	112	
<i>Average</i>	28	79	60	167	
<i>Very good</i>	9	49	63	121	
<i>Total</i>	60		188	152	400

Løsning:

1. Hypoteser:

H_0 : Præstation i træningsprogram og succes i jobbet er uafhængige

H_1 : Præstation i træningsprogram og succes i jobbet er afhængige

2. Signifikansniveau: $\alpha = 0.01$.

Eksempel 10.13 og 10.14, s. 320

3. **Kriterier:** Teststørrelsen $\chi^2_0 = \sum_{i=1}^r \sum_{j=1}^c \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$ er χ^2 fordelt med $(r - 1)(c - 1)$ frihedsgrader. Her er $r = c = 3$, så $v = 4$.

Vi forkaster H_0 , hvis $\chi^2_0 > \chi^2_\alpha$:

$$\begin{aligned}\chi^2_\alpha &= \text{qchisq}(0.99, 2 \cdot 2) \\ &= 13.277\end{aligned}$$

23	60	29	112
28	79	60	167
9	49	63	121
60	188	152	400

4. **Beregninger:** Vi beregner $e_{ij} = \frac{1}{n} \sum_{j=1}^c o_{ij} \sum_{i=1}^r o_{ij}$:

$$e_{11} = \frac{112 \cdot 60}{400} = 16.80 \quad e_{12} = \frac{112 \cdot 188}{400} = 52.64 \quad e_{13} = \frac{112 \cdot 152}{400} = 42.56$$

$$e_{21} = \frac{167 \cdot 60}{400} = 25.05 \quad e_{22} = \frac{167 \cdot 188}{400} = 78.49 \quad e_{23} = \frac{167 \cdot 152}{400} = 63.46$$

$$e_{31} = \frac{121 \cdot 60}{400} = 18.15 \quad e_{32} = \frac{121 \cdot 188}{400} = 56.87 \quad e_{33} = \frac{121 \cdot 152}{400} = 45.98$$

Eksempel 10.13 og 10.14, s. 320

4. Beregninger (fortsat):

Her er de forventede
antal vist i fed

Vi beregner test-
størrelsen:

$$\chi^2_0 = \frac{(23-16.80)^2}{16.80} + \frac{(60-52.64)^2}{52.64} + \dots \\ \dots + \frac{(63-45.98)^2}{45.98} = 20.179$$

		Performance in training program			Total
		Below average	Average	Above average	
Success in job (employer's rating)	Poor	23	60	29	112
	Average	16.80	52.64	42.56	167
	Very good	28	79	60	121
Total		25.05	78.49	63.46	400
		9	49	63	
		18.15	56.87	45.98	

5. Konklusion: Vi forkaster nulhypotesen, da

$$\chi^2_0 = 20.179 > 13.277 = \chi^2_{\alpha}$$

Succes i jobbet afhænger altså af, hvor godt man har præsteret i træningsprogrammet.

Eksempel 10.13 og 10.14, s. 320

- Vi kan undersøge mønsteret for afhængigheden mellem de to egenskaber ved at se på hver celles bidrag til teststørrelsen

2.288	1.029	4.320
0.347	0.003	0.189
4.613	1.089	6.300

		Performance in training program			Total
		Below average	Average	Above average	
Success in job (employer's rating)	Poor	23	60	29	112
	Average	16.80	52.64	42.56	167
	Very good	28	79	60	121
		25.05	78.49	63.46	
		9	49	63	
		18.15	56.87	45.98	
Total		60	188	152	400

- Hvert bidrag $(o_{ij} - e_{ij})^2 / e_{ij}$ er farvelagt efter størrelse på en skala fra grøn til rød. De største (rødeste) værdier er i hjørnerne
- Det største bidrag til χ^2_0 er 6.300. Det kommer fordi flere end forventet af dem, der har klaret sig 'Above average' i træningsprogrammet, klarer sig 'Very good' i jobbet. Der er 63 mod forventet 45.98
- Næststørste bidrag, 4.613, skyldes at der kun er 9 mod forventet 18.15, der har klaret sig 'Below average' i træning, men 'Very good' i job
- Træningsprogrammet virker efter hensigten.

Goodness of Fit – Eksempel

- Bogens eksempel 10.15 om Goodness of Fit er temmeligt omfattende, så jeg giver et simplere eksempel:
- I et programmeringskursus har 60 ingenørstuderende afleveret et computerprogram. Underviseren laver en opgørelse over antal fejl i de $n = 60$ programmer:

Antal fejl	Antal programmer
0	32
1	15
2	9
3	4

- Underviseren har en formodning om at antal fejl i programmet følger en Poisson fordeling.

Fra Kap. 4: Poisson-fordelingen

- En fabrik producerer i gennemsnit 4.2 defekte produkter om dagen. Hvad er sandsynligheden for at den producerer præcis 7 defekte i morgen?
- Poisson-fordelingen bruges, når man tæller antal ‘succes’er’ (x) og kender det forventede antal pr. enhed eller tidsrum (λ)
- Sandsynlighedsfunktionen for Poisson-fordelingen er:

$$f(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \quad \text{for } x = 0, 1, 2, \dots \text{ og } \lambda > 0$$

- Bemærk, at x ikke har en øvre grænse, i modsætning til binomialfordelingen, hvor $0 \leq x \leq n$
- Eksemplet: $f(7; 4.2) = \frac{(4.2)^7}{7!} e^{-(4.2)} = 0.0686$
- Middelværdi, varians og standardafvigelse for Poisson-fordelingen:

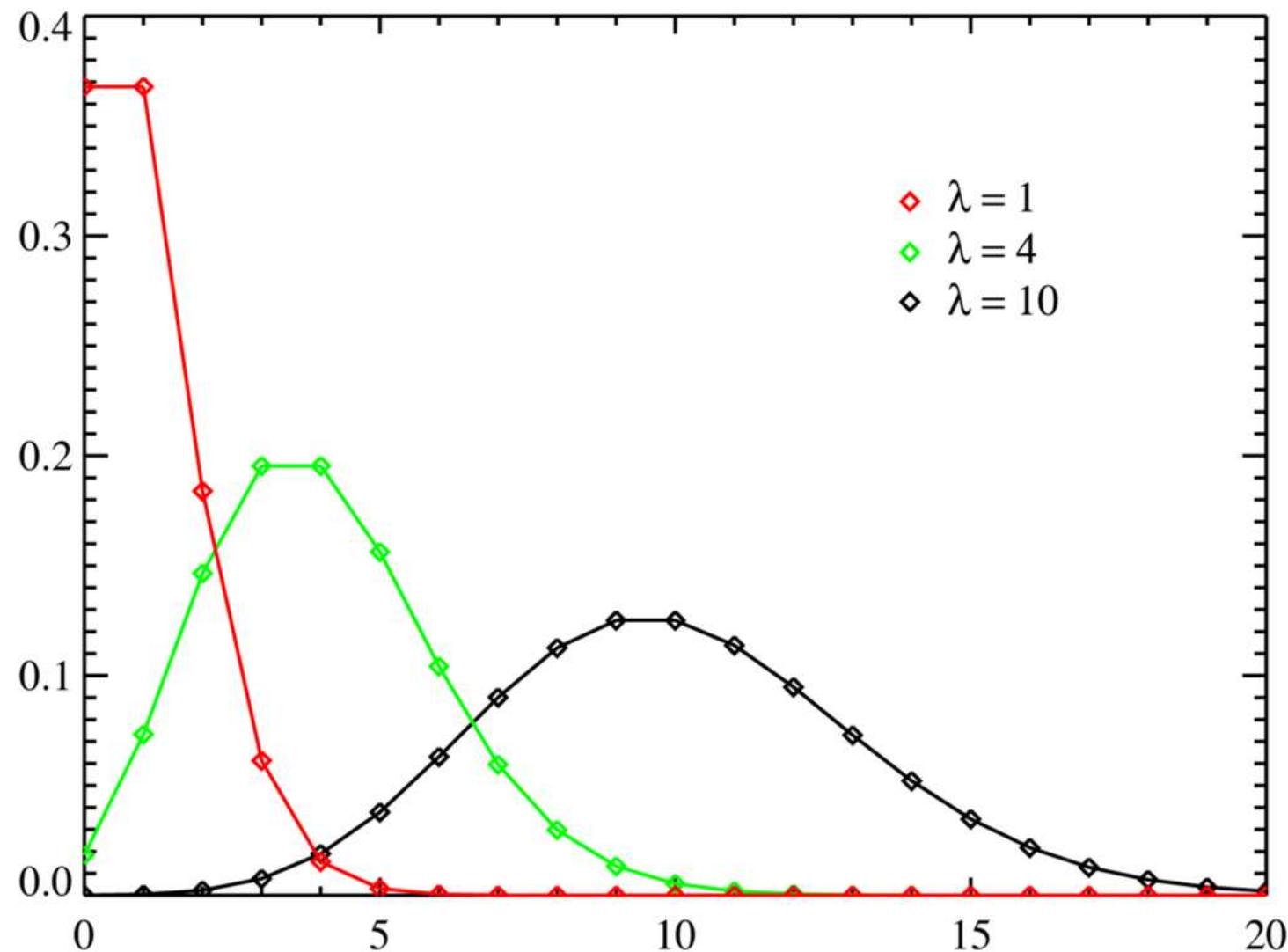
$$\mu = E(X) = \lambda$$

$$\sigma^2 = Var(X) = \lambda$$

$$\sigma = \sqrt{\lambda}.$$

Fra Kap. 4: Poisson-fordelingen

Afhængighed af λ :



Goodness of Fit – Eksempel

- En stikprøve på $n = 60$ programmer. Opgørelse over antal programmer efter antal fejl i programmerne:

Antal fejl	Antal programmer
0	32
1	15
2	9
3	4

- Man formoder, at antal fejl pr. program følger en Poisson fordeling
- Vi kender ikke gennemsnitligt antal fejl per program, λ , men vi kan estimere den:
$$\lambda = \frac{32 \cdot 0 + 15 \cdot 1 + 9 \cdot 2 + 4 \cdot 3}{60} = 0.75$$
- Nu kan vi beregne det forventede antal programmer med hhv. 0, 1, 2, 3 fejl, givet at antagelsen om Poisson fordeling er rigtig.

Goodness of Fit – Eksempel

- Sandsynligheden for, at et program har x fejl, hvis den kommer fra en Poissonfordelt stokastisk variabel X er:

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

- Vi har estimeret, at $\lambda = 0.75$, så:

$$- P(X = 0) = \frac{(0.75)^0}{0!} e^{-0.75} = 0.472$$

$$- P(X = 1) = \frac{(0.75)^1}{1!} e^{-0.75} = 0.354$$

$$- P(X = 2) = \frac{(0.75)^2}{2!} e^{-0.75} = 0.133$$

$$- P(X \geq 3) = 1 - 0.472 - 0.354 - 0.133 = 0.041$$

- Vi kan beregne det forventede antal ved at gange disse sandsynligheder med $n = 60$.

Goodness of Fit – Eksempel

Antal fejl	Sandsynlighed	Forventet antal	Observeret antal
0	0.472	28.34	32
1	0.354	21.26	15
2	0.133	7.97	9
3 (eller flere)	0.041	2.43	4

- Forventet antal (e_i) er sammenligneligt med observeret antal (o_i), men vi har brug for statistik for at afgøre, om observationerne følger Poisson fordelingen.

Goodness of Fit test generelt

- Vi har en stikprøve på n observationer, organiseret i k kategorier, så vi har antal observationer i hver kategori, o_i for $i = 1, 2 \dots, k$
- Vi har en formodning om at observationerne kommer fra en bestemt teoretisk fordeling (i eksemplet var det Poisson). Det er nulhypotesen
- Vi vil beregne det forventede antal observationer i hver kategori, e_i for $i = 1, 2 \dots, k$ for den formodede fordeling
- For at beregne det forventede antal skal vi estimere nogle parametre for fordelingen (f.eks. λ for Poisson fordelingen, μ og σ^2 for normalfordelingen). Lad m være antal parametre, der skal estimeres
- Vi beregner teststørrelsen χ^2_0 :

$$\chi^2_0 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

- Hvis H_0 er sand (o_i kommer fra fordelingen), så er teststørrelsen χ^2_0 χ^2 fordelt med $k - m - 1$ frihedsgrader
- Dermed har vi en statistisk metode til at forkaste eller acceptere nulhypotesen.

Tilbage til eksemplet

Antal fejl	Sandsynlighed	Forventet antal	Observeret antal
0	0.472	28.34	32
1	0.354	21.26	15
2	0.133	7.97	9
3 (eller flere)	0.041	2.43	4

- Tommelfingerregel: Hver kategori skal helst have et forventet antal på mindst 5, for at metoden er sikker
- For at opnå det, slår vi de sidste to kategorier sammen

Antal fejl	Sandsynlighed	Forventet antal	Observeret antal
0	0.472	28.34	32
1	0.354	21.26	15
2 (eller flere)	0.174	10.4	13

Hypotesetest for eksemplet

1. Hypoteser:

- H_0 : Observationerne kommer fra en Poisson fordeling
- H_a : Observationerne kommer *ikke* fra en Poisson fordeling

2. Signifikansniveau: $\alpha = 0.05$

3. Kriterier:

$$\chi^2_0 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

χ^2_0 er chi-i-anden fordelt med $k - m - 1$ frihedsgrader.

Vi har $k = 3$ kategorier. Vi estimerede 1 parameter (nemlig λ), så $m = 1$.

Dermed: $v = 3 - 1 - 1 = 1$

Vi forkaster H_0 , hvis χ^2_0 er større end χ^2_α :

$$\chi^2_\alpha = \text{qchisq}(1 - 0.05, 1) = 3.8415$$

Altså: Vi forkaster H_0 , hvis $\chi^2_0 > 3.8415$.

Hypotesetest for eksemplet

4. Beregninger:

Antal fejl	Sandsynlighed	Forventet antal	Observeret antal
0	0.472	28.34	32
1	0.354	21.26	15
2 (eller flere)	0.174	10.4	13

$$\chi^2_0 = \frac{(32-28.34)^2}{28.34} + \frac{(15-21.26)^2}{21.26} + \frac{(13-10.40)^2}{10.40} = 2.96$$

5. Konklusioner:

Da teststørrelsen $\chi^2_0 = 2.96 < 3.84$ kan vi ikke forkaste H_0

P-værdien er $1 - \text{pchisq}(2.96, 1) = 0.0852$

Vi tror altså på, at observationerne er Poisson fordelte

N.B. Antal frihedsgrader skal være mindst 1, så vi kunne ikke foretage hypotesesten med kun 2 kategorier.

Sandsynlighedsteori og statistik

Kapitel 11. Regressionsanalyse del 1 (afsnit 11.1-11.2)

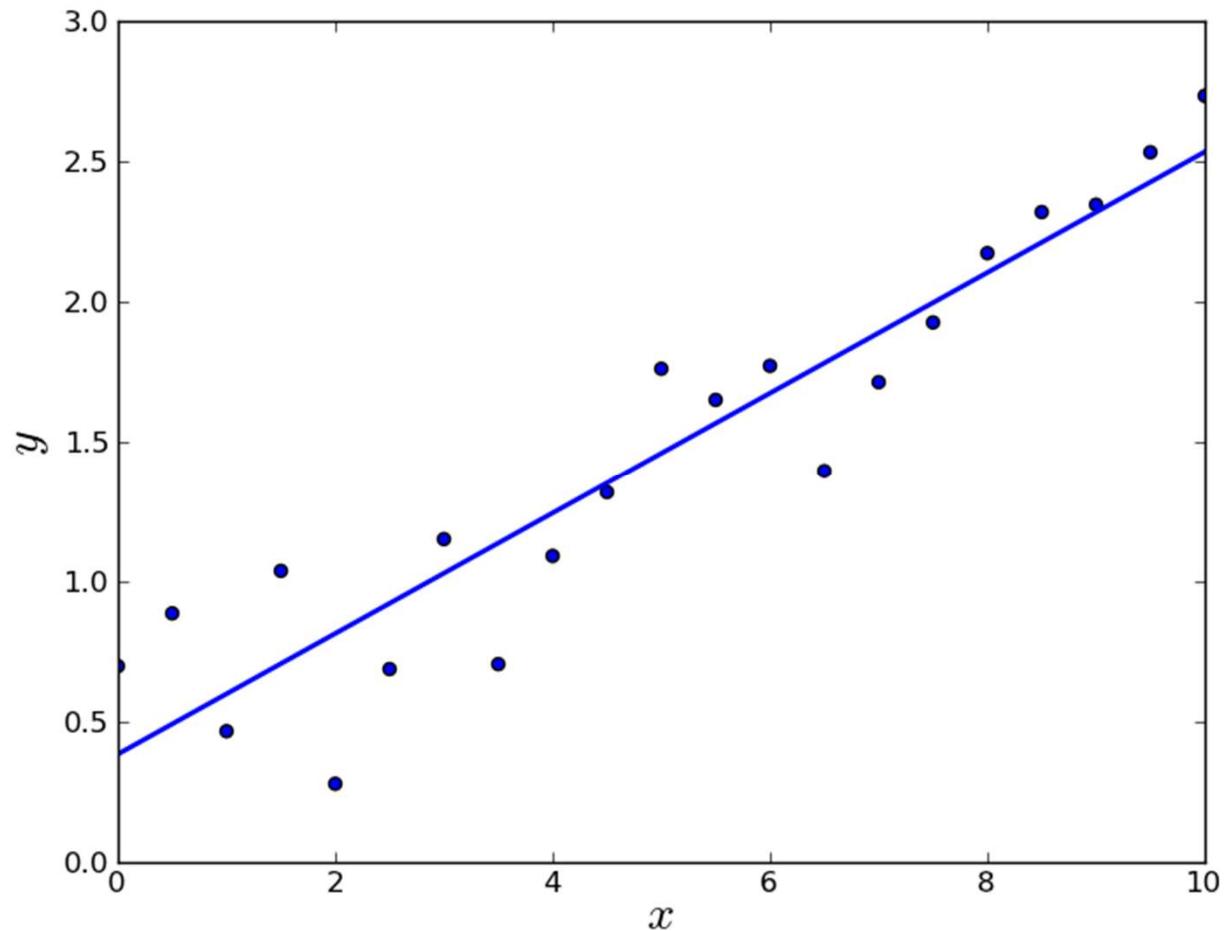
Allan Leck Jensen
alj@ece.au.dk

Overblik over kurset indtil nu

- Organisering og præsentation af data (kap. 1 og 2)
- Sandsynligheder til beskrivelse af usikkerhed (kap. 3)
- Stokastiske variable og deres fordelinger til at beregne sandsynlighed for udfald (kap. 4 og 5)
- Inferens om populationers middelværdi og varians fra stikprøver (kap. 6-9)
- Hypotesetests for egenskabers uafhængighed og for fordeling af stikprøver (Chi-i-anden tests) (kap. 10)
- Nu og resten af kurset: **Modellering** på baggrund af data:
 - Regression (kap. 11-12)
 - Faktorielle eksperimenter (kap. 13-14).

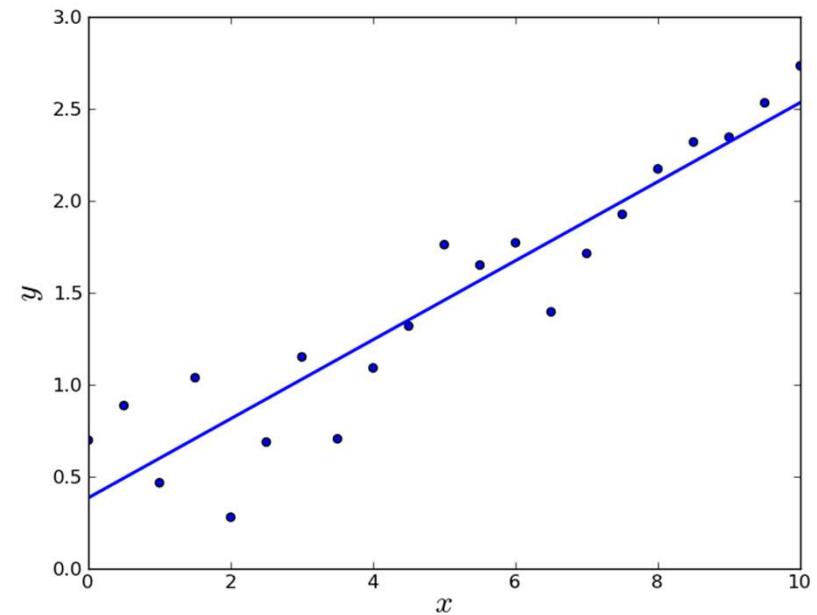
Sammenhænge mellem data (11.1)

- Vi forsøger at bestemme en matematisk sammenhæng mellem forskellige egenskaber, vi har målt
- Det svarer til en model $y = f(x)$, der bedst muligt beskriver et datasæt af sammenhængende målinger af (x, y) værdier
- Når vi har modellen kan vi forudsige y for værdier af x , som vi ikke har målt
- Her er modellen lineær:
$$f(x) = b_0 + b_1 x$$
- Her afhænger modellen kun af én variabel, nemlig x .



Sammenhænge mellem data

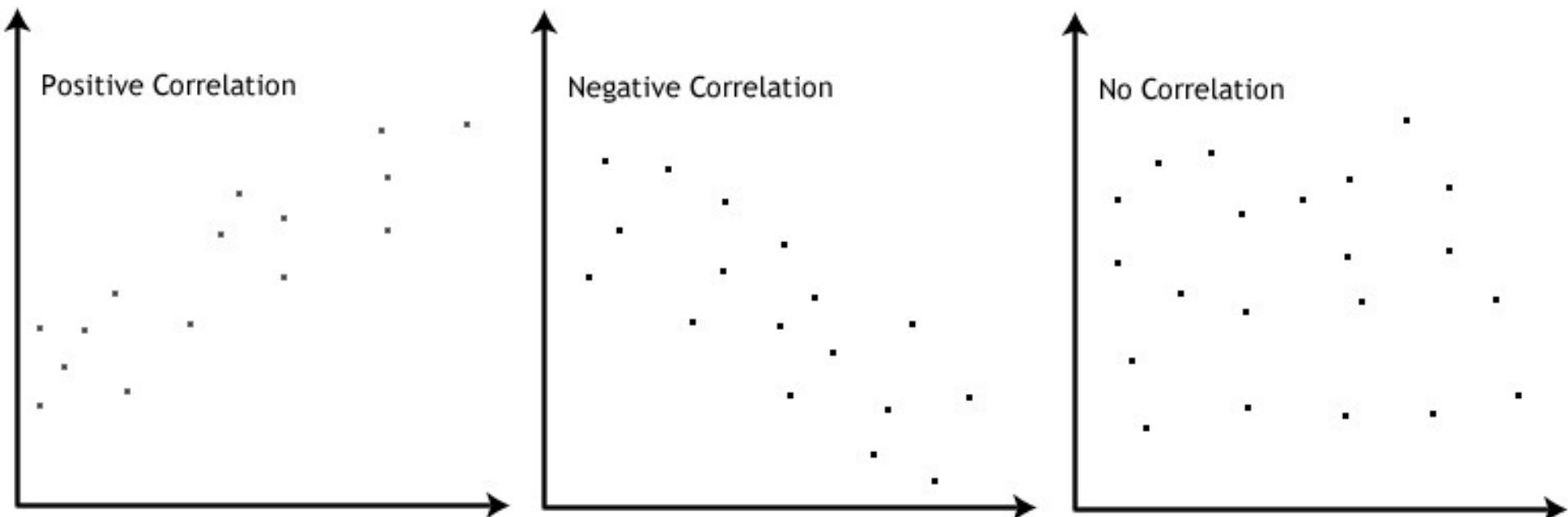
- Den variabel vi vil forudsige værdien af (her: y) kaldes:
 - Den **afhængige** variabel
 - **Responsvariablen**
- Den eller de variable vi bruger til at forudsige responsvariablen (her: x) kaldes:
 - De **uafhængige** variable
 - **Prædiktorer**
 - **Regressorvariable**
 - **Inputvariable**
 - **Faktorer.**



Sammenhænge mellem data

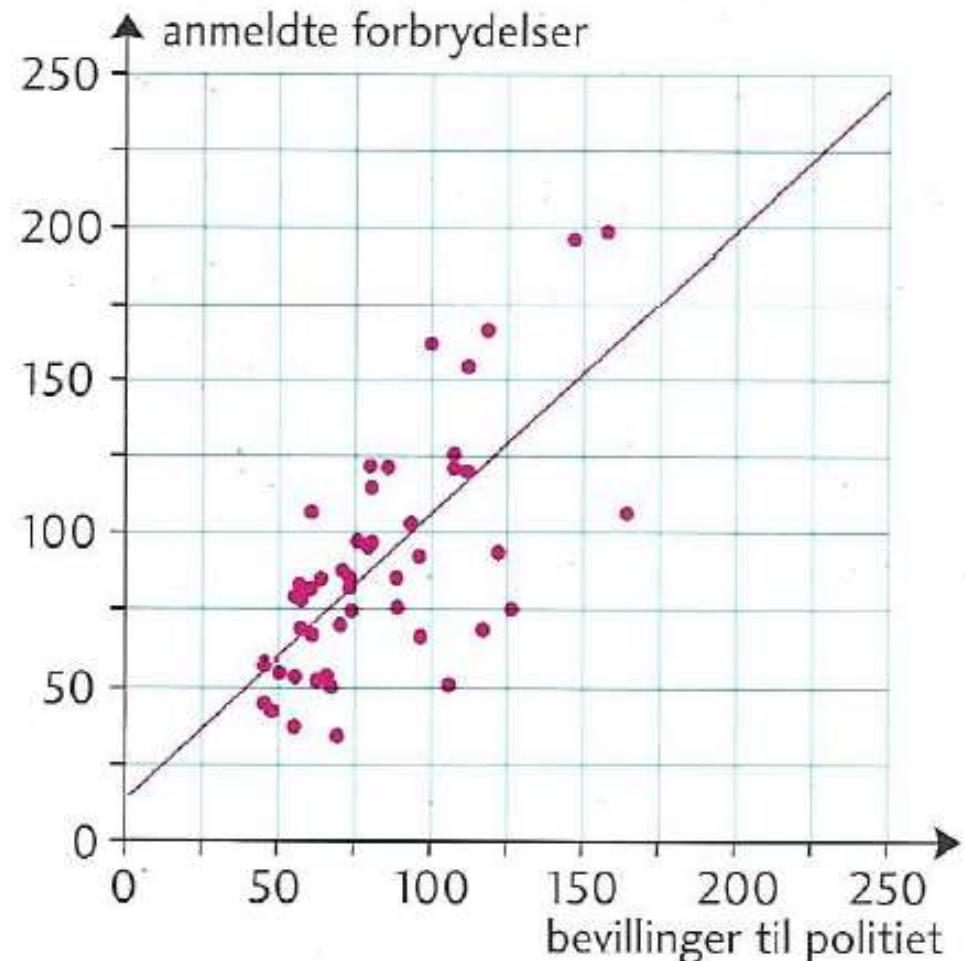
Korrelation

- Sammenhængen mellem y og x kan være
 - Positivt korreleret, hvis y typisk vokser, når x vokser
 - Negativt korreleret, hvis y typisk falder, når x vokser
 - Ukorreleret, hvis værdien af y ikke lader til at afhænge af værdien af x
- Korellation er ikke det samme som **kausalitet** (årsagssammenhæng)
- Statistik kan kun vise korrelation, ikke kausalitet.



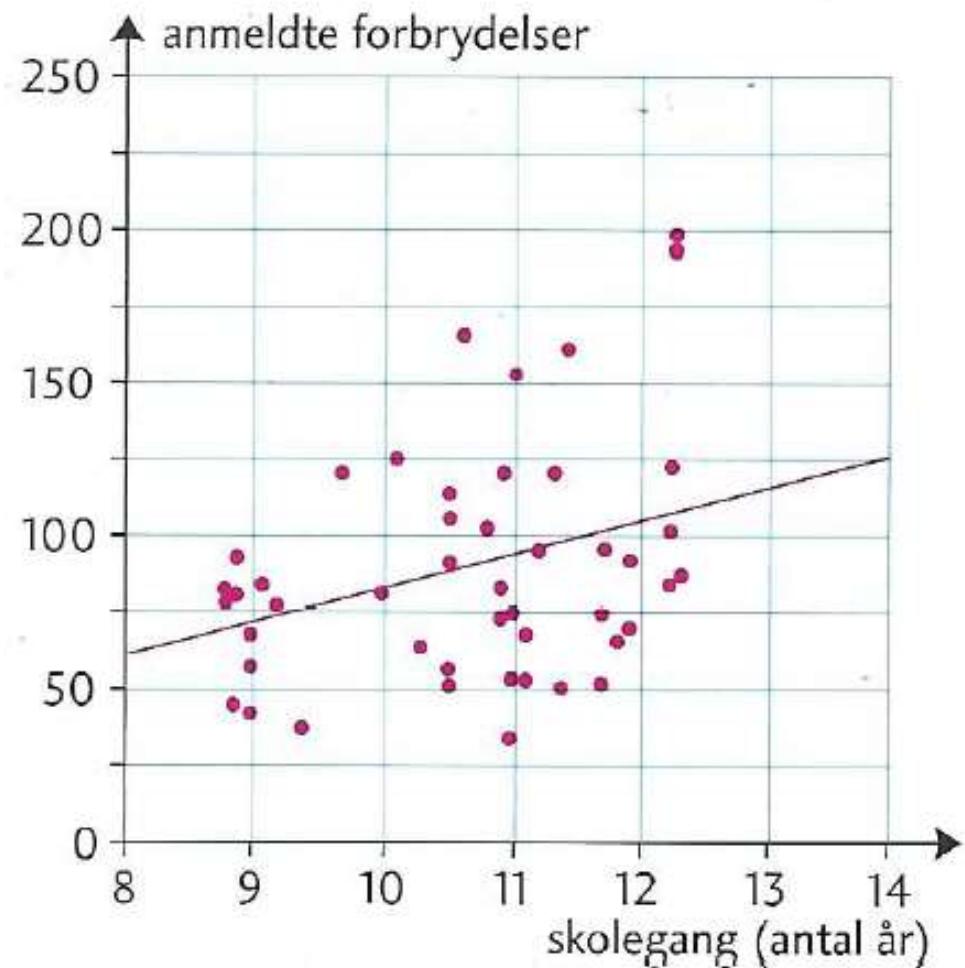
Eksempel: Kriminalitet

- Sammenhæng mellem bevillinger til politiet og anmeldte forbrydelser
- Statistik for 47 stater i USA over de årlige bevillinger i \$ pr. indbygger og antal anmeldte forbrydelser pr. million indbyggere for hver stat
- Resultatet:
Vi bør altså beskære politiet
for at reducere kriminaliteten ??



