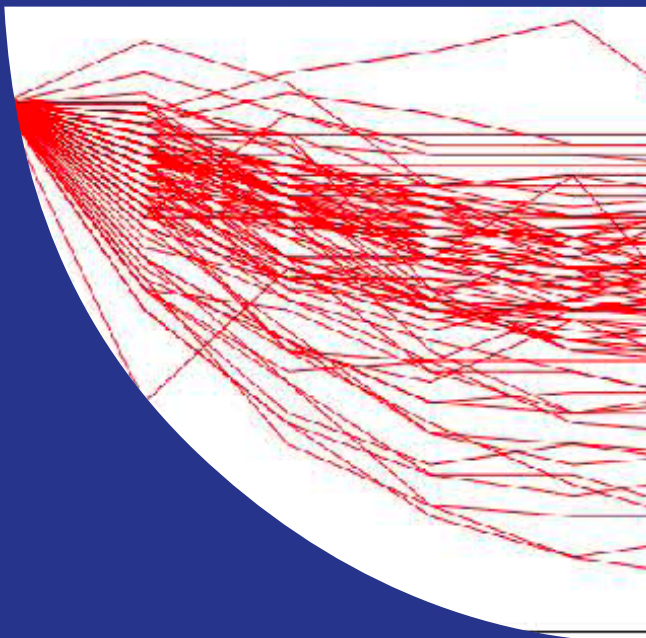
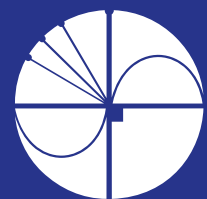


Matematikken bag ny medicin hos NOVO



UNDERVISNINGS OG PROJEKTMATERIALER TIL FILMEN
BJØRN GRØN (RED.), OLAV LYNDRUP, PER ROSENQUIST
OG MADS LETH



matematik
i arbejde

Matematikken bag ny medicin hos Novo Nordisk

Undervisnings og Projektmaterialer
til filmen

Bjørn Grøn (red.), Olav Lyndrup,
Per Rosenquist og Mads Leth

© 2020 Konceptet: *Træk virksomhederne ind i undervisningen* tilhører forfatterne til lærebogssystemet *Hvad er matematik?* Bjørn Grøn, Bodil Bruun og Olav Lyndrup.

© 2022 undervisningsmaterialer til filmen *Matematikken bag ny medicin hos Novo Nordisk* er produceret af *HEM – Produktion og Formidling*.

Grafisk koncept: Ulla Korgaard, Designeriet.

Filmene og de tilhørende projektmaterialer hostes af Lindhardt og Ringhof / Praxis på websitet:

[Træk virksomhederne ind i undervisningen \(praxis.dk\)](https://praxis.dk)

Filmene og de tilhørende projektmaterialer kan frit downloades og anvendes til selvstudium og i undervisningen. Hverken film eller projektmaterialer må gøres til genstand for kommerciel udnyttelse. Henvendelser om materialerne kan rettes til Bjørn Grøn: bjgro1@gmail.com.

Undervisningsmaterialer til filmen *Matematikken bag ny medicin hos Novo Nordisk*

Forfattere: Bjørn Grøn (red.), Olav Lyndrup, Per Rosenquist, Mads Leth

Vi har forsøgt at finde eventuelle rettighedsindehavere, som kan tilkomme honorar i henhold til loven om ophavsret. Skulle der mod forventning være rettighedsindehavere, som måtte have krav på vederlag, vil dette blive håndteret, som om der var indgået en aftale.

Projektet *Træk virksomhederne ind i undervisningen* er forankret på Rysensteen Gymnasium.

Film og tilhørende materialer er produceret med støtte fra Novo Nordisk Fonden

Indholdsfortegnelse




0. Indledning	7
1. Håndtering af store datasæt	8
1.1 Datasættet fra filmen	9
1.2 Projekt: Generer selv et datasæt om krop og sundhed	11
2. Grafiske metoder i beskrivende statistik: Histogram og sumkurve	13
2.1. Grafisk præsentation af grupperede datasæt – sumkurve	14
2.2. Grafisk præsentation af grupperede datasæt – histogram	17
3. Stikprøver og population	19
3.1 Forsøgsplanlægning: Repræsentativitet og selektionsbias	20
3.2 Skjulte variable, korrelation og kausalitet	23
3.2.1 Opdigtede avisnotitser	24
3.3 Case: Nedskydning af amerikanske kampfly under 2. VK	26
3.4 Case: Soldyrkere lever længere	28
4. Normalfordelingen	33
4.1 Hvor kommer normalfordelingen fra?	34
4.2 Normalfordelingen og binomialfordelingen	36
4.3 Normalfordelingens anvendelse breder sig til andre fag	37
4.4 Den generelle normalfordelings funktionsudtryk	38
4.5 Overlejlrede normalfordelinger	40
4.6 Undersøgelse af data ved hjælp af et QQ-plot	42
5. Ny medicin: Fra ide til kliniske lægemiddelforsøg	43
5.1 Difteri, stivkrampe og hesten Jim	44
5.2 Thalidomid	45
5.3 Helsinkideklarationen, apotekerloven og lægemiddelloven	46
5.4 Den prækliniske fase	46
5.5 Den kliniske fase	47
5.6 Hvis du vil vide mere om kliniske undersøgelser	49
5.7 Case: Skørbug – første skridt mod kontrollerede forsøg	50
6. Ethiske problemer i clinical trials	52
6.1. Case: Remdesivir og hydroxychloroquine	53
6.2. Case: Theis Lange og Frank Shan, <i>Perils of stopping trials early</i>	54
6.3. Case: Alzheimer – medicin godkendes, før vi kender sygdommens natur	57
7. Testteori 1. Standardfordelingerne	59
7.1 Projekt om det statistiske begrebsapparat: A lady Tasting Tea	60
7.2 Introduktion af hypotesetest for folkeskolen	62

7.3 Chi-i-anden test og Framingham Heart Study	62
7.3.1 Sandsynligheden for at begå fejl	65
7.4. Binomialmodel og binomialtest – beregning og simulering.....	66
7.4.1 Eksempel og øvelse om opinionsmålinger	69
7.4.2 Simulering af en binomialmodel	69
7.5. Normalfordelingen og konfidensintervaller	70
7.6. Hvorfor er $p=0.05$?	71
8. Case: Testet positiv – men er man syg?	72
9. Testteori 2. t-test og sammenligning af to dataserier.	76
Projekt 9.1. Eksperimentel undersøgelse af formelen for varians	80
Projekt 9.2. Analyse af unges adfærd med brug af chi-i-anden-test og t-test.....	80
10. Testteori 3. SRP-inspiration: ANOVA og F-fordelingen	81
10.1 One way ANOVA – (Enkeltfaktor-ANOVA).....	83
10.2 Two way ANOVA – (To faktor Anova)	86
10.3 ANCOVA	87
11. Matematisk modellering af medicins optag og forbrænding i kroppen 88	
11.1 Et lægemiddels rejse gennem kroppen – lidt baggrundsteori	89
11.1.1 Absorption	90
11.1.2 Distribution	90
11.1.3 Metabolisme.....	90
11.1.4 Elimination.....	91
11.2 Compartmentmodellering	91
11.3 1-compartmentmodeller	91
11.3.1 0. ordens elimination.....	92
11.3.2. 1. ordens elimination.....	93
11.3.3. 2. ordens elimination.....	93
11.4. 2-compartmentmodeller	94
11.4.1. Løsning vha. lineær 1. ordens differentialligning ('Panserformlen')	95
11.4.2. Løsning ved omskrivning til en 2. ordens differentialligning.....	96
11.5 Mere om compartmentmodeller	98
12. Matematisk modellering af diabetes	99
12.1 Introduktion til modellen	100
12.2 Opstilling af modellen.....	100
12.3 Løsning af differentialligningssystemet	101
12.4 Estimering af parametrene og tegning af grafiske billeder	106
Litteratur	109

0. Indledning

I projektet *Træk virksomhederne ind i undervisningen* vil der blive produceret 12 film med tilhørende undervisningsmaterialer til hver film. Det foreliggende er skrevet i tilknytning til filmen *Matematikken bag ny medicin hos Novo Nordisk*.

Materialet er opdelt i en række kapitler, der kan gennemgås hver for sig. Det enkelte kapitel bygger således ikke på de foregående. Med symboler er det markeret, at dette kapitel eller afsnit kan man arbejde med, når man er på pågældende niveau. Man kan naturligvis også have glæde af at se et stof på nye måder, selv om man selv nu befinder på et lidt højere niveau. Og man kan prøve at udfordre sig selv ved at gå i krig med et emne og øvelser, der er markeret til et højere niveau end ens eget. Niveauerne er markeret med symbolerne:

 9. klasse og 1.g	 Slut 1.g og 2.g	 3.g
---	--	--

En stor del af emnerne er med i de pågældende niveaus læreplaner / curriculum. Men de behandles givetvis her med en anden vinkel – og det kan være en fordel i læringsprocessen at se emner belyst og anvendt på forskellige måder.

Gå selv på opdagelse:

- Du vil finde mange forslag til mindre projekter som afleveringsopgaver, til SRP'er og til at eleverne kan arbejde eksperimenterende med stoffet.
- Der er forslag til at udbygge de statistiske emner fra læreplanerne, både mht teori og praksis. Der er fx en historisk introduktion til normalfordelingen med udgangspunkt i Gauss fejlfunktion, hvor man dels kan dykke ned i Gauss udledning af funktionsforskriften, og dels i den belgiske astronom og matematiker Quetelets argumentation for, at fejlfunktionen også kan anvendes i helt andre sammenhænge, en argumentation, der ikke er helt fin i kanten. I dag anvendes normalfordelingen ofte nærmest helt uden omtanke.
- Der vil være emner som anvendelse af ANOVA og ANCOVA med tilhørende F -fordelinger til statistiske analyser, og de to sidste kapitler om matematisk modellering af medicins og insulins optag i kroppen, som er udfordrende også for elever i 3.g. Men emnerne kan måske inspirere til studieretningsprojekter.

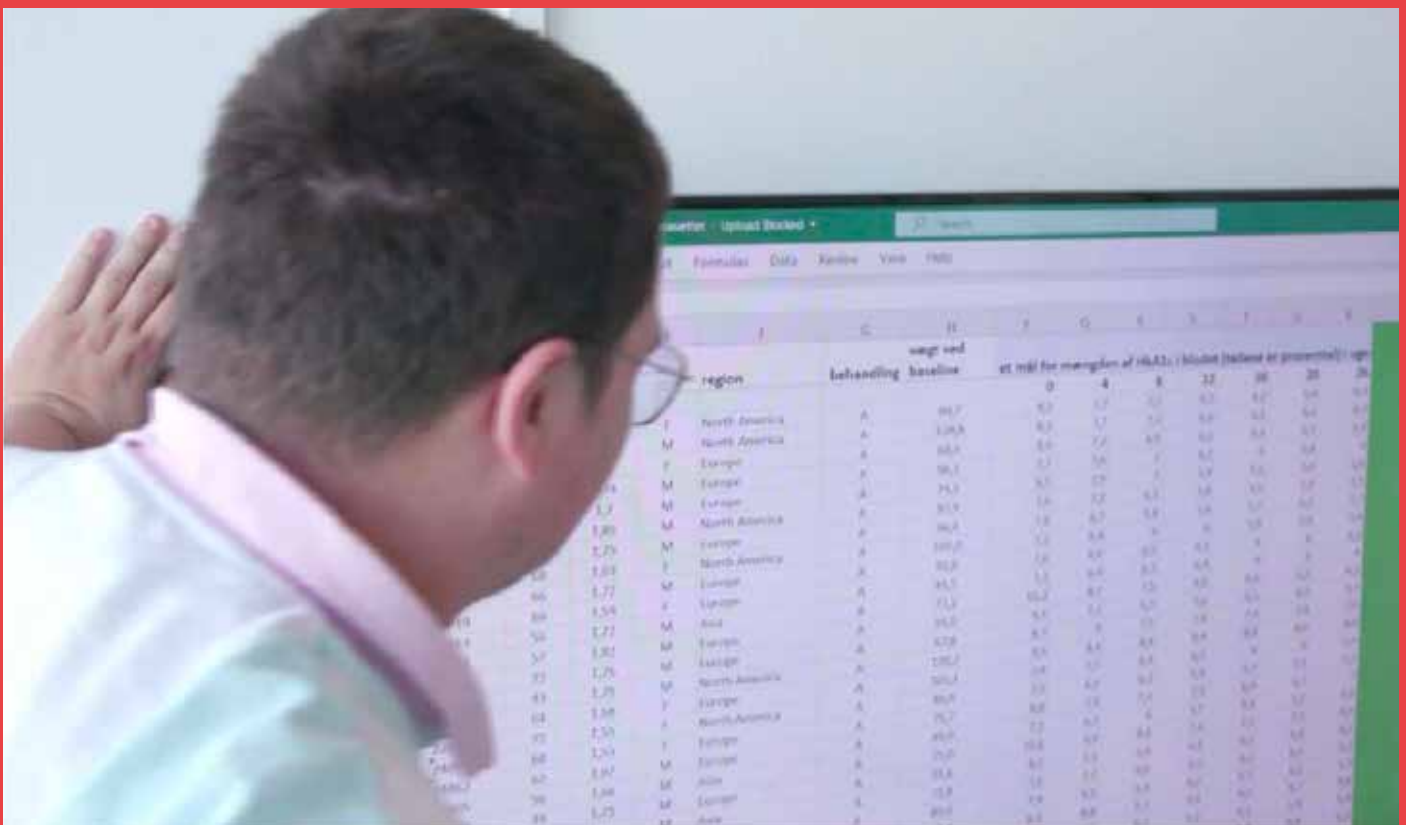
Film og materialer er til fri download og anvendelse i undervisning og selvstudier. Bliver undervisningsmaterialet downloadet og dele af det kopieret, skal der angives kilde.

Forkortelsen HEM står for lærebogssystemet: *Hvad er matematik?* Der kan frit hentes projekter og andet materiale fra HEM's website: [Hvad er matematik - LRU.dk \(praxis.dk\)](http://Hvad%20er%20matematik%20-%20LRU.dk%20(praxis.dk)). Hvor der er markeret **Link til 7** (og andre tal) angives, at her kan der hentes ekstra materialer ind. Disse tilgås via website for Novo-filmen. Find nummeret og klik.

Vi vil meget gerne have feedback med kommentarer og forslag, der kan forbedre de kommende samlinger af undervisningsmaterialer.

Film og undervisningsmaterialer produceres med støtte fra Novo Nordisk Fonden

1. Håndtering af store datasæt



I filmen præsenteres vi for et udtræk af et stort datasæt fra et fase-3 forsøg. Datasættet indeholder en række numeriske og kategoriske variable, og foreligger i et Excelformat. I kapitlet gennemgår vi en række metoder fra den deskriptive statistik, der kan give os et overblik over data.

1. Håndtering af store datasæt

1.1 Datasættet fra filmen

Det datasæt, vi har fået stillet til rådighed af Novo, og som du kan hente [her](#) <link til 1> vedrører et studie, hvor to insulinpræparater A og B sammenlignes. Filen rummer et udtræk på 200 individer ud af en meget større database, og der er valgt et begrænset antal variable ud af de mange, der indgår i undersøgelsen.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1										mængden af HbA1c i blodet (målt i en bestemt enhed) i uge:						
2	Patientnr	patientnr i studie	alder	højde	køn	region	behandl	vægt v baseline	bmi v baseline	uge 0	uge 4	uge 8	uge 12	uge 16	uge 20	uge 26
3	1	716	64	1,66	F	North America	A	99,7		8,5	7,7	7,2	6,5	6,2	6,4	6,3
4	2	923	47	1,72	F	Europe	B	82,5		7,9	7,3	7,1	6,8	6,8	6,9	6,2
5	3	948	50	1,88	M	North America	A	128,8		8,1	7,7	7,2	6,6	6,1	6,2	6,3
6	4	1118	49	1,65	F	Europe	B	94,3		7,4	6,9	6,4	6,3	6,1	6,2	6,3
7	5	1389	59	1,66	F	Europe	B	108,6		7,8	6,8	6,6	6,3	6,2	6	5,9
8	6	2272	65	1,6	F	Europe	A	68,6		8,6	7,2	6,9	6,5	6,3	6,1	5,9
9	7	2606	54	1,74	M	Europe	A	96,1		7,7	7,6	7	6,2	6	5,8	6
10	8	2681	53	1,7	M	Europe	A	74,5		8,5	7,9	7	5,9	5,5	5,6	5,8
11	9	3241	67	1,85	M	North America	A	97,9		7,6	7,2	6,1	5,8	5,6	5,6	5,5
12	10	3584	62	1,64	M	Europe	B	83,2		7,6	7	6,9	6,9	6,1	6,2	5,9

Raske personer producerer insulin, som signalerer til kroppen, at den skal optage glukose. Når man ikke producerer nok insulin, får man for høje glukose-niveauer i blodet, og det fører til forskellige kedelige følgevirkninger, hvis det står på over længere tid. Det kan f.eks. føre til følelsesløshed og sår i fingre og tæer, som kan føre til behov for amputationer, og blindhed.

En måde at behandle sukkersyge patienter på er altså at give dem den insulin, de ikke selv kan producere (nok af). Den variabel, der i excel-filen betegnes HbA1c er et udtryk for hvor højt blodsukker-niveau man har haft over de seneste uger/måneds tid. Hb står for hæmoglobin. Jo mere sukker (glukose) der er i blodet, jo mere bliver bundet til HbA1c. Ved høje blodsukker koncentrationer har man meget glukose bundet til hæmoglobinet. HbA1c-tallene i arket er *procent-tal*. Det er i denne sammenhæng ikke så vigtigt, hvad det er procent af.

Insulinen får blodsukkerniveauet til at falde. En rask person har et HbA1c typisk mellem 4% og 5,6%. Personer med HbA1c over 6,4% anses for **at have sukkersyge**. HbA1c mellem 5,6 og 6,4% anses for **at være prædiabetikere**, altså personer i fare for at udvikle sukkersyge. Når man behandler patienter, der har sukkersyge, er målet typisk at få HbA1c stabilt under 7%.

Udfordringen er, at man har brug for *et vist niveau af glukose i blodet*. Ellers risikerer man at få et hypoglykæmi anfald. Man kan blive klodset, svimmel, besvime og i sidste ende dø. *For meget insulin* kan føre til et hypoglykæmi anfald, og det frygter alle diabetikere. De fleste diabetikere oplever det ind imellem, især i mild form. En undersøgelse som den foreliggende havde også fokus på, om der optræder sådanne anfald og i givet fald, hvor alvorlige de er. Det er ikke med i vores datasæt.

I det følgende vil vi fokusere på, om de to præparater lever op til at bringe HbA1c stabilt under 7%. Og videre at sammenligne præparaterne: Hvilket er bedst! Undersøgelsen forløber over 26 uger med målinger hver uge. *Baseline* er et fagligt begreb for startværdien i uge 0. *Endpoint* er et fagligt begreb for slutværdien, her i uge 26

Vi starter med at stille en række spørgsmål, derefter skal du selv lege forsker og dermed selv stille spørgsmål til materialet. Materialet kan fx sorteres efter køn, efter region (verdensdel), efter patienternes vægt, efter alder, hvor man selv fastlægger intervaller.