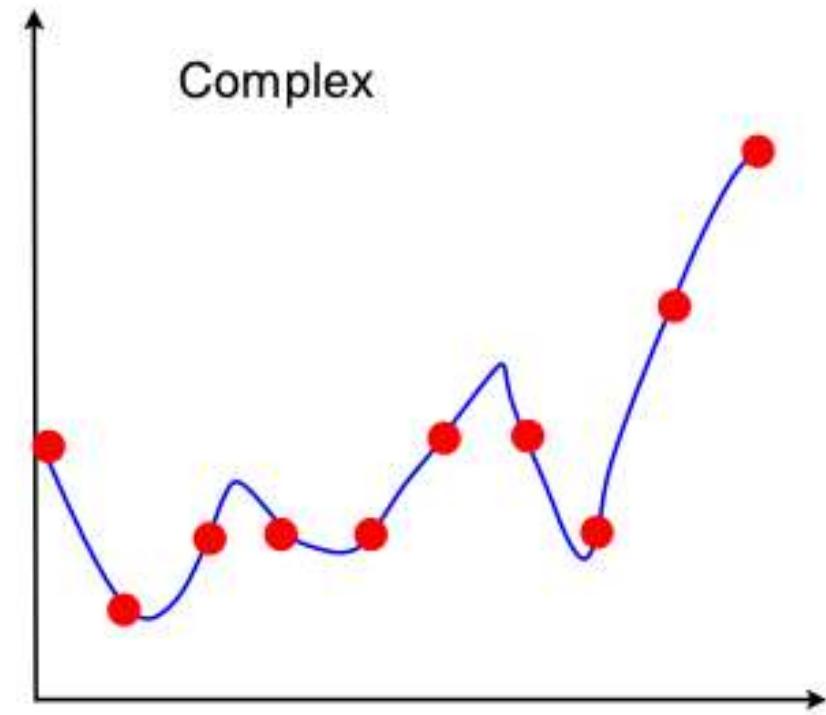
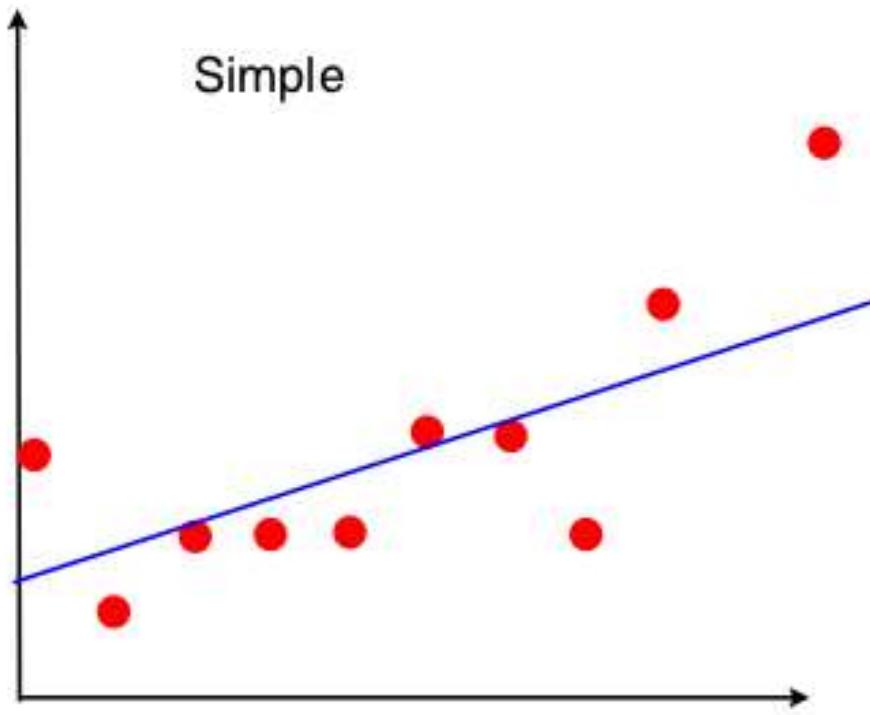
Eksempel: Kriminalitet

- Sammenhæng mellem borgernes uddannelsesniveau og anmeldte forbrydelser
- Resultatet:
Vi bør altså reducere
uddannelsesniveauet for at
reducere kriminaliteten ??



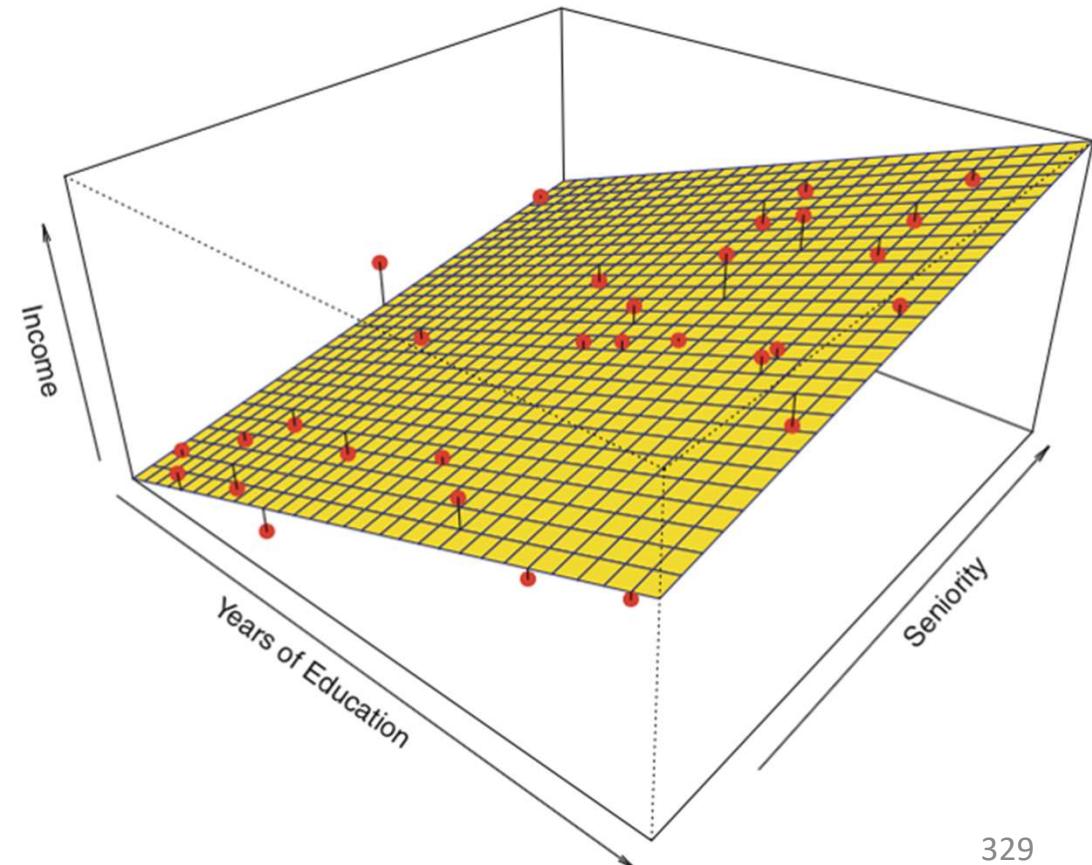
Sammenhænge mellem data

- Modellen kan være **for simpel**, så den ikke beskriver data særligt godt
- Modellen kan være **for kompleks**, så den beskriver data godt, men formodentlig ikke beskriver nye data særligt godt (**overfitting**).



Lineær regression

- Regression betyder ‘gå tilbage’ (forsimple)
- Simpel lineær regression har 1 uafhængig variabel
- Multipel lineær regression har flere uafhængige variable
- F.eks.: $\text{Income} = b_0 + b_1 \cdot \text{Education} + b_2 \cdot \text{Seniority}$
- Her er Income afhængig (respons) variabel
- Education og Seniority er uafhængige (regressor) variable
- b_0, b_1 og b_2 er koefficienter, som vi skal bestemme ‘bedst muligt’.



Eksempel: Nedkøling af en legering (s. 328)

Man vil undersøge om man kan speede nedkølingen af en legering op ved at tilsætte en ny komponent. Hurtig afkøling gør legeringen stærkere

- Data:

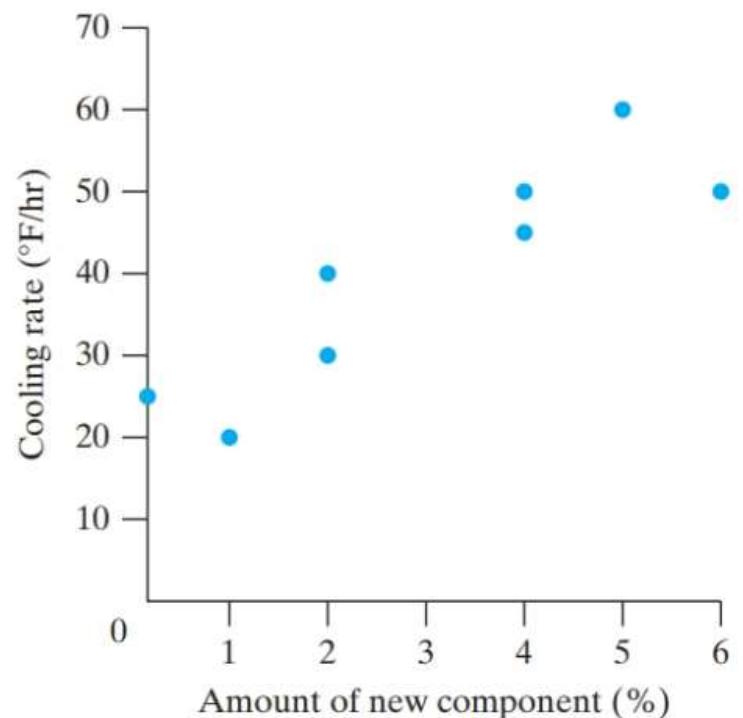
x	0	1	2	2	4	4	5	6
y	25	20	30	40	45	50	60	50

x : Procent af den nye komponent i legeringen (%)

y : Nedkølingshastighed ($^{\circ}\text{F}$ pr. time)

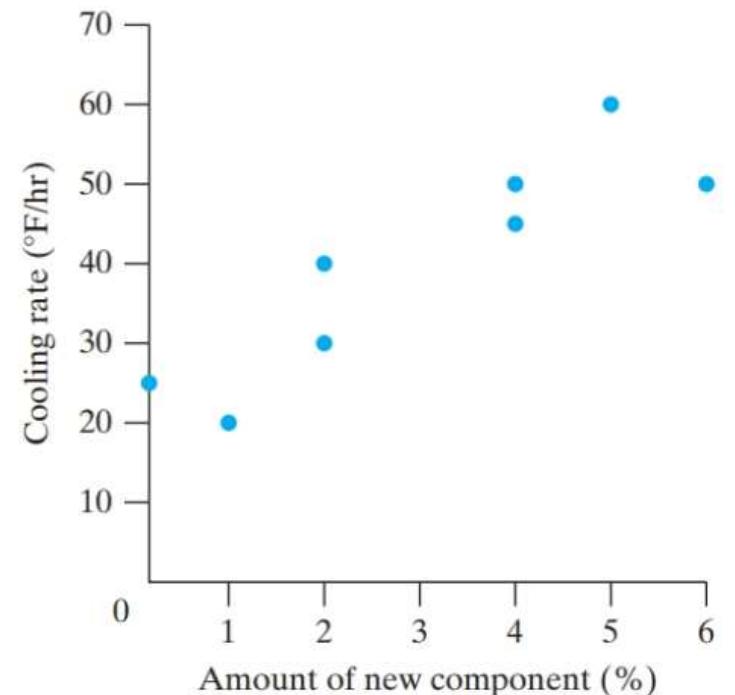
- Første skridt er at lave et **scatter plot** med x på den horizontale akse og y på den vertikale akse
- Plottet viser, at der lader til at være en **positiv korrelation** mellem x og y , og sammenhængen ser ud til at være lineær:

$$y = b_0 + b_1 x .$$



Lineær regression

- Typisk opfatter vi de uafhængige variable som målt uden eller med kun ubetydelig fejl, så al variabiliteten er på den afhængige variabel
- Vi forestiller os f.eks., at forskellen i nedkølingshastighed for de to målinger med 2 % komponent skyldes tilfældige forskelle i y og ikke måleusikkerhed på x
- Derfor opfattes den afhængige variabel som en stokastisk variabel Y , det har middelværdi $\beta_0 + \beta_1 x$, hvor vi ikke kender koefficienterne β_0 og β_1 (men vi vil estimere dem fra data). Generelt er $Y = \beta_0 + \beta_1 x + \varepsilon$
- Her er fejlen ε en stokastisk variabel med middelværdi 0. Vi antager ofte, at ε er normalfordelt, $N(0, \sigma)$
- **NB.** Bogen kalder koefficienterne for α og β i stedet for β_0 og β_1 .



Simpel lineær regression

- Vi har et datasæt med n sammenhørende observationer (x_i, y_i) for $i = 1, \dots, n$
- Vi ønsker at bestemme koefficienterne b_0 og b_1 , så vi kan forudsige værdierne y_i 'bedst muligt' ud fra x_i
- Vi kalder forudsigelsen af y_i for \hat{y}_i ('y hat'):

$$\hat{y}_i = b_0 + b_1 x_i$$

Desuden er b_0 og b_1 estimater for β_0 og β_1 : $b_0 = \widehat{\beta}_0$ og $b_1 = \widehat{\beta}_1$

- Residualet er forskellen på observeret og prædikteret værdi:

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

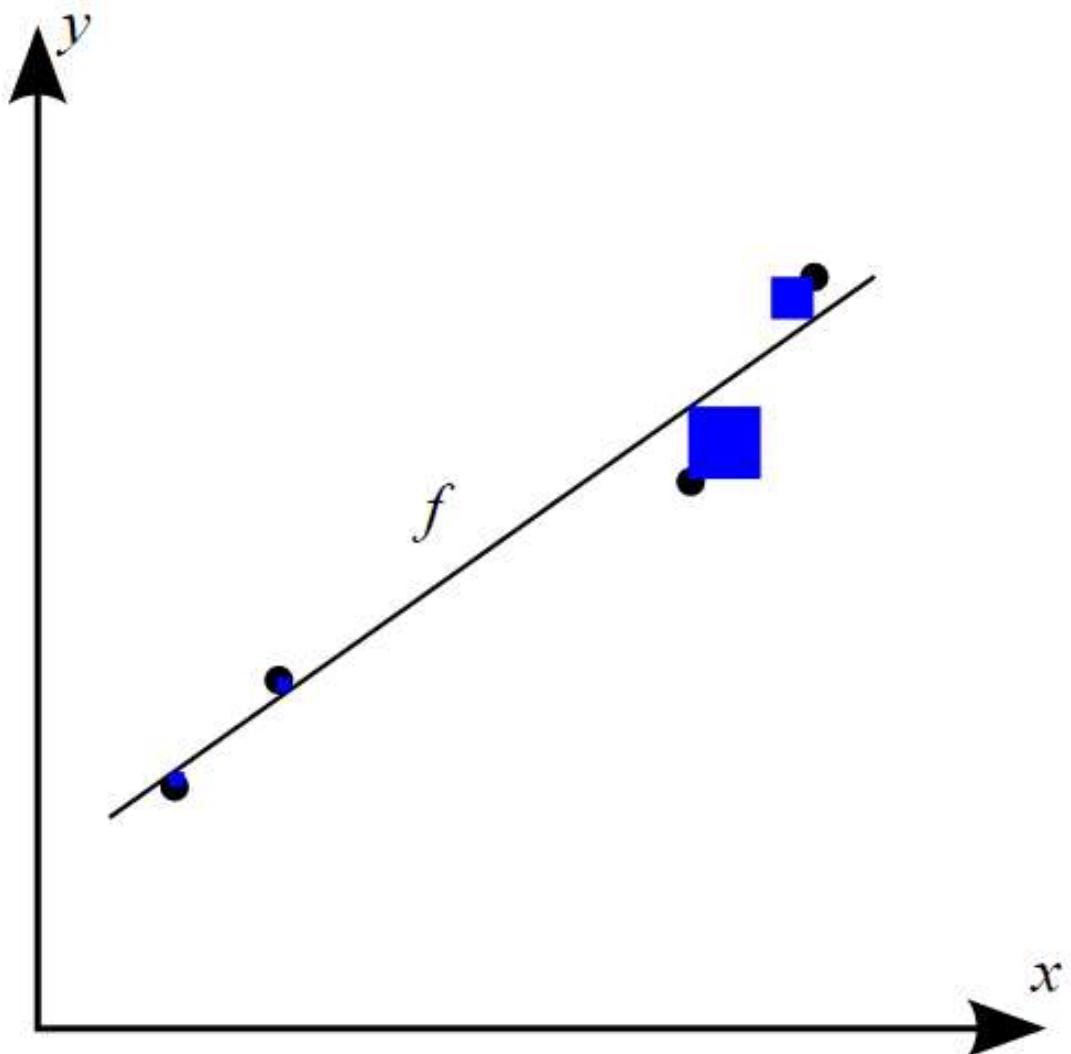
- Vi ønsker at minimere SSE , Sum of Squares for the residuals:

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2$$

- Denne metode kaldes Mindste Kvadraters Metode.

Eksempel (mindste kvadraters metode)

- Vi har 4 samhørende datapunkter (x_i, y_i) for $i = 1, \dots, 4$
- Vi har tegnet en ret linje ind, som ikke nødvendigvis er den ‘bedste’
- For hvert punkt (x_i, y_i) er residualet e_i den lodrette afstand mellem punktet og linjen, dvs.
$$e_i = y_i - (b_0 + b_1 x_i)$$
- Arealet på de blå kvadrater er således e_i^2
- SSE er det samlede areal af de blå kvadrater
- Vi ønsker altså at finde linjens skæring b_0 og hældning b_1 , så det samlede areal af blå kvadrater (SSE) er mindst muligt.



Mindste kvadraters metode

- De optimale værdier af b_0 og b_1 opfylder, at

$$\frac{\partial}{\partial b_0}(SSE) = 0$$

$$\frac{\partial}{\partial b_1}(SSE) = 0$$

- Det svarer til:

$$\frac{\partial}{\partial b_0}(\sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2) = 0$$

$$\frac{\partial}{\partial b_1}(\sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2) = 0$$

- Den partielle differentiering giver disse to ligninger:

$$nb_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

$$b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

Mindste kvadraters metode

- To ligninger med to ubekendte (b_0 og b_1):

$$nb_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

$$b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

- Disse ligninger kaldes normalligningerne (*normal equations*)
- Løsning:

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_{xx}}$$

S_{xx} , S_{yy} og S_{xy}

- Summerne S_{xx} , S_{yy} og S_{xy} dukker hyppigt op i formler:

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

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$$

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

- Beregningsformler til manuelle udregninger:

$$S_{xx} = \sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2$$

$$S_{yy} = \sum_{i=1}^n y_i^2 - \frac{1}{n} (\sum_{i=1}^n y_i)^2$$

$$S_{xy} = \sum_{i=1}^n x_i y_i - \frac{1}{n} (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)$$

- F.eks. kan man vise, at:

$$\text{SSE} = \sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}.$$

Eks. 11.1, s. 331 (nedkøling af legering)

- Vi vil beregne koefficienterne manuelt med mindste kvadraters metode (for første og eneste gang). Det gør bogen også, men de bruger ikke beregningsformlerne, derfor ser det forskelligt ud:

x_i	y_i	x_i^2	$x_i y_i$
0	25	0	0
1	20	1	20
2	30	4	60
2	40	4	80
4	45	16	180
4	50	16	200
5	60	25	300
6	50	36	300
$\sum x_i = 24$		$\sum x_i^2 = 102$	$\sum x_i y_i = 1140$

- Vi får:

$$S_{xy} = \sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) = 1140 - \frac{1}{8} \cdot 24 \cdot 320 = \mathbf{180}$$

$$S_{xx} = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 = 102 - \frac{1}{8} \cdot (24)^2 = \mathbf{30}$$

Eks. 11.1, s. 331 (nedkøling af legering)

- Linjens hældning:

$$b_1 = \frac{s_{xy}}{s_{xx}} = \frac{180}{30} = 6$$

- Linjens skæring med y -aksen:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{8} \cdot 24 = 3$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{8} \cdot 320 = 40$$

$$b_0 = \bar{y} - b_1 \bar{x} = 40 - 6 \cdot 3 = 22$$

- Regressionslinjens ligning:

$$\hat{y} = b_0 + b_1 x = 22 + 6x$$

Eks. 11.1, s. 331 (nedkøling af legering)

x_i	y_i	x_i^2	$x_i y_i$
0	25	0	0
1	20	1	20
2	30	4	60
2	40	4	80
4	45	16	180
4	50	16	200
5	60	25	300
6	50	36	300
$\sum x_i = 24$		$\sum x_i^2 = 102$	$\sum x_i y_i = 1140$

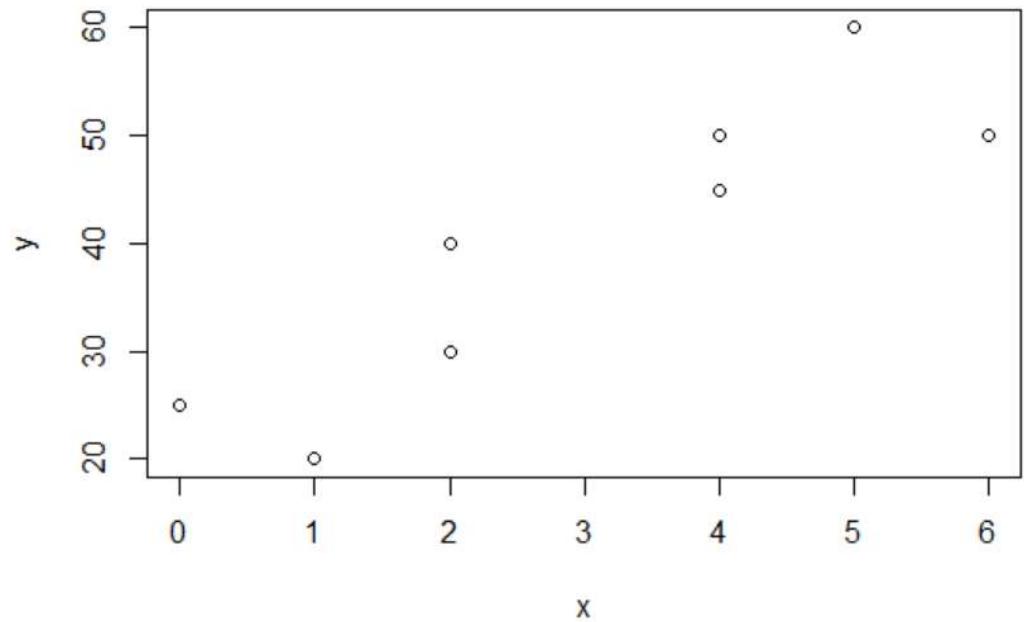
- Alternativt kan vi sætte ind i normalligningerne og løse for b_0 og b_1 :

$$\begin{bmatrix} nb_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \end{bmatrix} \Leftrightarrow$$

$$\begin{bmatrix} 8 \cdot b_0 + b_1 \cdot 24 = 320 \\ b_0 \cdot 24 + b_1 \cdot 102 = 1140 \end{bmatrix} \Leftrightarrow \begin{bmatrix} b_0 = 22 \\ b_1 = 6 \end{bmatrix}.$$

Hvordan regner vi dette i R?

- Scatter plot:
`plot(x, y, type="p")`



- Lineær regression:
`linmod = lm(y ~ x)`
`summary(linmod)`
- Output:

Call:

`lm(formula = y ~ x)`

Residuals:

Min	1Q	Median	3Q	Max
-8.0	-5.0	1.0	4.5	8.0

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	22.000	4.373	5.031	0
x	6.000	1.225	4.886	0

Signif. codes: 0 ‘***’

Residual standard error: 7.08 on 6 degrees of freedom
Multiple R-squared: 0.8, Adjusted R-squared: 0.7667
F-statistic: 24 on 1 and 6 DF, p-value: 0.002714

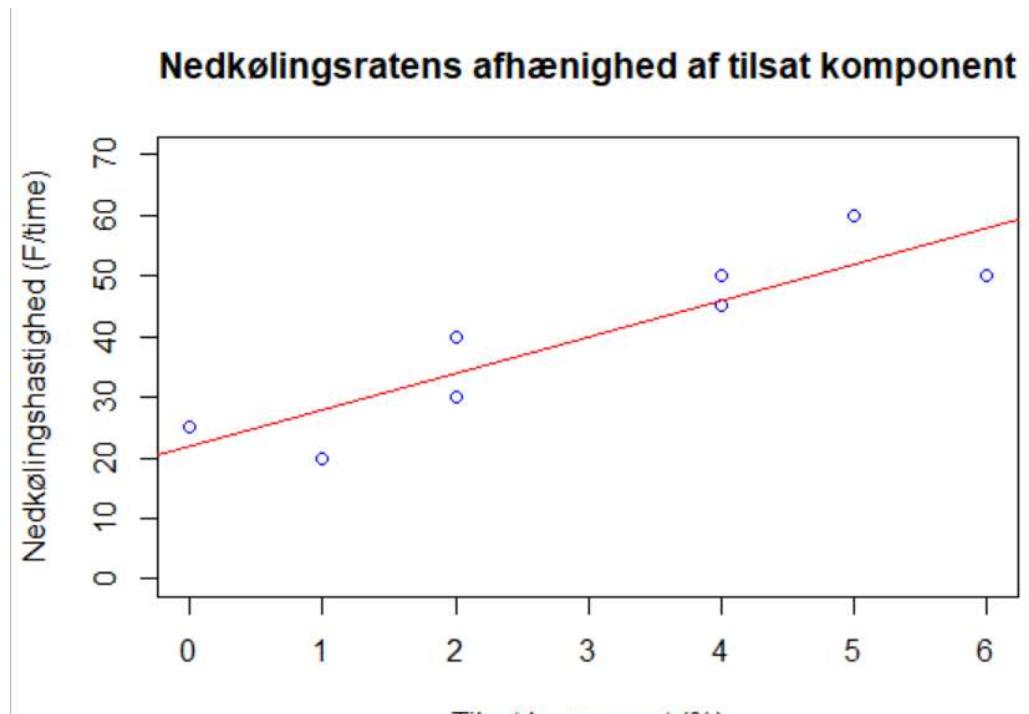
Vi ser på resten af outputtet lidt senere

Hvordan regner vi det i R?

- Plot af resultatet:

```
plot(x, y, type="p",
      main = "Nedkølingsratens afhænighed af tilsat komponent",
      xlab = "Tilsat komponent (%)",
      ylab = "Nedkølingshastighed (F/time)",
      ylim = c(0,70),
      col = "blue")
abline(reg=linmod, col="red")
```

- Vi kan tegne regressionslinjen ind i scatterplottet med funktionen abline().



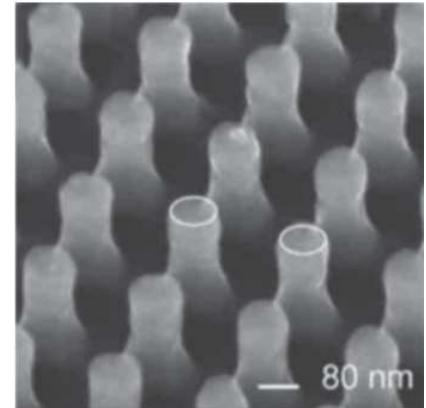
Eksempel 11.3, s. 333 (nanosøjler)

I nanoteknologi bruges nanosøjler af silicone. Man har målt bredde (x) og højde (y) af 50 nanosøjler:

$$n = 50 \quad \bar{x} = 88.34 \quad \bar{y} = 305.58$$

$$S_{xx} = 7239 \quad S_{xy} = 17840 \quad S_{yy} = 66976$$

- Lav en lineær regression, der forudsiger højde fra bredde
- Lav en lineær regression, der forudsiger bredde fra højde
- Vis begge regressionslinjer på et scatter plot



Løsning:

a) Hældning: $b_1 = \frac{S_{xy}}{S_{xx}} = \frac{17840}{7239} = 2.464$

Skæring: $b_0 = \bar{y} - b_1 \bar{x} = 305.58 - 2.464 \cdot 88.34 = 87.88$

Ligning: $y = b_0 + b_1 x = 87.88 + 2.464x$

b) Hældning: $c_1 = \frac{S_{xy}}{S_{yy}} = \frac{17840}{66976} = 0.266$

Skæring: $c_0 = \bar{x} - c_1 \bar{y} = 88.34 - 0.266 \cdot 305.58 = 6.944$

Ligning: $x = c_0 + c_1 y = 6.944 + 0.266y$.

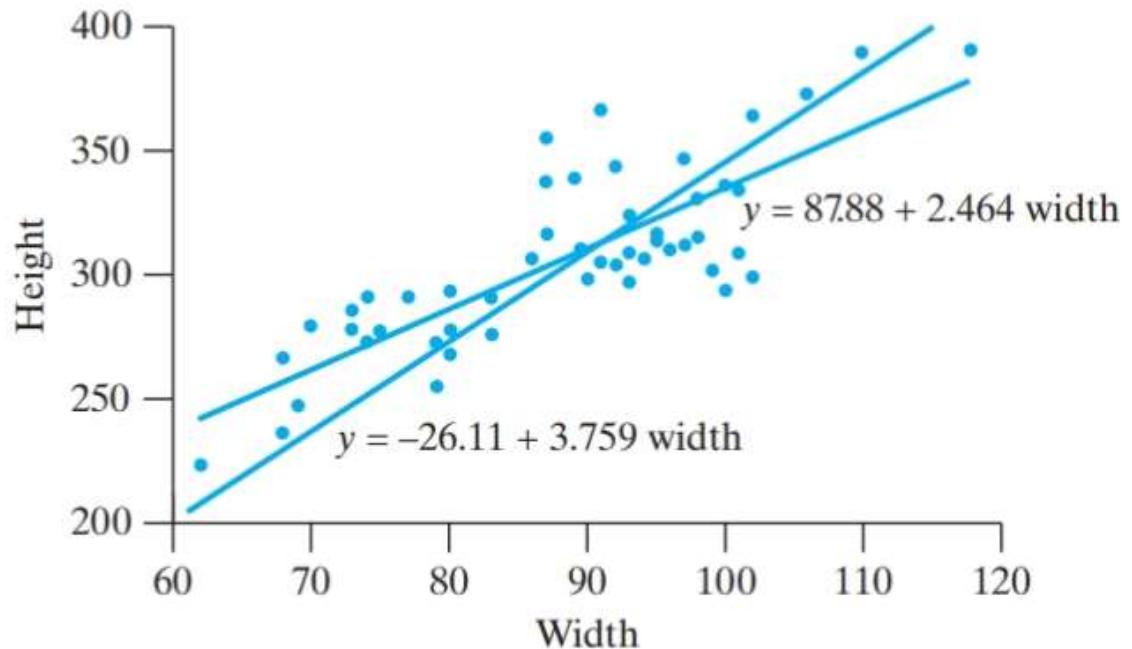
Eksempel 11.3, s. 333 (nanosøjler)

- c) Ligningen fra b) omformes, så y udtrykkes som funktion af x :

$$x = 6.944 + 0.266y \Leftrightarrow$$

$$y = \frac{x - 6.944}{0.266} = -26.11 + 3.759x$$

- Begge linjer går gennem (\bar{x}, \bar{y})
- Den optimale regressionslinje afhænger af, hvilken variabel, vi ønsker at prædiktere.



Statistisk teori (11.2)

- Vi antager, at der findes en lineær sammenhæng mellem responsvariablen y og regressorvariablen x :

$$y = \beta_0 + \beta_1 x,$$

men vi kender ikke koefficienterne β_0 og β_1

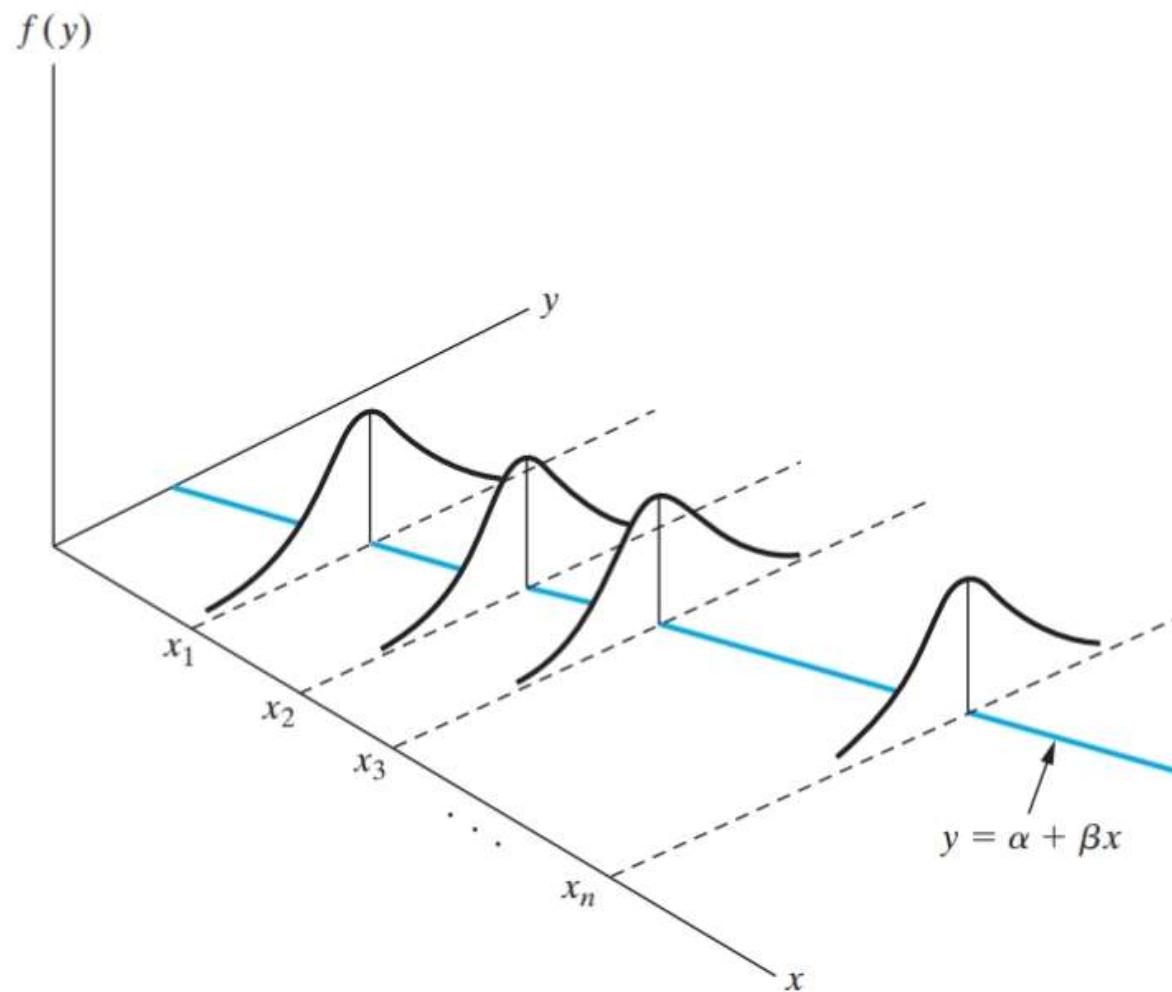
- Desuden er der variabilitet, så gentagne målinger giver ikke samme værdi af y for samme værdi af x . Derfor opfatter vi responsen y som resultatet af en stokastisk variabel Y :

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

- Vi antager, at Y_i er en **normalfordelt** stokastisk variabel for $i = 1, 2, \dots, n$.
 - Middelværdien af Y_i er $\beta_0 + \beta_1 x_i$, så den afhænger af x_i
 - Variansen af Y_i er den samme for alle i , nemlig σ^2
- Det svarer til, at ε_i er $N(0, \sigma^2)$ for $i = 1, 2, \dots, n$.

Statistisk teori

Illustration af modellen $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ med $\varepsilon_i \sim N(0, \sigma)$:



Statistisk teori

Vi kender ikke parametrene β_0 , β_1 og σ^2 , så dem estimerer vi fra vores datasæt:

- $\widehat{\beta}_1 = b_1 = \frac{S_{xy}}{S_{xx}}$, hvor S_{xy} og S_{xx} er defineret tidligere, sammen med S_{yy}
- $\widehat{\beta}_0 = b_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$
- σ estimeres med s_e , som beregnes med residualernes kvadrater:
$$\widehat{\sigma}^2 = s_e^2 = \frac{SSE}{n-2} = \frac{S_{yy} - S_{xy}^2 / S_{xx}}{n-2}$$
 (kan man vise)
- s_e kaldes ‘standard error of the estimate’ (i bogen) og ‘residual standard error’ (i R).

Eks. 11.1 (nedkøling af legering, forts.)

- Vi beregnede/kan beregne:

$$S_{xy} = 180$$

$$S_{xx} = 30$$

$$S_{yy} = 1350$$

- Dermed er

$$s_e^2 = \frac{S_{yy} - S_{xy}^2 / S_{xx}}{n-2} = \frac{1350 - 180^2 / 30}{8-2} = 45 \text{ og}$$

$$s_e = \sqrt{45} = 6.7082$$

- Beregningen af s_e findes i output af R's function summary():

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	22.000	4.373	5.031	0.00238 **
x	6.000	1.225	4.899	0.00271 **

Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1			

Residual standard error: 6.708 on 6 degrees of freedom

Multiple R-squared: 0.8, Adjusted R-squared: 0.7667

F-statistic: 24 on 1 and 6 DF, p-value: 0.002714

x_i	y_i	x_i^2	$x_i y_i$
0	25	0	0
1	20	1	20
2	30	4	60
2	40	4	80
4	45	16	180
4	50	16	200
5	60	25	300
6	50	36	300
$\sum x_i = 24$		$\sum x_i^2 = 102$	$\sum x_i y_i = 1140$

Statistisk teori

- Når vi har beregnet s_e kan den bruges til at beregne konfidensinterval for vores estimer af koefficienterne

- $(1 - \alpha)100\%$ konfidensinterval for β_0 :

$$\widehat{\beta}_0 \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}$$

- $(1 - \alpha)100\%$ konfidensinterval for β_1 :

$$\widehat{\beta}_1 \pm t_{\alpha/2} \cdot s_e \frac{1}{\sqrt{S_{xx}}}$$

- I begge former skal vi bruge $n - 2$ frihedsgrader til $t_{\alpha/2}$

- Bogen bruger α og β til at betegne koefficienterne, i stedet for β_0 og β_1 . Derfor indgår der *to forskellige betydninger af α* i udtrykkene for konfidensinterval (s. 338). Pinligt!!

Confidence limits for regression coefficients

$$\alpha: \widehat{\alpha} \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}$$

and

$$\beta: \widehat{\beta} \pm t_{\alpha/2} \cdot s_e \frac{1}{\sqrt{S_{xx}}}$$

Statistisk teori

- Når vi har beregnet s_e kan den bruges til at beregne konfidensinterval for vores estimer af koefficienterne
- $(1 - \alpha)100\%$ konfidensinterval for β_0 :

$$\widehat{\beta}_0 \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}$$

- $(1 - \alpha)100\%$ konfidensinterval for β_1 :

$$\widehat{\beta}_1 \pm t_{\alpha/2} \cdot s_e \frac{1}{\sqrt{S_{xx}}}$$

- I R kan disse konfidensintervaller beregnes med funktionen `confint()`:

```
# Konfidensinterval for koefficienterne  
confint(linmod, level=0.99)
```

- For eksempel 11.1 får vi
(med 99 % konfidens):

$$\beta_0 \in [5.8; 38.2]$$

$$\beta_1 \in [1.5; 10.5].$$

```
> confint(linmod, level=0.99)  
0.5 % 99.5 %  
(Intercept) 5.786624 38.21338  
x 1.459347 10.54065
```

Statistisk teori

- Nu vil vi undersøge, om koefficienterne β_0 og β_1 i virkeligheden kan være 0. Hvis hældningskoefficienten $\beta_1 = 0$ har vi en kedelig model, for den afhænger ikke af x : $y = \beta_0 + 0 \cdot x = \beta_0$
- For eksempel 11.1 er det usandsynligt at nogen af koefficienterne er 0, for 0 er langt udenfor 99 % konfidensintervallerne:
 $\beta_0 \in [5.8; 38.2]$
 $\beta_1 \in [1.5; 10.5]$
- Resultatet af regressionsanalysen i R indeholder en test af, om de estimerede koefficienter kunne være 0:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 22.000    4.373   5.031  0.00238 ***
x            6.000    1.225   4.899  0.00271 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.708 on 6 degrees of freedom
Multiple R-squared:  0.8,    Adjusted R-squared:  0.7667 
F-statistic: 24 on 1 and 6 DF,  p-value: 0.002714
```

Statistisk teori

- Normalt tjekker vi kun kolonnen ‘Pr(>|t|)’ længst til højre
- Tallene er p-værdier for hypotesetests med nulhypotesen, at koefficienten i virkeligheden er 0
- For hældningskoefficienten viser testen, at hvis nulhypotesen $H_0: \beta_1 = 0$ er sand, så vil man tilfældigvis kunne få et estimat på $|b_1| \geq 6.0$ med en sandsynlighed på 0.00271
- Det er altså meget usandsynligt, så vi forkaster nulhypotesen og tror på, at der er korrelation
- Koden med ** betyder, at vi kan forkaste nulhypotesen på 1 % signifikansniveau, men ikke på 0.1 %.

```
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 22.000    4.373   5.031  0.00238 **  
x           6.000    1.225   4.899  0.00271 **  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 6.708 on 6 degrees of freedom  
Multiple R-squared:      0.8,    Adjusted R-squared:  0.7667  
F-statistic: 24 on 1 and 6 DF,  p-value: 0.002714
```

Statistisk teori

- Baggrunden for de to hypotesetests er disse erkendelser fra ‘kloge statistikere’ (husk: bogen bruger α og β i stedet for β_0 og β_1):

Theorem 11.1 Under the assumptions given on page 336 the statistics

$$t = \frac{(\hat{\alpha} - \alpha)}{s_e} \sqrt{\frac{nS_{xx}}{S_{xx} + n(\bar{x})^2}}$$

and

$$t = \frac{(\hat{\beta} - \beta)}{s_e} \sqrt{S_{xx}}$$

are random variables having the t distribution with $n - 2$ degrees of freedom.

- For eks. 11.1 fandt vi $S_{xx} = 30$ og $s_e = 6.7082$. Derfor er teststørrelsen for β_1 :

$$t_0 = \frac{6.0 - 0}{6.7082} \sqrt{30} = 4.899$$

- Den tilhørende p-værdi:

$$P = 2 \cdot (1 - pt(4.899, 8 - 2)) = 0.00271$$

- ‘Std. Error’ er $\frac{s_e}{\sqrt{S_{xx}}} = \frac{6.7082}{\sqrt{30}} = 1.225$.

Std. Error	t value	Pr(> t)	
4.373	5.031	0.00238	**
1.225	4.899	0.00271	**

Statistisk teori

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	22.000	4.373	5.031	0.00238	**
x	6.000	1.225	4.899	0.00271	**

Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’	0.1 ‘ ’ 1

~~Residual standard error: 6.708 on 6 degrees of freedom~~

~~Multiple R-squared: 0.8, Adjusted R-squared: 0.7667~~

~~F-statistic: 24 on 1 and 6 DF, p-value: 0.002714~~

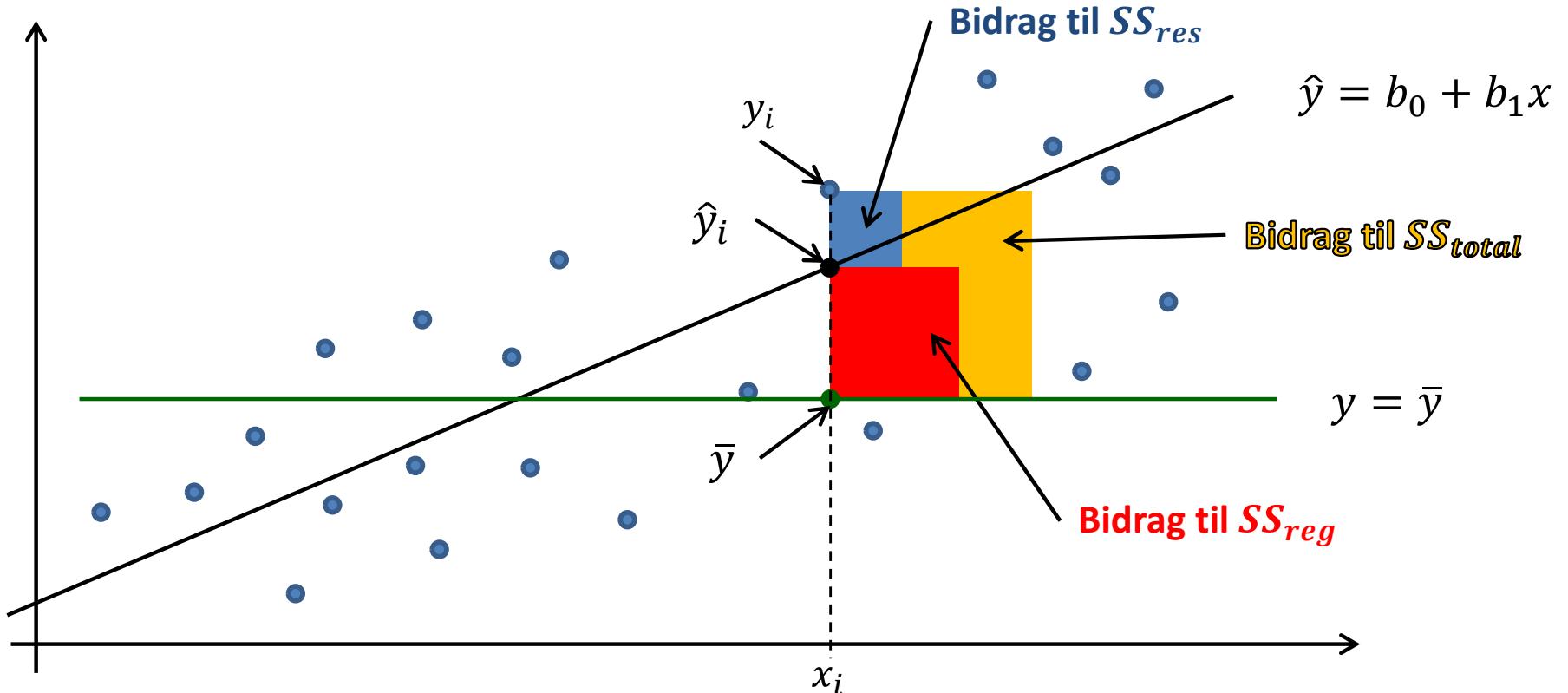
- For at forstå resten af outputted for regressionsanalysen, skal vi kende lidt til variansanalyse, der ellers er emnet for kapitel 12
- **Variansanalyse** eller **ANOVA (Analysis of Variance)** er en meget bredt anvendt statistisk metode til at undersøge, om der er forskelle mellem grupper af data
- Her bruges ANOVA til at se, hvor stor en del af variansen, der kan forklares af modellen, og dermed hvor god modellen er til at beskrive data.

ANOVA

Kilder	Frihedsgrader (DF)	Sum of Squares (SS)	Mean Squares (MS)	F
Regression	df_{reg}	SS_{reg}	MS_{reg}	
Residual	df_{res}	SS_{res}	MS_{res}	
Total	df_{total}	SS_{total}		

- SS_{res} er det samme som SSE fra mindste kvadraters metode

Forskellige Sums of Squares



$$SS_{total} = \sum_{i=1}^n (y_i - \bar{y})^2$$

Mål for samlet variation i data

$$SS_{res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

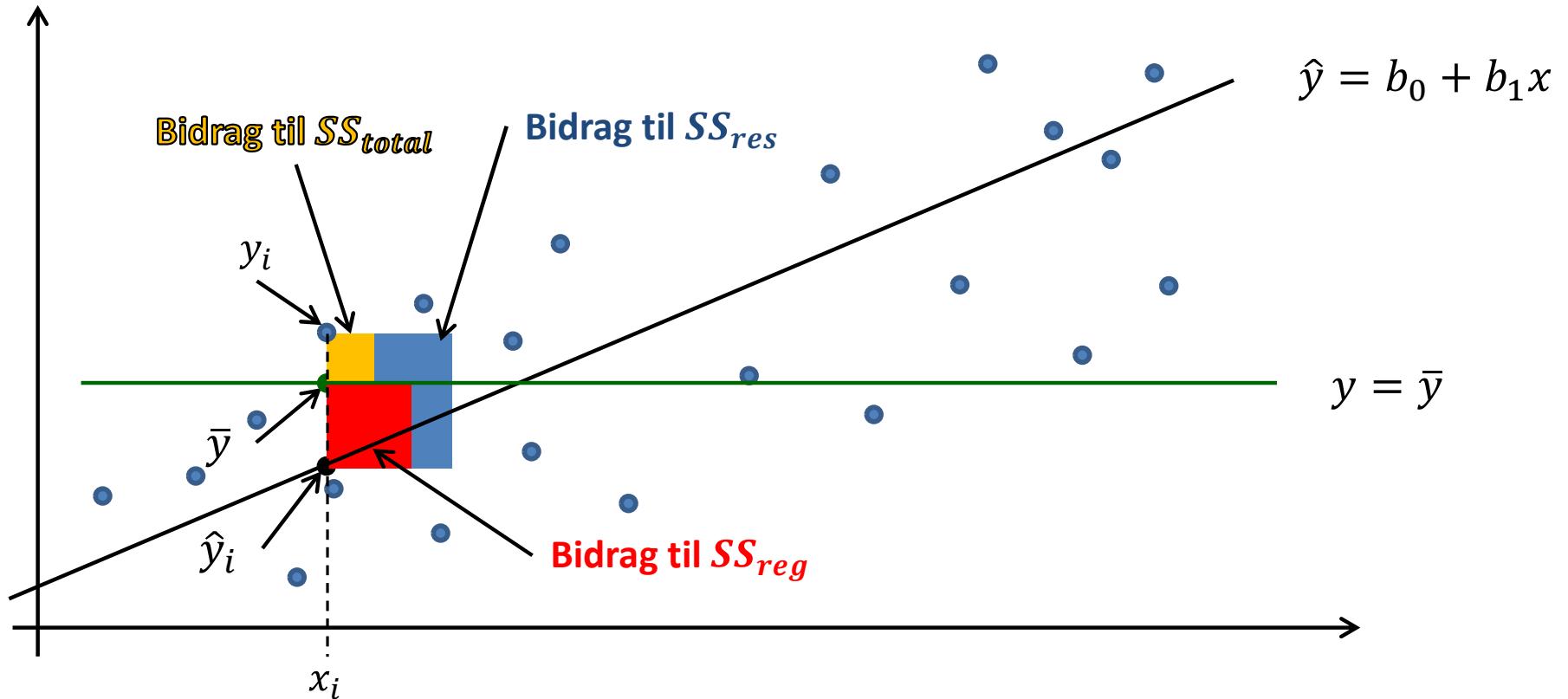
Mål for variation, som modellen ikke forklarer

$$SS_{reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Mål for variation, som modellen forklarer

$$SS_{total} = SS_{reg} + SS_{res}$$

Forskellige Sums of Squares



$$SS_{total} = \sum_{i=1}^n (y_i - \bar{y})^2$$

Mål for samlet variation i data

$$SS_{res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

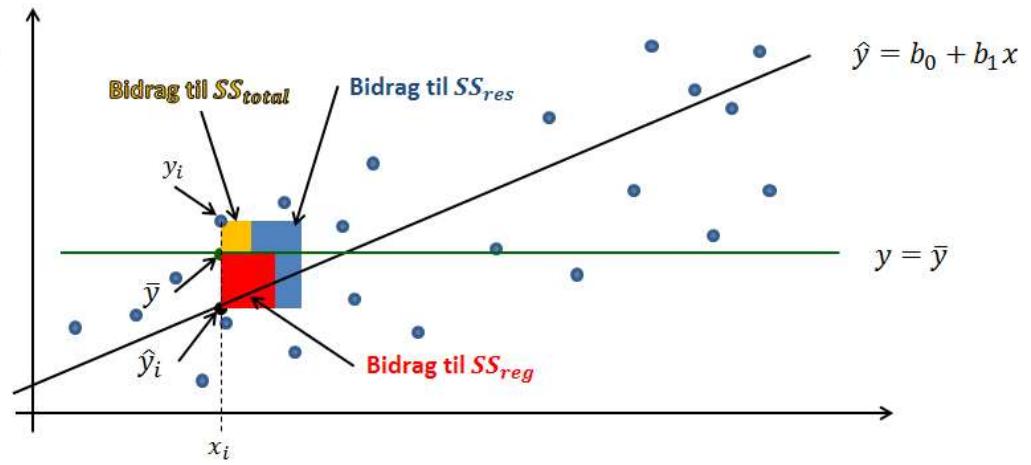
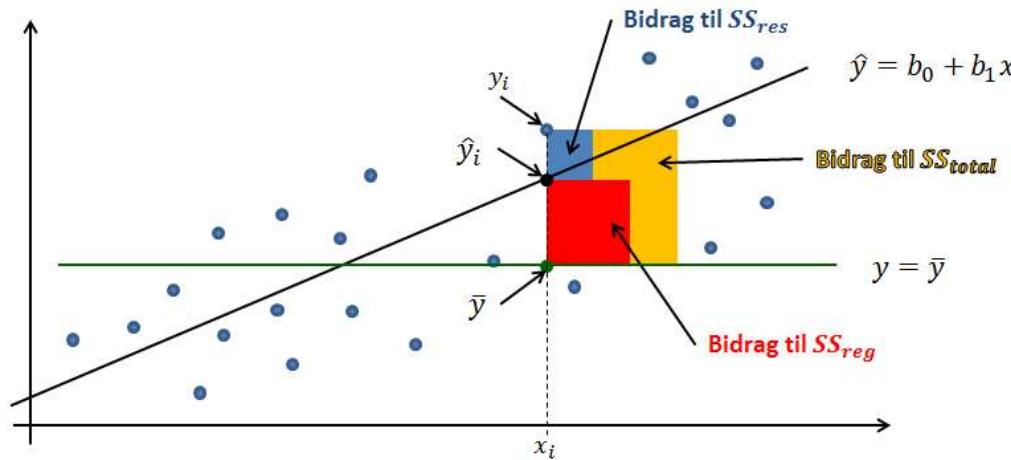
Mål for variation, som modellen ikke forklarer

$$SS_{reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Mål for variation, som modellen forklarer

$$SS_{total} = SS_{reg} + SS_{res}$$

Forskellige Sums of Squares



$$SS_{total} = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$SS_{res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$SS_{reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$SS_{total} = SS_{reg} + SS_{res}$$

Mål for samlet variation i data

Mål for variation, som modellen ikke forklarer

Mål for variation, som modellen forklarer

Forskellige Sums of Squares

- Residual Sum of Squares $SS_{res} =$

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2$$

$$SS_{res} = SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

- Total Sum of Squares

$$SS_{total} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} (\sum_{i=1}^n y_i)^2$$

$$SS_{total} = SS_{reg} + SS_{res}$$

- Explained Sum of Squares

$$SS_{reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$SS_{reg} = SS_{total} - SS_{res} = SS_{total} - (S_{yy} - \frac{S_{xy}^2}{S_{xx}}).$$

ANOVA

Kilder	Frihedsgrader (DF)	Sum of Squares (SS)	Mean Squares (MS)	F
Regression	df_{reg}	$SS_{reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	MS_{reg}	F
Residual	df_{res}	$SS_{res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	MS_{res}	
Total	df_{total}	$SS_{total} = \sum_{i=1}^n (y_i - \bar{y})^2$		

Mean Squares og frihedsgrader

- Mean Squares er Sum of Squares divideret med antal frihedsgrader:
$$MS = SS/df$$
- Antal frihedsgrader afhænger af hvilken gruppe vi ser på:
 - $df_{total} = (\text{Antal observationer} - 1) = n - 1$
 - $df_{reg} = (\text{Antal parametre i modellen} - 1) = (2 - 1) = 1$
(De 2 parametre er b_0 og b_1)
 - $df_{res} = (\text{Antal observationer}) - (\text{Antal parametre}) = n - 2$
- Bemærk: $df_{total} = df_{res} + df_{reg}$.

ANOVA

Kilder	Frihedsgrader (DF)	Sum of Squares (SS)	Mean Squares (MS)	F
Regression	df_{reg} $= \#parametre - 1$ $= 2 - 1 = 1$	SS_{reg} $= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	MS_{reg}	F
Residual	df_{res} $= \#obs. - \#parametre$ $= n - 2$	SS_{res} $= \sum_{i=1}^n (y_i - \hat{y}_i)^2$	MS_{res}	
Total	df_{total} $= \#observationer - 1$ $= n - 1$	SS_{total} $= \sum_{i=1}^n (y_i - \bar{y})^2$		

Mean Squares og frihedsgrader

- Mean Squares er Sum of Squares divideret med antal frihedsgrader:
$$MS = SS/df$$
- Antal frihedsgrader afhænger af hvilken gruppe vi ser på:
 - $df_{total} = \text{Antal observationer} - 1 = n - 1$
 - $df_{reg} = \text{Antal parametre i modellen} - 1 = 2 - 1 = 1$
(De 2 parametre er b_0 og b_1)
 - $df_{res} = (\text{Antal observationer}) - (\text{Antal parametre}) = n - 2$
- Bemærk: $df_{total} = df_{res} + df_{reg}$
- Mean Squares:
 - $MS_{reg} = \frac{SS_{reg}}{df_{reg}}$
 - $MS_{res} = \frac{SS_{res}}{df_{res}}$
 - $MS_{total} = \frac{SS_{total}}{df_{total}}$
 - Bemærk: MS_{total} skrives ikke i ANOVA tabellen.

ANOVA

Kilder	Frihedsgrader (DF)	Sum of Squares (SS)	Mean Squares (MS)	F
Regression	df_{reg} $= \#parametre - 1$ $= 2 - 1 = 1$	SS_{reg} $= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	MS_{reg} $= \frac{SS_{reg}}{df_{reg}}$	F
Residual	df_{res} $= \#obs. - \#parametre$ $= n - 2$	SS_{res} $= \sum_{i=1}^n (y_i - \hat{y}_i)^2$	MS_{res} $= \frac{SS_{res}}{df_{res}}$	
Total	df_{total} $= \#observationer - 1$ $= n - 1$	SS_{total} $= \sum_{i=1}^n (y_i - \bar{y})^2$		

F

- Værdien af F beregnes som forholdet mellem variansen forklaret af modellen og den uforklarede varians, d.v.s.:

$$F = \frac{MS_{reg}}{MS_{res}}$$

ANOVA

Kilder	Frihedsgrader (DF)	Sum of Squares (SS)	Mean Squares (MS)	F
Regression	df_{reg} $= \#parametre - 1$ $= 2 - 1 = 1$	SS_{reg} $= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	MS_{reg} $= \frac{SS_{reg}}{df_{reg}}$	F $= \frac{MS_{reg}}{MS_{res}}$
Residual	df_{res} $= \#obs. - \#parametre$ $= n - 2$	SS_{res} $= \sum_{i=1}^n (y_i - \hat{y}_i)^2$	MS_{res} $= \frac{SS_{res}}{df_{res}}$	
Total	df_{total} $= \#observationer - 1$ $= n - 1$	SS_{total} $= \sum_{i=1}^n (y_i - \bar{y})^2$		

ANOVA i R

- Funktionen `anova()` kan kaldes med en lineær regressionsmodel:

```
linmod = lm(y ~ x)  
anova(linmod)
```

- Resultatet er en tabel, der ligner vores ANOVA tabel, dog vises rækken for Total ikke. Her for eksempel 11.1:

Analysis of Variance Table

```
Response: y  
          Df Sum Sq Mean Sq F value    Pr(>F)  
x          1 1080   1080      24 0.002714 **  
Residuals 6  270     45  
---  
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA – med kaffeletter



	Frihedsgrader (DF)	Sum of Squares (SS)	Mean Squares (MS)	F
Regression	?	?	?	?
Residual	139	?	0.54339	
Total	?	126.8		

- Jeg har spilt kaffe på min ANOVA tabel, så nogle af tallene er ulæselige. Jeg kan heldigvis læse værdierne i tre af felterne
- Samarbejd evt. med sidemanden om at regne ud, hvad der stod i de kaffeplettede felter.