0. Hent datasættet, giv det et navn og gem det hos dig selv.

1. Scrol ned gennem regnearket og kommenter værdierne af HbA1c ved baseline for patienterne. Har de sukkersyge eller er de prædiabetikere?

2. Udregn hvad der skal stå i kolonnerne fra Q til W, der skal rumme information om *ændring af HbA1c i blodet ift baseline* i de enkelte uger. Fx i uge 16, dvs i kolonne U udregnes for patient nr 1 forskellen på værdien i uge 0 og værdien i uge 16, dvs: $6,2 - 8,5 = -2,3$. Dvs der har været et fald på 2,3. Udnyt igen regnearket til at gennemføre disse beregninger. Gem det under et nyt navn, så du bevarer det oprindelige.

	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1								HbA1c-niveauet i blodet i uge:					ændring af HbA1c-niveauet i blodet ift baseline i uge:						
2	køn	region	behandl	vægt v basel	bmi v baseline	uge 0	uge 4	uge 8	uge 12	uge 16	uge 20	uge 26	uge 0	uge 4	uge 8	uge 12	uge 16	uge 20	uge 26
3	F	North America	A	99,7		8,5	7,7	7,2	6,5	6,2	6,4	6,3	0	-0,8	-1,3	-1,8	-2,3	-2,1	-2,2
4	F	Europe	B	82,5		7,9	7,3	7,1	6,8	6,8	6,9	6,2	0	-0,6	-0,8	-1,1	-1,1	1	-1,7
5	M	North America	A	128,8		8,1	7,7	7,2	6,6	6,1	6,2	6,3	0	-0,4	-0,9	-1,5	-2	-1,9	-1,8
6	F	Europe	B	94,3		7,4	6,9	6,4	6,3	6,1	6,2	6,3	0	-0,5	-1	-1,1	-1,3	-1,2	-1,1
7	F	Europe	B	108,6		7,8	6,8	6,6	6,3	6,2	6	5,9	0	-1	-1,2	-1,5	-1,6	-1,8	-1,9
8	F	Europe	A	68,6		8,6	7,2	6,9	6,5	6,3	6,1	5,9	0	-1,4	-1,7	-2,1	-2,3	-2,5	-2,7
9	M	Europe	A	96,1		7,7	7,6	7	6,2	6	5,8	6	0	-0,1	-0,7	-1,5	-1,7	-1,9	-1,7
10	M	Europe	A	74,5		8,5	7,9	7	5,9	5,5	5,6	5,8	0	-0,6	-1,5	-2,6	-3	-2,9	-2,7

3. Hvis du er øvet i at bruge Excel kan du springe dette over og gå til punkt 4. Hvis ikke skal du lige lære nogle grundlæggende ting. [Hent link 2](#), og gennemfør følgende små øvelser:

3.1 Sorter hele datasættet efter følgende **numeriske variable**

- Alder
- Højde
- Vægt ved baseline

3.2 Sorter hele datasættet efter følgende **kategoriske variable**:

- Region, så Asien kommer først
- Region, så Sydamerika kommer først
- Behandlingstype, så alle data med type B behandlingen ligger øverst
- Køn, så alle mændenes data kommer først

4. Sorter nu datasættet efter behandlingstyperne A og B. Du kan evt. vælge at oprette to underark, et der vedrører præparat A, og et der vedrører præparat B. Gem det under et nyt navn!

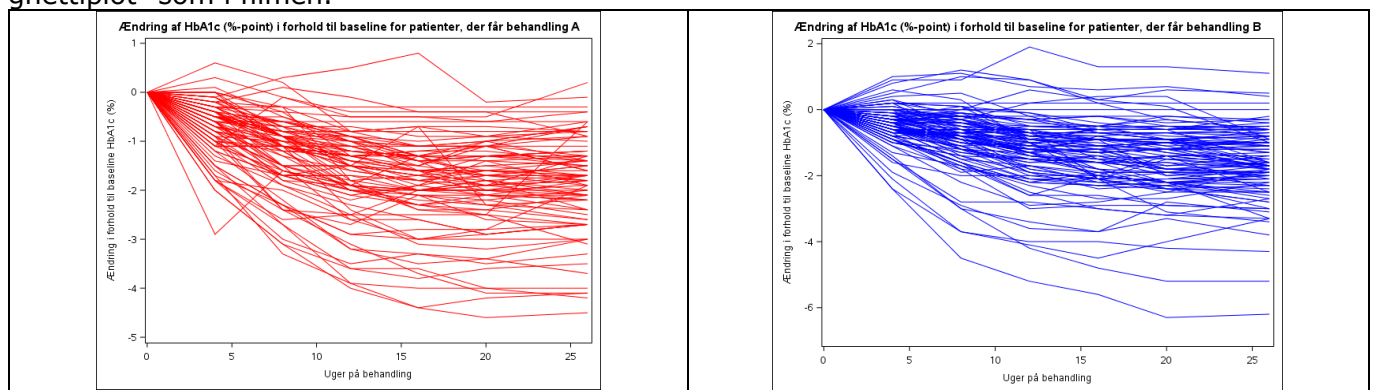
5. Scrol ned gennem de to præparater og kommenter værdierne af HbA1c ved endpoint for patienterne i de to grupper. Hvad er dit umiddelbare indtryk:

- Lykkes det for den overvejende part af patienterne at få HbA1c stabilt under 7%?
- Lykkes det lige godt for de to præparater? Hvis ikke, hvilket ser ud til at klare sig bedst?

6. Det er åbenlyst, at nogle patienter tager bedre imod behandlingen end andre. Prøv at underopdele de to ark efter køn, efter region og efter aldersintervaller (fordel gerne arbejdet imellem jer i en gruppe). Kan I se noget mønster i, hvem der tager dårligst imod medicinen?

7. Hvis du er øvet i at tegne grafer i excel, så kan du springe dette over og gå til punkt 8. Hvis ikke, så hent vejledningen her [- link til 3](#), og gennemfør de små øvelser der.

8. I skal nu for hver af de to præparater A og B tegne grafer for **ændringen** af værdierne af HbA1c som funktion af ugetallet. I må gerne eksportere tallene fra arket til det program, I plejer at arbejde i (Fx. Geogebra, Maple, TI Nspire). I skal tegne graferne i samme koordinatsystem, så I får et "spaghettiplot" som i filmen.



9. Kommenter forskelle og ligheder mellem de to spaghettigrafer. Understøtter graferne jeres svar på øvelse 5?

10. Hvis I fandt nogle sammenhænge i øvelse 6, så lav også et spaghettiplot af dette / disse. Kommenter graferne – understøtter de, jeres formodning fra øvelse 6.

11. Udregn nu for henholdsvis A og B gennemsnittet for værdierne i hver uge.

Hvis du er usikker på, hvordan man bestemmer middeltal, median og boksplot, så kan du i undervisningsmaterialerne til DMI-filmen i kapitel 2 finde en introduktion med øvelser

12. Udfør i samme koordinatsystem plots af disse gennemsnitstal.

- Kommenter ligheder og forskelle
- Hvilket præparat ser ud til at klare sig bedst

13. Vælg selv nogle underopdelinger og sammenlign på samme møde de grafiske forløb (af gennemsnittene) for fx: mænd-kvinder, to aldersgrupper, to regioner og andet.

14. Find formelen for BMI, og udregn denne for patienterne. Husk at udnytte regnearkets faciliteter, hvor en formel skrives med et lighedstegn, og med cellernes "adresser" eller "koordinater", fx kunne man i celle I8 skrive: $= \frac{H8}{(D8)^2}$; grunden til at man skriver adresserne, hvorfra tallene hentes og ikke

bare tallene selv er, at man derefter kan trække formelen ned gennem arket, så man kun skriver den én gang.

15. Opstil selv nogle hypoteser, der inddrager BMI, fx at HbA1c-niveauet ved baseline afhænger af BMI.

1.2 Projekt: Generer selv et datasæt om krop og sundhed

I dette projekt skal I indsamle en større mængde data fra så mange af jeres kammerater som muligt. Det vil være rigtig godt, hvis I indsamler data fra 2-3 andre klasser, så vi får et lidt større datasæt. Datasættet skal både bruges til opgaver om sortering som vist ovenfor, men skal også bruges til andre statistiske undersøgelser så gem det endelig! Hvis I deler dataindsamlingen op i flere grupper, kan I samle de enkelte delresultater til et stort samlet regneark til sidst. Der skal helst være over 50 forsøgspersoner, hvis det skal være lidt interessant.

Materialer: I skal bruge et målebånd og et Excel regneark, f.eks. på en bærbar computer, og indføre svarene for alle nedenstående spørgsmål for alle forsøgspersonerne i regnearket. Regnearket skal være udformet på følgende måde – gerne med flere kategorier I selv finder på:

D	E	F	G	H	I	J	K	L	M	N
forsøgs		omkreds	sko	køn		omkreds	kost			
person nr.		håndled	størrelse	mand/	antal	hoved	kød/	ryger		
nr.	højde/cm	cm	nr	kvind/andet	søskende	cm	vegetar	ikke ryger Find selv på flere	
1										
2										
3										
4										
5										

Spørgsmål i spørgeskemaet

Spørg og undersøg/mål for hver af forsøgspersonerne

- a) Hvad er personens højde i cm
- b) Mål omkredsen om personens håndled i cm (med målebåndet)
- c) Hvad er personens skonommer
- d) Hvilket køn har personen
- e) Hvor mange søskende har personen
- f) Mål personens omkreds om hovedet i cm
- g) Spørg personen om vedkommende er vegetar eller kødspiser. I kan også inkludere veganere

- h) Spørg om personen er ryger/ikkeryger... evt. kan I også inddrage festrygere
- i) Finde gerne selv på flere spørgsmål som f.eks. hvilket stjernetegn personen har eller om, hvor lang skærmtid de i gennemsnit bruger om dagen

Opgave 1 til datasættet

- a) Prøv at sortere datasættet efter hvor høje forsøgspersonerne er. Ser der ud til at være en sammenhæng med køn?
- b) Samme spørgsmål som a) men nu hvor datasættet er sorteret efter skostørrelse.
- c) Hvordan skal datasættet sorteres, hvis man vil have en fornemmelse af, om der er en sammenhæng mellem køn og omkreds om hovedet? Prøv at undersøge om der ser ud til at være det.
- d) Hvordan skal datasættet sorteres, hvis man vil undersøge, om der er en sammenhæng mellem hvilket køn forsøgspersonen har og hvor mange søskende, man har?
- e) Sorter datasættet så du kan se, om der er flest vegetarer blandt piger eller drenge. Hvad vil du forvente, før du sorterer det?

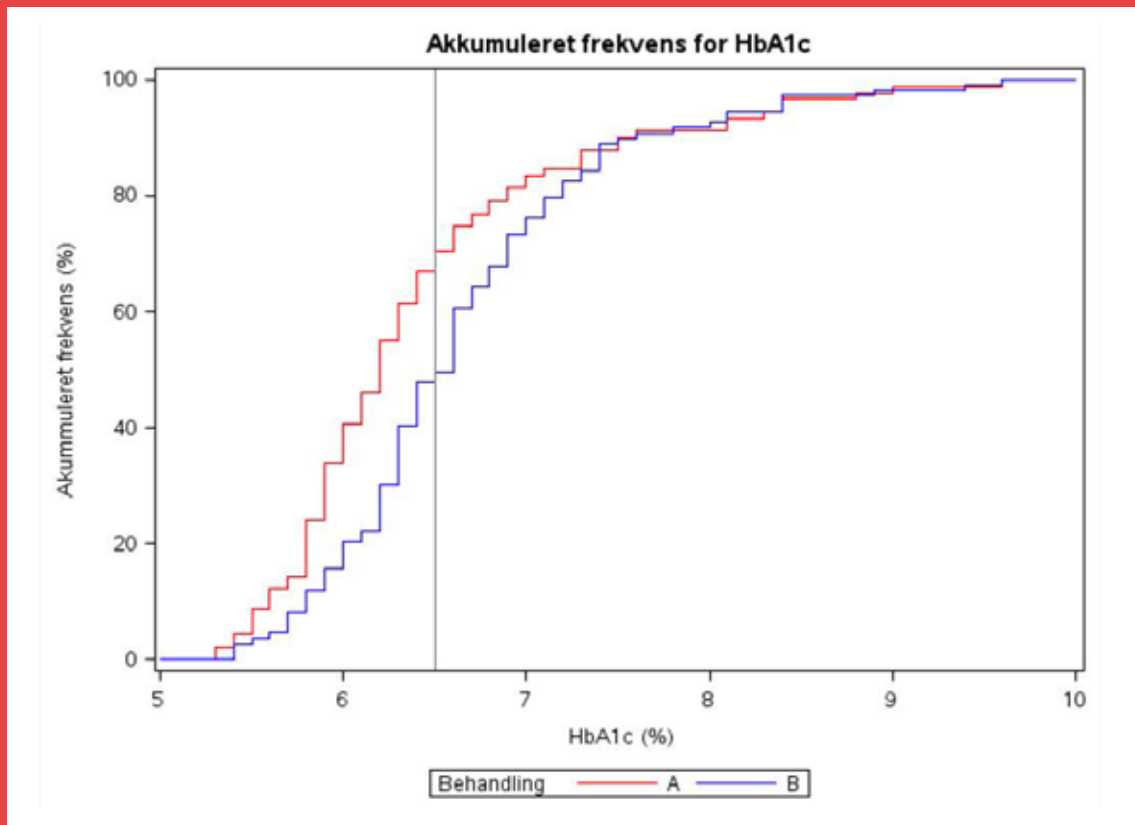
Opgave 2 til datasættet

Inddrag nu din viden om statistiske deskriptorer og boksplot og prøv at lave en boksplotsammenligning af pige og drengegruppen i forhold til

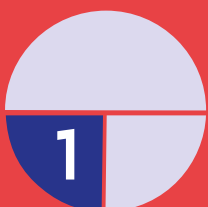
- a) Højdefordelingen af drenge og piger.
- b) Skonummerfordelingen af drenge og piger.
- c) Hvilken type boksplot skal du lave, hvis du vil undersøge om vegetarer er højere end kødspisere? Prøv at lave et boksplot der undersøger dette og se, hvad du kan konkludere.
- d) Prøv at undersøge med et boksplot, om vegetarer går i mindre sko end kødspisere.

2.

Grafiske metoder i beskrivende statistik: Histogram og sumkurve



I filmen præsenterer Lea en række grafiske metoder fra den deskriptive statistik, der skridt for skridt kan give os et stadigt mere klart billede af, om der er belæg for at gå videre i vores undersøgelse af de to præparater A og B.



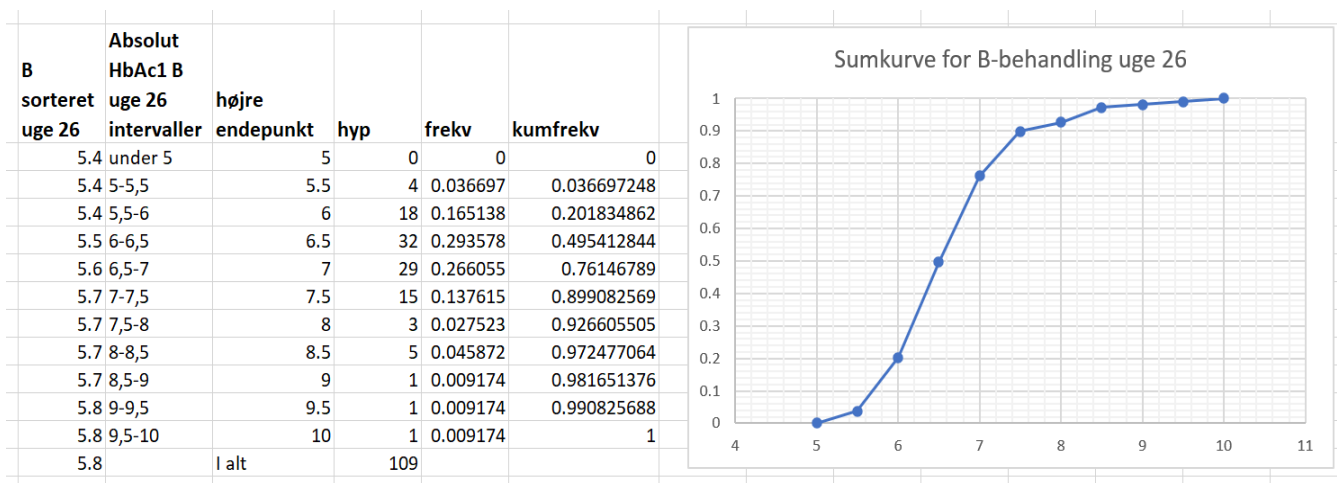
2. Grafiske metoder i beskrivende statistik:

Så længe vi bliver stående ved de individuelle målinger for hver patient, kan det være svært at se det store billede. Der er altid nogle individer, der "stikker ud". Det store billede får vi frem, når vi grupperer data. I filmen anvender Lea de grafiske værktøjer sumkurve og histogram til at visualisere, at det er rimeligt gå efter at påvise, at A er bedre end B. Selv om grafer kan være overbevisende, giver de ikke beviser for en påstand. Men de fortæller os, om det er en rimelig hypotese, eller om vi er på vildspor.

I den beskrivende statistik er der især to diagrammer, der anvendes i sådanne sammenhænge: Sumkurver og histogrammer. Man kunne godt vælge de rå data fra Excel arket som input, men når vi som her ønsker at *sammenligne* to præparater, omregner vi normalt *antal* til *procenttal*.

2.1. Grafisk præsentation af grupperede datasæt – sumkurve

I det følgende eksempel vil vi gerne vise, hvordan man kan skaffe sig en tabel over frekvenser og kumulerede frekvenser, når man vil betragte et stort grupperet datasæt. Vi viser også, hvordan man laver den sumkurve, der hører til grupperingen, så man har det samlede overblik over datasættet. Vi vil i eksemplet kigge på, hvordan data for HbA1c indholdet er fordelt i uge 26 for de patienter, der har modtaget B-behandlingen, og vi vil komme frem til følgende tabeller og diagram:



Metoden kan selvfølgelig bruges til at generere tilsvarende tabeller og sumkurver for alle mulige andre dele eller grupperinger af det samlede datasæt. Mere om det i opgaverne.

Det første man gør er, at sortere den del af datasættet, der skal undersøges som et grupperet datasæt. Metoden er gennemgået i kapitel 1. I vores tilfælde starter vi derfor med at sortere efter behandlingsmetode i uge 26 og får således et søjleområde med de 109 patienter der har fået B-behandlingen.

Næste punkt er at vælge størrelsen af vores grupperingsintervaller. Her har vi bare valgt intervaller med længde 0,5 og har derefter lavet en kolonne, der viser intervallerne. Derudover har vi tilføjet en vigtig kolonne, der angiver det højre intervalendepunkt for hvert af intervallerne – den skal bruges til at lave sumkurven med.