ANOVA – med kaffeletter



	Frihedsgrader (DF)	Sum of Squares (SS)	Mean Squares (MS)	F
Regression				
Residual	139		0.54339	
Total		126.8		

F

- Værdien af F beregnes som forholdet mellem variansen forklaret af modellen og den uforklarede varians, d.v.s.:

$$F = \frac{MS_{reg}}{MS_{res}}$$

- F er faktisk teststørrelsen for en hypotesetest med
 - H_0 : Data er ukorrelerede (d.v.s. $\beta_1 = 0$)
 - H_1 : Data er korrelerede (d.v.s. $\beta_1 \neq 0$)
- Når vi har multipel regression med regressorvariable x_1, x_2, \dots, x_r , så er hypoteserne:
 - H_0 : Data er ukorrelerede (d.v.s. $\beta_1 = \beta_2 = \dots = \beta_r = 0$)
 - H_1 : Data er korrelerede (d.v.s. $\beta_i \neq 0$ for mindst én i)
- Ifølge ‘de klogte statistikere’ er F -teststørrelsen F -fordelt med df_{reg} frihedsgrader i tælleren og df_{re} frihedsgrader i nævneren.

P-værdi

- P-værdien er sandsynligheden for at få en værdi som F eller større, hvis den modellerede korrelation *ikke* er korrekt (d.v.s. H_0 er sand)
- Husk: H_0 : Data er ukorrelerede (d.v.s. $\beta_1 = 0$)
- I eksempel 11.1 med nedkøling af en legering er $F = 24$, $df_{reg} = 1$ og $df_{res} = n - 2 = 6$:

$$\begin{aligned}P(F > 24) &= 1 - P(F < 24) \\&= 1 - \text{pf}(24, 1, 6) \\&= 0.002714\end{aligned}$$

- Det stemmer:
- Bemærk at p-værdien er 0.00271 for både F -test og t -test af β_1 .

```
Coefficients:  
Estimate Std. Error t value Pr(>|t|)  
(Intercept) 22.000 4.373 5.031 0.00238 **  
x 6.000 1.225 4.899 0.00271 **  
---  
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 6.708 on 6 degrees of freedom  
Multiple R-squared: 0.8, Adjusted R-squared: 0.7667  
F-statistic: 24 on 1 and 6 DF, p-value: 0.002714
```

Det gælder, når vi har simpel regression (kun 1 regressorvariabel).

R^2 og R^2_{adj}

- R^2 (Coefficient of determination, determinationskoefficienten) er et mål for hvor stor en del af variationen i data, der forklares af modellen:

$$R^2 = \frac{SS_{reg}}{SS_{total}} = \frac{SS_{total} - SS_{re}}{SS_{total}} = 1 - \frac{SS_{re}}{SS_{total}}$$

- Bemærk at $0 \leq R^2 \leq 1$. En perfekt model har $SS_{res} = 0$ og dermed $R^2 = 1$. Normalt er man tilfreds med modeller med $R^2 > 0.9$, men det afhænger af domænet, der modelleres
- R^2 er et mål for *Goodness of fit*, dvs. hvor godt modellen fitter observationer
- En ulempe ved R^2 er at den stiger med antal parametre i modellen (multipel regression). For at mindske risikoen for overfitting har man indført adjusted R^2 , hvor der justeres for antal parametre i modellen

$$R^2_{adj} = 1 - \frac{SS_{res}/df_{res}}{SS_{total}/df_{total}} = 1 - \frac{MS_{res}}{MS_{total}}$$

R^2 og R^2_{adj}

- I outputtet fra regressionsanalysen med R kaldes determinationskvotienten R^2 for 'Multiple R-squared', her for eksempel 11.1:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	22.000	4.373	5.031	0.00238 **
x	6.000	1.225	4.899	0.00271 **

Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1			

Residual standard error: 6.708 on 6 degrees of freedom

Multiple R-squared:	0.8,	Adjusted R-squared:	0.7667
F-statistic:	24 on 1 and 6 DF,	p-value:	0.002714

- Værdien $R^2 = 0.8$ er beregnet med SS-værdier fra ANOVA tabellen:

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}$$

$$= 1 - \frac{270}{1080+270} = 0.8$$

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	1080	1080	24	0.002714 **
Residuals	6	270	45		

- Modellen forklarer altså 80 % af variationen i data. Modellen er ikke overfitted, for R^2_{adj} er kun lidt mindre end R^2 :

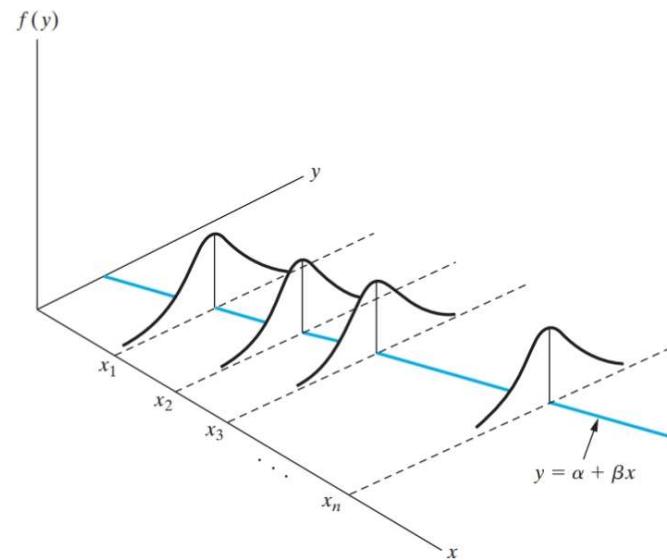
$$R^2_{adj} = 1 - \frac{SS_{res}/df_{res}}{SS_{total}/df_{total}} = 1 - \frac{270/6}{(1080+270)/(1+6)} = 0.7667.$$

Konfidensinterval og prædiktionsinterval af $y(x)$

- Vi vil gerne bruge modellen til at prædiktere responsværdien svarende til vilkårlige værdier af regressoren, f.eks. $x = x_0$
$$y(x_0) = \beta_0 + \beta_1 x_0$$
- Vi er interesserede i konfidensintervallet omkring $\hat{y}(x_0)$, f.eks. på 95 % niveau, generelt på $(1 - \alpha) \cdot 100\%$ niveau:

$$\hat{y}(x_0) \pm t_{\alpha/2} \cdot s_e \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)}, \text{ hvor } t_{\alpha/2} \text{ er med } n - 2 \text{ d.f.}$$

- Konfidensintervallet giver et interval, for middelværdien af den normalfordelte variabel, der svarer til $x = x_0$
- Hvis vi i stedet vil forudsige fremtidige værdier af $y(x_0)$, hvor variationen fra den normalfordelte støj tages med, så skal vi beregne et prædiktionsinterval.



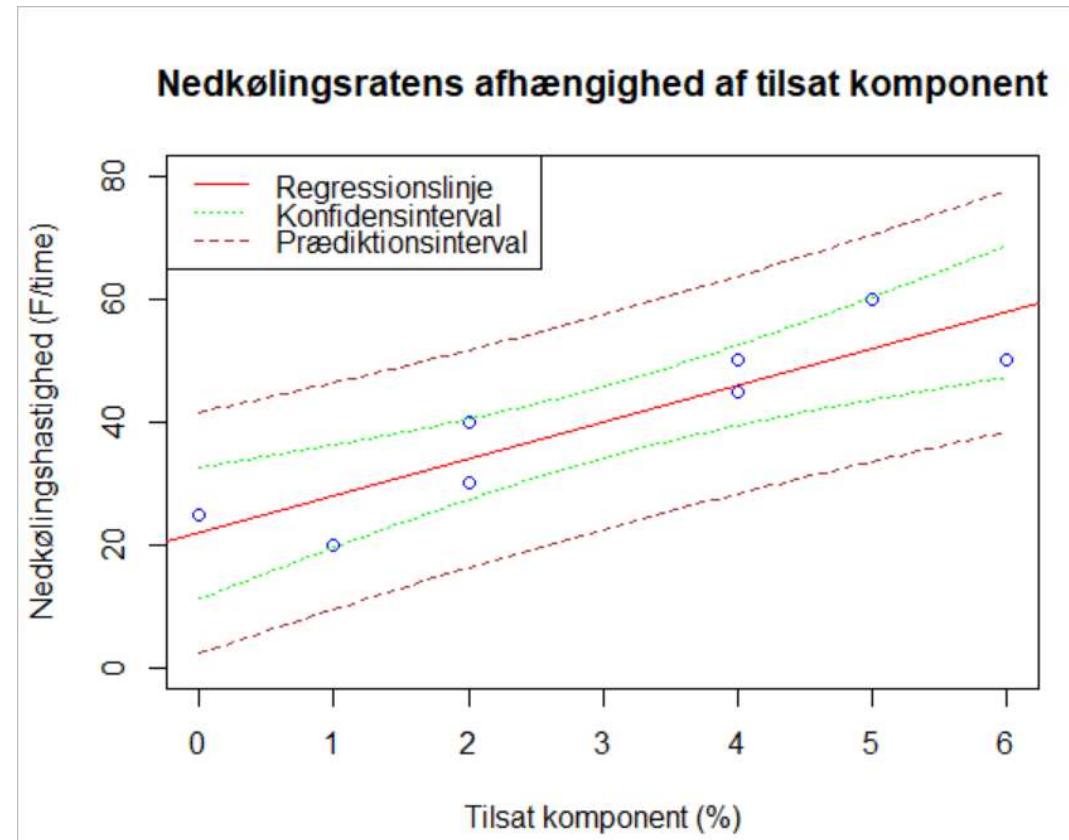
Konfidensinterval og prædiktionsinterval af $y(x)$

- Prædiktionsintervallet omkring $\hat{y}(x_0)$ beregnes med denne formel:

$$\hat{y}(x_0) \pm t_{\alpha/2} \cdot s_e \sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)}, \text{ hvor } t_{\alpha/2} \text{ er med } n - 2 \text{ d.f.}$$

Regression i R

- Scatter plot:
`plot(x, y, type="p")`
- Lineær regression:
`linmod = lm(y ~ x)`
`summary(linmod)`
- Regressionslinje på plot:
`abline(linmod)`
- Anova:
`anova(linmod)`
- Konfidensinterval for
koefficienterne β_0 og β_1 :
`confint(linmod, level=0.99)`
- Konfidens- og prædiktionsinterval:
`predict(linmod, level=0.95, interval='confidence')`
`predict(linmod, level=0.95, interval='prediction').`



Sandsynlighedsteori og statistik

Kapitel 11. Regressionsanalyse 2. del (afsnit 11.3-11.7)

Allan Leck Jensen
alj@ece.au.dk

Residualanalyse (11.5)

- Vi har udviklet lineære modeller til at prædiktere den afhængige variabel (responsvariablen) ud fra en eller flere uafhængige variable (regressorvariable)
- Residualet er forskellen på den observerede og modellens prædikterede værdi:

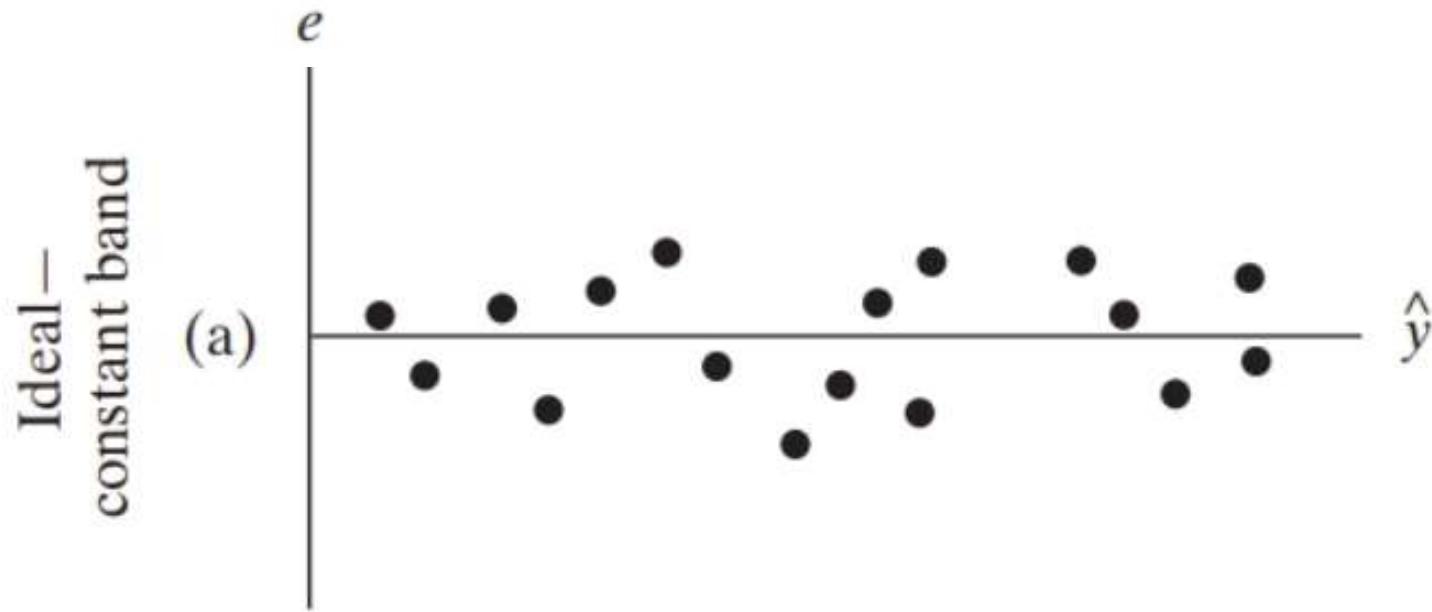
$$e_i = y_i - \hat{y}_i$$

- Residualet er således et mål for hvor meget modellen fejler i prædiktionen i et givet punkt
- Hvis et punkt har et stort residual, skyldes det enten at modellen er for dårlig, eller at punktet er en outlier (f.eks. en fejlmåling)
- Vi kan bruge residualerne til at teste om
 - modellens antagelser holder
 - modellen kan forbedres, f.eks. med en transformation
 - et punkt er en outlier
- En god metode er at lave residualplots.

Residualplots

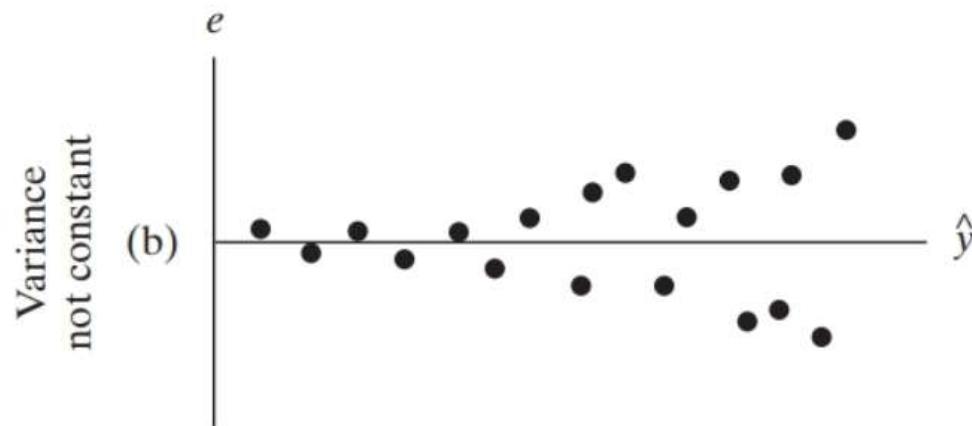
En type residualplots er residualer e plottet mod estimerater \hat{y}

- Residualerne skal gerne være tilfældige og ikke f.eks. afhænge af størrelsen af \hat{y}
- Her er ingen åbenlys sammenhæng – residualerne ser ud til at være tilfældigt fordelt i et bånd omkring 0. Det er idealt.

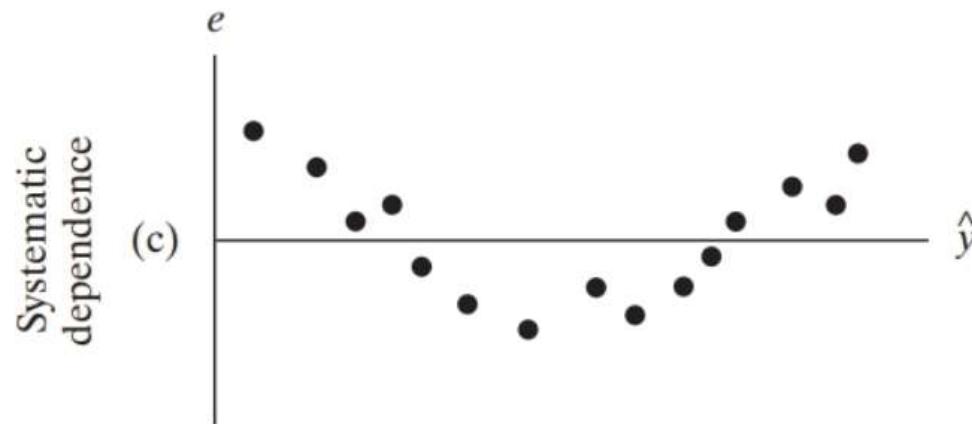


Fordeling af residualer

- Her stiger residualets numeriske værdi med prædiktionerne (funnel pattern)



- Her er en tydelig kurveformet sammenhæng (systematic curvature)



- I begge tilfælde tyder det på, at modellen kan forbedres.

Residualanalyse

- Simpel lineær regressionsmodel:

$$y = \beta_0 + \beta_1 x$$

- Vi har n sammenhørende observationer (x_i, y_i) for $i = 1 \dots n$
- Hvis vi forestiller os, at vi kender koefficienterne β_0 og β_1 kan vi skrive observationerne:

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

hvor ε_i er en tilfældig afvigelse (**random error**)

- Det antages, at afvigelserne ε_i er uafhængige og kommer fra en stokastisk variabel med middelværdi 0 og varians σ^2 .
Ofte antages det, at de kommer fra normalfordelingen $N(0, \sigma)$
- Vi kender ikke β_0 og β_1 , så vi ønsker at estimere dem ud fra vores observationer (x_i, y_i) :

$$y_i = b_0 + b_1 x_i + e_i$$

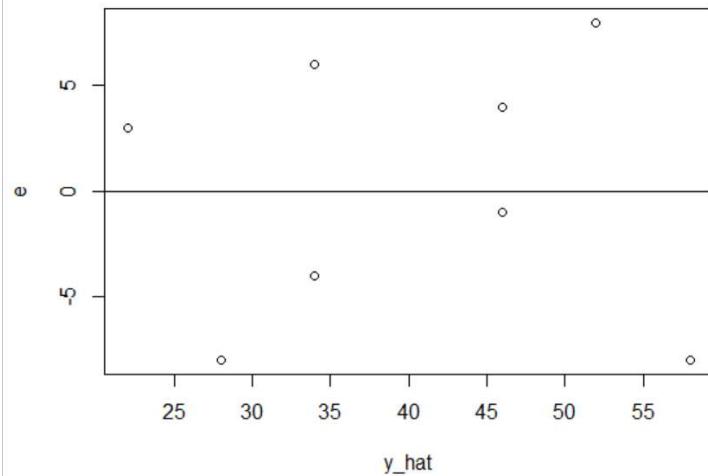
- Her er b_0 , b_1 og e_i vores bedste **estimater** for hhv. β_0 , β_1 og ε_i .
Flere observationer (en større stikprøve) kunne forbedre estimaterne
- Vi kan bruge vores residualer e_i til at estimere variansen σ^2 .

Kontrol af modelantagelser

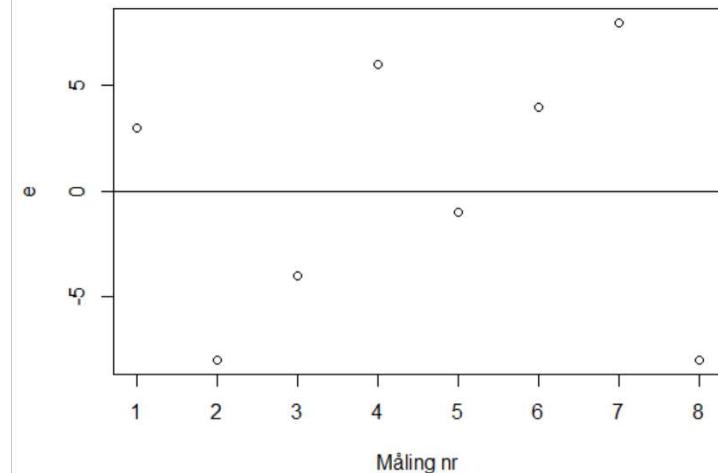
- Modelantagelser om random error ε : ε er $N(0, \sigma^2)$. M.a.o.:
 1. Middelværdi er 0
 2. Har konstant varians σ^2
 3. ε_i er uafhængige
 4. ε_i kommer fra normalfordeling
- Metoder til at kontrollere modelantagelser vha. residualer:
 - Mindste kvadraters metode sikrer antagelse 1 for residualerne
 - Plot residualer mod estimeret værdi (e_i mod \hat{y}_i) (2)
 - Plot residualer mod hver regressor (e_i mod x_{ji}) (2)
 - Plot residualer i tidsrækkefølge (e_i mod i) (3)
 - Stem-and-leaf plot eller histogram af residualer (4)
 - Normalfordelingsplot af residualer (4).

Eksempel 11.1, residualanalyse

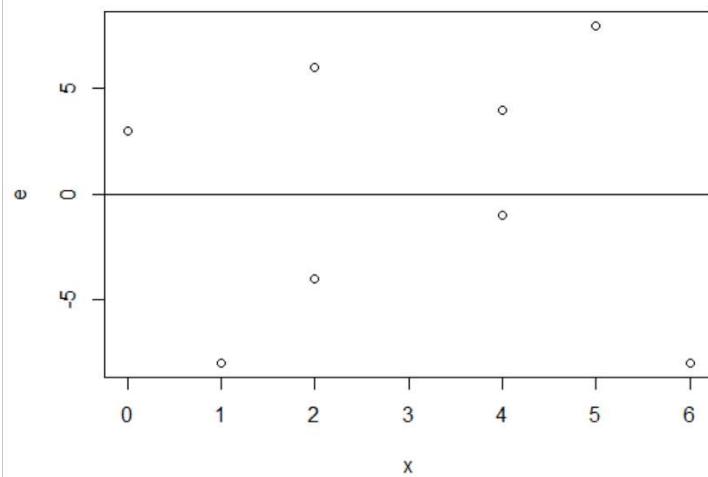
Eksempel 11.1: Residualplot for y_{hat}



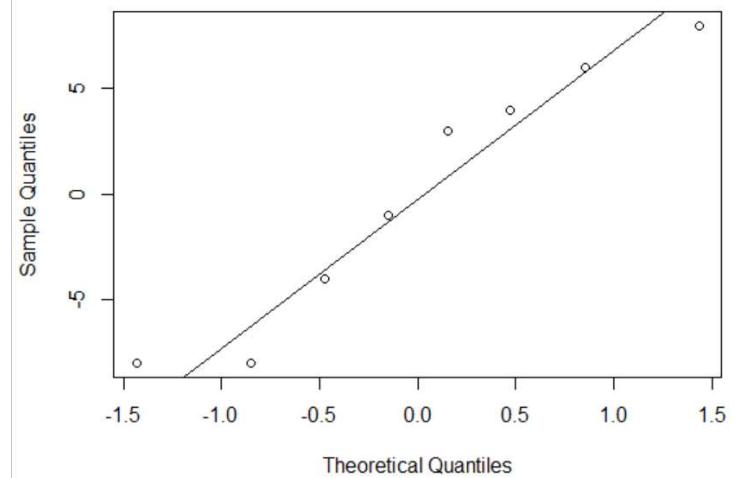
Eksempel 11.1: Residualplot for rækkefølge



Eksempel 11.1: Residualplot for x



Normal Q-Q Plot



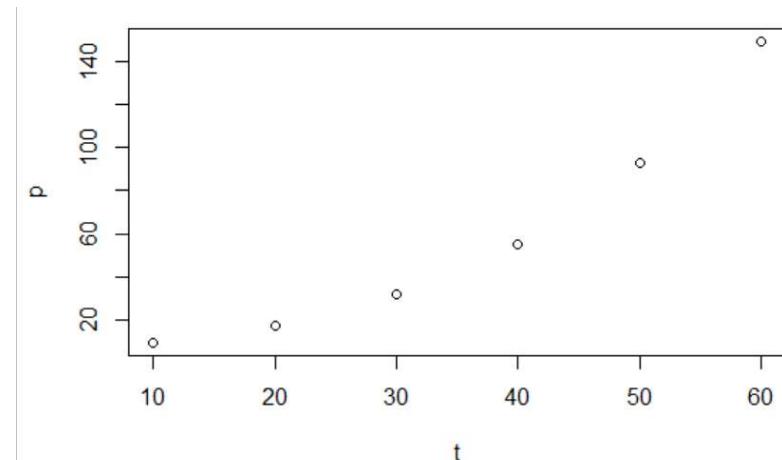
Kurvelineær regression (11.3)

- Vi kan bruge lineær regression også for ikke-lineære sammenhænge ved at transformere data
- Årsager til at transformere data:
 - En kendt/forventet sammenhæng
 - Plots viser en lineær sammenhæng mellem de transformerede data
 - Residualanalyse viser, at modelantagelser ikke holder, f.eks. ‘funnel pattern’ eller ‘systematic curvature’.