Hyppigheder

Nu skal vi have talt intervalhyppighederne op, dvs. hvor mange patienter der er i hvert interval. Det gøres med en "TÆL.HVISER"-kommando. F.eks. tæller vi op, hvor mange der ligger i intervallet]6;6,5] med kommandoen, du ser her:

Kommandoen skal have udpeget det område, der skal tælles op over (W3:W111) og derudover hvilke kriterier, der skal være opfyldt, for at cellerne skal tælles med. Her er kriterierne, at cellen skal tælles med, hvis indholdet ligger over 6 og højst kan være 6,5.

=TÆL.HVISER(W3:W111;">6";W3:W111;"<=6.5")

	Y	Z	AA	AB
højre				
r endepunkt		hyp	frekv	kumfrekv
	5	0	0	0
	5.5	4	0.036697	0.036697248
	6	18	0.165138	0.201834862
	6.5	32	0.293578	0.495412844
	7	29	0.266055	0.76146789

Bemærk notationen med *anførselstegn*: Vi skriver ">6" og "<=6,5", som det kan ses herover.

På samme måde tælles de andre intervalhyppigheder op. Antallet af observationer under den mindst observation er 0, som her bare er indtastet manuelt.

Frekvenser

Omregningen fra antal til procenter foregår, som vi kender det: Hyppighederne divideres med det samlede antal observationer.

Derefter "trækkes denne beregning ned" gennem søjlen

	Y	Z	AA	AB
Absolut HbAc1 B				
uge 26				
intervaller	højre endepunkt	hyp	frekv	ku
4 under 5	5	0	0	
4 5-5,5	5.5	4	0.036697	=Z4/109
4 5,5-6	6	18	0.165138	

Kumulerede frekvenser

Endelig skal vi have beregnet de *kumulerede intervalfrekvenser*. Disse kaldes også af og til "summerede frekvenser". Forestil dig, at du bevæger dig på tallinjen, fra venstre mod højre. For hvert interval, du når frem til, angiver den kumulerede frekvens, hvor stor en andel der ligger i intervallet eller til venstre for det.

Eksempelvis er den kumulerede frekvens hørende til intervallet fra 6 til 6,5 lig med den samlede procentdel op til og med 6,5.

=AB3+AA4

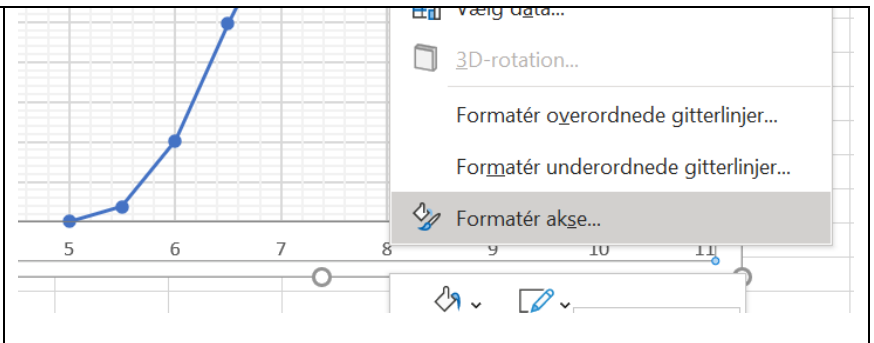
	Y	Z	AA	AB
højre				
endepunkt	hyp	frekv	kumfrekv	
	5	0	0	0
	5.5	4	0.036697	=AB3+AA4
	6	18	0.165138	0.201834862
	6.5	32	0.293578	0.495412844

Den første er selvfølgelig 0, da der er 0 procent under den mindste observation, og dette tages igen bare ind manuelt. De resterende kumulerede frekvenser kan let beregnes ved blot at lægge frekvenserne til den forrige kumulerede frekvens og så igen "trække denne beregning ned".

Du kan tjekke, at det er gået godt ved at sikre dig at de kumulerede frekvenser ender på 1 – dvs. 100%.

Tegning af sumkurven

Nu har vi genereret de data, der skal bruges til at tegne sumkurven. Sumkurven tegnes som et punktplot forbundet med linjer (se ovenfor), hvor vi plotter kolonnen med højre intervalendepunkt ud ad x-aksen og søjlen med de kumulerede frekvenser op ad y-aksen.

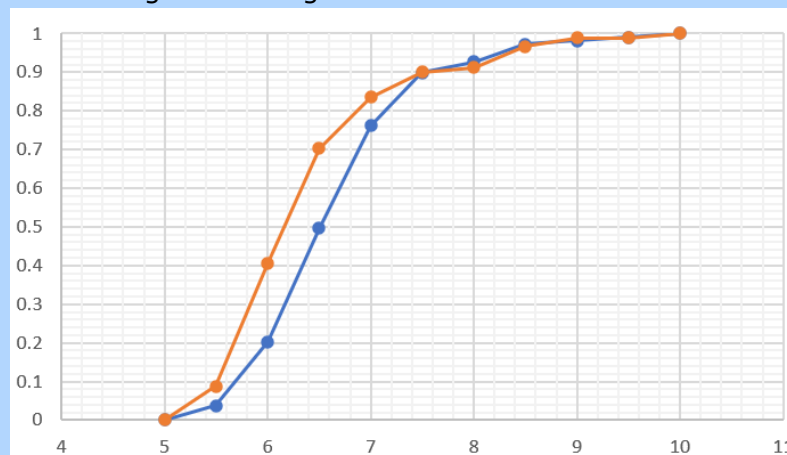


Ofte passer akserne ikke helt med det plot man gerne vil lave og man kan så højreklikke på akserne og vælge "formatér akse":

Der dukker herefter en menu op, hvor man kan vælge maksimum og minimum på akserne. På samme måde kan man redigere i antallet og visningen af gitter linjer, så man endelig får diagrammet til at se ud, som man ønsker det.

Øvelse 2.1

- Tegn sumkurven for B som beskrevet herover.
- Lav en tilsvarende tabel over hyppigheder, frekvenser og kumulerede frekvenser for behandling A i uge 26.
- Tegn en sumkurve for A-behandlingen uge 26 i samme diagram som den du lavede for B-behandlingen – det skulle gerne se nogenlunde således ud:



- Hvordan kan man på grafen over de to sumkurve se hvilken behandling, der ser ud til at virke bedst?

På samme måde som i denne øvelse kan man vælge forskellige udsnit af datasættet og sammenligne virkningen af medicinen på disse underkategorier. Det kigger vi på i følgende opgaver.

Tip: For hver opgave kan det være en god idé at bruge en "frisk" version af regnearket med data, da sortering i efter forskellige kategorier kan ødelægge nogle af de diagrammer, du allerede har lavet.

Øvelse 2.2

- Hent et eksemplar af regnearket, hvor patienterne er sorteret efter A og B
- Beregn i kolonne W faldet i HbA1c indholdet i blodet for alle patienter, dvs indholdet i uge 0 minus indholdet i uge 26
- Gennemfør en intervalsortering af kolonne W for henh. A og B (vælg samme interval-inddeling!)

- d) Tegn de to sumkurver for henh. A og B, og sammenlign A og B patienternes fald i HbA1c indholdet.
- e) Giver disse kurver det samme indtryk som sumkurverne i opgaven ovenover? Hvorfor er kurverne i øvelse 2.1 og i denne øvelse ikke ens?

Øvelse 2.3

I denne opgave skal vi undersøge, om *startværdierne* af HbA1c indholdet for de to behandlingsgrupper A og B er ens.

- Gruppér de A-behandlede i uge 0 i intervaller af bredde 0,5, og beregn en tabel med frekvenser og kumulerede frekvenser.
- Tegn en sumkurve for de kumulerede frekvenser for dem, der skal have A-behandlingen
- Udfør den samme øvelse for de forsøgsparticipanter, der skal modtage B-behandlingen og tilføj en sumkurve for dem i samme diagram
- Kommentér på resultatet – ser det ud på samme måde som efter 26 ugers behandling, eller er de to grupper mere "ens" i udgangspunktet i uge 0?

2.2. Grafisk præsentation af grupperede datasæt – histogram

Vi minder lige om hvordan man tegner et histogram over sine observationer:

Praxis: Tegning af et histogram

- Et histogram tegnes i et koordinatsystem, hvor dataværdierne er afsat på førsteaksen.
- Histogrammet består af søjler, tegnet over hvert af de intervaller, som datasættet er grupperet i.
- Søjleens bredde bør svare til bredden af det underliggende interval.
- Intervalbredden skal så vidt muligt være den samme i alle intervaller.
- Intervalhøjden repræsenterer antallet af observationer i det pågældende interval eller, hvis man regner i procent, andelen af observationer i intervallet.

Hvis vi har grupperet vore observationer så intervallerne er lige brede er det ret let at lave et histogram over observationerne i Excel. Lad os kigge på eksemplet i starten af afsnit 2.1 ovenfor, hvor vi allerede har beregnet interval-frekvenserne:

Læg mærke til, at vi har tilføjet en søjle med intervallerne lige før søjlen med frekvenser – den bliver vist under vores førsteakse når vi plotter histogrammet, da Excel automatisk danner en "tekstakse" når man laver et histogram.

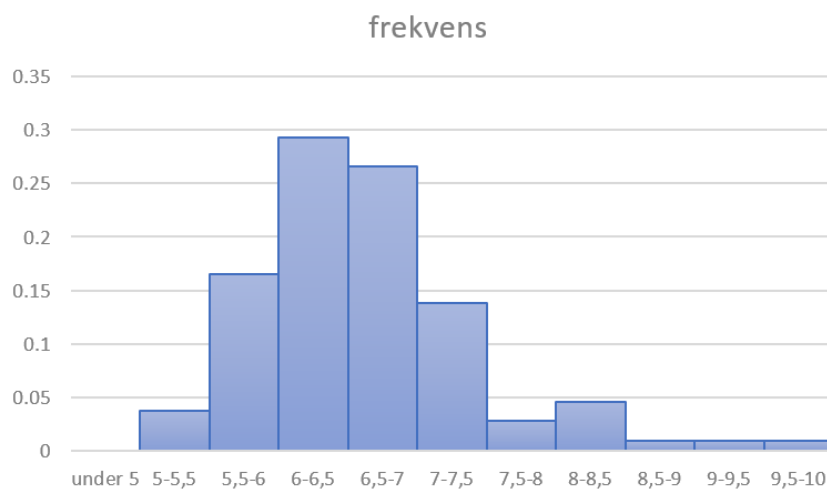
Interval	frekvens
under 5	0
5-5,5	0.036697
5,5-6	0.165138
6-6,5	0.293578
6,5-7	0.266055
7-7,5	0.137615
7,5-8	0.027523
8-8,5	0.045872
8,5-9	0.009174
9-9,5	0.009174
9,5-10	0.009174

Hvis du ikke er vant til at tegne histogrammer i excel, så hent en vejledning her – [link til 4](#)

Vi markerer nu interval og frekvens søjlerne og klikker på indsæt diagram. Vi vælger det simpleste 2-D diagram. I diagramområdet, kan man vælge mellem forskellige typografier oppe i værktøjsbjælken og finde en, hvor søjlerne er markeret med linjer.

Gør dernæst søjlerne lige så brede som intervallerne. Det gøres ved at højreklikke på en søjle og vælge "**Formatér dataserie**". Derved fremkommer en menu, hvor man kan vælge mellemrummet mellem søjlerne til at være 0%.

Og nu skulle histogrammet gerne se sådan her ud:



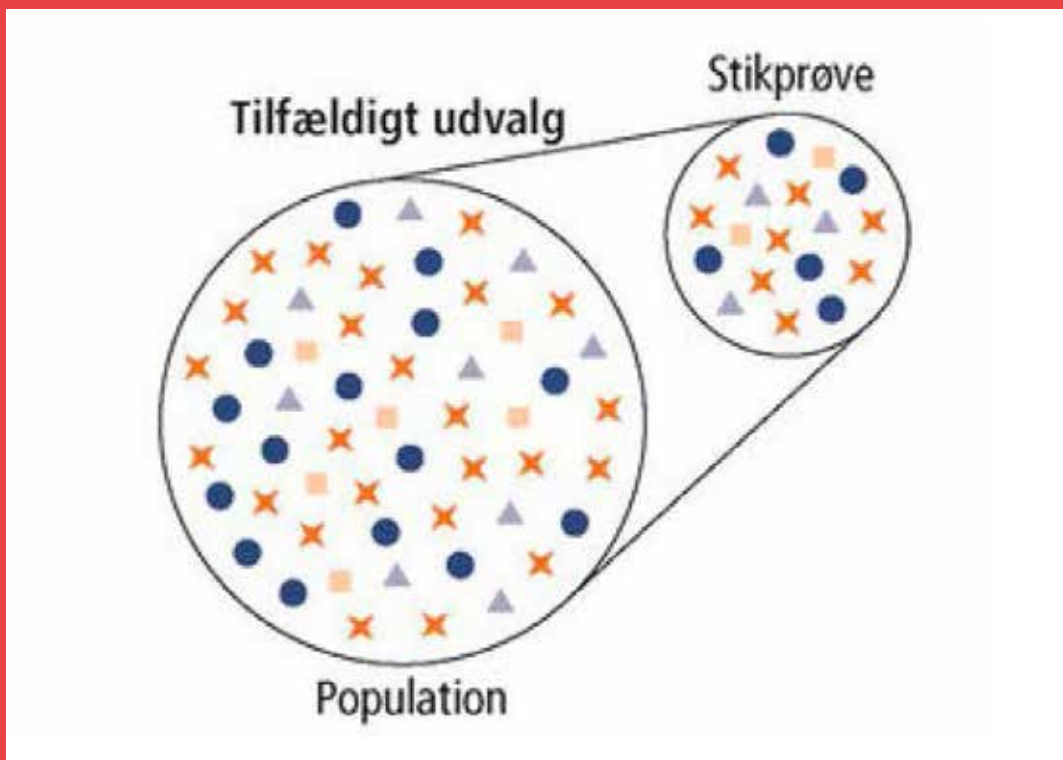
Øvelse 2.4

- Lav et tilsvarende histogram for A-behandlingen, uge 26
- Lav et histogram over *faldet* i HbA1c indholdet for de patienter, der har modtaget A-behandlingen
- Lav et tilsvarende histogram over *faldet* i HbA1c indholdet for de patienter, der har modtaget B-behandlingen

Hvis der er brug for træning i opgaver om sumkurver og histogrammer, kan man hente nogle ekstra opgaver her – [link til 5](#)

3.

Stikprøver og population



Når man skal afprøve lægemidler som de to insulinpræparater A og B, er det afgørende for troværdigheden, at testpersoner er udvalgt, så de er repræsentative ift. en række faktorer. Og når man skal afprøve to præparater som her, er det vigtigt, at det er tilfældigt, hvilke personer der får det ene og det andet præparat. Når man har gennemført en undersøgelse og skal drage konklusioner ud fra data, er det vigtigt at inddrage begrebet "skjulte variable", og at forstå forskellen på kausalitet og korrelation. En sammenhæng kan ikke afsløres alene af statistik – viden fra andre fag skal inddrages. I kapitlet gennemgår vi i en række eksempler og cases disse begreber og illustrerer, hvor let det er at begå fejl!

3. Stikprøver og population

I NOVO-filmen forklares, hvordan udvikling af nye lægemidler tager lang tid og kræver adskillige testfaser. Man skal dels sikre sig, at medicinen ikke har store bivirkninger, man skal finde frem til den rette dosering, og frem for alt skal man undersøge medicinens virkning i forhold til en lang række baggrundsvariable. Men hvordan sikrer man sig egentlig, at man har udvalgt de rette personer og opnået stikprøverepræsentativitet? I det følgende skal vi gennem forskellige cases undersøge nogle af de faldgruber, der eksisterer i statistikkens verden.

3.1 Forsøgsplanlægning: Repræsentativitet og selektionsbias

Øvelse 3.1

På en skole med 700 elever ønsker en af de politiske ungdomsorganisationer at få mulighed for at stille et bord op, hvor eleverne i spiseafdelingen kan hente materialer og få information. Da skolens ledelse siger nej, opfordrer organisationen alle elever til at tilkendegive om de er for eller imod dette. 127 afgav deres stemme og heraf støttede 92 forslaget.

- Hvad er populationen og hvad er stikprøven?
- Hvor stor en andel stemte ja til forslaget?
- Kommenter undersøgelsen.
- Rådgiv organisationen om, hvorledes de kunne foretage en mere kvalificeret undersøgelse.

Øvelse 3.2

En amerikansk politiker udtaler til en Tv-station, at han er interesseret i at høre vælgernes holdning til en lov om våbenkontrol. Hans sekretær opgør efter en uges tid, at de har modtaget breve om spørgsmålet fra 361 vælgere. 323 var imod loven.

- Hvad er populationen og hvad er stikprøven?
- Kommenter undersøgelsen.
- Rådgiv politikerens om, hvorledes der kunne foretages en mere kvalificeret undersøgelse.

Øvelse 3.3

Et sundhedsmagasin ønsker at undersøge om store doser vitamintilskud forbedrer sundhedstilstanden. Bladet anmoder de af læserne, som gennem længere tid har taget store doser vitamintilskud om at skrive ind og fortælle om positive og negative erfaringer med dette. 2754 læsere skriver ind. 93% fortæller, at de kan spore en vis forbedring af helbredet.

- Hvad er population og hvad er stikprøve.
- Kommenter undersøgelsen.
- Giv en vurdering af, om andelen af hele befolkningen, der vil få forbedret sundhedstilstanden ved at indtage store doser vitamintilskud, er større, den samme eller mindre end 93%. Be- grund dit svar.

Øvelse 3.4

En artikel i Politiken omhandlede fænomenet "Mænd der får tæsk", og giver en række overraskende oplysninger om voldelige mænd og voldelige kvinder. Bl.a. fortælles det, at "Australian Institute of Health and Welfare" har opgjort, at 40% af børnemishandlinger i landet står de enlige mødre for. De enlige fædre står kun for 5%.

Kommenter disse oplysninger. Du må gerne selv indføre taleksempler i din argumentation.

Øvelse 3.5

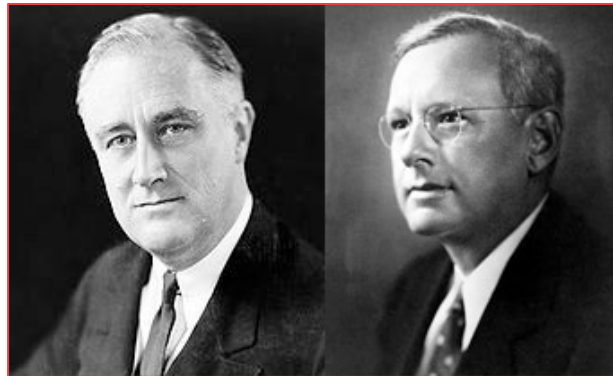
En kvindelig redaktør af et stort amerikansk kvindemagasin spurgte engang sine læsere, om de ville stille sig tilfreds med mænd, der gav dem kærlighed og hengivenhed, men ingen sex.

90.000 kvinder skrev ind og 72.000 svarede ja.

- Hvad er populationen og hvad er stikprøven.
- Giv en vurdering af metoderne i denne statistiske undersøgelse og af hvilke konklusioner, der kan drages.

Øvelse 3.6 Det amerikanske præsidentvalg 1936

Ved præsidentvalget i 1936 skulle amerikanerne vælge mellem demokraten Franklin D. Roosevelt og republikaneren Alfred Landon. I månederne op til præsidentvalget foretog det største amerikanske ugeblad *The Literary Digest* (et magasin med baggrundsartikler, noveller, rejseberetninger og meget andet) en gigantisk meningsmåling blandt bladets egne læsere. Det var den største politiske meningsmåling i USA nogensinde, idet man spurgte 2,4 mio. af læserne, hvem de ville stemme på til valget.



Franklin D. Roosevelt & Alfred Landon

Resultatet af meningsmålingen var en overvældende sejr til Alfred Landon. I virkeligheden gik det dog helt anderledes, idet Roosevelt vandt valget med 61% af stemmerne.

- a) Hvad gik galt for *The Literary Digest*? Hvordan kunne de tage så meget fejl? Hvad betyder begrebet stikprøverepræsentativitet?

Øvelse 3.7

Et konsulentfirma bliver bedt om at sammenligne kvaliteten af behandlingen på to hospitaler A og B. En lille arbejdsgruppe indhenter et materiale, som de opstiller i følgende tabel:

Situationen 2 måneder efter operativt indgreb		
	Hospital A	Hospital B
Døde	63	16
Overlevede	2037	784
I alt	2100	800

Tabellen dannede grundlag for arbejdsgruppens vurdering af de to hospitaler. *Gruppen fremhævede B som det bedste.* En af arbejdsgruppens medlemmer er imidlertid ikke tilfreds med materialet og indhenter supplerende oplysninger, som stilles op i følgende tabel:

Situationen 2 måneder efter operativt indgreb				
	God helbredstilstand før operation		Dårlig helbredstilstand før operation	
	Hospital A	Hospital B	Hospital A	Hospital B
Døde	6	8	57	8
Overlevede	594	592	1443	192
I alt	600	600	1500	200

Skriv din egen konklusion om kvaliteten af behandlingen på hospital A og B. Kommenter samtidig undersøgelsen og materialet i de to forskellige tabeller.

Øvelse 3.8: Grønt lys er farligere end rødt lys

Denne avisartikel blev bragt i *MetroXpress* og citeret i adskillige dagblade i maj.

Undersøgelsen er et rigtig godt eksempel på, at man skal passe meget på med at drage forhastede konklusioner ud fra statistiske data.

Hvorfor kan man ikke drage den konklusion, som journalisten gør i overskriften og "mellem linjerne": Nemlig, at grønt lys er farligere end rødt?

Grønt lys er farligere end rødt



Selvom det burde være børnelærdom, er det ikke ufarligt at køre over for grønt lys. Tværtimod sker der fire gange så mange ulykker, når cyklister kører over for grønt lys, end når de kører overfor rødt.

Fire gange så mange cyklister kommer til skade ved at cykle over for grønt end over for rødt lys

25 maj 2009 kl. 11:20



Vibeke Toft-Nielsen

Selvom det burde være børnelærdom, er det ikke ufarligt at køre over for grønt lys. Tværtimod sker der fire gange så mange ulykker, når cyklister kører over for grønt lys, end når de kører overfor rødt.

Det skriver *MetroXpress* i dag.

Tallene er fra den seneste landsdækkende opgørelse fra 2005 til 2007 fra politiet, der har opgjort antallet af cyklistulykker, hvor personer er kommet til skade.

- Der er ikke 100 procents sikkerhed for, at man bare kan køre over, fordi der er grønt, selvom vi forventer, at så er det okay at køre. Der kan hurtigt komme en bil snigende over for rødt, siger direktør i Dansk Cyklist Forbund, Jens Toft Rasmussen til *MetroXpress*.

Han råder alle cyklister til at have forskellige ruter at skifte mellem til skole eller arbejde, så man ikke bliver alt for uopmærksom, når man cykler på sin vante vej.

Øvelse 3.9

På et site på nettet kan man læse følgende:

STEM NU

Ville du have ret til at gøre en ende på dit liv, evt. med lægelig bistand, hvis du fik at vide, at du var uhelbredelig syg.

Hvis ja: klik her

Hvis nej: klik her

Giv en vurdering af metoden i denne statistiske undersøgelse.

Øvelse 3.10

Antallet af studenter på et bestemt institut er vokset betydeligt, uden der er blevet mere plads eller flere undervisningslokaler. Instituttet vil undersøge studenternes syn på holdstørrelse og andre spørgsmål vedr. de fysiske rammer. De sætter ressourcer af til at interviewe 25 ud af i alt 450 studenter.

Rådgiv dem om, hvorledes du synes de skal udvælge de 25, så der kan drages bedst mulige konklusioner ud af materialet.

3.2 Skjulte variable, korrelation og kausalitet

Øvelse 3.11

Vil indtagelse af urtete styrke helbredet hos de ældre? Dette ønsker en gruppe studerende at undersøge. Over en periode på 6 måneder besøger de nogle tilfældigt udvalgte beboere på et plejehjem og serverer urtete for dem. Efter 6 måneder viser det sig, at de beboere, der fik serveret urtete faktisk har færre sygedage, end de som ikke fik serveret noget.

Giv en vurdering af undersøgelsen og af troværdigheden af resultatet

Øvelse 3.12 Matematik og kriminalitet

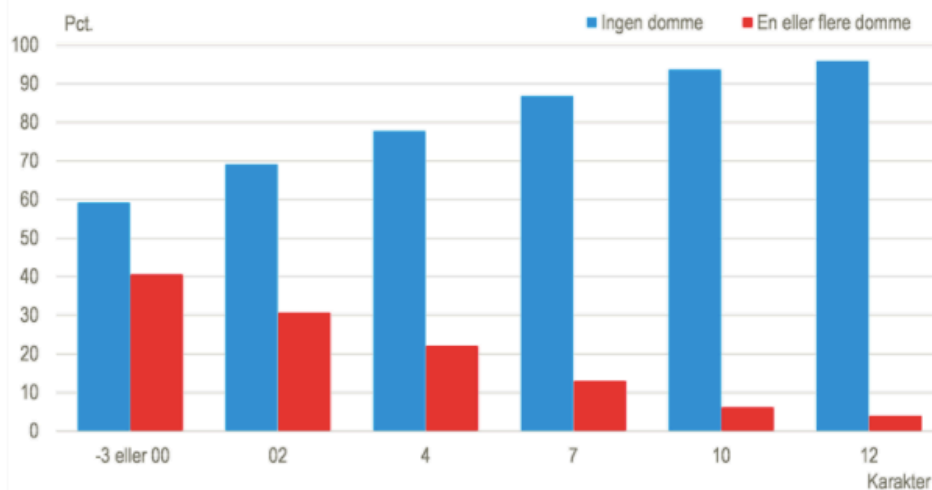
Den følgende artikel kan man finde på Danmarks Statistiks hjemmeside:

40 pct. af mænd der dumper i matematik får dom

Elever i grundskolen (tillæg) 2014 karakterer for unge med dom

Andelen af unge, der har begået kriminalitet, stiger i takt med, at karakteren for grundskolens afgangsprøve i skriftlig matematik falder. Blandt mænd i alderen 25-29 år, der ikke har bestået den skriftlige afgangsprøve i matematik, har mere end fire ud af ti (41 pct.) fået én eller flere domme efter straffeloven, våbenloven eller loven om euforiserende stoffer som 25-årig. Knap hver tredje (31 pct.) af mændene med karakteren 02 har fået en dom, og blandt mænd med karakteren 12 er kun 4 pct. blevet dømt.

Mænd i alderen 25 til 29 år fordelt på domme og grundskolens afgangsprøve i skriftlig matematik. 2014



Anm.: Dømte personer har en eller flere domme efter straffeloven, våbenloven og lov om euforiserende stoffer.

Kilde: <https://www.dst.dk/da/Statistik/nyheder-analyser-publ/nyt/NytHtml?cid=26418>

- a) Hvad kan man konkludere ud fra denne undersøgelse? Hvad kan man ikke konkludere? Ville statistikken mon se anderledes ud for andre fag ved grundskolens afgangsprøve? Inddrag begreberne kausalitet og korrelation i din forklaring

Øvelse 3.13 Vaccinestik dræber

I en folder fra "Foreningen mod Covid-19 vaccination" kan man læse følgende fakta:

"... udviklingen er meget foruroligende: Siden årsskiftet er 2.423 mennesker **døde** inden for 30 dage efter 1. **vaccine** stik, mens 1.992 er **døde** inden for 30 dage efter 2. **vaccine** stik. Disse tal viser klart, at vaccinen er en dræber..."

Lyder det overbevisende? Så overvej følgende:

- Hvilke mennesker er det fortrinsvist, der blev vaccineret til at starte med?
- Hvad kan dødsårsagerne have været (ud over vaccinen) for denne gruppe?
- Hvor mange ville være døde af disse årsager alligevel, selvom de ikke var blevet vaccineret?

3.2.1 Opdigtede avisnotitser

Formålet med denne aktivitet, er bl.a. at introducere/formalisere/diskutere statistiske begreber, så eleverne får nogle sproglige og matematiske værktøjer, der kan sætte dem i stand til at kritisere og forklare misbrug/forkert brug af sammenhæng i tabeller. Samt specielt at klargøre forholdet mellem statistisk og kausal afhængighed.

Udgangspunktet for arbejdet er en række opdigtede avisnotitser. De er selvfølgelig opdigtede, men realistiske om end på en overdreven måde. Den stokastiske variation er taget væk for at gøre pointerne tydeligere. Ved at beskæftige sig med situationer, som de har erfaring med og derfor intuition omkring ledes eleverne frem til at operationalisere en række af de statistiske begreber, population, stikprøver, repræsentativitet, skjulte variable, bias, samt se sammenhængen med de teoretiske begreber statistisk og kausal afhængighed.

Eleverne arbejder i grupper med en række af de små cases, og de skal fx med brug af taleksempler eller diagrammer, gennemføre analyser af situationerne og være i stand til at præsentere et ræsonnement, der afdækker fejlene i "avisnotitserne". Endelig skal de selv formulere et antal tilsvarende notitser – eller endnu bedre: finde eksempler fra aviserne.

Ideen er hentet fra et materiale, Inge Henningsen udarbejdede til den indledende undervisning i statistik på Matematisk institut, KU.

Øvelse 3.14 Føddernes størrelse afgørende for valg af uddannelse

Ved en undersøgelse i efteråret 2020, hvor man sammenlignede en gruppe ingeniørstuderende og en gruppe fransk-studerende opdagede man en meget overraskende forskel. De studerende på ingeniørstudiet brugte i gennemsnit sko, der var 1,6 numre større end de fransk-studerende. Forskerholdet bag undersøgelsen mener, at der måske er en genetisk forklaring. En talsmand for Polyteknisk Forening siger: "Jeg har også altid syntes, at humanister gik i små sko."



- Hvorfor bruger de franskstuderende generelt mindre sko end de ingeniørstuderende. Ofte taler man i statistik om "forstyrrende baggrundsvariable". Hvad har det med denne undersøgelse at gøre? Hvad er mon den forstyrrende baggrundsvariable i artiklen?

Øvelse 3.15 Rollatorer er farligere end rulleskøjter

En omfattende analyse fra 2020 foretaget af skadestuen på Rigshospital har vist, at tre gange så mange kommer til skade ved at bruge rollator som ved at løbe på rulleskøjter. Socialforvaltningen i Københavns Kommune har derfor taget kontakt med rulleskøjte-producenten Bauer for at få en aftale om levering af "inlinere" til alle borgere i kommunen på over 70 år. På sigt forventer forvaltningen at omlægningen vil medføre en nedgang på 30 % i indlæggelserne i den pågældende aldersgruppe.



- a) Hvorfor er der flere, der kommer til skade ved brug af rollator end rulleskøjter? Hvilken baggrundsvariabel er på spil her

Øvelse 3.16 Hoftebrud medfører forhøjet kræftisiko.

Ved en undersøgelse på Rigshospitalet har det vist sig, at patienter indlagt for brud på lårhalsknoglen har en 10 gange så høj risiko for at få kræft inden for en periode på 5 år, som patienter indlagt med henblik på en meniskoperation. Lægerne undersøger nu, hvorfor en forskellig traumelokalisering giver en så markant forskel i risikoen for få kræft. På baggrund af undersøgelsen påpeger bandedagist NN, at systematisk brug af firmaets nyudviklede underbukser med stødabsorberende indlæg til beskyttelse af hoftepartiet, vil kunne reducere den alt for høje danske kræftisiko.

- a) Kommenter!

Øvelse 3.17 Det er livsforlængende at komme på plejehjem

Ifølge Danmarks Statistik er der ca. 45000 personer, der bor på plejehjem i Danmark. Plejehjemsbeboerne i Danmark bliver i gennemsnit 86 år gamle, før de dør. I resten af den danske befolkning er den gennemsnitlige levealder imidlertid kun 81 år. En konklusion kunne være: Det er livsforlængende at komme på plejehjem!



- a) Hvorfor er konklusionen ovenfor helt forkert?

Øvelse 3.18 Amtssygehus nedlægger intensiv afdeling.

Amtssygehuset i XX har besluttet at nedlægge den intensive afdeling, idet en undersøgelse har vist, at dødeligheden på denne afdeling ligger langt over hospitalets gennemsnit. I fremtiden vil alle alvorligt syge patienter blive indlagt på ortopædkirurgisk afdeling. Hospitalsdirektøren tror, at den nye organisering vil blive en væsentlig sundhedsmæssig gevinst for amtets borgere.

Øvelse 3.19 Storke og fødsler

I 2000 skrev Professor Robert Matthews fra Aston University en videnskabelig artikel med titlen: *Storks deliver babies*. I artiklen påviser han en overraskende stor statistisk sammenhæng mellem storkepopulationer og fødselsrater blandt mennesker i forskellige europæiske lande. Artiklen har til formål at fortælle læseren om forskellen på *korrelation* og *kausalitet*.

Artiklen kan hentes i en pdf-version her – [link til 6](#)

- a) Læs artiklen og diskutér hvad forskellen er på begreberne korrelation og kausalitet. Hvad kan man konkludere ud fra Matthews' undersøgelse? Hvad kan man ikke konkludere?

3.3 Case: Nedskydning af amerikanske kampfly under 2. VK

(Dette afsnit er baseret på: Jordan Ellenberg: *How not to be wrong - The hidden maths of everyday life*, kapitel 1. Eksemplet kan foldes ud til et større projekt ved at inddrage artikler af og om Walds arbejde, som der linkes til i slutningen af afsnittet)

Luftangrebene mod Tyskland intensiveres i 1943. Men tyskerne har stadig et effektivt luftforsvar, og sammen med deres kampfly påfører de USA store tab i form af nedskudte fly. På dette tidspunkt er der etableret et samarbejde mellem hæren og en række grupper af specialiserede forskere på universiteterne, og man beslutter at bede en af disse grupper, *The Statistical Research Group at Columbia University, SRG* om hjælp.

En af forskerne i SRG-gruppen er Abraham Ward, der er matematiker og statistiker, og som i 30'erne var flygtet fra Nazi-Tyskland til USA. Han havde jødisk baggrund og de i hans familie, der ikke slap ud, blev alle udryddet. Han får en hovedrolle i arbejdet med at finde nogle modtræk til nedskydningen.

Det amerikanske militær forestiller sig, at de skal have hjælp til at forstærke armeringen af deres kampfly.

Der var flere problemstillinger Wald skulle tage stilling til. Dels hvor meget armering man kunne give kampflyene uden at hæmme deres manøvredegytighed for meget og uden at bruge for meget yderligere brændstof. Og dels hvilke dele af kampflyet, der var særligt udsat for tysk beskydning og derfor havde behov for særlig beskyttelse:



Abraham Wald (1902-1950)

Så hvor på kampflyet skulle man ofre særlig tyk armering, og hvor kunne man nøjes med tyndere armering?

Til støtte for Walds analyser havde det amerikanske militær fremskaffet data om skudhuller, for de kampfly, der vendte tilbage til baserne, og der var lavet forskellige visualiseringer af det:

	Udsnit af flyet	Antal skudhuller pr kvadratfod
	Motoren	1.11
	Skroget	1.73
	Brændstoftankene	1.55
	Resten af flyet (vinger mm)	1.8

Data er de autentiske data fra anden verdenskrig. Visualiseringen er ikke autentisk, men en nyere tegning. Den illustrer dog udmærket, hvilket materiale Wald havde.

1. del: Data fremlægges og diskuteres

Data præsenteres, så man er sikker på, alle forstår problemstillingen. *Er der elever, der kender problemstillingen på forhånd, kan de ikke deltage i diskussionen.*

Klassen deles op i grupper, fx parvis. Grupperne diskuterer, hvad de ville anbefale vedrørende armeringen af flyene: Hvilken del af flyet ville de ofre mest armering på.

Derefter tages en åben diskussion grupperne indbyrdes, man hører de forskellige overvejelser, de måtte have gjort og de forskellige anbefalinger, det måtte føre til. Hvis der er konsensus om et af

forslagene kan man notere det, men på dette stadie gælder det om at holde alle muligheder åbne og ikke 'røbe' det 'rigtige svar'.

2. del: Datastrukturen analyseres nøjere

Vii skal nu nøje overveje strukturen af de indsamlede data, og udnytte dette til en mere præcis dataanalyse. Denne del kan evt. føres som en samlet diskussion med klassen, eller grupperne kan igen blive bedt om at overveje fx de følgende spørgsmål:

- Hvad er **stikprøven** og hvad er **populationen** i de data som Wald fik præsenteret? Hvorfor udgør de ikke data, for alle de kampfly, som Amerikanerne sendte mod tyskerne, og hvad er det særlige kendetegn ved de kampfly, der er med i stikprøven?
- Hvorfor er stikprøven **ikke** repræsentativ? Hvis den skulle have været repræsentativ, hvilke data skulle man så yderligere have indsamlet?
- Hvordan kunne en **nulhypotese** om skudhullernes fordeling se ud? Luftkampe mellem amerikanske kampfly og tyske kampfly er ret kaotiske og der er ikke tid til præcisionsindstilling af maskingeværerne! Hvilken fordeling følger skudhullerne ifølge nulhypotesen?
- Hvorfor afviger skudhullernes *observerede* fordeling i stikprøven fra den *forventede* fordeling?
- Hvad fortæller det om skudhullernes fordeling på de fly, der blev skudt ned af tyskerne?
- Hvad fortæller det om sårbarheden af de forskellige dele på et fly? Hvilken anbefaling ville du nu give det amerikanske militær, vedrørende armeringen af de amerikanske kampfly: Hvor på flyet skal de ofre særligt meget armering?

3. del: Videre arbejde med problemstillingen – et oplæg til srp

I de foregående afsnit har vi undersøgt Walds problemstilling med elementære metoder fra især den deskriptive statistik. Det gør projektet tilgængeligt allerede på C-niveau. Men hvis man arbejder med projektet på A-niveau og gerne vil forankre det i en sandsynlighedsteoretisk ramme, er der flere muligheder for at uddybe projektet.