Eksempel: Damptryk

Sammenhørende målinger af temperatur og damptryk i en lukket beholder (analogt til bogens eksempel 11.10, men jeg synes dette illustrerer emnet bedre)

Temp (°C) t	Damptryk (mmHg) p
10	9.2
20	17.5
30	31.8
40	55.3
50	92.5
60	149.4



- Vi laver et scatter plot
- Selvom det ikke ser lineært ud, laver vi en lineær regression
- God model med R^2 over 90 % og signifikant hældning.

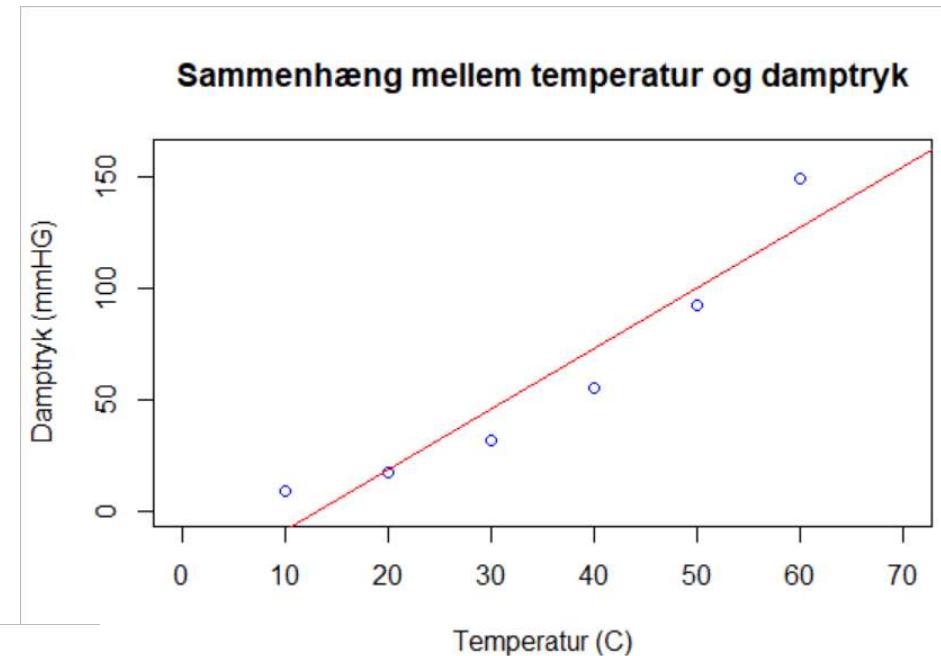
Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -35.6667    17.2317  -2.070  0.10725
t             2.7129     0.4425   6.131  0.00359 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

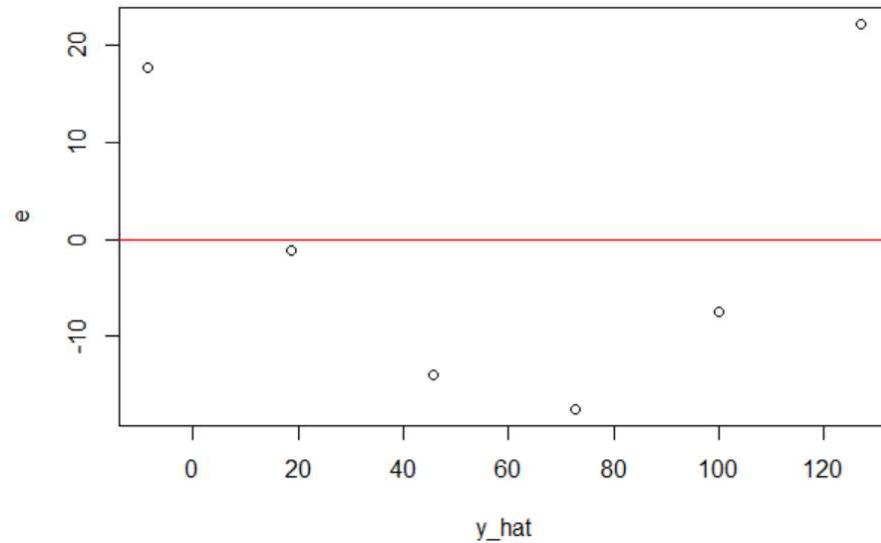
Residual standard error: 18.51 on 4 degrees of freedom
Multiple R-squared:  0.9038,    Adjusted R-squared:  0.8798
F-statistic: 37.59 on 1 and 4 DF,  p-value: 0.003586
```

Eksempel: Damptryk

- Problemer med modellen:
 - Fysisk: Modellen prædikterer negativt damptryk for $t < 10$
 - Statistisk: Data følger en ikke-lineær kurve
 - Statistisk: Et residualplot viser systematisk kurveform.



Eksempel om damptryk: Residualplot for $y_{\hat{}}$



Eksempel: Damptryk

- Clausius-Clapeyron ligningen siger, at logaritmen til damptrykket er omvendt proportionalt med temperaturen i Kelvin:

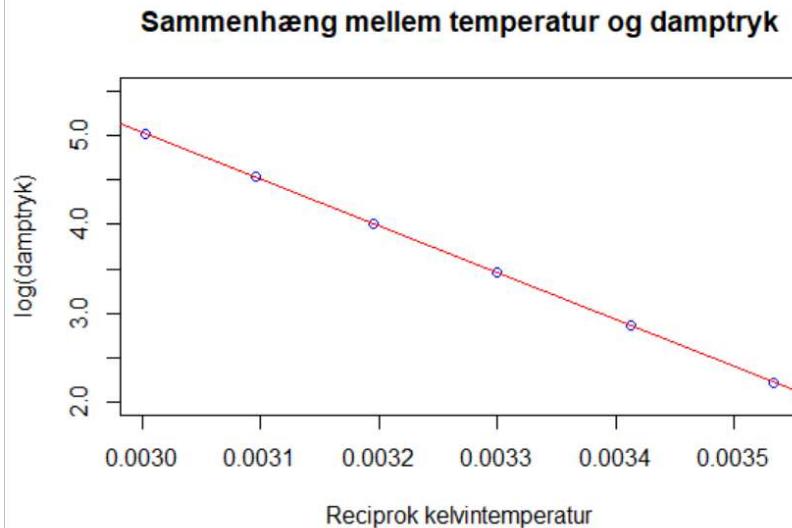
$$\ln(p) = b_0 + b_1 \frac{1}{T} = b_0 + b_1 \frac{1}{t+273}$$

- Ved at transformere vores data bør vi få en lineær sammenhæng:
 - $t \rightarrow RT = \frac{1}{(t+273)}$
 - $p \rightarrow \ln p = \ln(p)$

Temp (°C) t	Damptryk (mmHg) p	Temp Kelvin T	1/Temp RT	ln(damptryk) ln p
10	9.2	283	0.003534	2.2192
20	17.5	293	0.003413	2.8622
30	31.8	303	0.003300	3.4595
40	55.3	313	0.003195	4.0128
50	92.5	323	0.003096	4.5272
60	149.4	333	0.003003	5.0066

Eksempel: Damptryk

Resultatet er en perfekt model:



```
lm(formula = lnp ~ RT)
```

Residuals:

1	2	3	4	5
-0.0076268	0.0016464	0.0070176	0.0062507	0.0009209
				-0.0

Coefficients:

(Intercept)	RT	Estimate	Std. Error	t value	Pr(> t)
2.079e+01	-5.255e+03	5.418e-02	1.661e+01	383.8	2.77e-10 ***
				-316.3	5.99e-10 ***

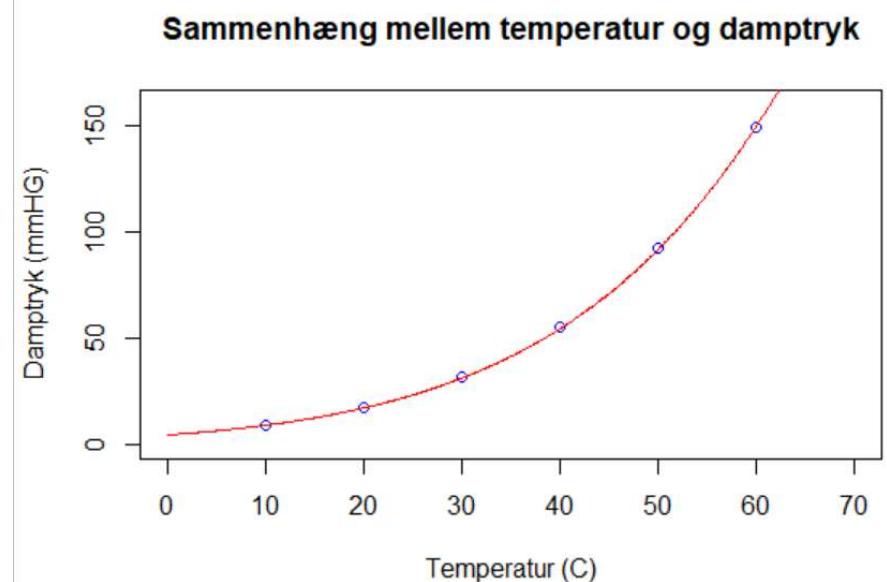
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 0.007373 on 4 degrees of freedom

Multiple R-squared: 1, Adjusted R-squared: 1

F-statistic: 1.001e+05 on 1 and 4 DF, p-value: 5.991e-10

- Den lineære model:
 $\ln p = 20.79 - 5255RT$
- Vi kan transformere tilbage til de oprindelige variable:
 $p = \exp(20.79 - \frac{5255}{T+273})$



Kurvelineær regression

- Vi kan bruge lineær regression også for ikke-lineære sammenhænge ved at transformere data
- Årsager til at transformere data:
 - En kendt/forventet sammenhæng (f.eks. damptrykseksemplet)
 - Plots viser en lineær sammenhæng mellem de transformerede data
 - Residualanalyse viser, at modelantagelser ikke holder, f.eks. ‘funnel pattern’ eller ‘systematic curvature’
- Sammenhænge, hvor transformation giver linearitet:
 - **Eksponentiel:** $y = \beta_0 \cdot \beta_1^x$
Transformation af y med log: $\log y = \log \beta_0 + x \cdot \log \beta_1$
 - **Reciprok:** $y = \frac{1}{\beta_0 + \beta_1 x}$
Transformation af y med reciprok: $\frac{1}{y} = \beta_0 + \beta_1 x$
 - **Potens:** $y = \beta_0 \cdot x^{\beta_1}$
Transformation af x og y med log: $\log y = \log \beta_0 + \beta_1 \cdot \log x$.

Polynomiel regression

- Hvis man ikke kender den underliggende sammenhæng i data kan man som regel lave en god model med polynomiel regression
- Vi har n punkter (x_i, y_i) . Vi vil fitte til et polynomium af grad p til vores data ($p < n$):

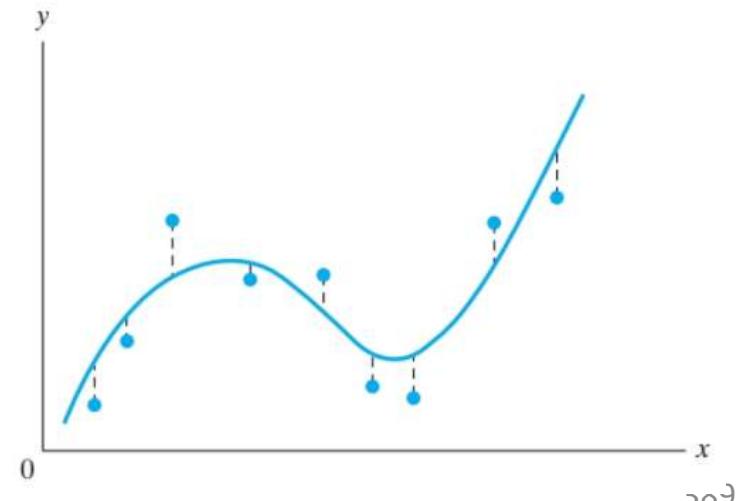
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p$$

- Det svarer til at bestemme estimeret b_0, b_1, \dots, b_p ‘bedst muligt’. Ifølge mindste kvadraters metode, skal vi minimere SSE :

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - (b_0 + b_1 x_i + b_2 x_i^2 + \cdots + b_p x_i^p))^2$$

- Det giver $p + 1$ normalligninger:

$$\begin{aligned}\sum y &= n b_0 + b_1 \sum x + \cdots + b_p \sum x^p \\ \sum xy &= b_0 \sum x + b_1 \sum x^2 + \cdots + b_p \sum x^{p+1} \\ &\vdots \\ \sum x^p y &= b_0 \sum x^p + b_1 \sum x^{p+1} + \cdots + b_p \sum x^{2p}\end{aligned}$$



Eksempel 11.11, s. 354 om tørretid af lak

Tørretiden af en lak afhænger af, hvor meget af et bestemt additiv, der er tilsat

- Lav et scatterplot
- Fit et andengrads-polynomium til data
- Hvad er den forventede tørretid, hvis der tilsættes 6.5 g additiv?

Amount of varnish additive (grams) <i>x</i>	Drying time (hours) <i>y</i>
0	12.0
1	10.5
2	10.0
3	8.0
4	7.0
5	8.0
6	7.5
7	8.5
8	9.0

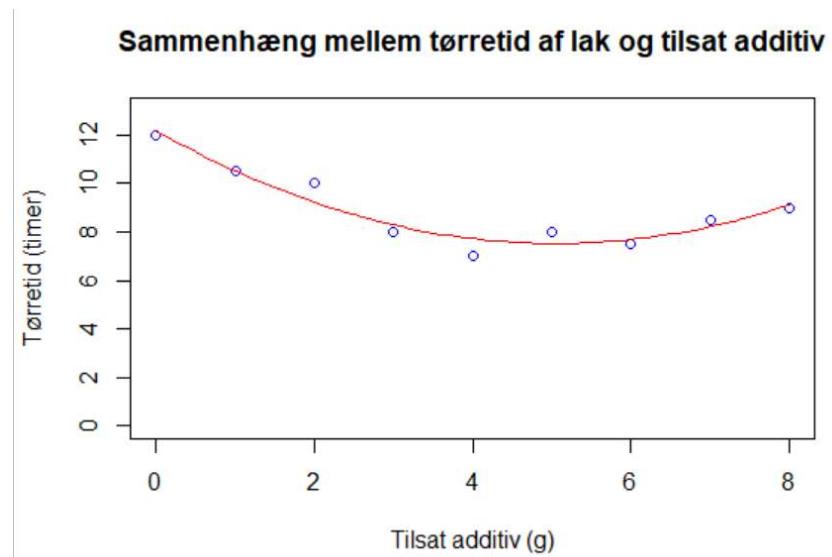
Løsning i R giver:

$$y = 12.185 - 1.847x + 0.183x^2$$

Når $x = 6.5$ sættes ind i formlen, fås en estimeret tørretid på

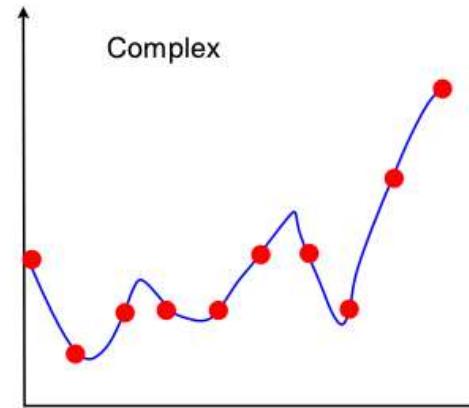
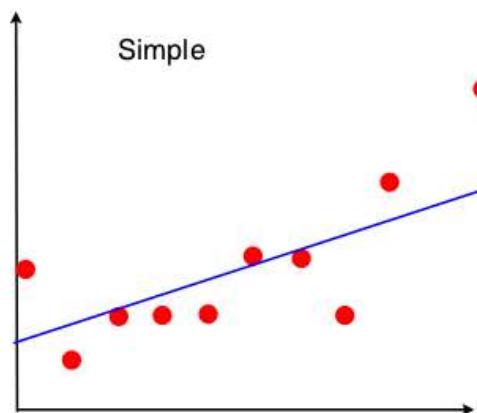
$$y = 7.9 \text{ timer}$$

Modellen estimerer min. tørretid på 7.5 timer for $x = \frac{1.847}{2 \cdot 0.183} = 5.05$ g additiv.



Eksempel 11.11, s. 354 om tørretid af lak

- Bemærk: Vi har $n = 9$ datapunkter, så vi kan i princippet fitte et 8-grads-polynomium perfekt til data
- Forskel på polynomiel interpolation og regression:
 - **Polynomiel interpolation:** Vi opfatter datapunkter som ‘sandheden’ og fitter et polynomium, så vi kan interpolere
 - **Polynomiel regression:** Vi opfatter data som kommende fra en underliggende polynomial sammenhæng, men med støj. Derfor søger vi en simpel model, der ‘beskriver data godt’.



Eksempel 11.11, s. 354 om tørretid af lak

- Hvis vi fitter et tredjegrads-polynomium til data fra eksempel 11.11:
 - R^2 stiger lidt (fra 0.9227 til 0.9237): Lidt mere variation i data forklares
 - Adjusted R^2 falder lidt (fra 0.8969 til 0.8779): Vi bliver straffet for en ekstra parameter til at forklare variationen
 - Det er sandsynligt, at den nye koefficient β_3 er nul, for p-værdien er høj (0.8094)
- Med andre ord er andengrads-polynomiet en bedre model for data.

```
lm(formula = y ~ x + I(x^2) + I(x^3))

Residuals:
    1      2      3      4      5      6      7
-0.12121 -0.05303  0.71753 -0.33225 -0.72511  0.51623 -0.13095
    9
-0.18182

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.121212   0.521159 23.258 2.73e-06 ***
x           -1.709416   0.601493 -2.842  0.0362 *
I(x^2)       0.137446   0.181606  0.757  0.4833
I(x^3)       0.003788   0.014896  0.254  0.8094
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5624 on 5 degrees of freedom
Multiple R-squared:  0.9237,    Adjusted R-squared:  0.8779
F-statistic: 20.17 on 3 and 5 DF,  p-value: 0.003188
```

Multipel lineær regression (11.4)

- **Multipel**: Der er mere end én regressor. Vi bruger r til at betegne antal regressorer: x_1, x_2, \dots, x_r
- **Lineær**: Modellen er at responsvariablen kan udtrykkes som en linearkombination af regressorerne:
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r$$
- Vi fitter modellen til data, der består af n $(r + 1)$ -tupler $(x_{i1}, x_{i2}, \dots, x_{ir}, y_i)$ for $i = 1 \dots n$.

Multipel lineær regression

- For $r = 2$ har vi 2 uafhængige variable (regressorer): x_1, x_2

Modellen er: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

- Vi minimerer SSE :

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i}))^2$$

- Planen i rummet (x_1, x_2, y) har forskriften

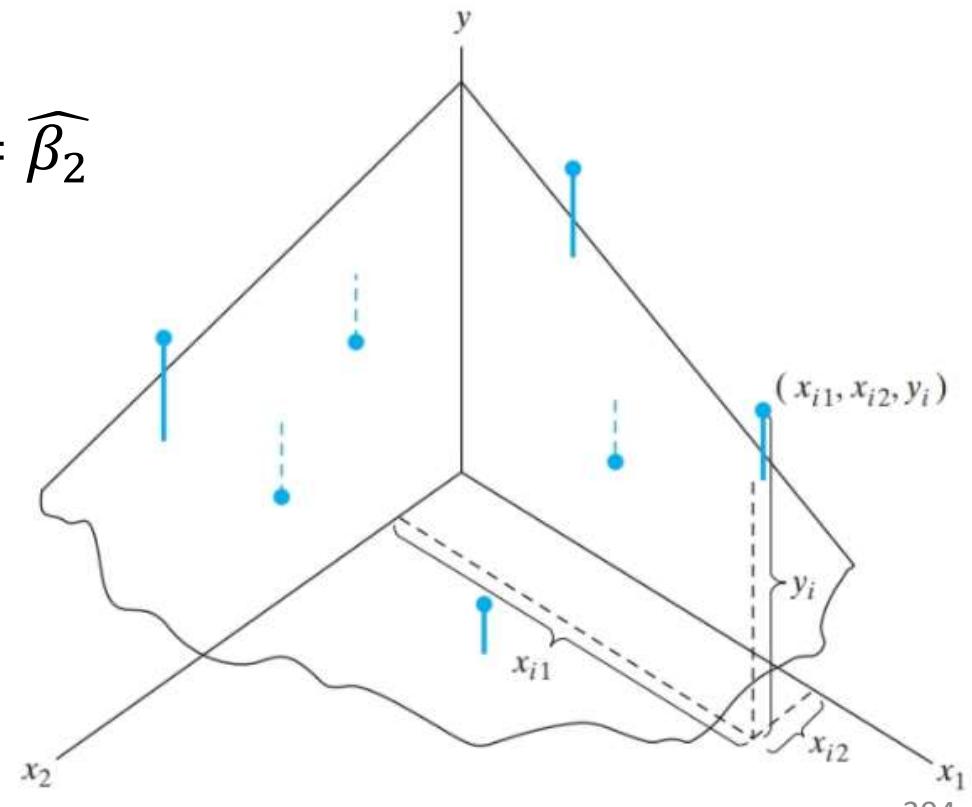
$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

- Her er $b_0 = \widehat{\beta}_0$, $b_1 = \widehat{\beta}_1$ og $b_2 = \widehat{\beta}_2$ bestemt vha. mindste kvadraters metode:

$$\sum y = n b_0 + b_1 \sum x_1 + b_2 \sum x_2$$

$$\sum x_1 y = b_0 \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2$$

$$\sum x_2 y = b_0 \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2$$



Eks. 11.12, s. 357 om knæk af en stang

Antal knæk, en stang kan klare, inden den brækker, afhænger af hvor meget, der er tilsat af to stoffer A og B

Multipel regression i R:

```
linmod = lm(y ~ x1 + x2)
summary(linmod)
```

Resultat:

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max 
 -4.938 -3.744 -1.050  2.825  7.513 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 46.4375   3.5172 13.203 6.59e-09 ***
x1          7.7750   0.9485  8.197 1.71e-06 ***
x2         -1.6550   0.1897 -8.724 8.55e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 4.242 on 13 degrees of freedom
Multiple R-squared: 0.9168, Adjusted R-squared: 0.904
F-statistic: 71.65 on 2 and 13 DF, p-value: 9.546e-08

Number of twists <i>y</i>	Percentage of element A		Percentage of element B
	<i>x₁</i>	<i>x₂</i>	
41	1		5
49	2		5
69	3		5
65	4		5
40	1		10
50	2		10
58	3		10
57	4		10
31	1		15
36	2		15
44	3		15
57	4		15
19	1		20
31	2		20
33	3		20
43	4		20

Regressionsligning:

$$\hat{y} = 46.4 + 7.78x_1 - 1.66x_2$$

Eks. 11.12, s. 357 om knæk af en stang

Er det en god model?

Ja, det er, fordi:

- Alle tre koefficienter har meget lave p-værdier. Dermed er det ekstremt usandsynligt, at koefficienterne i virkeligheden er 0
- Modellen forklarer knap 92% af variationen i data ($R^2 = 0.9168$). Adjusted $R^2 = 0.904$ er tæt på 0.9168, så modellen overfitter ikke
- F-statistikken på 71.65 har en tilhørende p-værdi på $9.5 \cdot 10^{-8}$. Det er ekstremt usandsynligt, at begge hældningskoefficienter er 0 samtidig
- F-statistikken kommer fra en variansanalyse (ANOVA).

```
Call:  
lm(formula = y ~ x1 + x2)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-4.938 -3.744 -1.050  2.825  7.513  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 46.4375   3.5172 13.203 6.59e-09 ***  
x1          7.7750   0.9485  8.197 1.71e-06 ***  
x2         -1.6550   0.1897 -8.724 8.55e-07 ***  
---  
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 4.242 on 13 degrees of freedom  
Multiple R-squared:  0.9168,    Adjusted R-squared:  0.904  
F-statistic: 71.65 on 2 and 13 DF,  p-value: 9.546e-08
```

ANOVA

Konceptuelt det samme som for simpel regression, det er blot en generalisering fra 1 til r regressorvariable (ændringer vist i **rød skrift**):

Kilder	Frihedsgrader (DF)	Sum of Squares (SS)	Mean Squares (MS)	F
Regression	df_{reg} $= \#parametre - 1$ $= (r + 1) - 1 = r$	SS_{reg} $= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	MS_{reg} $= \frac{SS_{reg}}{df_{reg}}$	F $= \frac{MS_{reg}}{MS_{res}}$
Residual	df_{res} $= \#obs. - \#parametre$ $= n - (r + 1)$	SS_{res} $= \sum_{i=1}^n (y_i - \hat{y}_i)^2$	MS_{res} $= \frac{SS_{res}}{df_{res}}$	
Total	df_{total} $= \#observationer - 1$ $= n - 1$	SS_{total} $= \sum_{i=1}^n (y_i - \bar{y})^2$		

ANOVA i R

- Jeg har ikke fundet en R-funktion, der direkte beregner ANOVA tabellen for multipel regression, som for simpel regression
- Det er ikke alvorligt, for F-teststørrelsen og tilhørende p-værdi kan aflæses i output fra summary():

```
Residual standard error: 4.242 on 13 degrees of freedom
Multiple R-squared:  0.9168,   Adjusted R-squared:  0.904
F-statistic: 71.65 on 2 and 13 DF,  p-value: 9.546e-08
```

- Hvis man ønsker det, kan ANOVA tabellen dannes med denne R-kode:

```
# Alternativ ANOVA beregning af F og tilhørende p-værdi:
constmod = lm(y ~ 1) # Her laves den konstante model uden hældning: y = b_0
summary(constmod) # Resultatet af den konstante model
anova(constmod, linmod) # F beregnes ved sammenligningen af de to modeller
```

- Resultat:

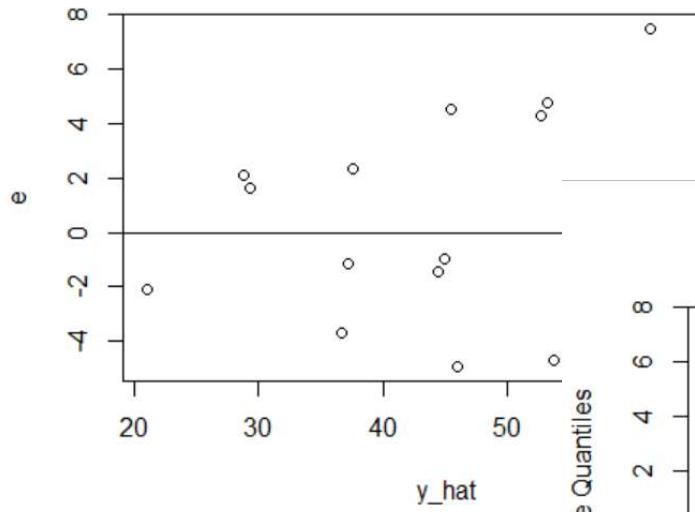
	df	SS	MS	F
Reg	2	2578	1289	71.6
Res	13	234	18	
Tot	15	2812		

```
Analysis of Variance Table

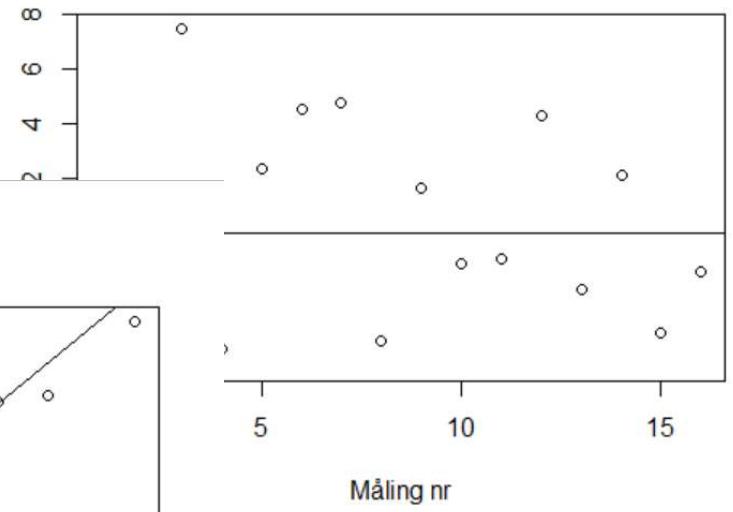
Model 1: y ~ 1
Model 2: y ~ x1 + x2
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
  1      15 2812.44
  2      13 233.91  2     2578.5 71.653 9.546e-08 ***
---
signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Eksempel 11.12, residualanalyse