Dels kan man læse en moderne statistisk analyse af Walds originale rapport: *Abraham Walds work on aircraft vulnerability*. [Link til 7](#)

Og i tilknytning hertil læse en artikel fra AMS-Feature Column: *The Legend of Abraham Wald*. [Link til 8](#)

Opgaven vil først og fremmest være at forklare den teori, der ligger bagved.

Endelig kan man kaste sig ud i den store udfordring, nemlig at læse Walds originale rapport: *A method of estimating plane vulnerability based upon damage of survivors*. [Link til 9](#)

Hele denne problemstilling, vi har analyseret ovenfor, er et eksempel på det, man kalder *survival-bias*. Det var ikke Wald, der indførte dette begreb, det var faktisk kendt før ham, men eksemplet her er i dag *klassikeren indenfor dette felt*.

Prøv at forklare med almindelige ord, hvad der menes med begrebet *survival-bias*.

3.4 Case: Soldyrkere lever længere

Artiklerne fra Politiken og fra tidsskriftet *International Journal of Epidemiology* kan hentes [via Link til 10.](#)

I oktober 2013 kunne man i Politiken læse nedenstående artikel. Den har et ret sensationelt indhold og blev da også bragt på forsiden.

Øvelse 3.20

Læs artiklen og præsenter informationerne heri med anvendelse af de statistiske begreber, du har lært:

- Kan du indkredse, hvad populationen er? Angiv fx nogle individer, der er med i populationen og nogle der ikke er med.
- Kan du lokalisere en stikprøve? Hvad er denne gruppe en stikprøve af?
- I hvilken forstand er den repræsentativ for populationen
- I artiklen tales om gennemsnitlig levealder. Hvem er det man tager gennemsnittet af, og hvordan udregnes et sådant?

1. ARTIKEL OM SOL OG HUDKRÆFT

15. OKT. 2013

Soldyrkere lever meget længere

Ny forskning blandt 4,4 millioner danskere viser, at soldyrkere i gennemsnit lever seks år længere. Kræftens Bekæmpelse finder tallene spændende.

Henrik Larsen

Et hold danske forskere er på vej med en videnskabelig artikel, som rejser spørgsmålet: Er der særlige livsforlængende 'sager' i solens stråler?

Artiklen står foran offentliggørelse i videnskabstidsskriftet *International Journal of Epidemiology* og viser, at mennesker, som har været ivrige soldyrkere – og har fået såkaldt almindelig hudkræft, den ikke-dødelige form for hudkræft – i gennemsnit lever seks år længere end befolkningen som helhed.

Overdreven solforskrækkelse

Gennemsnitsdanskeren – kvinder og mænd under ét – bliver i dag 80 år. Men når det gælder denne gruppe soldyrkere, kan vi altså se, at de i snit når at fejre 86-års fødselsdagen. Og at de i øvrigt har en lavere forekomst af både blodpropper i hjertet og knogleskørhed end resten af befolkningen, siger en af forskerne bag undersøgelsen, professor Børge Nordestgaard, Herlev Hospital.

Forskerne kan ikke påpege den direkte årsagssammenhæng mellem soldyrkning og fundene: *Men tallene som sådan lyver ikke. Det vil være oplagt gennem en række forskningsprojekter at prøve at blive klogere på, om solen har nogle helbedsbedørende egenskaber, vi hidtil ikke har kendt til, siger han og tilføjer: Man skal bestemt have respekt for solen – og undgå solskoldninger, ikke mindst hos børn og unge, for på den måde især at mindske forekomsten af den slemme form for hudcancer, malignt melanom. Men vi har de seneste år set en vis grad af solforskrækkelse, og det er efter min mening at gå i den anden grøft.*

Ingen egentlig årsagssammenhæng

Hos Kræftens Bekæmpelse – der står bag de såkaldte solråd, som blandt andet opfordrer danskerne til at omgås solen med større forsigtighed – finder Inge Haunstrup Clemmensen, overlæge i Kræftens

Bekæmpelse, undersøgelsen spændende: *Det vil være oplagt gennem en række forskningsprojekter at prøve at blive klogere på, om solen har nogle helbredsbedørende egenskaber, vi hidtil ikke har kendt til.*

Børge Nordestgaard, professor, Herlev Hospital fortæller om undersøgelsen: *Den rejser en mistanke om, at der kan være en sammenhæng mellem solens stråler og de positive helbredsfind, der påvises hos gruppen af mennesker, som har haft almindelig hudkræft. Men det er bare vigtigt at holde fast i, at den ikke viser en egentlig årsagssammenhæng.* Børge Nordestgaard har sammen med tre kolleger – læge Peter Brøndum-Jacobsen og seniorforsker Sune Nielsen, begge Herlev Hospital, og overlæge Marianne Benn, Gentofte Hospital – analyseret helbredsrelaterede data fra 4,4 millioner danskere i alderen 40-100 år.

Analysen dækker perioden 1980-2006, og forskerne havde adgang til en meget bred vifte af registre, hvorfra de i anonymiseret form kunne trække oplysninger om hver eneste: om blandt andet sygdomme, uddannelsesniveau, bopæl, erhverv, dødsårsag og dødstidspunkt.

Forskerne gik en form for bagvej

Selv om danske registre tager meget med om den enkelte, indeholder de ingen oplysninger om sol- dyrkningsvaner. Så her måtte forskerne gå en form for bagvej, fortæller Børge Nordestgaard: *Det er velkendt, at mennesker, der får almindelig hudkræft, som hovedregel har udviklet de celleforandringer, der er tale om, fordi de har opholdt sig ekstraordinært meget i solen. For at finde eventuelle helbredsbevarende sammenhænge, der muligvis kunne skyldes solen – vores egentlige mål – sammenlignede vi derfor 130.000 borgere, der havde haft almindelig hudkræft, med resten af befolkningen.*

Og de positive opdagelser, vi gjorde blandt disse 130.000 mennesker i form af længere gennemsnitslevetid og lavere forekomst af blodpropper i hjertet og af knogleskørhed, viste sig at være helt uafhængige af forhold som køn, alder, om man bor i byen eller på landet – ligesom uddannelse og erhverv heller ikke spiller nogen rolle. Derfor er det nærliggende at stille spørgsmålet: Kan dette i en eller anden udstrækning være knyttet til solens aktivitet, mener Børge Nordestgaard.

Hos Kræftens Bekæmpelse mener Inge Haunstrup Clemmensen dog også, at en mulig forklaring kan være, at mennesker, der har råd og tid til at være meget i solen, ofte vil dyrke udendørs sport eller motion – og at de på den måde styrker helbredet.

Artiklen blev dagen efter kommenteret af Kræftens Bekæmpelse, og gav i det hele taget anledning til en ganske ophedet debat.

Og så 10 dage efter tager sagen en dramatisk vending, da to statistikere fra afdelingen for Biostat på Københavns Universitet retter en sønderlemmende kritik mod den statistiske metode, og ikke nok med det: De sender deres indvendinger til samme tidsskrift, hvor artiklen netop er blevet offentliggjort.

Dansk forskning: Soldyrkere lever meget længere

Danske forskere vil offentliggøre ny undersøgelse der viser, at danskere der kan defineres som soldyrkere lever seks år længere.

Foruden den længere levetid viser undersøgelsen også, at soldyrkere har en lavere forekomst af både blodpropper i hjertet og knogleskørhed.

Stadig sol med omtanke

- Vi har de seneste år set en vis grad af solforskækkelse, og det er efter min mening at gå i den anden grøft, udtaler professor Børge Nordestgaard, der er en af forskerne bag den nye undersøgelse til Politiken, og han understreger i den forbindelse, at man stadig skal solbade med omtanke og undgå solskoldninger, ikke mindst hos børn og unge.

Kræftens Bekæmpelse ikke overbevist

Selvom den nye forskning omfatter 4,4 millioner danskere og i øvrigt

Ikke overraskende blandede Dansk Solarieforening sig med nogle stærke udfald mod Kræftens Bekæmpelse

Øvelse 3.21

Læs artiklen og præsenter informationerne heri med anvendelse af de statistiske begreber, du har lært:

- Hvad er problemet i forholdet mellem population og stikprøve, sådan som den oprindelige forskergruppe har behandlet det?
- Hvad går den centrale indvending mht. gennemsnitlig levealder for de to grupper ud på?

3. ARTIKEL OM SOL OG HUDKRÆFT

25. OKT. 2013

Solstrid brudt ud: Forskere strides om levealder

To danske forskergrupper strides om levealder for folk med hudkræft.
Henrik Larsen

Lever såkaldte soldyrkere, som har fået ikkedødelig hudkræft, i gennemsnit seks år længere end resten af den danske befolkning? Det spørgsmål bliver nu genstand for en strid mellem to danske forskergrupper.

Den ene gruppe, under ledelse af professor Børge Nordestgaard fra Herlev Hospital og Københavns Universitet, siger ja. Den anden gruppe, professor Niels Keiding og lektor Theis Lange, begge fra Afdeling for Biostatistik ved Københavns Universitet, siger nej.

Ikke muligt at bevise sammenhæng

Stridens kerne er en videnskabelig artikel, som Børge Nordestgaard sammen med tre danske kolleger har publiceret i det meget anerkendte videnskabstidsskrift *International Journal of Epidemiology*. En videnskabelig undersøgelse, som Politiken skrev om i sidste uge.

Den videnskabelige undersøgelse bygger på helbredsrelaterede oplysninger om 4,4 millioner danskere mellem 40-100 år i perioden 1980-2006. På basis af undersøgelsen konkluderer Børge Nordestgaard og hans kolleger blandt andet, at mennesker, der har fået ikkedødelig hudkræft – en lidelse, der som regel skyldes, at en person har været udsat for store mængder sollys – i gennemsnit lever seks år længere end resten af befolkningen. Og bliver 86 år, hvor gennemsnitslevetiden for resten af befolkningen er 80 år. I undersøgelsen gør de dog klart opmærksom på, at det ikke er muligt at påvise nogen årsagssammenhæng mellem sollys og levealder.

Personer har i forvejen høj alder

Metoden bag undersøgelsens afsnit om forskellen i levealder på seks år mellem de to grupper holder imidlertid ikke, mener Niels Keiding og Theis Lange, der har bedt *International Journal of Epidemiology* om at trykke deres argumenter for, at Nordestgaard-gruppen har brugt *en analysemetode, der er absolut forkert*, siger Theis Lange: *Problemet med deres analyse er, at for at komme ind i hudkræftgruppen skal man – naturligvis – have levet længe nok til at udvikle hudkræft – og de fleste, der diagnosticeres med ikkedødelig hudkræft, er over 50 år. For gruppen af personer uden hudkræft forholder det sig lige omvendt, her kan personer af alle aldre indgå. Konsekvensen bliver, at personer i gruppen med hudkræft dør i en højere alder end den anden gruppe, men dette har absolut intet at gøre med deres diagnose. Det er alene fordi gruppen med hudkræft er udvalgt på en måde, så den hovedsageligt indeholder personer, der i forvejen har en høj alder*, siger Theis Lange.

Forskere er klar til at svare på kritikken

Børge Nordestgaard siger, at han og de tre medforfattere ser frem til at modtage Keiding og Langes

henvendelse fra International Journal of Epidemiology: *Når det sker, vil vi svare detaljeret på deres kommentarer og kritik – sådan er proceduren. Vi er glade for den store interesse og medieomtale, undersøgelsen har affødt – det viser, at dette emne er vigtigt. Og vi er glade for de kommentarer, vi har modtaget, både de konstruktive og de kritiske, som peger på, hvordan en sådan undersøgelse kan laves endnu bedre, siger professor Børge Nordestgaard, og tilføjer: Man må håbe, at andre forskere gennem nye undersøgelser vil vise, om solskin kan have andre positive effekter på livslængde og helbred.*

International Journal of Epidemiology oplyser, at man har modtaget og nu vil studere henvendelsen fra Niels Keiding og Theis Lange.

Øvelse 3.22

Antag vi har et land, hvor den demografiske struktur er meget stabil over tid, og fordeler sig således, hvis vi fokuserer på levealder:

Andel af befolkningen	Gennemsnitlig levealder
15%	under 45 år
10%	45-55
20%	55-65
30%	65-75
20%	75-85
5%	85-95

For nyfødte børn gælder altså, at 20% af dem bliver mellem 75 og 85, mens 15% vil dø inden de bliver 45.

- Hvad skal vi forstå ved gennemsnitlig levealder? Udregn denne!
- Hvad er den gennemsnitlige levealder for dem, der er blevet 55? Og for dem der er blevet 75?

Øvelse 3.23

Antag du i morgenavisen læser følgende overskrift på en artikel: *Paver og biskopper lever længere end almindelige præster.* Hvad vil din umiddelbare kommentar være?

(Hint: Hvornår i livet bliver man normalt udnævnt som biskop?)

Sagen fik også en dramatisk afslutning, idet tidsskriftet gav statistikerne fra Biostat ret. Det skete ikke i form af en lille notits, men via hele to artikler, nemlig én artikel, der var indsendt uden viden om den danske kontrovers, og som var en generel analyse af den type fejl, den oprindelige undersøgelse var behæftet med, og hvor redaktøren derfor overtalte forfatterne til at inkludere det danske eksempel som et typisk eksempel på fejlslutninger. Samt en artikel skrevet af redaktøren selv, og som du ser første side af her.

Øvelse 3.24

Læs redaktørens artikel og giv et sammendrag af hans vurdering af den videnskabelige kvalitet af den oprindelige artikel.

Studieretningsprojekt eller anden for projektarbejde

Hele forløbet, som det er beskrevet ovenfor, inklusiv de forskellige øvelser, samt materialet i mappen du kan tilgå via hjemmesiden, hvor du blandt finder de originale tidsskriftartikler, kan danne grundlag for forskellige typer af studieretningsprojekter. Man kan koncentrere sig om substansen vedr. soldyrkning og kræft, og hvorledes man med statistiske metoder kan svare på opstillede spørgsmål. Man kan også vælge at lægge vægten på de videnskabsteoretiske sider, der blev afdækket i diskussionen.



Editor's choice

Sun exposure and longevity: a blunder involving immortal time

Jane E Ferrie* and Shah Ebrahim

*Corresponding author. E-mail: jane.ferrie@bristol.ac.uk

Unfortunately we have to start this Editor's Choice with an acknowledgment that we have fallen prey to a common, perennial problem; immortal time bias.

To illustrate the concept we borrow an example from William Farr, as used by James Hanley and Bethany Foster in a full and entertaining exposition of the problem in this issue of the journal.¹ Generals and bishops live longer than corporals and curates—but this is not necessarily because an elevated occupational status makes you live longer—it may simply be because you have to reach a certain age before it is possible to hold such positions. People become generals and bishops in middle age so their deaths arise after this point in time, whereas corporals and curates can die at any age above 20 or so.² This difference in time during which an event can occur to one group but not the other produces a bias favouring longer life expectancy—immortal time bias. In the figure on the next page, the problem is evident at a glance (Figure 1).³

In the October issue of the *International Journal of Epidemiology (IJE)* last year, we published a paper by Peter Brøndum-Jacobsen and colleagues in which they examined the effects of sunlight exposure on mortality among the whole population of Denmark aged above 40 years, using linked data from national registries.⁴ They used non-melanoma skin cancer as a proxy for sun exposure, which is a clever idea but it should have been obvious that the findings were 'too good to be true'—an apparent halving of all-cause mortality and reductions in myocardial infarction and hip fracture. The authors concluded: 'Causal conclusions cannot be made from our data. A beneficial effect of sun exposure per se needs to be examined in other studies'.

The Danish media picked up the story and it became front page news—'Sunbathers live longer'.⁵ Although the authors never made this claim in their published paper,

their interviews with the press did not appear to emphasize their non-causal conclusion. The Danish Cancer Association claims that this paper has undone all their good work in persuading Danes to keep out of the sun to avoid skin cancers.

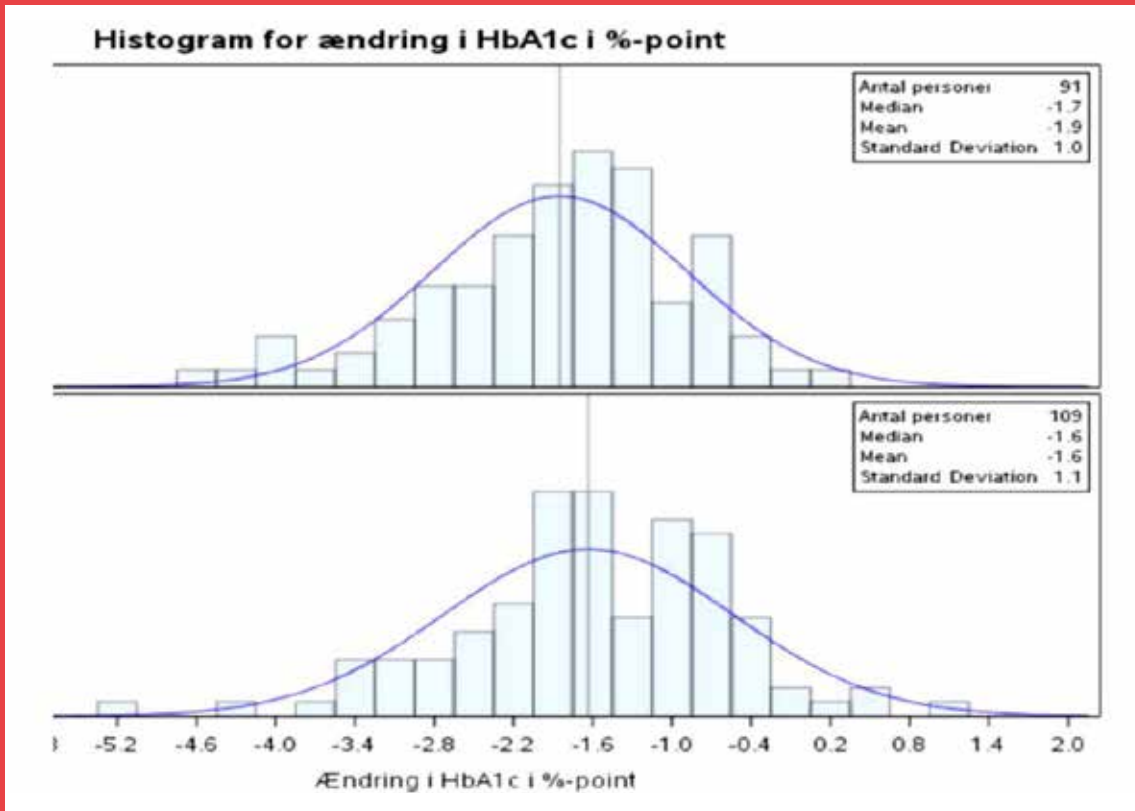
Commentators on the story identified a likely problem of immortal time bias. People in the 'sun exposure' group had to live long enough to be diagnosed with skin cancer but the comparison group only had to be over 40 years old—the design of the study had built in a potential bias in favour of longevity among those presumed to be more highly exposed to sunlight. Theis Lange and Neils Keiding, in a letter commenting on the paper, pose questions about how such highly improbable findings got through the editorial process at *IJE*.⁶

In response to this criticism, Brøndum-Jacobsen and colleagues argue that their paper used both cohort and case-control analyses, and that the latter should be free from immortal time bias as cases and controls were matched on age.⁷ They acknowledge that the case-control analyses—which showed much smaller survival advantage [odds ratio (OR): 0.97, 95% confidence interval (CI) 0.96 to 0.99; vs hazard ratio (HR): 0.52, 95% CI 0.52 to 0.53]—should have been included in their abstract. In addition, they conducted a revised Cox proportional hazards analysis stratified by 10-year, 5-year and 2-year age strata in an attempt to control for immortal time bias, and interpret these findings as similar to those in their original paper. However, they fail to stress that the effect sizes become increasingly attenuated as the age matching becomes more exact, suggesting that the apparent effect of sun exposure may indeed be produced by immortal time bias.