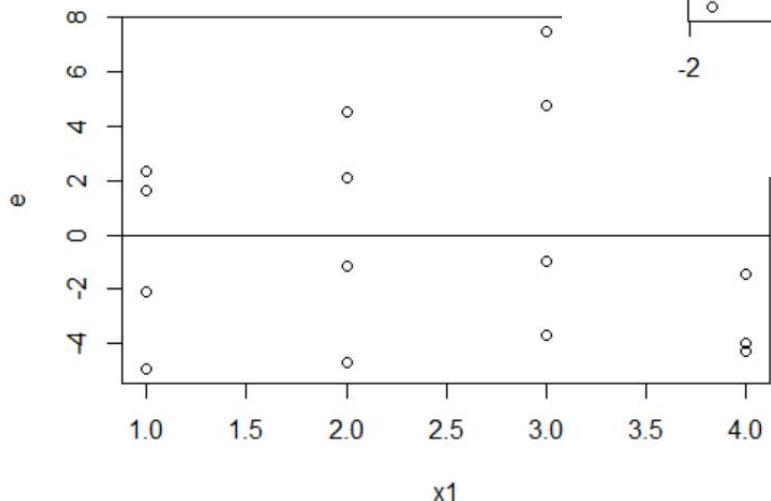
Eksempel 11.12: Residualplot for y_{hat}



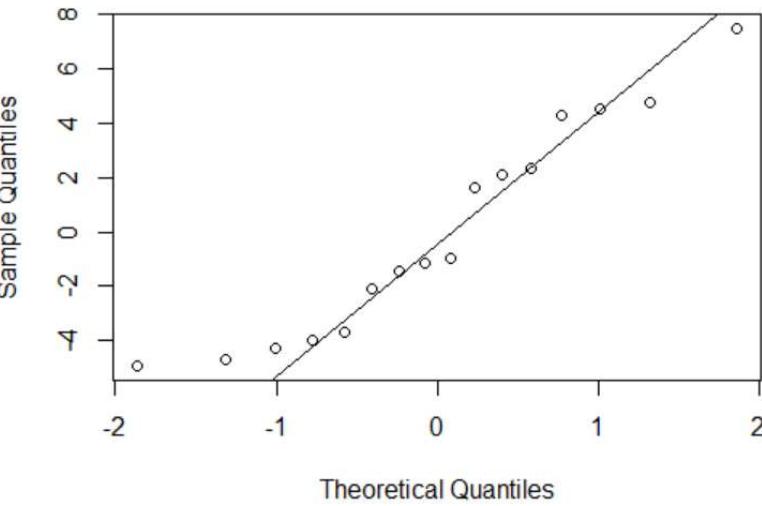
Eksempel 11.12: Residualplot for rækkefølge



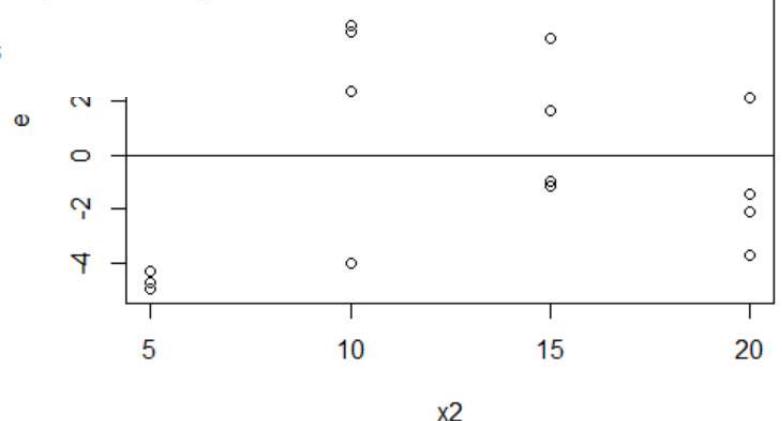
Eksempel 11.12: Residual



Normal Q-Q Plot



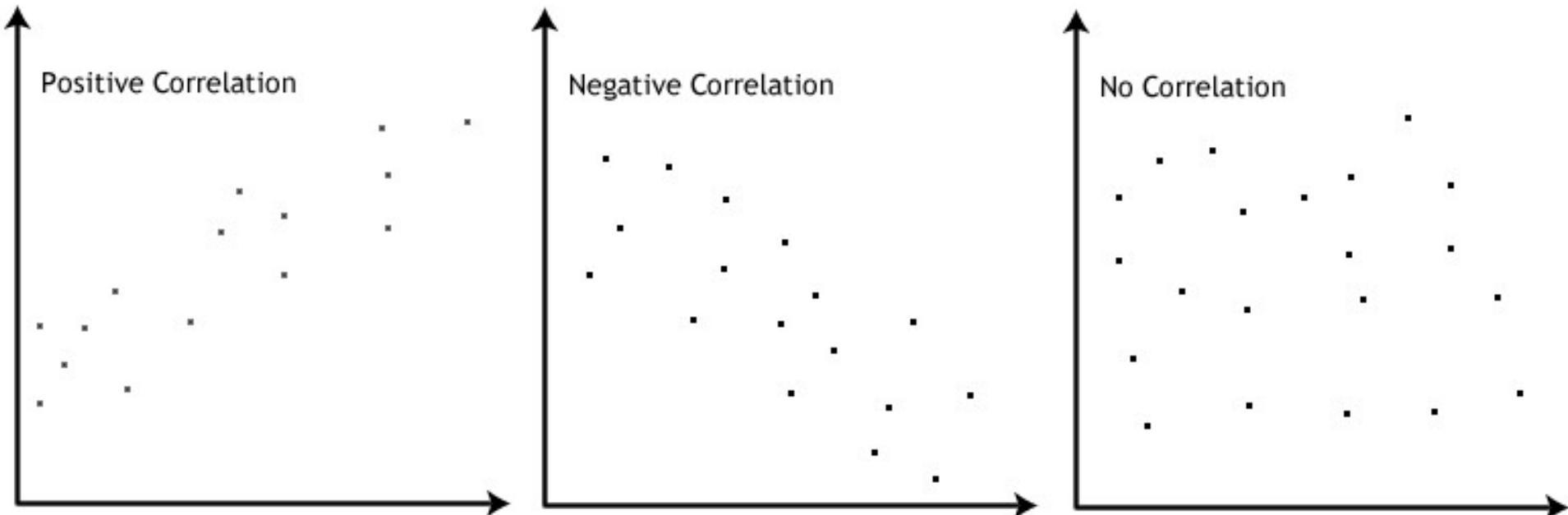
Eksempel 11.12: Residualplot for x_2



Korrelation (11.6)

Korrelation

- Sammenhængen mellem y og x kan være
 - Positivt korreleret, hvis y typisk vokser, når x vokser
 - Negativt korreleret, hvis y typisk falder, når x vokser
 - Ukorreleret, hvis værdien af y ikke lader til at afhænge af værdien af x
- Korrelation er ikke det samme som **kausalitet** (årsagssammenhæng)
- Statistik kan kun vise korrelation, ikke kausalitet.



Korrelationskoefficienten r

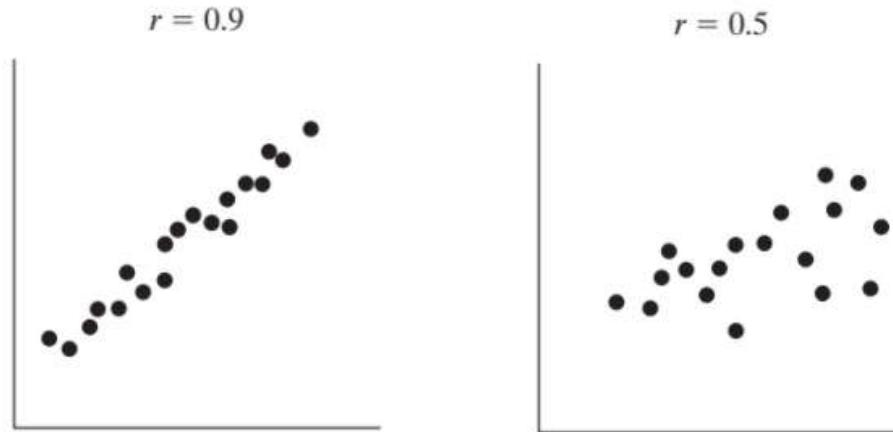
- Vi har n observationer (x_i, y_i) . I hvor høj grad er der korrelation mellem x og y ?
- Vi kan standardisere hver måling af x ved at trække middelværdien \bar{x} fra målingen og dele med standardafvigelsen s_x :
$$x_i^* = \frac{x_i - \bar{x}}{s_x}$$
- Vi kan gøre tilsvarende for målingerne af y :
$$y_i^* = \frac{y_i - \bar{y}}{s_y}$$
- Vores standardiserede målinger x_i^* og y_i^* er uden enhed og i sammenlignelig størrelsesorden, fordelt omkring 0
- Datasættets **korrelationskoefficient r** defineres som:

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

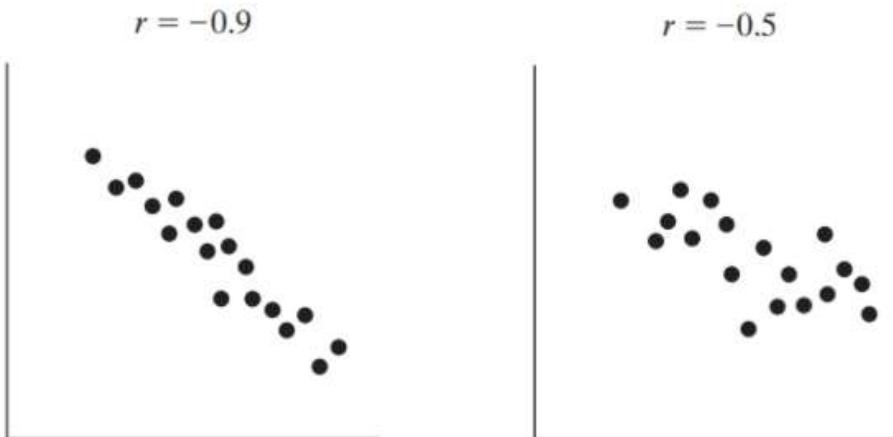
- Hvis data er sådan, så x_i^* og y_i^* typisk har samme fortegn, så er $r > 0$
- Det svarer til, at hvis x_i er over/under \bar{x} , så er y_i også over/under \bar{y}
- Hvis x_i^* og y_i^* typisk har modsat fortegn, så er $r < 0$.

Korrelationskoefficienten r

- Man kan vise, at $-1 \leq r \leq 1$
- Hvis $r > 0$: Positiv korrelation mellem x og y . Jo større værdi desto tættere på en ret linje ligger punkterne

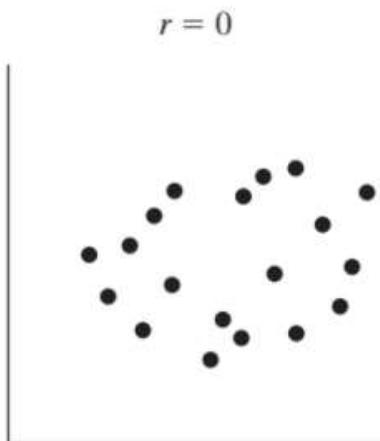


- Hvis $r < 0$: Negativ korrelation mellem x og y .

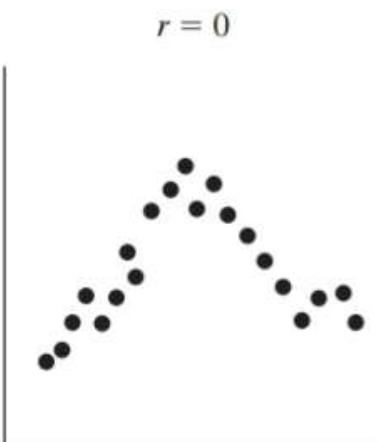


Korrelationskoefficienten r

- Hvis $r \approx 0$: Der er ikke en lineær sammenhæng mellem x og y



- Der kan dog godt være en ikke-lineær sammenhæng mellem x og y .



Korrelationskoefficienten r

- Man kan også vise:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

- ... og:

$$r = \sqrt{R^2}$$

hvor R^2 er *determinationskoefficienten*, der er angivet i output fra lineær regression. R^2 betegner, hvor stor en del af variationen i data, der kan forklares af den lineære sammenhæng

- Beregning af korrelationskoefficienten mellem x og y i R:

$$r = \text{cor}(x, y)$$

Mindste kvadraters metode (Matrix) (11.7)

- Simpel: To partielle differentialligninger:

$$\frac{\partial}{\partial b_0} (\sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2) = 0$$

$$\frac{\partial}{\partial b_1} (\sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2) = 0$$

- Multipel: $r + 1$ partielle differentialligninger:

$$\frac{\partial}{\partial b_0} (\sum_{i=1}^n [y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_r x_{ir})]^2) = 0$$

$$\frac{\partial}{\partial b_1} (\sum_{i=1}^n [y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_r x_{ir})]^2) = 0$$

⋮

$$\frac{\partial}{\partial b_r} (\sum_{i=1}^n [y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_r x_{ir})]^2) = 0$$

Mindste kvadraters metode

- Simpel: To ligninger med to ubekendte (normalligningerne)

$$nb_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

$$b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

- Multipel: $r + 1$ ligninger med $r + 1$ ubekendte:

$$nb_0 + b_1 \sum_{i=1}^n x_{i1} + b_2 \sum_{i=1}^n x_{i2} + \cdots + b_r \sum_{i=1}^n x_{ir} = \sum_{i=1}^n y_i$$

$$b_0 \sum_{i=1}^n x_{i1} + b_1 \sum_{i=1}^n x_{i1}^2 + b_2 \sum_{i=1}^n x_{i1} x_{i2} + \cdots + b_r \sum_{i=1}^n x_{i1} x_{ir} = \sum_{i=1}^n x_{i1} y_i$$

$$b_0 \sum_{i=1}^n x_{i2} + b_1 \sum_{i=1}^n x_{i1} x_{i2} + b_2 \sum_{i=1}^n x_{i2}^2 + \cdots + b_r \sum_{i=1}^n x_{i2} x_{ir} = \sum_{i=1}^n x_{i2} y_i$$

⋮

$$b_0 \sum_{i=1}^n x_{ir} + b_1 \sum_{i=1}^n x_{i1} x_{ir} + b_2 \sum_{i=1}^n x_{i2} x_{ir} + \cdots + b_r \sum_{i=1}^n x_{ir}^2 = \sum_{i=1}^n x_{ir} y_i$$

Løsning med matrix notation

- Det kan vises, at de $r + 1$ ligninger med $r + 1$ ubekendte kan skrives som (X' er den transponerede matrix til X):

$$X'Xb = X'y$$

hvor

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1r} \\ 1 & x_{21} & x_{22} & \ddots & x_{2r} \\ \vdots & & & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nr} \end{bmatrix}$$

$$\mathbf{b} = [b_0, b_1, \dots, b_r]'$$

$$\mathbf{y} = [y_1, y_2, \dots, y_n]'$$

- Hvis $X'X$ er ikke-singulær er det muligt at finde den inverse matrix og dermed bestemme koefficienterne \mathbf{b} :

$$\mathbf{b} = (X'X)^{-1}X'y.$$

Eksempel 11.20, s. 380

Fitting a straight line using the matrix formulas

Use the matrix relations to fit a straight line to the data

x	0	1	2	3	4
y	8	9	4	3	1

Here $k = 1$ and, dropping the subscript 1, we have

\mathbf{X}'	\mathbf{y}	$\mathbf{X}'\mathbf{X}$	$(\mathbf{X}'\mathbf{X})^{-1}$	$\mathbf{X}'\mathbf{y}$
$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \end{bmatrix}$	$\begin{bmatrix} 8 \\ 9 \\ 4 \\ 3 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 5 & 10 \\ 10 & 30 \end{bmatrix}$	$\begin{bmatrix} 0.6 & -0.2 \\ -0.2 & 0.1 \end{bmatrix}$	$\begin{bmatrix} 25 \\ 30 \end{bmatrix}$

Consequently,

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} 0.6 & -0.2 \\ -0.2 & 0.1 \end{bmatrix} \begin{bmatrix} 25 \\ 30 \end{bmatrix} = \begin{bmatrix} 9 \\ -2 \end{bmatrix}$$

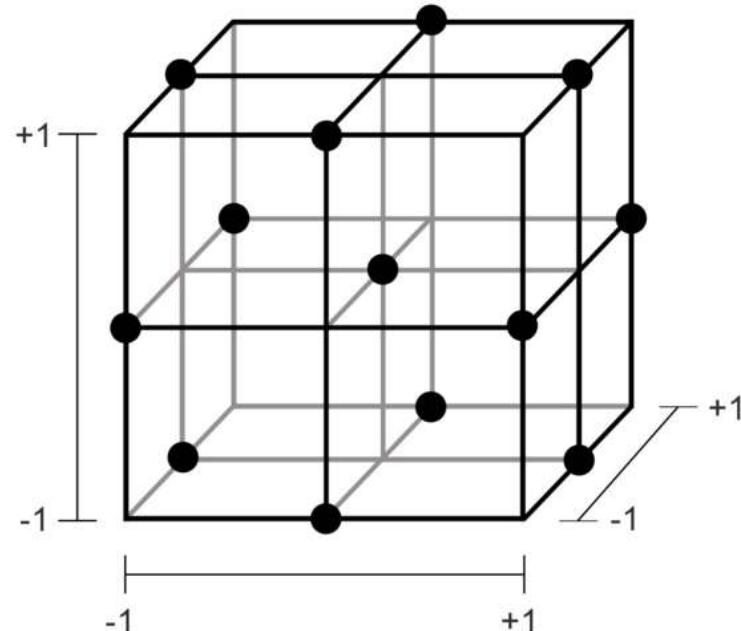
and the fitted equation is

$$\hat{y} = 9 - 2x$$

- Lad os prøve i R.

Eksempel (optimeret popcorn-popning)

- Minimering af antal uspiselige popcornkerner, enten brændte eller upoppedede (y)
- Tre faktorer med hver tre niveauer:
 - Kogepladens varme (x_1 : Temp.): 5, 6, 7
 - Oliemængde (x_2 : Oil): 2, 3, 4
 - Popningstid (x_3 : Time): 75, 90, 105
- I alt $3^3 = 27$ kombinationer, her reduceret til 15 med et såkaldt 'Box-Behnken design':



Temp.	Oil	Time	y
7	4	90	24
5	3	105	28
7	3	105	40
7	2	90	42
6	4	105	11
6	3	90	16
5	3	75	126
6	2	105	34
5	4	90	32
6	2	75	32
5	2	90	34
7	3	75	17
6	3	90	30
6	3	90	17
6	4	75	50

Box-Behnken har 13 observationer, vi har 15. Hvorfor?
Der er 3 gentagelser af kombinationen 6, 3, 90 (centrum i boksen)

Eksempel ('poptimering')

- Minimering af antal uspiselige popcornkerner, enten brændte eller upoppedede (y)
- Tre faktorer med hver tre niveauer:
 - Kogepladens varme (x_1 : Temp.): 5, 6, 7
 - Oliemængde (x_2 : Oil): 2, 3, 4
 - Popningstid (x_3 : Time): 75, 90, 105
- I alt $3^3 = 27$ kombinationer, her reduceret til 15 med Box-Behnken design

- Lineær model:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

- Lineær model med *kvadratled*:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{11}x_1^2 + b_{22}x_2^2 + b_{33}x_3^2$$

- Lineær model med *kvadratled og interaktioner* mellem faktorer:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{11}x_1^2 + b_{22}x_2^2 + b_{33}x_3^2 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 .$$

Temp.	Oil	Time	y
7	4	90	24
5	3	105	28
7	3	105	40
7	2	90	42
6	4	105	11
6	3	90	16
5	3	75	126
6	2	105	34
5	4	90	32
6	2	75	32
5	2	90	34
7	3	75	17
6	3	90	30
6	3	90	17
6	4	75	50

Eksempel ‘poptimering’ regnet i R

- Hvordan laver man model med kvadratled og interaktioner?
- Almindelig lineær model:

$$m = lm(y \sim x_1 + x_2 + x_3)$$

- Model med interaktioner:

$$m = lm(y \sim x_1 + x_2 + x_3 + x_1:x_2 + x_1:x_3 + x_2:x_3)$$

- Model med interaktioner og kvadratled:

Dette virker ikke:

$$\begin{aligned} m = lm(y \sim & x_1 + x_2 + x_3 \\ & + x_1^2 + x_2^2 + x_3^2 \\ & + x_1:x_2 + x_1:x_3 + x_2:x_3) \end{aligned}$$

Dette virker (sæt ‘I()’ omkring hvert kvadratled):

$$\begin{aligned} m = lm(y \sim & x_1 + x_2 + x_3 \\ & + I(x_1^2) + I(x_2^2) + I(x_3^2) \\ & + x_1:x_2 + x_1:x_3 + x_2:x_3) \end{aligned}$$

- **Skridtvis heuristisk metode:** Fjern det mindst signifikante led, med mindre det reducerer Adj. R² for meget. Demo!

Sandsynlighedsteori og statistik

Kapitel 12. Variansanalyse (ANOVA) og eksperimenter med en faktor (afsnit 12.1-12.2, 12.4)

Allan Leck Jensen

alj@ece.au.dk

Eksperimenter

- De fleste processer påvirkes samtidig af mange forskellige forhold, f.eks. tryk, temperatur, tid, støkiometri i en produktionsproces
- Ofte spiller kendte og ukendte faktorer ind og giver anledning til usikkerhed (variation i processens udbytte)
- Ofte ved man ikke nok om, hvordan man skal forbedre/optimere processen. Så kan man eksperimentere
- Sekvens af aktiviteter i forbindelse med et eksperiment:
 1. **Formodning** – hvad er det man vil undersøge? F.eks. om temperatur påvirker processen
 2. **Eksperiment** – man afprøver formodningen struktureret ved at måle responsen med forskellige niveauer af en eller flere variable, mens alle andre forhold forsøges at holdes ensartet
 3. **Analyse** – de målte resultater af eksperimentet analyseres statistisk
 4. **Konklusion** – resultatet af eksperimentet gøres op, hvad har vi lært?
 5. Evt. ny formodning og nyt eksperiment – hvis temperatur påvirker processen ønsker vi at finde det optimale temperatur-område.

Eksempel 12.3, s. 397 (styrke af resiner)

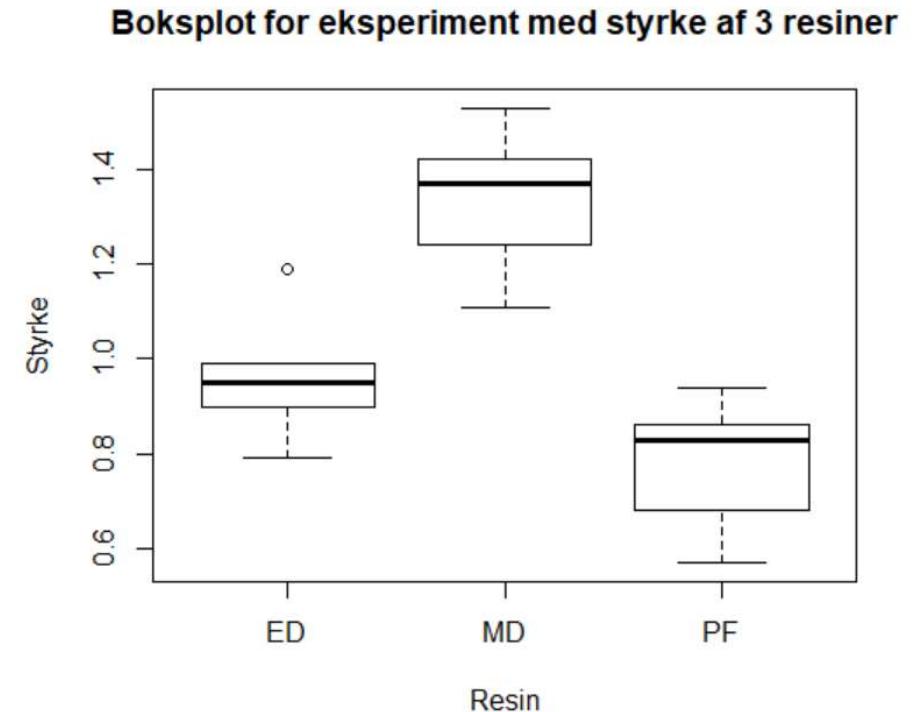
- 'Resin' betyder harpiks, men der findes også syntetiske resiner, som bruges i epoxylim. **Formodningen** er, at der er forskel i styrken af de indre, kemiske bindinger for 3 udvalgte resiner, EM, MD og PF
- Vi laver et *fuldstændigt randomiseret enkelt-faktor eksperiment*
- 1 faktor: Type af resin
 - 3 niveauer (treatments, behandlinger): EM, MD og PF
 - 5 gentagelser (replications)
- Antal prøver: $1 \cdot 3 \cdot 5 = 15$
- Respons: Resinens bindingsstyrke måles for hver prøve
- Hver prøve fremstilles og måles i tilfældig rækkefølge (**randomiseret**).

Resin	Strength					Mean
ED	0.99	1.19	0.79	0.95	0.90	0.964
MD	1.11	1.53	1.37	1.24	1.42	1.334
PF	0.83	0.68	0.94	0.86	0.57	0.776

Eksempel 12.3, s. 397 (styrke af resiner)

Mean
0.964
1.334
0.776

- Der er forskel på gns. styrke for de tre resiner. Men er forskellene signifikante?
- Parallelt boksplot for hvert niveau viser ligeledes, at styrken er forskellig. MD lader til at være bedst. ED er måske bedre end PF, men forskellen kan være tilfældig
- For hvert niveau er fordelingerne nogenlunde ensartede: symmetriske og med ensartet varians, men det er et tyndt grundlag at vurdere ud fra (5 observationer pr. boksplot).



Generelt enkelt-faktor eksperiment

- 1 faktor med k niveauer (behandlinger)
- For den i 'te behandling har vi n_i gentagelser
- Antal prøver, der skal måles respons på er $N = \sum_{i=1}^k n_i$
- y_{ij} er responsen af j 'te gentagelse af i 'te behandling
- Middelværdi af den i 'te behandling:

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \quad \text{for } i = 1, 2, \dots, k$$

- Overordnet middelværdi:

$$\bar{y} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}$$

- Et eksperiment med samme antal gentagelser for hver behandling, altså

$n_1 = n_2 = \dots = n_k$
kaldes *balanceret*.

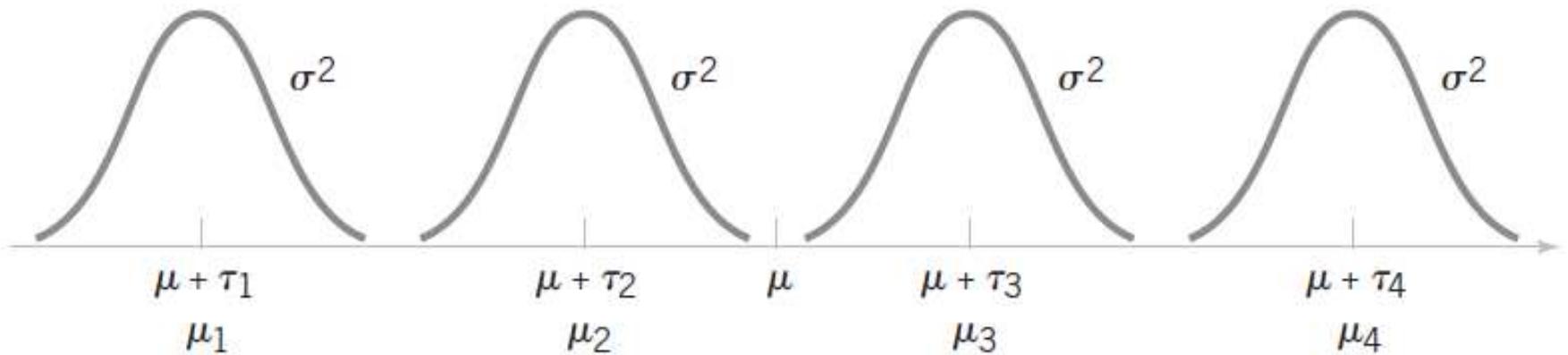
<i>Observations</i>	<i>Means</i>
<i>Sample 1 :</i> $y_{11}, y_{12}, \dots, y_{1j}, \dots, y_{1n_1}$	\bar{y}_1
<i>Sample 2 :</i> $y_{21}, y_{22}, \dots, y_{2j}, \dots, y_{2n_2}$	\bar{y}_2
⋮	⋮
<i>Sample i :</i> $y_{i1}, y_{i2}, \dots, y_{ij}, \dots, y_{in_i}$	\bar{y}_i
⋮	⋮
<i>Sample k :</i> $y_{k1}, y_{k2}, \dots, y_{kj}, \dots, y_{kn_k}$	\bar{y}_k

Lineær statistisk model

- y_{ij} kommer fra en stokastisk variabel:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1, \dots, k; \quad j = 1, \dots, n_i$$

- μ er den forventede respons (her styrke) på tværs af behandlinger
- α_i er effekten af den i 'te behandling. $\sum \alpha_i = 0$
- ε_{ij} er tilfældig afvigelse (random error). Antages normalfordelt $N(0, \sigma^2)$



- Det svarer til, at hver behandling i er normalfordelt $N(\mu + \alpha_i, \sigma^2)$
- Vi antager varianshomogenitet, altså σ^2 er uafhængig af behandling.

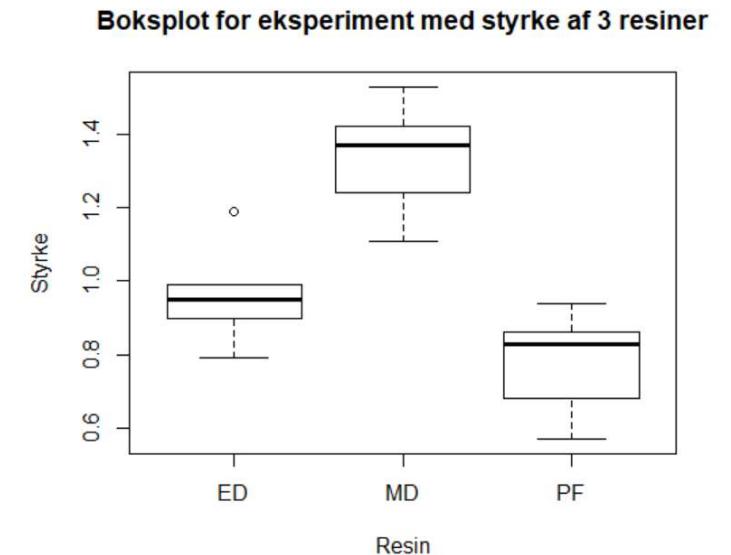
Hypotesetest

- Er der en effekt af behandlingerne?

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$$

$$H_1: \alpha_i \neq 0 \text{ for mindst een behandling } i$$

- Hvis H_0 er sand: $y_{ij} = \mu + \varepsilon_{ij}$ $i = 1, \dots, k; j = 1, \dots, n_i$
dvs. alle N observationer er $N(\mu, \sigma)$
- Hvis H_0 er sand for eksemplet med resiners styrke, så betyder det, at alle tre resiner har samme styrke. Så er forskellene i det parallelle boksplot tilfældige
- Vi undersøger hypotesen med variansanalyse, ANOVA. Vi undersøger, om hvor meget af variationen, der kan forklares med behandlingerne.



ANOVA fra regression (kap. 11)

Kilder	Frihedsgrader (DF)	Sum of Squares (SS)	Mean Squares (MS)	F
Regression	df_{reg} $= \#parametre - 1$ $= (r + 1) - 1 = r$	SS_{reg} $= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	MS_{reg} $= \frac{SS_{reg}}{df_{reg}}$	F $= \frac{MS_{reg}}{MS_{res}}$
Residual	df_{res} $= \#obs. - \#parametre$ $= n - (r + 1)$	SS_{res} $= \sum_{i=1}^n (y_i - \hat{y}_i)^2$	MS_{res} $= \frac{SS_{res}}{df_{res}}$	
Total	df_{total} $= \#observationer - 1$ $= n - 1$	SS_{total} $= \sum_{i=1}^n (y_i - \bar{y})^2$		

ANOVA i eksperimenter

- Samme koncept, men forskelle i notation:
- Regression: $SS_{total} = SS_{reg} + SS_{res}$
- Eksperimenter: $SST = SS(Tr) + SSE$
- Konceptuelt det samme: Den samlede variation (SS_{total} eller SST) kan opdeles i 2 dele:
 - den del, der forklares af regressionsmodellen / behandlingen (SS_{reg} eller $SS(Tr)$) og
 - den del, der ikke kan forklares (SS_{res} eller SSE), og som opfattes som tilfældig støj.

ANOVA med notation fra eksperimenter

	Frihedsgrader (DF)	Sum of Squares (SS)	Mean Squares (MS)	F
Treatments	$df(Tr)$ $= \# \text{behandlinger}$ $- 1$ $= k - 1$	$SS(Tr)$ $= \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2$	$MS(Tr)$ $= \frac{SS(Tr)}{df(Tr)}$	$F_0 =$ $\frac{MS(Tr)}{MSE}$
Error	dfE $= \# \text{obs.} - \# \text{behandl.}$ $= N - k$	SSE $= \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	MSE $= \frac{SSE}{dfE}$	
Total	dfT $= \# \text{obs.} - 1$ $= N - 1$	SST $= \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$		

Eksempel, s. 389 (hærdning af lim)

- Tre metoder til at hærde en lim er prøvet. Der er målt hærdningstid:

<i>Formula A:</i>	13	10	8	11	8
<i>Formula B:</i>	13	11	14	14	
<i>Formula C:</i>	4	1	3	4	2 4

- Antal behandlinger: $k = 3$
- Antal prøver: $N = n_1 + n_2 + n_3 = 5 + 4 + 6 = 15$
- Overordnet middelværdi: $\bar{y} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} = \frac{120}{15} = 8$
- Behandlingsmiddelværdier: $\bar{y}_1 = \frac{50}{5} = 10$; $\bar{y}_2 = \frac{52}{4} = 13$; $\bar{y}_3 = \frac{18}{6} = 3$
- Model: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$:

$$y_{ij} = y_{ij} + (\bar{y} - \bar{y}) + (\bar{y}_i - \bar{y}_i)$$

(lægger \bar{y} og \bar{y}_i til og trækker fra igen)

$$= \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

(omorganiserer)

$$(\text{observation}) \quad y_{ij} = \bar{y}$$

(overordnet middelværdi)

$$+ (\bar{y}_i - \bar{y})$$

(afvigelse pga. behandling)

$$+ (y_{ij} - \bar{y}_i)$$

(afvigelse pga. støj).

Eksempel, s. 389 (hærdning af lim)

- Tre metoder til at hærde en lim er prøvet. Der er målt hærdningstid:

<i>Formula A:</i>	13	10	8	11	8
<i>Formula B:</i>	13	11	14	14	
<i>Formula C:</i>	4	1	3	4	2

- Antal behandlinger: $k = 3$
 - Antal prøver: $N = n_1 + n_2 + n_3 = 5 + 4 + 6 = 15$
 - Overordnet middelværdi: $\bar{y} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} = \frac{120}{15} = 8$
 - Behandlingsmiddelværdier: $\bar{y}_1 = \frac{50}{5} = 10$; $\bar{y}_2 = \frac{52}{4} = 13$; $\bar{y}_3 = \frac{18}{6} = 3$
 - (observation) $y_{ij} = \bar{y}$ (overordnet middelværdi)
 $+ (\bar{y}_i - \bar{y})$ (afvigelse pga. behandling)
 $+ (y_{ij} - \bar{y}_i)$ (afvigelse pga. støj)
 - F.eks. for $i = 2, j = 3$:
 $y_{23} = 14 = \bar{y} + (\bar{y}_2 - \bar{y}) + (y_{23} - \bar{y}_2) = 8 + (13 - 8) + (14 - 13)$
 $y_{23} = 14 = 8 + 5 + 1.$

Eksempel, s. 389 (hærdning af lim)

- Generelt for alle observationerne:

$$\begin{array}{c} \text{observation} \\ y_{ij} \\ \hline \begin{bmatrix} 13 & 10 & 8 & 11 & 8 \\ 13 & 11 & 14 & 14 & \\ 4 & 1 & 3 & 4 & 2 & 4 \end{bmatrix} = \begin{bmatrix} 8 & 8 & 8 & 8 & 8 \\ 8 & 8 & 8 & 8 & \\ 8 & 8 & 8 & 8 & 8 & 8 \end{bmatrix} \end{array}$$
$$\begin{array}{c} \text{grand mean} \\ \bar{y} \\ \hline \end{array}$$
$$\begin{array}{c} \text{treatment effects} \\ \bar{y}_i - \bar{y} \\ \hline + \begin{bmatrix} 2 & 2 & 2 & 2 & 2 \\ 5 & 5 & 5 & 5 & \\ -5 & -5 & -5 & -5 & -5 \end{bmatrix} + \begin{bmatrix} 3 & 0 & -2 & 1 & -2 \\ 0 & -2 & 1 & 1 & \\ 1 & -2 & 0 & 1 & -1 & 1 \end{bmatrix} \end{array}$$
$$\begin{array}{c} \text{error} \\ y_{ij} - \bar{y}_i \\ \hline \end{array}$$

- Sum of squares:
 - Treatment: $SS(Tr) = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 = 5(2)^2 + 4(5)^2 + 6(-5)^2 = 270$
 - Error: $SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = (3)^2 + (0)^2 + \dots + (1)^2 = 32$
 - Total: $SST = SS(Tr) + SSE = 270 + 32 = 302$
- Nu kan det sættes op i en ANOVA tabel.

Eksempel, s. 389 (hærdning af lim)

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Treatments	$k - 1$	$SS(Tr)$	$MS(Tr) = SS(Tr)/(k - 1)$	$\frac{MS(Tr)}{MSE}$
Error	$N - k$	SSE	$MSE = SSE/(N - k)$	
Total	$N - 1$	SST		

Analysis of Variance Table for Cure Times				
Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Treatment	2	270	135	50.6
Error	12	32	2.667	
Total	14	302		

- $F_0 = 50.6$ er F-fordelt med 2 frihedsgrader i tælleren og 12 i nævneren, så den tilhørende p-værdi er
$$1 - \text{pf}(50.6, 2, 12) = 1.42 \cdot 10^{-6} \approx 0$$
- Vi kan forkaste nulhypotesen om at alle tre metoder til hærdning af lim har samme effekt. Mindst én metode adskiller sig fra de andre.

Simplere beregning af Sums of Squares

- **Total:**
$$\begin{aligned} SST &= \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2 - C \end{aligned}$$
- **Treatment:**
$$\begin{aligned} SS(Tr) &= \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 = \\ &= \sum_{i=1}^k \frac{T_i^2}{n_i} - C \end{aligned}$$
- hvor
 - $C = \frac{T^2}{N}$ (kaldes korrektionsleddet)
 - $N = \sum_{i=1}^k n_i$, (det samlede antal observationer)
 - $T_i = \sum_{j=1}^{n_i} y_{ij}$, (summen af observationer for i'te behandling)
 - $T = \sum_{i=1}^k T_i$ (summen af alle observationer)
- **Error:**
$$\begin{aligned} SSE &= \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \\ &= SST - SS(Tr) . \end{aligned}$$

Eksempel, s. 389 (hærdning af lim) i R

- Løsning i R:

```
1 ## Eksempel fra kap.12. Hærdningstiden målt med tre
2 ## hærdningsmetoder A, B og C. Er der forskel?
3
4 D = read.delim("Kap12/C12_lim_haerdning.txt", header=TRUE)
5 metode = D$Formula
6 tid = D$Time
7
8 boxplot(tid~metode)
9
10 anova(lm(tid~metode))
11
12 # Alternativ metode:
13 summary(aov(tid~metode))
```

- Resultat:

Analysis of Variance Table

Response: tid

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
metode	2	270	135.000	50.625	1.415e-06 ***
Residuals	12	32	2.667		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Tilbage til eks. 12.3 (styrke af resiner)

Resin	Strength					Mean
ED	0.99	1.19	0.79	0.95	0.90	0.964
MD	1.11	1.53	1.37	1.24	1.42	1.334
PF	0.83	0.68	0.94	0.86	0.57	0.776

- Den tilhørende ANOVA tabel:

Source of variation	Degrees of freedom	Sum of squares	Mean square	F	P value
Resin	2	0.8060	0.4030	17.2	0.000
Error	12	0.2810	0.0234		
Total	14	1.0870			

- Igen forkaster vi H_0 , fordi p-værdien er ≈ 0 , så mindst een af resinerne har en effekt på styrken. Det er relevant at undersøge for hvert par af resiner, om der er forskel på dem
- Det kan vi undersøge ved at se på konfidensinterval for forskellen på to behandlingsers middelværdier, $\mu_i - \mu_l$, for behandling i og l .

Konfidensinterval for forskel mel. to beh.

- Det svarer til at bruge to stikprøver til at finde konfidensinterval for forskellen på de to populationers middelværdier (*t*-test, kap. 8)
- $100(1 - \alpha)\%$ konfidensintervallet for forskellen på to behandlings-mittelværdier $\mu_i - \mu_l$ er:

$$\bar{y}_i - \bar{y}_l \pm t_{\alpha/2} \cdot \sqrt{s^2 \left(\frac{1}{n_i} + \frac{1}{n_l} \right)}$$

hvor s^2 er *MSE* fra ANOVA tabellen og $t_{\alpha/2}$ er med $df = N - k$ frihedsgrader.

Tilbage til eks. 12.3 (styrke af resiner)

Source of variation	Degrees of freedom	Sum of squares	Mean square	F	P value	Mean
Resin	2	0.8060	0.4030	17.2	0.000	0.964
Error	12	0.2810	0.0234			1.334
Total	14	1.0870				0.776

- 95 % konfidensintervaller for parvise forskelle:

$$s^2 = 0.0234, df = 12, \alpha = 0.05, t_{\alpha/2} = qt(1 - 0.025, 12) = 2.179$$

$$\text{MD} - \text{ED}: 1.334 - 0.964 \pm 2.179 \sqrt{0.0234 \left(\frac{1}{5} + \frac{1}{5}\right)} \text{ eller } [0.159; 0.581]$$

$$\text{MD} - \text{PF}: 1.334 - 0.776 \pm 2.179 \sqrt{0.0234 \left(\frac{1}{5} + \frac{1}{5}\right)} \text{ eller } [0.347; 0.769]$$

$$\text{ED} - \text{PF}: 0.964 - 0.776 \pm 2.179 \sqrt{0.0234 \left(\frac{1}{5} + \frac{1}{5}\right)} \text{ eller } [-0.023; 0.399]$$

- Der er forskel på MD og ED, for 0 er ikke i konfidensintervallet
- Der er forskel på MD og PF, for 0 er ikke i konfidensintervallet
- Der er ikke forskel på ED og PF, for 0 er i konfidensintervallet.

Konfidensinterval for forskel mel. to beh.

- $100(1 - \alpha)\%$ konfidensintervallet for forskellen på to behandlings-middelværdier $\mu_i - \mu_l$ er:

$$\bar{y}_i - \bar{y}_l \pm t_{\alpha/2} \sqrt{s^2 \left(\frac{1}{n_i} + \frac{1}{n_l} \right)}$$

- Denne sammenligningsmetode kaldes **Least Significant Difference (LSD)**. Der findes utallige andre metoder, da 'de kloge statistikere' siger, at sammenligningerne ikke er uafhængige fra par til par
- **Benferroni** metoden: I stedet for $t_{\alpha/2}$ bruges $t_{\alpha/k(k-1)}$:

$$\bar{y}_i - \bar{y}_l \pm t_{\alpha/k(k-1)} \sqrt{s^2 \left(\frac{1}{n_i} + \frac{1}{n_l} \right)}$$

- **Tukey honest significant difference (Tukey HSD)** metoden:

$$\bar{y}_i - \bar{y}_l \pm \frac{q_\alpha}{\sqrt{2}} \sqrt{\frac{2s^2}{n}}$$

hvor beregningen af q_α er for langhåret for os (men R kan).

Bemærk at metoden kræver, at alle behandlinger har samme antal gentagelser, n , altså et balanceret forsøg.

Eks. 12.3 (Benferroni konf.intervaller)

Source of variation	Degrees of freedom	Sum of squares	Mean square	F	P value	Mean
Resin	2	0.8060	0.4030	17.2	0.000	0.964
Error	12	0.2810	0.0234			1.334
Total	14	1.0870				0.776

- Benferroni 95 % konfidensintervaller for parvise forskelle: I stedet for $t_{\alpha/2}$ skal vi bruge

$$t_{\alpha/k(k-1)} = t_{\alpha/3 \cdot 2} = t_{\alpha/6} = qt(1 - 0.05/6, 12) = 2.779$$

MD – ED: $1.334 - 0.964 \pm 2.779 \sqrt{0.0234 \left(\frac{1}{5} + \frac{1}{5}\right)}$ eller [0.101; 0.639]

MD – PF: $1.334 - 0.776 \pm 2.779 \sqrt{0.0234 \left(\frac{1}{5} + \frac{1}{5}\right)}$ eller [0.289; 0.827]

ED – PF: $0.964 - 0.776 \pm 2.779 \sqrt{0.0234 \left(\frac{1}{5} + \frac{1}{5}\right)}$ eller [-0.081; 0.457]

- Bemærk at bogen beregner 94 % intervaller.

Eks. 12.3 (Tukey HSD konf.intervaller)

- Tukeys 95 % konfidensintervaller for parvise forskelle: Vi kan beregne intervallerne, for der er 5 observationer i hver behandling (balanceret)
- Vi bruger R:

```
# Tukey HSD:  
fm1 = aov(styrke~f_resin)  
summary(fm1)  
TukeyHSD(fm1,"f_resin", conf.level=0.95)
```
- Resultat:

	\$f_resin	diff	lwr	upr	p adj
2-1	0.370	0.1118184	0.62818158	0.0063383	
3-1	-0.188	-0.4461816	0.07018158	0.1692986	
3-2	-0.558	-0.8161816	-0.29981842	0.0002438	
- Da numrene 1, 2, 3 svarer til hhv. ED, MD og PF, får vi:
MD – ED: 2 – 1: [0.112; 0.628]
MD – PF: 2 – 3: [0.300; 0.816]
ED – PF: 1 – 3: [-0.070; 0.446].

Eks. 12.3 (oversigt over metoder)

- $100(1 - \alpha)\%$ konfidensintervallet for forskellen på to behandlings-middelværdier $\mu_i - \mu_l$ er:

$$\bar{y}_i - \bar{y}_l \pm B$$

- Den halve bredde af konfidensintervallet, B, afhænger af metoden:
 - LSD: $B = t_{\alpha/2} \sqrt{s^2(1/n_i + 1/n_l)} = 0.211$
 - Benferroni: $B = t_{\alpha/k(k-1)} \sqrt{s^2(1/n_i + 1/n_l)} = 0.269$
 - Tukeys HSD: $B = q_\alpha / \sqrt{2} \sqrt{s^2 \cdot 2/n} = 0.258$
- Konfidensintervaller:

	LSD	Benferroni	Tukeys HSD
MD – ED:	[0.159; 0.581]	[0.101; 0.639]	[0.112; 0.628]
MD – PF:	[0.347; 0.769]	[0.289; 0.827]	[0.300; 0.816]
ED – PF:	[-0.023; 0.399]	[-0.081; 0.457]	[-0.070; 0.446]

- Vi kommer til at bruge **Tukeys HSD** pga. funktionen i R.

Eks. 12.3 (test af antagelser)

- Vores statistiske model: y_{ij} kommer fra en stokastisk variabel:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1, \dots, k; \quad j = 1, \dots, n_i$$

Vi har antaget, at støjen ε_{ij} er normalfordelt $N(0, \sigma)$ med konstant σ , så variansen er uafhængig af behandlingerne ([varianshomogenitet](#))

- Vi tester antagelserne med residualanalyse:
 - Normalfordelingsplot af residualerne
 - Scatterplot af residualer mod behandlinger
 - Test for varianshomogenitet (Bartlett's test).

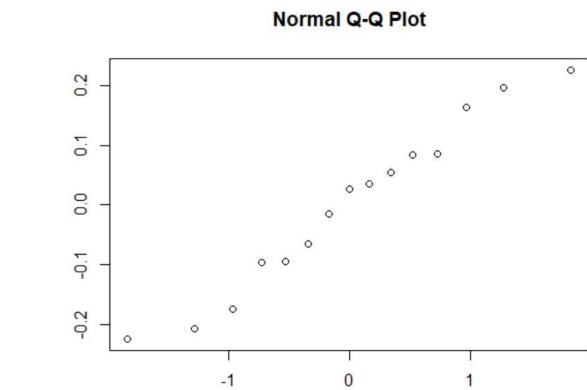
Eks. 12.3 (test af antagelser i R)

- Beregning af residualer:

```
# Vi får beregnet y_hat og residualerne med aov funktionen:  
e = fm1$residuals  
y_hat = fm1$fitted.values
```

- Test af om residualerne er normalfordelte:

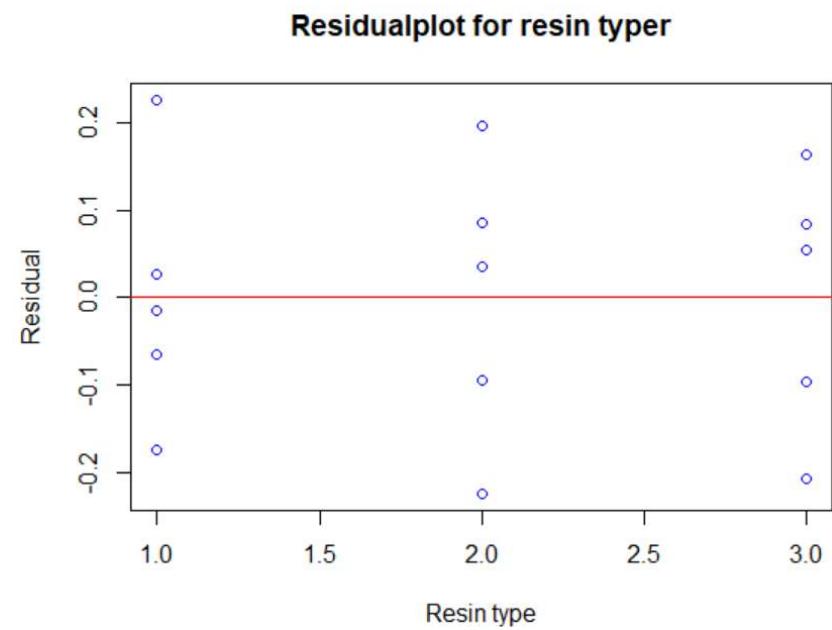
```
# Test for normalfordelte residualer:  
qqnorm(e)
```



- Grafisk test af om residualerne er uafhængige af behandlinger:

```
# Residualplot  
plot(resin, e, type="p",  
      main="Residualplot for resin typer",  
      xlab="Resin type",  
      ylab="Residual",  
      col="blue")  
abline(h=0, col="red")
```

- Bartlett's test for varianshomogenitet (ikke i bogen).



Eks. 12.3 (test af antagelser i R)

- **Bartlett's test** for varianshomogenitet er en hypotestest, hvor nulhypotesen er:
$$H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 \text{ for alle } k \text{ behandlinger}$$
- Teststørrelsen for Bartlett's test er χ^2 fordelt med $k - 1$ frihedsgrader

```
# Test af varianshomogenitet med Bartlett's test  
bartlett.test(styrke,f_resin)
```

- Resultat:

```
Bartlett test of homogeneity of variances
```

```
data: styrke and f_resin  
Bartlett's K-squared = 0.046583, df = 2, p-value = 0.977
```

- P-værdi: $p = 1 - \text{pchisq}(0.046583, 3 - 1) = 0.976978$
- Konklusion: Da p-værdien på 0.98 er langt fra 0 kan vi ikke afvise nulhypotesen, at variansen for hver behandling er ens.
Antagelsen om varianshomogenitet holder.

Sandsynlighedsteori og statistik

Kapitel 13. Eksperimenter med to og flere faktorer (afsnit 13.1-13.2)

Allan Leck Jensen

alj@ece.au.dk

Eksperiment med 2 faktorer

- Faktor A har a niveauer og faktor B har b niveauer.
 - Det giver ab forskellige kombinationer af A's og B's niveauer
 - For hver kombination er der r gentagelser (NB. Figuren herunder benævner antal gentagelser med n). Der er ialt abr observationer
- Observationen y_{ijk} er den k 'te gentagelse af det i 'te niveau af A og det j 'te niveau af B
- Bemærk at bogen har ændret notation: I kap. 12 havde faktor A k niveauer (nu a). Nu er k indeks for gentagelsesnummeret. Dårlig stil!

		Factor B				
		1	2	...	b	Averages
Factor A	1	$y_{111}, y_{112}, \dots, y_{11n}$	$y_{121}, y_{122}, \dots, y_{12n}$		$y_{1b1}, y_{1b2}, \dots, y_{1bn}$	$\bar{y}_{1..}$
	2	$y_{211}, y_{212}, \dots, y_{21n}$	$y_{221}, y_{222}, \dots, y_{22n}$		$y_{2b1}, y_{2b2}, \dots, y_{2bn}$	$\bar{y}_{2..}$
	:					
	a	$y_{a11}, y_{a12}, \dots, y_{a1n}$	$y_{a21}, y_{a22}, \dots, y_{a2n}$		$y_{ab1}, y_{ab2}, \dots, y_{abn}$	$\bar{y}_{a..}$
Totals		$y_{..1}$	$y_{..2}$		$y_{..b}$	
Averages		$\bar{y}_{..1}$	$\bar{y}_{..2}$		$\bar{y}_{..b}$	$\bar{y}...$

Eksperiment med 2 faktorer

- Notation:

- $\bar{y}_{ij\cdot} = \frac{1}{r} \sum_{k=1}^r y_{ijk}$ (snit for i,j'te behandling)
- $\bar{y}_{i..} = \frac{1}{br} \sum_{j=1}^b \sum_{k=1}^r y_{ijk}$ (snit for i'te behandling af A)
- $\bar{y}_{\cdot j\cdot} = \frac{1}{ar} \sum_{i=1}^a \sum_{k=1}^r y_{ijk}$ (snit for j'te behandling af B)
- $\bar{y}_{...} = \frac{1}{abr} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r y_{ijk}$ (overordnet snit)

		Factor B				
		1	2	...	b	Averages
Factor A	1	$y_{111}, y_{112}, \dots, y_{11n}$	$y_{121}, y_{122}, \dots, y_{12n}$		$y_{1b1}, y_{1b2}, \dots, y_{1bn}$	$\bar{y}_{1..}$
	2	$y_{211}, y_{212}, \dots, y_{21n}$	$y_{221}, y_{222}, \dots, y_{22n}$		$y_{2b1}, y_{2b2}, \dots, y_{2bn}$	$\bar{y}_{2..}$
	:					
a	$y_{a11}, y_{a12}, \dots, y_{a1n}$	$y_{a21}, y_{a22}, \dots, y_{a2n}$		$y_{ab1}, y_{ab2}, \dots, y_{abn}$		$\bar{y}_{a..}$
Totals	$y_{\cdot 1\cdot}$	$y_{\cdot 2\cdot}$		$y_{\cdot b\cdot}$		
Averages	$\bar{y}_{\cdot 1\cdot}$	$\bar{y}_{\cdot 2\cdot}$		$\bar{y}_{\cdot b\cdot}$		$\bar{y}_{...}$

Lineær statistisk model med 2 faktorer

- Det ville være oplagt at udvide den statistiske model for 1 faktor til 2 faktorer ved at lægge et led til for bidrag fra faktor B:

- Observationer y_{ijk} kommer fra en stokastisk variabel:

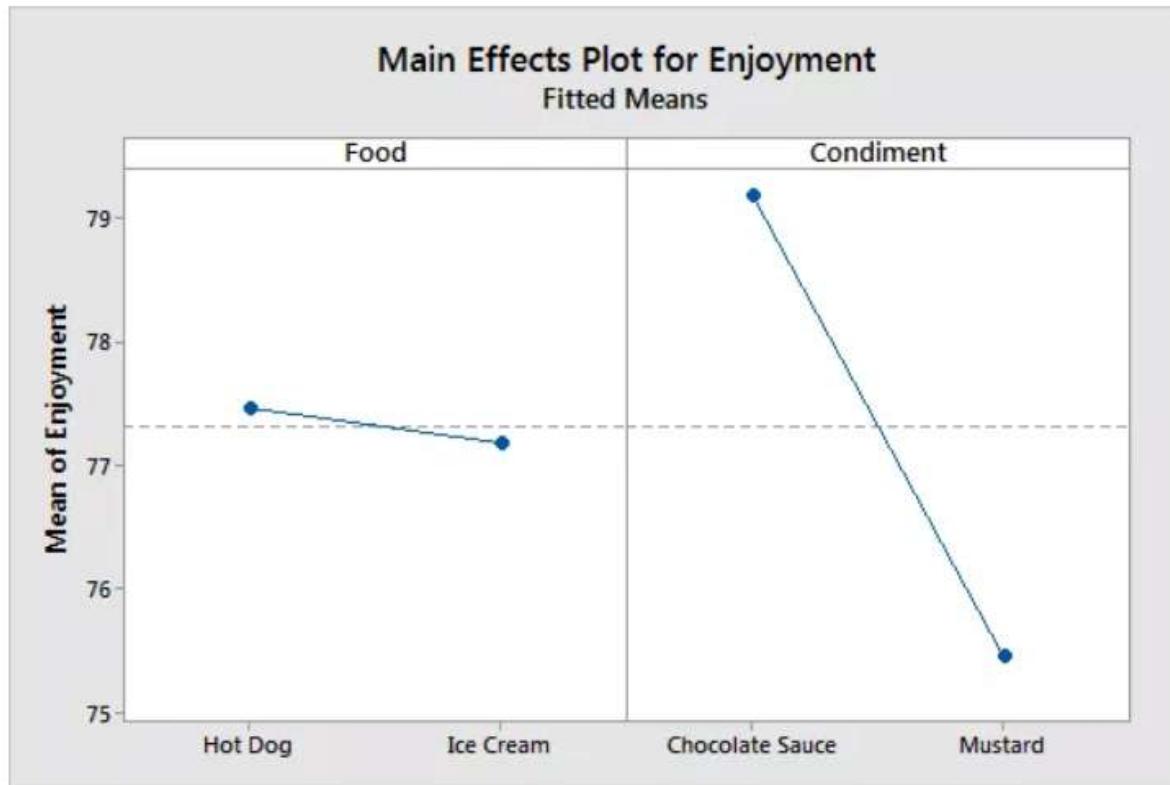
$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \quad i = 1, \dots, a; j = 1, \dots, b; k = 1, \dots, r$$