Ironically, in parallel with the review and publication of this paper we had commissioned an 'Education Corner'

4.

Normalfordelingen



Den grafiske analyse af datasættet slutter med tegning af et histogram over ændringerne for henholdsvis A og B. Sammen med diagrammet har Lea tegnet en graf, som hun forklarer, er en overlejet normalfordelingskurve, der bl.a. skal give statistikerne et indtryk af, om det er rimeligt at bruge de værktøjer i en hypotesetest, som der lægges op til. Men hvad er normalfordelingen egentlig. Det går vi i dybden med i dette kapitel

4. Normalfordelingen

4.1 Hvor kommer normalfordelingen fra?

Når man foretager målinger af vægt, tid, længde eller andre størrelser i et naturvidenskabeligt fag, vil det aldrig blive en 100% præcis måling, uanset udstyr. Skal vi eksempelvis opløse 5 g af et stof i 1 dl af en bestemt syre, og dernæst måle opløsningens pH-værdi, så findes der ikke en vægt eller et rumfangsmål, der er uendelig nøjagtig. Dertil kommer vores egen unøjagtighed. Og aflæsningen af pH-værdien kan også kun være tilnærmet. Men selvfølgelig har den resulterende opløsning en ganske bestemt pH-værdi. Tilsvarende: Vi ved, at længden af diagonalen i et kvadrat med sidelængde 1 er $\sqrt{2}$, men det tal vil man i sagens natur aldrig kunne måle sig frem til helt præcist. Det er jo et uendeligt decimaltal. Det samme gælder, hvis man i fysik ønsker at bestemme tyngdeaccelerationen g ud fra en række forsøg med pendulsvingninger.

Men vi er jo ikke i tvivl om, at der i hvert tilfælde findes en "sand værdi", som ens målinger er en tilnærmelse til. $\sqrt{2}$ og størrelsen g findes! Så det er simpelthen en betingelse vi må leve med, at der er fejl i vores empiriske målinger, i forhold til den sande værdi. Derfor foretager man normalt flere målinger. Det gør man ved forsøg i klassen, og det gør forskere, hvis det er muligt umiddelbart at gentage en måling.

Man kan spørge: Hvor vigtigt er det at få meget præcise målinger? Og det er rigtigt, at til mange praktiske forhold er tilnærmede værdier med måske 1 decimal gode nok. Behovet for at få styr på sine målefejl opstod heller ikke først i praktiske sammenhænge, men indenfor astronomien. Går vi nogle hundrede år tilbage – og endnu længere – så var astronomien stort set *ren videnskab*. Formålet var erkendelse: At forstå universet, solsystemet og planeternes gang, vores plads i det osv. Der var også en spin-off i form af støtte til navigation, men det centrale var, både for videnskab og religion, at forstå himmelrummet. Og hertil var præcise målinger afgørende.

Tycho Brahe formåede omkring år 1600 med det blotte øje – kikkerten blev først opfundet omkring 1608 – at bestemme så nøjagtige koordinater for planeternes bevægelser, at det satte Kepler i stand til at fastslå, at planetbanerne ikke kunne være cirkler. Observationerne af Mars bane passede med stor nøjagtighed til en ellipse. Men der var trods alt fejl, og vi ved, at Kepler og andre efter ham udviklede egne metoder til at udligne disse. Men hvad metoderne gik ud på, ved vi faktisk ikke.

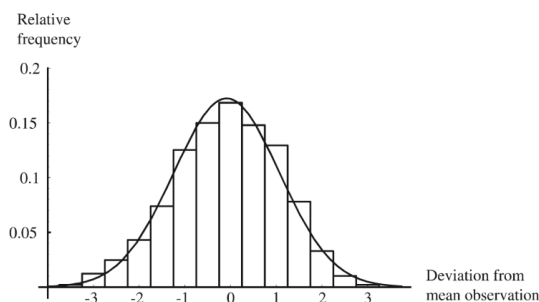
Det var den tyske matematiker Gauss (1777-1855), der først i 1800-tallet udviklede det matematiske værktøj til håndtering af tilfældige fejl. Dengang kaldte man værktøjet for *fejlfunktionen* (og fejlgraf). Senere fik det navnet normalfordelingsfunktionen og -graf. Men stadig anvendes især i engelsksproget litteratur begrebet *error function*. Normalfordelingskurven var blot et af flere store bidrag til statistikkens udvikling fra Gauss i denne periode. Allerede som 24 årig var han blevet verdensberømt, da han demonstrerede en spektakulær anvendelse af sin netop udviklede "*mindste kvadraters metode*".

1. januar 1801 havde en italiensk amatør-astronom opdaget en ny planet, men efter få ugers observationer forsvandt den ind bag solen. Gauss fik de få data, og beregnede ud fra disse den bedste elliptiske bane som himmellegemet vil følge. Han skrev til "himmelpolitiet", hvor de skulle lede – og de fandt den! Det var ikke en planet, men den store asteroide Ceres.

Her var observationerne rimeligt præcise, men i andre situationer søger man som beskrevet at kompensere, ved at foretage flere målinger. Hvordan håndterer man et datasæt med flere målinger? Tag gennemsnittet, vil vi nok sige i dag. Men hvorfor ikke medianen? Kan vi regne med, at alle målinger er af samme kvalitet? Og kan vi håndtere usikkerheden på gennemsnittet? Der var mange diskussioner om dette på Gauss tid. Efter succesen med Ceres vælger han blandt andet at studere de data, der lå til grund for Keplers beregning af Mars-banen. Her forelå en række observationer, foretaget

gennem mange år, af den samme position for Mars. Men hvordan Kepler lige valgte sine data ud ved vi som sagt ikke - nogle mener han simpelthen prøvede sig frem, da han fik ideen om ellipsen.

Men er det mest rimelige ikke at tage hensyn til alle observationerne? Kan vi tillade os at smide nogle væk? Gauss skalerer observationerne ned til overskuelige tal, udregner middeltallet og udregner så, hvordan de fordeler sig omkring middeltallet. Han regner i højere grad end han tegner, det var åbenbart naturligt for ham, der kunne uddrage kvadratrødder i hovedet med 8-10 decimaler. Så han tegner ikke selv, men i en artikel om Gauss opdagelse af normalfordelingen har forfatteren tegnet eksempler på, hvordan observationerne fordeler sig omkring deres middeltal. Det er rimeligt overbevisende, som forfatteren skriver.



Når man har set den slags kurver en del gange, synes man nok, at man kan genkende den "klokkeformede" graf. Som Lea fortæller i filmen, er det en *normalfordelingskurve*, der overlejrer histogrammet, og hvis konkrete forløb er beregnet ud fra data i regnearket. Men der er jo mange klokkeformede kurver, så hvor kommer lige præcis denne fra? Gauss udleder forskriften for denne kurve ud fra nogle ganske få kriterier, der i almindelig prosa lyder:

- små fejl er mere almindelige end store fejl
- for ethvert positivt tal ε er sandsynligheden for en fejl af størrelsesorden ε (eller større) lig med sandsynligheden for en fejl af størrelsesorden $-\varepsilon$ (eller mindre)
- Har vi mange målinger af det samme, så er den mest sandsynlige værdi, denne størrelse har, lig med middeltallet af målingerne.

Disse tre kriterier oversætter Gauss til matematik. Og ud fra disse udleder han nu, at funktionsudtrykket for en fejlfunktion, der opfylder kriterierne må være:

$$\varphi(x) = \frac{h}{\sqrt{\pi}} \cdot e^{-h^2 \cdot x^2}$$

Den "klokkeformede" kurve er altså ikke bare klokkeformet, men er graf for en funktion af præcis denne type. Det kaldes for **normalfordelingens tæthedsfunktion**. Udledningen er demonstreret i en historisk artikel om Gauss, som er kilde til ovenstående, og som du kan hente her – [link til 11](#)

Øvelse 4.1

Gauss tre punkter giver også en vejledning til at vurdere, om vi har et gøre med en normalfordeling, når vi har en samling af data. Giv et bud på, om følgende (med tilnærmelse) vil kunne beskrives ved en normalfordeling, og argumenter hver gang for det eller imod det:

1. I en matematiktime skal eleverne bestemme højden af en bygning ved hjælp af en vinkelmåler og et længdemål. Skitser situationen. Hvert hold foretager flere målinger, og til sidst puljes alle målinger. Er resultaterne normalfordelt?
2. Ringer man op til en telefonlinje, hvor der er stort pres på, får man beskeden: Du er nu nr xxx i køen. Er ventetiden på at komme igennem normalfordelt?
3. Kaster man en almindelig terning kan man få tallene 1 .. 6. Hvad er det gennemsnitlige resultat? Vi kaster mange gange og tæller op, hvor mange vi får af hver. Giver dette en normalfordeling?
4. Vi planter 200 tulipanløg ud i en nedlagt græsplæne. Der er en spiringsgaranti på 85%. Hvor mange kan vi forvente vil spire. Er antallet af tulipaner der spirer normalfordelt? Hvordan var situationen, hvis spiringsgarantien var 50%?

Øvelse 4.2. Gauss udledning af normalfordelingen som funktion

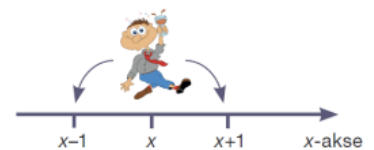
Orienter dig i den historiske artikel, specielt siderne 9-12. Overvej om det ville kunne give dig et grundlag for en SRP.

4.2 Normalfordelingen og binomialfordelingen

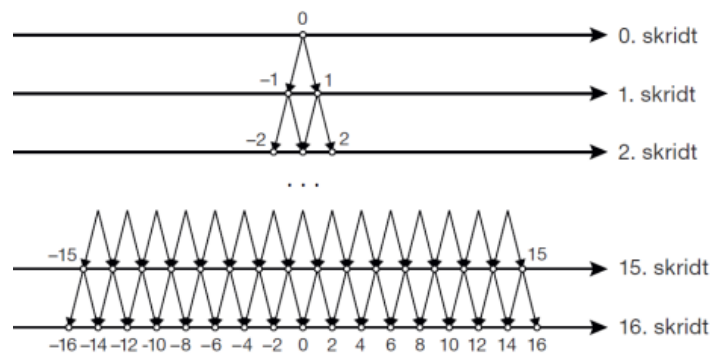
Har man allerede lært om normalfordelingen, har man givetvis berørt slægtskabet med binomialfordelingen. Binomialfordelingen bliver ofte introduceret ud fra kast med terninger eller med mønter: Kaster man 1000 gange med en mønt, og udregner man sandsynlighederne for at få følgende antal af krone: 0, 1, 2, ..., 1000 så vil det tilhørende pindediagram følge en normalfordeling. Men hvorfor egentlig, når vi tænker over Gauss beskrivelse?

Det kan man indse, ved at tage udgangspunkt i en såkaldt *random walk*. Det følgende er gennemført i stor detalje i HEM3, hvor normalfordelingen introduceres som grænseværdi for en random walk med uendeligt mange skridt. Du kan finde en tilsvarende grundig gennemgang her - [link til 12](#).

Hvad er en "random walk"? En person starter i 0 på tallinjen og bevæger sig så tilfældigt et skridt frem eller tilbage (af og til beskrevet som en fuld mands slingren frem og tilbage). Det tilfældige beskrives matematisk med et møntkast: 50% chance for frem, 50% for tilbage. Den proces kan visualiseres med et såkaldt *Galtonbræt*:



På et bræt er der sat søm i, så de danner horisontale rækker med samme mellemrum mellem sømmene som mellem linjerne. Kugler triller ned ad brættet én ad gangen, hver kugle rammer først det øverste midterste søm lige midt på. Halvdelen ryger til den ene side, den anden halvdel til den anden – og her falder de ned i et hul og rammer et søm som før, osv. Efter at have passeret n rækker lander en kugle et sted på tallinjen, og vi tæller op, hvor mange der lander hvert sted. Det hele er søgt illustreret her: [The Galton Board — A Desktop Probability Machine — Tools and Toys](#)



Øvelse 4.3 Random walk og binomialfordelingen

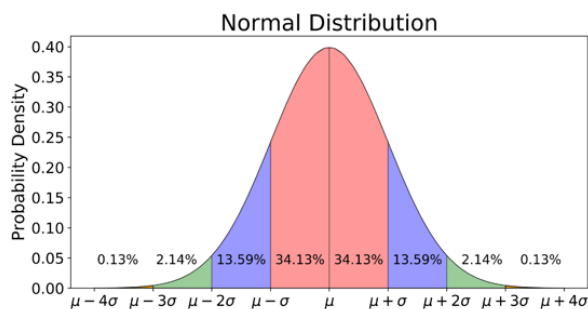
- Argumenter for, at den beskrevne proces må resultere i en binomialfordeling, $b(n, 0.5)$.
- Diskuter, om det du ser på filmstrimlen, svarer til beskrivelsen.

Øvelse 4.4 Random walk og Gauss 3 betingelser

Lad os betegne *resultatet* af en kugles løb ned ad brættet som en *måling*. Når vi lader N kugler løbe ned, svarer det til at foretage N målinger. De fleste kugler lander relativt tæt ved 0, men nogle lander længere væk. Hvis resultatet for en måling er tallet ϵ , kan vi betragte dette tal som en *fejl* ift tallet 0. Argumenter for, at med disse definitioner lever resultatet af en random walk (med mange skridt) op til Gauss tre betingelser

Har vi en random walk med n skridt, kan vi beregne middelværdi og spredning med de kendte formler. Når n går mod uendelig og random walk-processen nærmer sig en normalfordeling, så vil også

middelværdi og spredning nærme sig to tal, der er henh. middelværdi μ og spredning σ for normalfordelingen. σ kaldes også for *standardafvigelsen*. I HEM3 er der argumenteret for, at dette betyder, at vi får denne opdeling af arealet under grafen for normalfordelingen:



Der gælder altså:

Sætning om normalfordelte stokastiske variable

- 1) ca. 68% af observationerne forventes at ligge inden for én standardafvigelses afstand fra middelværdien.
- 2) ca. 95% af observationerne forventes at ligge inden for to standardafvigelser (dvs. de er normale).
- 3) ca. 0,25% af observationerne forventes at ligge uden for tre standardafvigelser (dvs. de er exceptionelle).

4.3 Normalfordelingens anvendelse breder sig til andre fag

Gennem 1800-tallet var der bestræbelser indenfor samfundsfaglige, sundhedsfaglige og biologiske fagområder på at højne det videnskabelige niveau og dermed opnå større anerkendelse. En af metoderne hertil var anvendelsen af *matematiske* begreber, symboler og metoder, ikke altid med lige stor indsigt, men hensigten var god nok. Centralt i dette stod netop normalfordelingen, og hvis man i dag søger på nettet efter eksempler på brug af normalfordelingen, er stort set alle opslag fra disse fag.

Det betyder også, at alle i et land som Danmark møder normalfordelingen mange gange i sit liv, startende med det første besøg af en fra sundhedsplejen, der måler om spædbarnets vægt og højde ligger indenfor normalområdet på den normalfordelingskurve, de har som værktøj.

Men hvorfra ved man egentlig, at sådanne data vil følge en normalfordelingskurve? Opfylder de fx Gauss' tre kriterier? Eller kan de realiseres i en random walk model? Lad os se på et af de historisk set tidligste eksempler. I den omtalte artikel fortælles om den belgiske samfundsforsker Quetelet og hans bestræbelser på at inddrage naturvidenskabelige metoder i sociologiske og biologiske fag.

Øvelse 4.5. Quetelets brug af Gauss metoder på biologiske data.

Hent artiklen [- link til 11](#) - og orienter dig i siderne 13 - 15 (tidsskriftnummereringen: 108-112).

- a) Beskriv det eksempel, vi præsenteres for.
- b) Hvordan argumenterer Quetelet for, at man kan analysere data med Gauss' metode?
- c) Bag Gauss fejlkurver er der altid en "sand værdi". Hvad er Quetelets bud på den sande værdi af soldaternes brystmåle?
- d) Quetelet indfører et begreb, han kalder "det gennemsnitlige menneske". Diskuter, hvad man skal forstå ved begrebet.
- e) Hvordan håndterer Quetelet data, og hvordan når han frem til at de er normalfordelte? (Du skal ikke selv regne på data, men blot referere, hvad artiklens forfatter er nået frem til.

Quetelets bog, hvor disse data er gengivet, får enorm indflydelse. Han var selv oprindelig astronom, og kendte normalfordelingen derfra. Det gav ham givetvis en vis autoritet, selv om det er svært at forstå, at man - som han gør - kan påstå, at data kan opfattes som 5.738 forskellige målinger af samme persons brystvidde. Der er et spænd på over 12 tommer! Det må have været tydeligt misvisende med disse data at fortsætte med at bruge betegnelsen "fejlkurve", og i løbet af få år blev det

da også erstattet af ordet og begrebet *normalfordelingen*. En af de mest berømte statistikere omkring år 1900 Karl Pearson fortrød, at han selv havde bidraget til at udbrede denne betegnelse, fordi det indikerede, at andre fordelinger var *unormale*. Men det var umuligt at ændre.

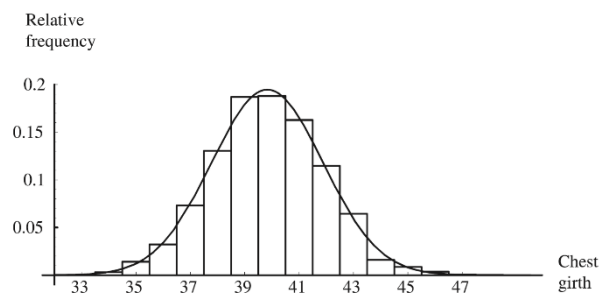


Figure 8 Is this data normally distributed?

Og det var faktisk heller ikke helt forkert. Quetelets grundlæggende opfattelse, at normalfordelingen er velegnet til at beskrive mange fænomener indenfor sociologi, sundhed og sociale forhold rummer en vis sandhed. Hans argumenter og beregninger var, som vi har set, ikke i orden, men hans intuition mht. brugen af normalfordelingen var. Anvendelsen retfærdiggøres af **”den centrale grænseværdisætning”**, der kort fortalt siger, at har man en række stokastiske variable, der fx måler brystvidden af skotske soldater, og er disse nogenlunde ens fordelt – dvs. der er ikke stor forskel i de genetiske, sociale, sundhedsmæssige osv. variable, der bestemmer den enkeltes brystvidde – så vil gennemsnittet af de stokastiske variable nærme sig en normalfordeling, når vi tager flere og flere med. Det var først 50 år efter Quetelets bog, at man beviste de første versioner af denne helt fundamentale sætning i statistik og sandsynlighedsregning. Vi vil ikke gå længere ind i det her.

Et af de næste store skridt mht. videnskabeliggørelsen og matematiseringen af de biologiske og samfundsvidenskabelige fag blev taget af den danske genetiker Wilhelm Johansen omkring år 1900. Danmark havde i første halvdel af 1900-tallet en så stærk position indenfor medicinsk og biologisk forskning, at vi hørte til verdens bedste. Det illustreres af antallet af Nobelpristagere, hvor Danmark i perioden havde 4, kun overgået af England, Tyskland og USA. Dertil kom flere ‘Nobelpristabere’, der burde have haft prisen, bl.a Wilhelm Johannsen (1857-1927).

De helt store spørgsmål og emner for biologisk forskning i årene omkring 1900 var: Hvordan nedarves egenskaber? Og: Hvoraf kommer artsrigdommen? Wilhelm Johannsen giver et stort bidrag til at løse disse spørgsmål. Det er ham, der indfører begrebet *gen*, og ikke mindst begrebsapparatet *genotype* og *fænotype* giver et nyt værktøj til at forstå forholdet mellem arv og miljø.

Projekt: Wilhelm Johannsens eksakte arvelighedslære

Du kan [her](#) – link til 13 - hente et kapitel fra HEM3, hvor Wilhelm Johannsens bidrag er gennemgået i større detalje, og hvor du kan finde de data, som han arbejdede med. Han anvendte Galtonbrættet til at demonstrere, at data fordelte sig efter en normalfordelingskurve, og gjorde i det hele taget et stort arbejde for at videnskabeliggøre *den eksperimentelle arvelighedslære*. Husk på, det er lang tid før opdagelsen af DNA. Der var heller ikke stor viden om normalfordelingen, eller om statistik i det hele taget blandt hans kolleger, så han brugte den del plads i sine bøger til at give en indføring heri.



4.4 Den generelle normalfordelings funktionsudtryk

Gauss udledte funktionsudtrykket for sin fejlkurve ud fra et datasæt, der grupperede sig omkring en middelværdi på 0. I hans funktionsudtryk:

$$\varphi(x) = \frac{h}{\sqrt{\pi}} \cdot e^{-h^2 \cdot x^2}$$

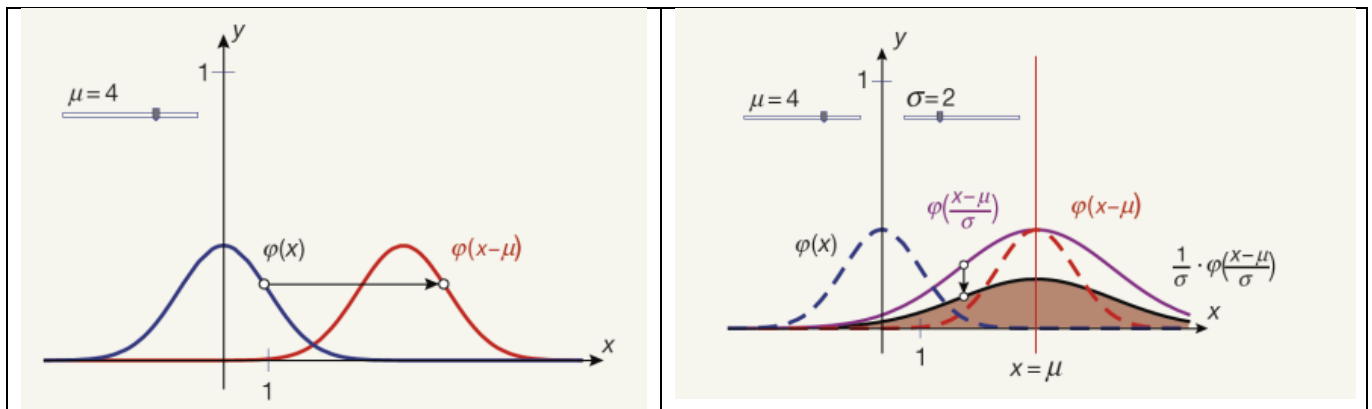
er *h* et udtryk for *præcisionen* af vores målinger. Da anvendelsen af normalfordelingen bredte sig til mange andre fag, forlod man efterhånden betegnelsen *fejlkurve* og det gav heller ikke så megen

mening at tale om *præcision* af fx data for spædbørns højde og vægt. I stedet skiftede fokus til *spredning* af data: En stor præcision svarer til en lille spredning på data, og omvendt: En stor spredning på måleresultater svarer til en beskeden eller direkte ringe præcision.

Det funktionsudtryk for normalfordelingens tæthedsfunktion, vi i dag anvender indeholder derfor spredningen σ i stedet for h . Som omtalt i afsnit 4.2 kan normalfordelingen fremkomme som grænseværdi for en ideel random walk, og betragter vi en sådan med middelværdi 0 og spredning 1, så bliver funktionsudtrykket:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$$

En normalfordeling med middelværdi μ fremkommer ved at parallelforskyde grafen, som du ser på illustrationen til venstre. En normalfordeling med en spredning på σ fremkommer ved skalering af grafen i y-aksens retning, hvor vi dividerer med σ . Da det er sandsynlighedsfordelinger, skal arealet under grafen være 1, og derfor må vi gange funktionsudtrykket med $\frac{1}{\sigma}$.



Øvelse 4.6 Normalfordelingens funktion undersøgt med brug af et værktøjsprogram

(Du kan oprette egne ark eller anvende de geogebra-ark, du finder her - [Link til 14](#).)

a) Vis at funktionsudtrykket for en normalfordeling med middelværdi μ og spredning σ bliver:

$$\varphi(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

b) Sammenlign dette udtryk med Gauss' funktionsudtryk, hvor middelværdien er 0, og vis, at sammenhængen mellem h og spredningen er: $\sigma = \frac{1}{\sqrt{2} \cdot h}$.

c) Undersøg i dit værktøjsprogram med brug af "skydere", hvorledes det grafiske forløb for $\varphi(x)$ afhænger af parametrene σ og μ . Giv en sproglig formulering af denne sammenhæng.

d) Med et vilkårligt valg af de to parametre ser kurven ud til at være symmetrisk. Opstil en ligning, der udtrykker dette! Indsæt funktionsudtrykket i ligningen og bevis symmetrien, dvs. vis, at din ligning er sand.

e) Da der er tale om grafer for sandsynlighedsfordelinger, skal arealet under grafen være 1. Vis med anvendelse af numerisk integration, at: $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} dx = 1$

f) Vis, ved at anvende integration ved substitution, samt resultatet i e), at

$$\int_{-\infty}^{\infty} \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1$$

Bemærkning: Integralet i e) kan faktisk udregnes. Det er gennemført i HEM3, projekt 8.12, der kan hentes fra website, eller direkte [her](#) – [link til 15](#).

Øvelse 4.7. Normalfordelingsfunktionen undersøgt med brug af differentialregning

(Du kan oprette egne ark eller anvende de geogebra-ark, du finder her – [link til 16](#))

- Bestem $\varphi'(x)$
- Løs ligningen $\varphi'(x) = 0$
- Bestem monotoniforholdene for f . Stemmer dette med det grafiske billede?
- Bestem $\varphi''(x)$
- Løs ligningen $\varphi''(x) = 0$
- Bestem monotoniforholdene for $\varphi'(x)$, og krumningsforholdene for $\varphi(x)$. Stemmer dette med det grafiske billede?
- Beskriv med dine egne ord, hvad $\varphi''(x)$ fortæller om grafen for $\varphi(x)$?
(Hint: Tænk på vendepunkter fra logistisk vækst og tredjegradspolynomier).

Funktionen $\varphi(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, hvis graf er den "klokkeformede" kurve, kaldes *tæthedsfunktionen* hørende til normalfordelingen med den angivne middelværdi og spredning. Grafen for $\varphi(x)$ "svare" til *histogrammet* over dataværdierne. Hvad svarer *sumkurven* til?

Øvelse 4.8. Tæthedsfunktionen og fordelingsfunktionen

(Du kan oprette egne ark eller anvende det geogebra-ark, du finder her – [link til 17](#))

Antag vi har et bestemt datasæt.

a) Slå op i din grundbog eller i formelsamlingen og beskriv, hvad en sumkurve over det givne datasæt er.

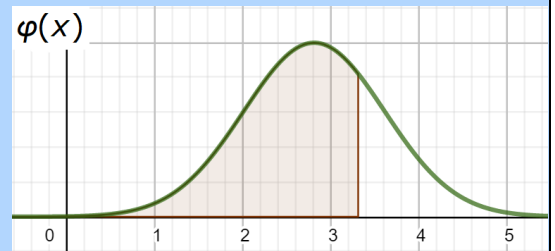
b) Tæthedsfunktionens graf kan opfattes som *grænseværdien af histogrammet*, hvor *intervalbredden går mod 0*. Argumenter for, at *sumkurvens værdi* i et bestemt x kan opfattes som arealet under grafen for tæthedsfunktionen, i området fra nedre grænse $-\infty$ til øvre grænse x .

c) Opret denne funktion $\Phi(x) := \int_{-\infty}^x \varphi(x) dx$ i dit værktøjs-

program og tegn grafen af den. Funktionen kaldes for normalfordelingens *fordelingsfunktion*

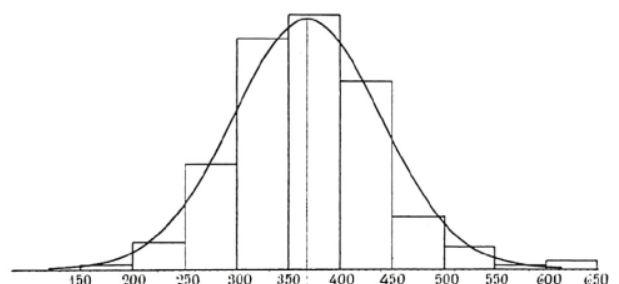
d) Beskriv grafen, når du varierer μ og vælger $\sigma = 1$. Tegn grafen for $\Phi(x)$ og for $\varphi(x)$ i samme billede, og argumenter for monotoniforholdene for $\Phi(x)$ ud fra definitionen (uden at regne!)

e) Beskriv grafen, når du varierer σ og vælger $\mu = 0$.



4.5 Overlejrrede normalfordelinger

I Quetelets graf over de skotske soldater brystvidde, og i Wilhelm Johannsen graf over vægten af de fynske prinsessebønner, som er gengivet her ud fra hans bog, er histogrammerne *overlejret af normalfordelinger*. Det samme så vi i filmen, hvor Lea forklarer, at grafen er tegnet på basis af en udregning af de to parametre, der fastlægger en normalfordeling: middelværdien og spredningen.

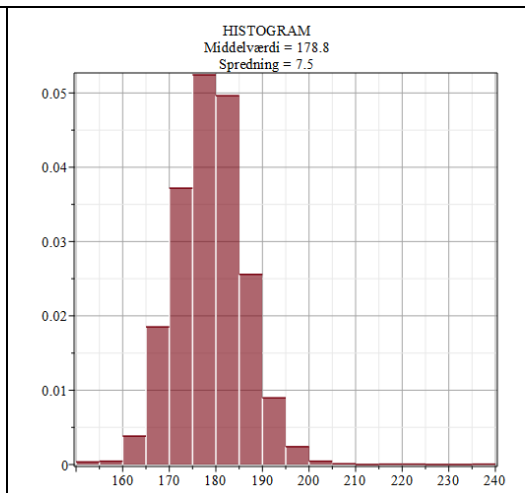


Quetelets og Johannsens grafer er selvfølgelig ikke egentlige beviser på, at data er normalfordelt. Som vi skrev, så er Quetelets data faktisk *ikke* normalfordelt. Men graferne kan give en indikation på, om der er grund til at gå videre i undersøgelsen.

Lad os i dette afsnit illustrere det, vi har lært, med et større datasæt over henholdsvis 4082 mænd og 1986 kvinder, alle med tilknytning til den amerikanske hær. Datasættet blev offentliggjort i 2017. En detaljeret beskrivelse af de 93 variable, der inkluderer både numeriske og kategoriale variable, kan findes [her](#) – [link til 18](#). Den officielle website findes her:

[Anthropometric Database - Army Public Health Center](#)

En af de mange numeriske variable i datasættet er højden. Hvis vi importerer regnearket for mændene, så kan et histogram over fordelingen af mændenes højde se sådan ud, *efter at vi har omregnet inches til cm*:



Histogrammet kan beskrives ved, at mange af mændenes højde ligger omkring middelværdien på 178,8 cm, og højderne til hver side af middelværdien fordeles sig symmetrisk. Denne beskrivelse af fordelingen af data er et eksperimentelt syn på normalfordelingen. I næste afsnit demonstrerer vi, hvordan man mere systematisk kan undersøge en mulig normalfordeling af data. Hent de to Excelfiler [her](#) – [link til 19...](#) og [link til 20](#).

Øvelse 4.9

a) Orienter dig i Excel-filen med data for mændene. Højden er en af de sidste kolonner. Opret en kolonne, hvor du omregner fra inches til cm: 1 inch = 2,54 cm.
 b) Importer data i dit værktøjsprogram. Vi arbejder her videre i Maple. I filen, der kan hentes [her](#) – [link til 21](#) – kan du se kommandoerne.

c) Grupper data ud fra en selvvalgt gruppering. Vi har valgt intervalbredde på 5.
 d) Tegn et histogram over mændenes højde. Histogrammet giver os de empiriske værdier af middelværdien $\mu = 178,8$ og spredningen $\sigma = 7,5$. Dermed har vi parametrene til forskriften for den tæthedsfunktion f , hvis graf kan approksimere histogrammets form. Dette bud er

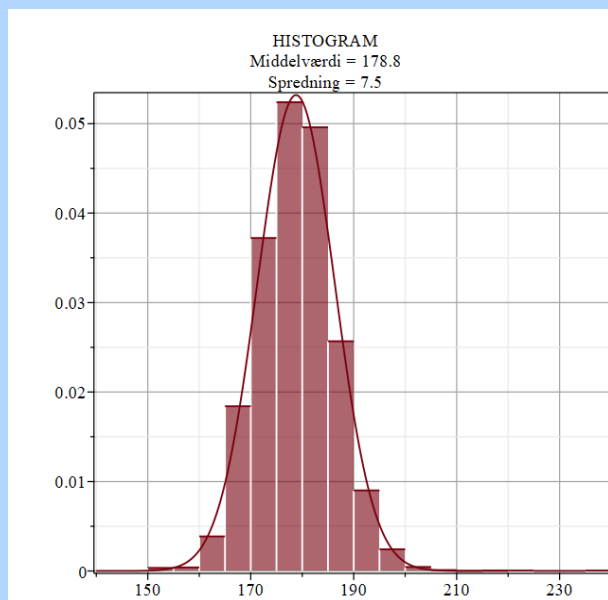
$$f(x) = \frac{1}{\sqrt{2\pi} \cdot 7,5} \cdot e^{-\frac{1}{2} \left(\frac{x-178,8}{7,5} \right)^2}.$$

Vi kalder denne tæthedsfunktion for *den til datasættet overvejrede normalfordeling*.

e) Tegn histogrammet og grafen for tæthedsfunktionen i samme billede. Du skal få et grafisk billede, som det du ser her.

f) Gennemfør selv punkterne a) til e) med data for kvindernes højde.

g) Vælg selv en af de andre variable, og gennemfør en tilsvarende undersøgelse.



Bemærk: Når histogrammet er konstrueret over frekvenserne, er det samlede areal naturligvis 1 eller 100%. Hvis det i stedet var absolutte tal, hvor det samlede areal var A , så måtte vi skalere tæthedsfunktionen op med faktoren A .

Øvelse 4.10:

Udgangspunktet i denne øvelse er novo datasættet med data fra 200 personer. ([Link til 1](#))

C	D	E	F	G	H	I	J	K	L	M	N	O	P
										mængden af HbA1c i blodet (målt i en bestemt enhed) i uge:			
alder	højde	køn	region	behandl	vægt v basel	bmi v baseline	uge 0	uge 4	uge 8	uge 12	uge 16	uge 20	uge 26
64	1,66	F	North America	A	99,7		8,5	7,7	7,2	6,5	6,2	6,4	6,3
47	1,72	F	Europe	B	82,5		7,9	7,3	7,1	6,8	6,8	6,9	6,2
50	1,88	M	North America	A	128,8		8,1	7,7	7,2	6,6	6,1	6,2	6,3
49	1,65	F	Europe	B	94,3		7,4	6,9	6,4	6,3	6,1	6,2	6,3
59	1,66	F	Europe	B	108,6		7,8	6,8	6,6	6,3	6,2	6	5,9
65	1,6	F	Europe	A	68,6		8,6	7,2	6,9	6,5	6,3	6,1	5,9
54	1,74	M	Europe	A	96,1		7,7	7,6	7	6,2	6	5,8	6
53	1,7	M	Europe	A	74,5		8,5	7,9	7	5,9	5,5	5,6	5,8
67	1,85	M	North America	A	97,9		7,6	7,2	6,1	5,8	5,6	5,6	5,5

- Undersøg fordelingen af mændenes og kvindernes højder hver for sig efter samme fremgangsmåde som i øvelse 4.9.
- Undersøg fordelingen af mændenes og kvindernes vægt ved baseline efter samme fremgangsmåde som i øvelse 4.9.
- Undersøg fordelingen af ændringen i HbA1c ift. baseline ved uge 26 efter samme fremgangsmåde som i øvelse 4.9.

4.6 Undersøgelse af data ved hjælp af et QQ-plot

I HEM3 er der givet en indføring i teorien bag qq-plots. Du kan hente afsnittet her – [link til 22](#).

Ved at tegne et histogram kan vi vurdere, om vi synes det ligner grafen for en normalfordeling (tæthedsfunktionen). Eller vi kan tegne en sumkurve og vurdere, om det ligner grafen for en normalfordelings fordelingsfunktion. Men det er selvfølgelig meget løst. Det er ikke let med øjet at give en sikker vurdering af krumme kurver.

De eneste kurver, vi med rimelig sikkerhed kan udtale os om, er de lineære! Derfor er der, i lighed med det vi kender fra potensmodeller og eksponentielle modeller, udviklet en metode til linearisering af fordelingsfunktionen for en normalfordeling. Metoden går ud på at fremstille såkaldte qq-plots (der tidligere blev kaldt 'probit-plots').