- μ er det overordnede gennemsnit på tværs af behandlinger
- α_i er **effekten** af den i 'te behandling af A. $\sum \alpha_i = 0$
- β_j er **effekten** af den j 'te behandling af B. $\sum \beta_j = 0$
- ε_{ij} er tilfældig afvigelse (random error). Antages normalfordelt $N(0, \sigma)$
- Men denne model er for simpel, fordi den ikke tager højde for **interaktion** mellem de to faktorer A og B. Derfor:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad \begin{cases} i = 1, \dots, a \\ j = 1, \dots, b \\ k = 1, \dots, r \end{cases}$$

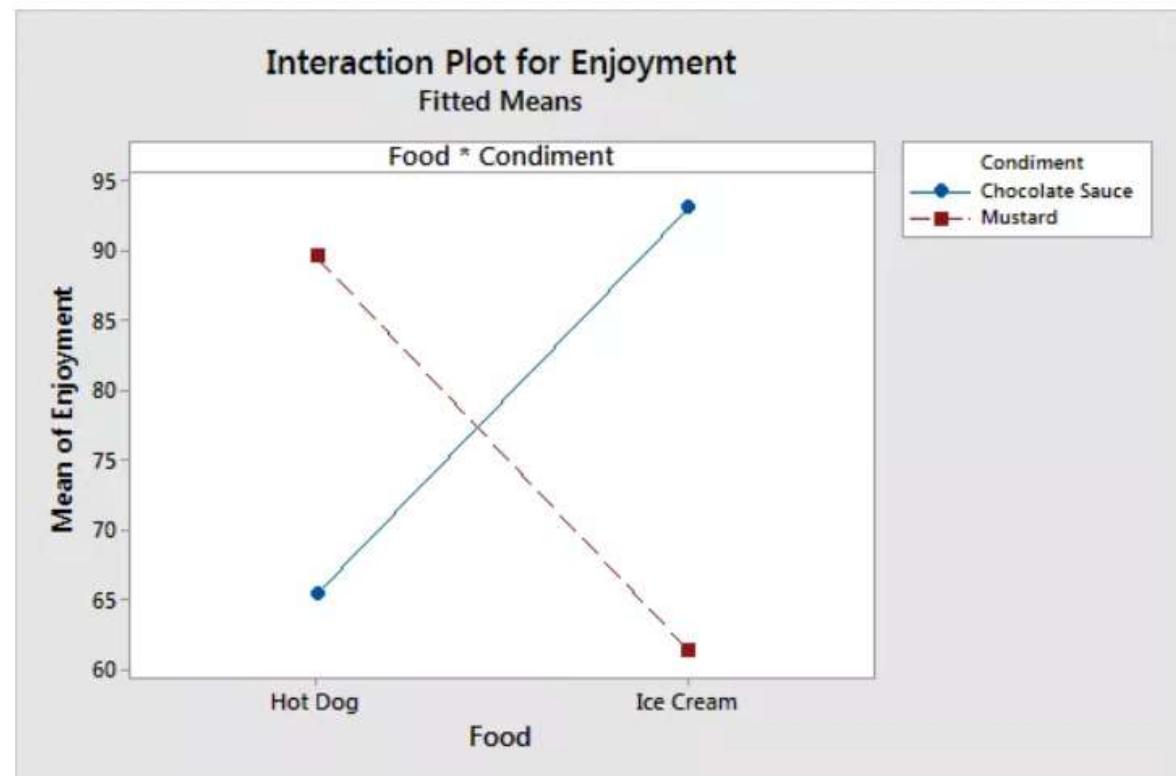
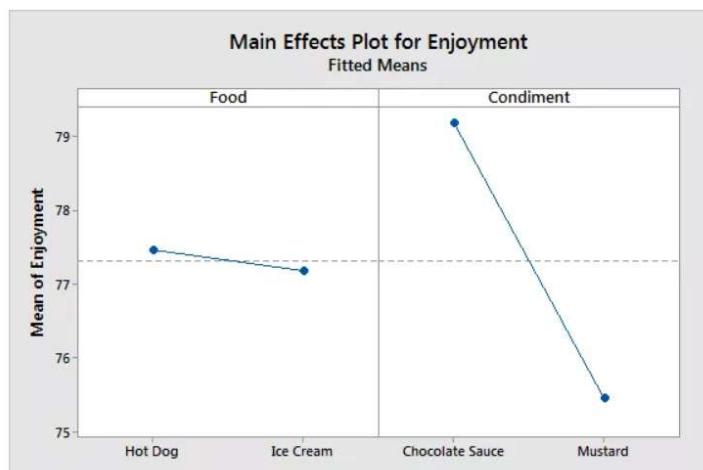
Interaktion eksempel

- Y: Tilfredshed med måltid
- Faktor A: Måltid; 2 Niveauer: Hotdog + Is
- Faktor B: Sovs; 2 niveauer: Sennep + chokoladesovs
- Model uden interaktion: $y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$
Største smagsoplevelse opnås med hotdog med chokoladesovs



Interaktion eksempel

- “Vil du have sennep eller chokoladesovs på dit måltid?”
 - Hvis svaret er: “Chokoladesovs”, så er der *ikke interaktion*
 - Hvis svaret er: “Det afhænger af måltidet”, så er der *interaktion*
- Model med interaktion: $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$
Største smagsoplevelse opnåes med is med chokoladesovs.



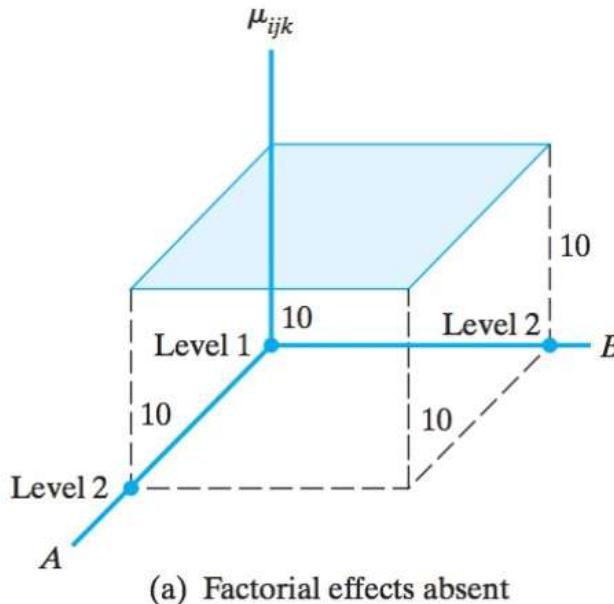
Lineær statistisk model med 2 faktorer

- a) $Y_{ijk} = \mu$
(ingen effekt af faktorerne)

- b) $Y_{ijk} = \mu + \alpha_i$
(kun effekt af faktor A)

- c) $Y_{ijk} = \mu + \alpha_i + \beta_j$
(kun effekt af faktor A og B)

- d) $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$
(effekt af A, B og interaktion.
Responsoverfladen er ikke
længere plan).



3 hypotesetests

1. $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_a = 0$ (ingen direkte effekt af faktor A)
 $H_1: \text{Mindst en } \alpha_i \neq 0$

2. $H_0: \beta_1 = \beta_2 = \dots = \beta_b = 0$ (ingen direkte effekt af faktor B)
 $H_1: \text{Mindst en } \beta_j \neq 0$

3. $H_0: (\alpha\beta)_{11} = (\alpha\beta)_{12} = \dots = (\alpha\beta)_{ab} = 0$ (ingen interaktioner)
 $H_1: \text{Mindst en } (\alpha\beta)_{ij} \neq 0$

Sum of Squares til ANOVA

- Total Sum of Squares kan opdeles i komponenter:

$$\begin{aligned} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r (y_{ijk} - \bar{y}_{...})^2 &= br \sum_{i=1}^a (\bar{y}_{i..} - \bar{y}_{...})^2 + ar \sum_{j=1}^b (\bar{y}_{.j.} - \bar{y}_{...})^2 \\ &\quad + r \sum_{i=1}^a \sum_{j=1}^b (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j..} + \bar{y}_{...})^2 \\ &\quad + \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r (y_{ijk} - \bar{y}_{ij.})^2 \end{aligned}$$

- ... eller skrevet med symboler:

$$SST = SS(A) + SS(B) + SS(AB) + SSE$$

- Den totale variation (*SST*) består af :

- den del, der kan forklares med faktor A (*SS(A)*),
- den del der kan forklares med faktor B (*SS(B)*),
- den del der kan forklares med interaktionen mellem A og B (*SS(AB)*),
- og resten, der ikke kan forklares med faktorerne (*SSE*).

ANOVA for 2-faktor eksperimenter

	Frihedsgrader (df)	Sum of Squares (SS)	Mean Squares (MS)	F
Faktor A	$a - 1$	$SS(A)$	$MS(A) = \frac{SS(A)}{a - 1}$	$\frac{MS(A)}{MSE}$
Faktor B	$b - 1$	$SS(B)$	$MS(B) = \frac{SS(B)}{b - 1}$	$\frac{MS(B)}{MSE}$
Interaktion	$(a - 1)(b - 1)$	$SS(AB)$	$MS(AB) = \frac{SS(AB)}{(a - 1)(b - 1)}$	$\frac{MS(AB)}{MSE}$
Error	$ab(r - 1)$	SSE	$MSE = \frac{SSE}{ab(r - 1)}$	
Total	$abr - 1$	SST		• Bemærk: Hvis der kun er 1 gentagelse ($r = 1$), så er der 0 frihedsgrader for Error, så MSE kan ikke beregnes og dermed kan hypoteserne ikke testes.

Eksempel 13.1, s. 428 (vejbelægning)

- Gamle byggematerialer genbruges til vejbelægning. Kvaliteten af materialet måles som 'modulus of resilience' (~elasticiteten). I et eksperiment blev resiliensen målt på to typer materialer fordelt på tre lokaliteteter og med tre gentagelser. Man ønsker 1 % signifikansniveau
- Faktor A, lokalitet. 3 niveauer:
 - MN (Minnesota)
 - CO (Colorado)
 - TX (Texas)
- Faktor B, type af materiale. 2 niveauer:
 - RCA (Recycled Concrete Aggregate) ~ Beton
 - RPA (Recycled Pavement Aggregate) ~ Fliser
- Vi har $\alpha = 0.01$, $a = 3$, $b = 2$, $k = 3$. Dermed $N = abr = 18$. Resultat:

		Faktor A: Type af materiale								Snit
		RCA				RPA				
Faktor B: Lokalitet	MN	707	632	604	647.7	652	669	674	665.0	656.3
	CO	522	554	484	520.0	630	648	610	629.3	574.7
	TX	450	545	474	489.7	845	810	682	779.0	634.3
Snit		552.4				691.1				621.8

Eksempel 13.1, s. 428 (vejbelægning)

- Data kan præsenteres på forskellige, mere eller mindre bearbejdede (og overskuelige) måder:

```

C13Ex1.TXT - Notepad
File Edit Format View Help
A      B      resilmord
MN    RCA    707
MN    RPA    652
CO    RCA    522
CO    RPA    630
TX    RCA    450
TX    RPA    845
MN    RCA    632
MN    RPA    669
CO    RCA    554
CO    RPA    648
TX    RCA    545
TX    RPA    810
MN    RCA    604
MN    RPA    674
CO    RCA    484
CO    RPA    610
TX    RCA    474
TX    RPA    682

```

Factor A Location	Factor B Type of Mat.			Rep. 1	Rep. 2	Rep. 3
	RCA	RPA	CO			
MN	RCA	707		707	632	604
MN	RPA	652		652	669	674
CO	RCA	522		522	554	484
CO	RPA	630		630	648	610
TX	RCA	450		450	545	474
TX	RPA	845		845	810	682

		Faktor A: Type af materiale								Snit
		RCA				RPA				
Faktor B: Lokalitet	MN	707	632	604	647.7	652	669	674	665.0	656.3
	CO	522	554	484	520.0	630	648	610	629.3	574.7
	TX	450	545	474	489.7	845	810	682	779.0	634.3
	Snit	552.4				691.1				621.8

Eksempel 13.1, s. 428 (vejbelægning)

- Konklusion af testene:

- Effekt af faktor A (lokalitet)

På 1 % signifikansniveau kan vi *ikke* forkaste nulhypotesen om, at der ikke er forskel på niveauerne i faktor A.

P-værdien for testen er 0.035, som er over 0.01. Der er ikke signifikant forskel på lokaliteterne

- Effekt af faktor B (materiale)

Som følge af F-værdien på 36.1 og den tilhørende p-værdi på næsten 0 kan vi forkaste nulhypotesen. Der er forskel på de to materialer

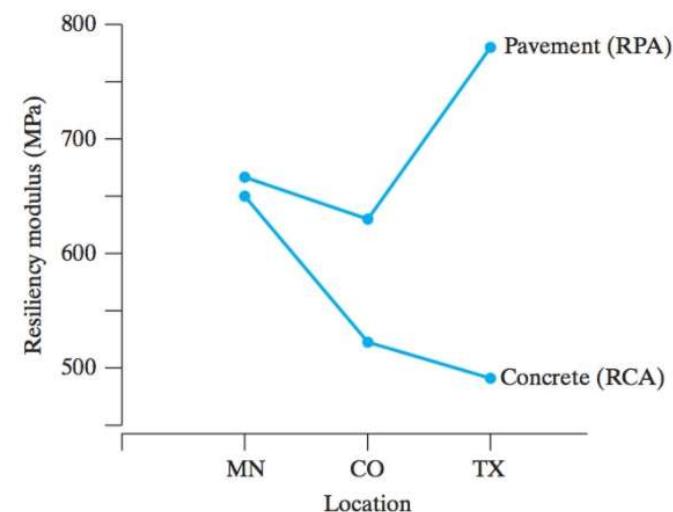
- Effekt af interaktion mellem lokalitet og materiale

Der er signifikant interaktion mellem de to faktorer.

Derfor kan vi ikke ignorere faktor A (lokalitet)

- Grafens seks punkter er gennemsnitlig resiliens for de seks kombinationer af lokalitet og materiale. Forskel på de to materialers resiliens er afhængig af lokaliteten. Hvis der ikke var interaktion, ville vi se to parallelle kurver (med samme forskel for de tre lokaliteter).

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	P
Main Effects:					
A	2	21427	10714	4.48	0.035
B	1	86528	86528	36.1	0.000
Interaction	2	57424	28712	12.0	0.001
Error	12	28724	2394		
Total	17	194103			



Hvordan i R? (Eksempel 13.1)

- Vi får R til at beregne ANOVA tabellen med kommandoen
`fm1 = aov(res ~ lok + mat + lok:mat)`
`summary(fm1)`

hvor 'lok' er lokalitet (faktor A), 'mat' er materiale (faktor B) og 'res' er den målte resiliens. Vi får taget højde for interaktion i modelen med leddet 'lok:mat'

- Resultat:
- Vi kan få beregnet konfidensinterval for forskellen på hvert par af lokalitet:

`TukeyHSD(fm1, "lok", conf.level=0.99)`

- For hvert af de tre par indeholder konfidensintervallet 0 og p-værdien er over 0.01
- Vi kan få beregnet konf.intervall for alle kombinationer af par og lokalitet:

`TukeyHSD(fm1, conf.level=0.99)`

- F.eks. er forskellen på RCA fra MN og CO mellem -44.7 og 300.0 (række 1). P-værdien for, om forskellen er 0, er 0.066.

```
> summary(aov(res ~ lok + mat + lok:mat))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
lok	2	21427	10714	4.476	0.03530 *
mat	1	86528	86528	36.149	6.1e-05 ***
lok:mat	2	57424	28712	11.995	0.00137 **
Residuals	12	28724	2394		

```
$`lok`
```

	diff	lwr	upr	p adj
MN-CO	81.66667	-19.11877	182.45211	0.0336685
TX-CO	59.66667	-41.11877	160.45211	0.1289165
TX-MN	-22.00000	-122.78544	78.78544	0.7225246

```
$`lok:mat`
```

	diff	lwr	upr	p adj
MN:RCA-CO:RCA	127.66667	-44.671457	300.00479	0.0656147
TX:RCA-CO:RCA	30.33333	202.671457	142.00479	0.3634733
CO:RPA-CO:RCA	109.33333	-63.004791	281.67146	0.1379956
MN:RPA-CO:RCA	145.00000	-27.338124	317.33812	0.0316951
TX:RPA-CO:RCA	259.00000	86.661876	431.33812	0.0003350
TX:RCA-MN:RCA	-158.00000	-330.338124	14.33812	0.0182870
CO:RPA-MN:RCA	-18.33333	-190.671457	154.00479	0.9967855
MN:RPA-MN:RCA	17.33333	-155.004791	189.67146	0.9975302
TX:RPA-MN:RCA	131.33333	-41.004791	303.67146	0.0563238
CO:RPA-TX:RCA	139.66667	-32.671457	312.00479	0.0396989
MN:RPA-TX:RCA	175.33333	2.995209	347.67146	0.0088230
TX:RPA-TX:RCA	289.33333	116.995209	461.67146	0.0001162
MN:RPA-CO:RPA	35.66667	-136.671457	208.00479	0.9410862
TX:RPA-CO:RPA	149.66667	-22.671457	322.00479	0.0260171
TX:RPA-MN:RPA	114.00000	-58.338124	286.33812	0.1146698

Eksperimenter med flere faktorer

- Her med **tre** faktorer, A, B og C:

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k \\ + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} \\ + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl}$$

for $\begin{cases} i = 1, \dots, a \\ j = 1, \dots, b \\ k = 1, \dots, c \\ l = 1, \dots, r \end{cases}$

- Der er
 - tre **primære effekter**,
 - tre **to-faktor interaktioner** og
 - en **tre-faktor interaktion**.

ANOVA for 3-faktor eksperimenter

	Frihedsgrader (<i>df</i>)	Sum of Squares (<i>SS</i>)	Mean Squares (<i>MS</i>)	<i>F</i>
Faktor A	$a - 1$	$SS(A)$	$MS(A) = \frac{SS(A)}{a - 1}$	$\frac{MS(A)}{MSE}$
Faktor B	$b - 1$	$SS(B)$	$MS(B) = \frac{SS(B)}{b - 1}$	$\frac{MS(B)}{MSE}$
Faktor C	$c - 1$	$SS(C)$	$MS(B) = \frac{SS(C)}{c - 1}$	$\frac{MS(C)}{MSE}$
A:B	$(a - 1)(b - 1)$	$SS(AB)$	$MS(AB) = \frac{SS(AB)}{(a - 1)(b - 1)}$	$\frac{MS(AB)}{MSE}$
A:C	$(a - 1)(c - 1)$	$SS(AC)$	$MS(AC) = \frac{SS(AC)}{(a - 1)(c - 1)}$	$\frac{MS(AC)}{MSE}$
B:C	$(b - 1)(c - 1)$	$SS(BC)$	$MS(BC) = \frac{SS(BC)}{(b - 1)(c - 1)}$	$\frac{MS(BC)}{MSE}$
A:B:C 7 hypotesetests!	$(a - 1)(b - 1)(c - 1)$	$SS(ABC)$	$MS(ABC) = \frac{SS(ABC)}{(a - 1)(b - 1)(c - 1)}$	$\frac{MS(ABC)}{MSE}$

Eksempel 13.3, s. 436 (Detonator)

- Som jeg forstår opgaven, handler den om undersøgelse af en detonator (ignitor). Sprængstoffer er som regel svære at antænde. Derfor bruger man en tændsats til at antænde et booster sprængstof, som antænder det primære sprængstof.
- Forsinkelsestiden imellem antænding af tændsats til ekspllosion måles. Den skal helst være under 30 ms
- Faktor A, 3 typer tændsats
- Faktor B, 2 typer booster
- Faktor C, 4 typer primær sprængstof
- To gentagelser. Derfor er
$$N = abcr = 3 \cdot 2 \cdot 4 \cdot 2 = 48.$$

A	B	C	Delay Time (milliseconds)	
			Rep. 1	Rep. 2
Initiator 1	Powder	Mc 1	10.70	9.82
Initiator 1	Pellet	Mc 1	10.02	13.66
Initiator 1	Powder	Mc 2	14.46	20.86
Initiator 1	Pellet	Mc 2	11.44	13.76
Initiator 1	Powder	Mc 3	15.04	16.02
Initiator 1	Pellet	Mc 3	27.26	21.42
Initiator 1	Powder	Mc 4	20.82	14.46
Initiator 1	Pellet	Mc 4	24.56	36.48
Initiator 2	Powder	Mc 1	18.42	18.62
Initiator 2	Pellet	Mc 1	22.80	25.14
Initiator 2	Powder	Mc 2	33.40	20.62
Initiator 2	Pellet	Mc 2	31.86	19.78
Initiator 2	Powder	Mc 3	22.94	31.12
Initiator 2	Pellet	Mc 3	32.92	21.38
Initiator 2	Powder	Mc 4	27.92	59.86
Initiator 2	Pellet	Mc 4	31.94	28.32
Initiator 3	Powder	Mc 1	7.14	7.98
Initiator 3	Pellet	Mc 1	24.32	10.26
Initiator 3	Powder	Mc 2	8.30	7.86
Initiator 3	Pellet	Mc 2	7.00	8.40
Initiator 3	Powder	Mc 3	8.40	10.94
Initiator 3	Pellet	Mc 3	17.82	15.28
Initiator 3	Powder	Mc 4	9.56	19.04
Initiator 3	Pellet	Mc 4	19.98	18.46

Eksempel 13.3, s. 436 (Detonator)

- ANOVA tabellen viser, at kun faktor A og C er signifikante:

```
> fm1 = aov(y ~ A + B + C + A:B + A:C + B:C + A:B:C);  
> summary(fm1)  
             Df Sum Sq Mean Sq F value    Pr(>F)  
A            2 1973.2  986.6  22.215 3.46e-06 ***  
B            1   74.9   74.9   1.687  0.20640  
C            3  864.4  288.1   6.488  0.00227 **  
A:B          2  141.8   70.9   1.597  0.22333  
A:C          6  201.0   33.5   0.754  0.61237  
B:C          3  122.1   40.7   0.917  0.44764  
A:B:C        6  319.6   53.3   1.199  0.34042  
Residuals    24 1065.9   44.4
```

- Da B ikke har effekt, hverken direkte eller gennem interaktioner, kan vi fjerne den:

```
> fm2 = aov(y ~ A + C + A:C);  
> summary(fm2)  
             Df Sum Sq Mean Sq F value    Pr(>F)  
A            2 1973.2  986.6  20.598 1.09e-06 ***  
C            3  864.4  288.1   6.016  0.00198 **  
A:C          6  201.0   33.5   0.699  0.65187  
Residuals    36 1724.3   47.9
```

- Interaktionen mellem A og C er ikke signifikant, så den kan også fjernes:

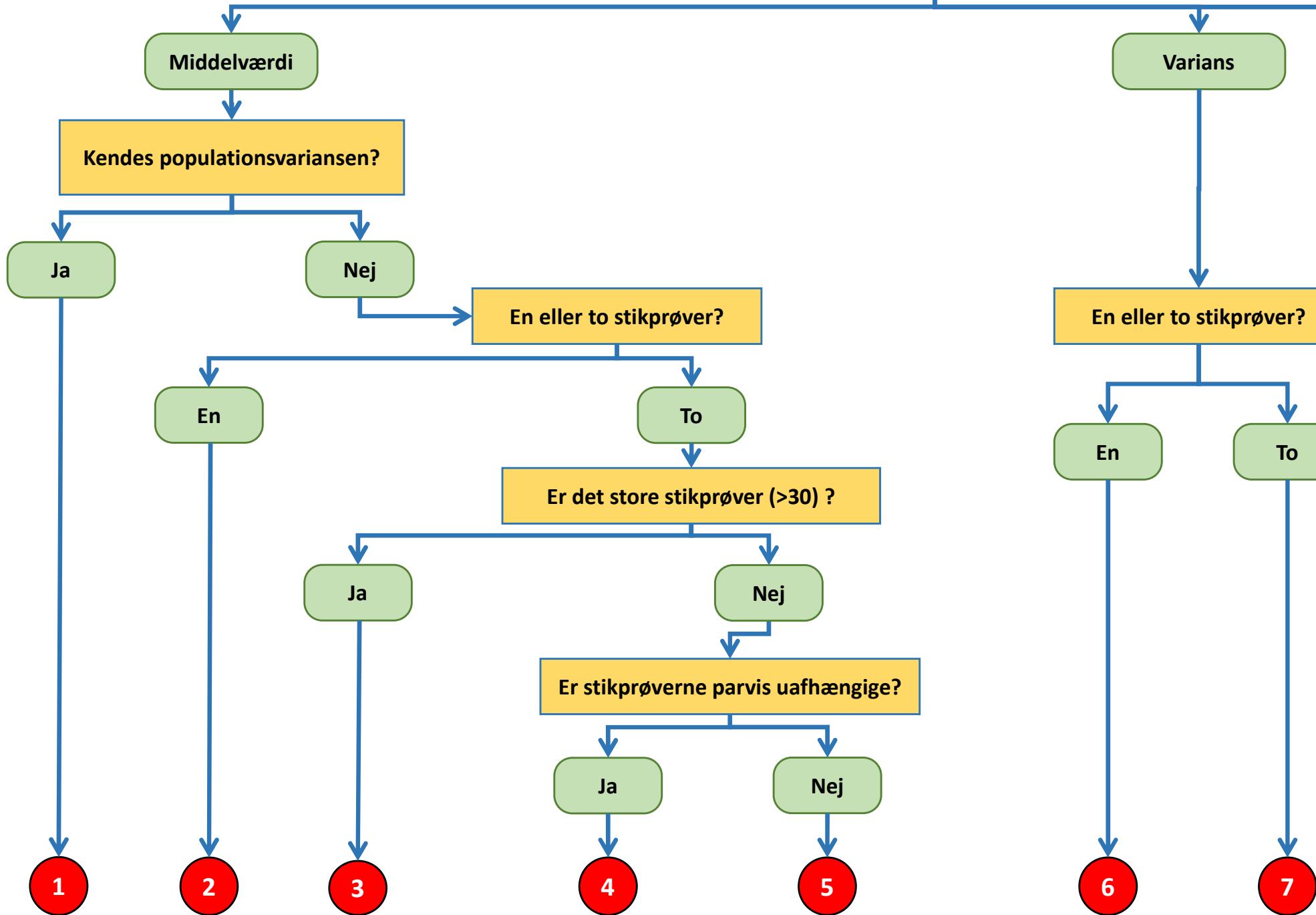
```
> fm3 = aov(y ~ A + C);  
> summary(fm3)  
             Df Sum Sq Mean Sq F value    Pr(>F)  
A            2 1973.2  986.6  21.523 3.68e-07 ***  
C            3  864.4  288.1   6.286  0.00127 **  
Residuals    42 1925.3   45.8
```

Eksempel 13.3, s. 436 (Detonator)

- Test af antagelser:
 - Den tilfældige støj ε_{ijk} er normalfordelt $N(0, \sigma)$
 - Variansen σ^2 er uafhængig af faktorerne værdi
- Hvordan?
 - Vi kan estimere ε_{ijkl} med residualerne e_{ijkl} :
$$e_{ijkl} = y_{ijkl} - \hat{y}_{ijk} = y_{ijkl} - \bar{y}_{ijk}.$$
 - Vi kan teste om residualerne er normalfordelte med normalfordelingsplot
 - Vi kan teste om residualernes variation er ensartet, uafhængigt af faktorerne, med residualplots.

Oversigt over hypotesetests

Hvad handler hypotesesten om?



Nr	Lektion	Afsnit	Teststørrelse	Fordeling	Frihedsgrader	Eksempel
1	L7	7.7	$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	z	-	7.? s. 249
2	L7	7.7	$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	t	$n - 1$	7.20 s. 254
3	L8	8.2	$z_0 = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$	z	-	8.5 s. 271
4	L8	8.3	$t_0 = \frac{\bar{x} - \bar{y} - \delta_0}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	t	$n_1 + n_2 - 2$	8.7 s. 274
5	L8	8.4	$t_0 = \frac{\bar{d} - \delta_0}{s_d/\sqrt{n}}$	t	$n - 1$	8.12 s. 281
6	L9	9.2	$\chi^2_0 = \frac{(n - 1)s^2}{\sigma_0^2}$	χ^2	$n - 1$	9.3 s. 294
7	L9	9.3	$F_0 = \frac{s_1^2}{s_2^2}$	F	$n_1 - 1, n_2 - 1$	9.6 s. 298
8	L9	10.4	$\chi^2_0 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$	χ^2	$(r - 1)(c - 1)$	10.13 s. 320
9	L10	10.5	$\chi^2_0 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$	χ^2	$k - p - 1$	Eks. i ppt K10