De fleste værktøjsprogrammer kan producere et sådant plot ud fra et datasæt. Lad os igen se på eksemplet med mændenes højder fra ANSURII. Du kan finde en gennemgang af, hvordan plottet kommer frem [her – link til 23](#)

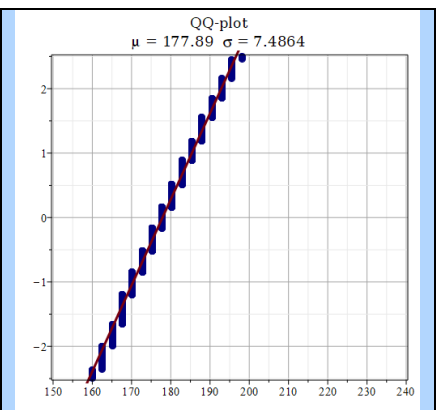
Øvelse 4.11: Punktplot af de transformerede data

- Læg listen med højderne ind i dit værktøjsprogram.
- Anvend QQ-kommandoen.

I det teoretiske afsnit vises det, at hvis det er en normalfordeling,

så vil ligningen for lineariseringen være:
$$y = \frac{1}{\sigma} \cdot x - \frac{\mu}{\sigma}$$

- Tegn selv grafen for den lineære funktion ind i samme billede, og kontroller. Det skal se ud nogenlunde som på billedet her: *Bemærk: Ved tegningen af histogrammet blev parametrene estimeret til $\mu = 178,8$ og spredningen $\sigma = 7,5$. Forskellen skyldes den lidt grovere intervalinddeling dengang.*



Øvelse 4.12: QQ-plot ud fra novo datasættet

- Lav et QQ-plot for A og B grupperne fra novodatasættet. ([Link til 1](#))
- Understøtter det en påstand om, at datasættene er normalfordelt?

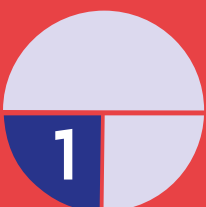
Et QQ-plot kan understøtte en argumentation om, at et datasæt er normalfordelt. Men det kan ikke bevise det.

5.

Ny medicin: Fra ide til kliniske lægemiddelforsøg



I filmen forklarer Lea lidt om faserne i de kliniske undersøgelser, der ligger til grund for udviklingen af ny medicin. Hvordan kommer man fra "den gode idé", som en forskergruppe har fået, til det tidspunkt, hvor medicinen er at finde på apotekernes hylder eller på lægeklinikkerne? Hvor lang tid går der, og hvilke trin skal man igennem? Hvilke lægefaglige, farmakologiske, statistiske, praktiske osv. overvejelser indgår? Hvordan kan man eksempelvis være sikker på, at medicinen virker? Og hvad med bivirkninger? Hvordan udvælges testpersoner? Hvordan randomiserer man, og hvordan er man sikker på, at testpersonen ikke kan afkode, hvad vedkommende får? Det er nogle af de spørgsmål, vi går i dybden med i dette kapitel.



5. Ny medicin: Fra idé til kliniske lægemiddelforsøg

5.1 Difteri, stivkrampe og hesten Jim

I dag er lægemiddeludvikling en lang proces, og vejen fra idé til produkt er både besværlig og beko-stelig. Kun 1 ud af 10000 idéer til lægemidler ender med at blive til noget, og ofte går der 10-15 år fra den oprindelige idé til det færdige medicinske produkt, mens den samlede pris for processen kan være op mod 20 milliarder danske kroner. Lægemiddeludvikling eller 'drug hunting' er 'big busi-ness'!

Men sådan har det ikke altid været, og rent faktisk skal man ikke længere tilbage end 1950'erne for at finde en tid, hvor der ikke var faste regler for udvikling og godkendelse af nye lægemidler, hvilket naturligvis kunne resultere i tragiske hændelser. Et godt eksempel på dette er historien om sygdom-mene difteri, stivkrampe og hesten Jim fra 1901.

I 1800-tallet var difteri, der er en slags alvorlig halsbetændelse, et alvorligt sundheds-mæssigt problem i det meste af verden. Sygdommen er en infektionssygdom, der skyldes den patogene bakterie *Corynebacterium Diphtheriae*, som angriber slimhinderne og produ-cerer et kraftigt toksin, der i nogle tilfælde kan medføre blodforgiftning og død. Hvis sygdommen ikke behandles, er dødeligheden op mod 33%. I New York oplevede man bl.a. to difteriepandemier i 1881 og 1887 som kostede over 10.000 mennesker livet.



Corynebacterium Diphtheriae set gennem mikroskop, https://en.wikipedia.org/wiki/Corynebacterium_diphtheriae

Øvelse 5.1

Hvad er et toksin? Hvad betyder det, at en bakterie er patogen? Undersøg om difteri også er et problem i verden i dag.

I 1890 udviklede bl.a. den tyske læge Emil Von Behring en smart metode til behandling af difteri. Den gik ud på at sprøjte relativt små doser af difte-ribakterier ind i store dyr som heste eller får, hvor-ved dyrenes immunsystemer dannede antitoksiner, som kunne aftappes via blodet.

Disse antitoksiner kunne nu injiceres i mennesker, som derved opnåede en høj modstandsdygtighed over for de toksiner, som difteribakterien dannede. Behandlingen kaldtes serumterapi, og den var en stor succes allerede fra starten, og et stort produktionsapparat blev sat i gang.

Laboratorier skød op mange steder i både Nordame-rika og i Europa, og der blev fremstillet antitoksiner fra hesteblood i stor stil.



Emil Von Behring, der udviklede en behandling af difteri

1901 fik Emil Von Behring Nobelprisen blandt andet for sit arbejde med bekæmpelse af difteri.

Kvalitetssikringen af det antitoksin, der blev produceret, var imidlertid mangelfuld, og i 1901 gik det galt på et laboratorium i St. Louis, USA.

Her havde man gennem nogle år med stor succes benyttet blod fra den pensionerede arbejdshest 'Jim' til fremstilling af antitoksin, som man benyttede til serumterapi af områdets børn. Men i løbet af efteråret døde mindst 12 børn umiddelbart efter at have modtaget serumterapi mod difteri. Obduktionerne viste, at de alle var døde af sygdommen stivkrampe, som skyldes en helt andet bakterie *Clostridium tetani*, og efter yderligere undersøgelser fandt man frem til, at bakterien stammede fra Jims blod, der ikke var blevet undersøgt, før det blev givet til børnene i serumterapi-behandlingen.



Antitoksinproduktion fra hesteblood omkring år 1900

<https://www.science.org/content/article/scientists-find-way-make-diphtheria-treatment-without-injecting-horses-toxin>

Den tragiske historie førte bl.a. til dannelsen af FDA – *U.S. Food and Drug Administration*, der bl.a. har til opgave at sikre, at udviklingen og anvendelsen af samt kontrollen med lægemidler foregår efter særligt givne retningslinjer.

Øvelse 5.2

I artiklen "*Hero Horses in the Fight Against Disease*" som du kan hente her [- link til 24-](#) kan du læse lidt mere om difteri-skandalen fra St. Louis og de efterfølgende høringer og undersøgelser.

- Hvilken rolle spillede forskerne Amand Ravold og Martin Schmidt?
- Diskutér fordele og ulemper ved skærpet kontrol med udviklingen af lægemidler.

5.2 Thalidomid

Et berøgt eksempel på nødvendigheden af skrappe regler og streng kontrol med lægemiddeludviklingen er stoffet *Thalidomid*, der blev lanceret af det tyske medicinalfirma Grünenthal i 1950'erne som et middel mod svimmelhed, søvnløshed og kvalme. Det blev betragtet som harmløst, og på få år blev det overalt i verden populært at ordinere til kvinder mod kvalme i starten af deres graviditet.

I 1959 begyndte der imidlertid at komme rapporter ind om børn født med misdannelser på grund af mødrenes brug af thalidomid. Børnene var født mere eller mindre uden arme og ben. Det tyske medicinalfirma afviste hårdnakket enhver sammenhæng med deres produkt. Men det blev værre, for i 1960 påviste tyske og australske læger en markant stigning i antallet af børn født med misdannede lemmer.



Thalidomidbørn i en svømmehal (ca. 1967)

Efter hårdt pres og stærke beviser måtte Grünenthal i 1961 trække thalidomid ud af markedet. Senere holdt man en verdenskongres i København om thalidomid, og på verdensplan var omkring 8.000 børn blevet født med misdannelser, og børnene blev ligefrem kaldt *thalidomidbørn*. De mest almindelige misdannelser var korte lemmer ofte med manglende tæer og fingre.

Rent fysiologisk modvirker thalidomid ganske rigtigt kvalme, men stoffet har også en uheldig bivirkning: det hæmmer fosterets blodforsyning til ekstremerne – specielt i fjerde og femte graviditetsuge, hvor netop lemmerne udvikles.



De fleste personer med thalidomid-relaterede misdannelser er født i 1959. Et dansk "kendis-eksempel" er forsangeren i rockbandet Sort Sol Steen Jørgensen, der er født uden nederste del af venstre arm, fordi hans mor under graviditeten fik ordineret thalidomid.

Historien om thalidomid bliver sjældent præsenteret i hele sin grusomhed. Stoffet blev opdaget af den tyske kemiker Otto Ambros under 2. verdenskrig. Han arbejdede for IG Farben, hvor han havde udviklet giftstoffet Sarin, som tyskerne planlagde at bruge i deres krigsførelse. Men før de slap det løs, skulle de have en modgift til egne soldater, og det var i det arbejde, han opdagede stoffet thalidomid. Han fik adgang til at afprøve det på jødiske og andre kz-fanger, bl.a. i Bergen Belsen. Her blev født børn med samme misdannelser som man så omkring 1960. Otto Ambros blev dømt som krigsforbryder ved domstolen i Nürnberg, men under den kolde krig havde USA brug for hans og mange andre tyske videnskabsmænds ekspertise, så han blev løsladt efter få år. I 1956 ansættes han Grünenthal, hvor han bringer thalidomid på banen igen. Du kan læse yderligere om historien i en DR-dokumentar, som du kan finde her [– link til 25](#).

5.3 Helsinkideklarationen, apotekerloven og lægemiddeloven

I 1964 afholdtes i Helsinki en international konference, hvor man diskuterede etik og patientretigheder i forbindelse med kliniske undersøgelser og udvikling af lægemidler. Konferencen mandede ud i den såkaldte Helsinkideklaration, som alle lande i dag er forpligtet til at overholde. I Danmark fik vi endvidere i 1954 'Apotekerloven' og i 1975 'Lægemiddeloven' som danske medicinalproducenter skal overholde.

Øvelse 5.3

Undersøg på nettet, hvad 'Helsinkideklarationen', 'Apotekerloven' og 'Lægemiddeloven' omfatter. Benyt f.eks. disse links:

Helsinkideklarationen: [Helsinki-deklarationen | læger.dk \(laeger.dk\)](#)

Apotekerloven: [Apotekerloven \(laegemiddelstyrelsen.dk\)](#)

Lægemiddeloven: [Lægemiddeloven \(laegemiddelstyrelsen.dk\)](#)

5.4 Den prækliniske fase

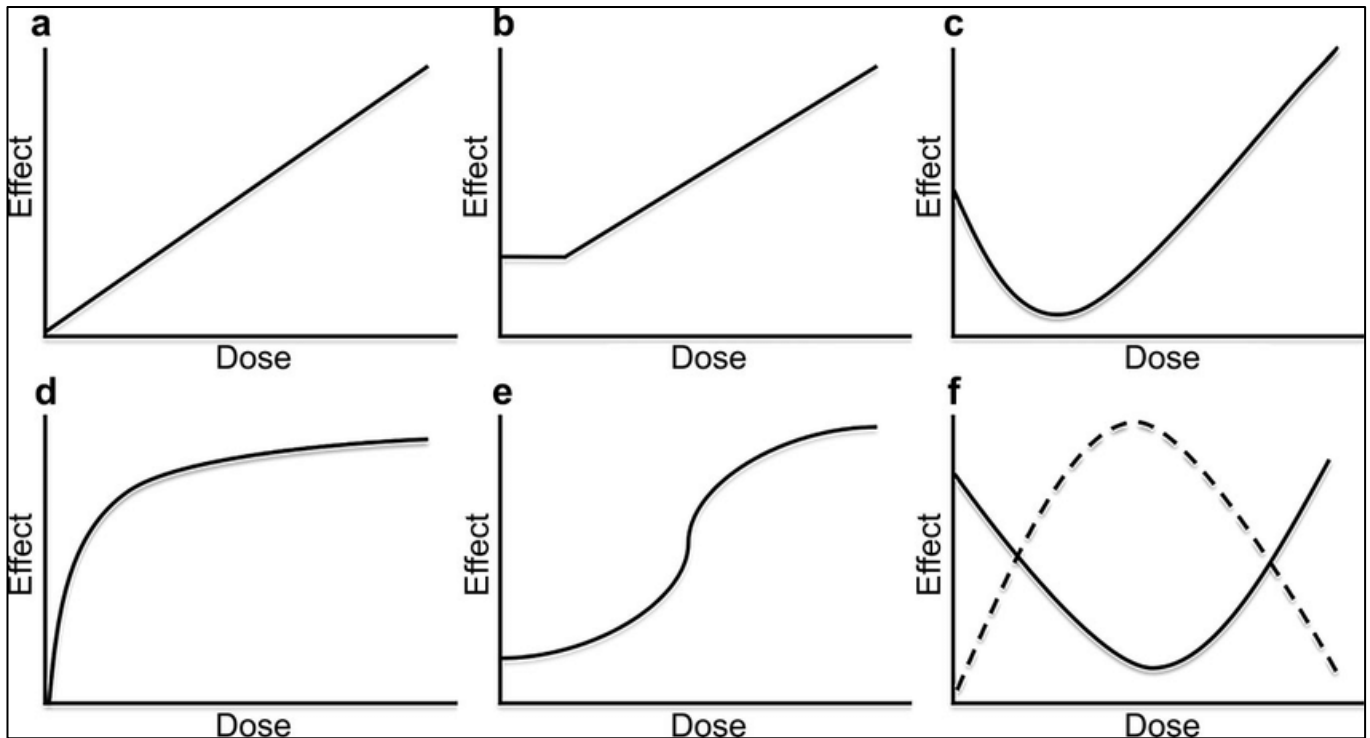
Man inddeler typisk udviklingen af nye medicintyper i to faser: *den prækliniske* og *den kliniske* fase. Den prækliniske fase handler om at få gode idéer. Ved kemisk eller biologisk syntese designer man molekyler, som man formoder har gavnlige virkninger i forhold til bestemte sygdomme. Herefter afprøves molekylerne i simple laboratoriemodeller.

Dette foregår først '*in vitro*' – dvs. uden for levende organismer og senere '*in vivo*' – i form af dyreforsøg. Her undersøger man dels molekylets farmakokinetik (hvordan optages, distribueres og nedbrydes molekylet i forsøgsdyrenes organismer) samt farmakodynamikken (dosis-respons sammenhænge). Derudover ses på toksikologiske effekter.

Øvelse 5.4

Hvad betyder 'in vitro' og 'in vivo' hvis man oversætter det direkte fra latin?

Når man laver dosis-respons undersøgelser på forsøgsdyr, resulterer de ofte i kvantitative data, der kan benyttes til såkaldte dosis-responskurver eller dosis-effektkurver. Disse kan f.eks. se således ud:



Eksempler på dosis-responskurver: a) Lineær respons, b) Begrænset respons, c) J-formet respons, d) Mættet respons, e) S-formet respons, f) U-formet eller omvendt U-formet respons

[Illustration of common dose-response curves. \(a\) Linear-no-threshold... | Download Scientific Diagram \(researchgate.net\)](#)

Øvelse 5.5

Hvad er der på akserne på dosis-responskurverne? Diskutér for hvert af de seks tilfælde, hvor området for optimal dosering ligger.

Den prækliniske fase tager typisk op til fire år, og man anslår, at kun ca. 1‰ af de potentielle lægemidler, der testes præklinisk, slipper gennem nåleøjet til selve den kliniske fase.

5.5 Den kliniske fase

Fase I

Den kliniske fase er underinddelt i fire faser: I-IV. I den første fase - Fase I - er det primære formål at teste lægemidlets tolerance. Ofte testes kun på en ganske lille gruppe frivillige personer typisk i intervallet fra 10 til 50 og oftest testet kun på raske, yngre, mandlige personer. Hvis det giver mening ift. forsøgets design, inddrages her og i de følgende faser placebo-grupper, og der foretages randomisering, dvs patienternes placering i behandlingsgrupper foregår ved lodtrækning.

Øvelse 5.6

Hvorfor tror du, man starter de kliniske forsøg med primært at teste på en meget lille gruppe personer. Og hvorfor er det at foretrække, at der er tale om raske, yngre mænd?

I Fase I forøger man gradvist dosis af lægemidlet startende med kun ca. 1% af den dosis, der viste sig at have effekt på forsøgsdyr, hvorefter man gradvist forøger. I slutningen af fasen, som typisk varer 1-2 år, har man et billede af, om lægemidlet har en egnet profil til at gå videre til Fase II. Man vurderer bl.a. om lægemidlets effekt er høj, om bivirkningerne er få og ikke-alvorlige, om halveringstiden er lang og om stoffet har en høj oral biotilgængelighed. Ca. 40% af de lægemidler, der testes i Fase I kvalificerer sig til næste fase.

Øvelse 5.7

Hvorfor foretrækker man mon lægemidler med lang halveringstid og høj oral biotilgængelighed, når man skal vurdere om de er egnede til Fase II?

Fase II

I Fase II gives lægemidlet for første gang til egentlige patienter, der lider af den pågældende sygdom. Normalt testes på et sted mellem 100 og 300 personer af varierende alder, køn osv. Formålet med testfasen er at undersøge langtidsvirkninger, lave dosis-effekt kurver for derved at bestemme områder for optimal dosering. Normalt inddrages i Fase II også placebokontrolgrupper.

Øvelse 5.8

Ud over alder og køn, hvilke baggrundsvariable er så vigtige at tage hensyn til, når man udvælger personer til Fase II?

Fase III

I Fase III afprøves lægemidlet på en stor gruppe patienter – som regel op til ca. 2000. Formålet med fasen er at be- eller afkræfte, om lægemidlet er brugbart mod sygdommen og om det har terapeutiske fordele i forhold til allerede eksisterende behandlinger. Som omtalt i starten er det vigtigt, at anvende såkaldte "randomiserede forsøg" – dvs. forsøg, hvor man ved ren og skær lodtrækning udvælger patienter til henholdsvis behandlings- og kontrolgrupper. Fase III er omfattende og tidskrævende og foregår på tværs af hospitaler og geografiske regioner. Det er ligeledes i denne fase man undersøger den farmakoøkonomiske betydning af lægemidlet. Hvis forsøget muliggør det, vil der ofte blive indraget placebo-grupper.

Øvelse 5.9

Hvad tror du der ligger i begrebet et lægemiddels 'farmakoøkonomiske betydning'? Hvordan vurderer man mon denne betydning? Hvilke overvejelser indgår?

Sammenlagt tager Fase II og Fase III 4-6 år, og ca. 10% af de lægemidler, der introduceres i Fase II og Fase III forsøg, går videre til den sidste fase - Fase IV.

Fase IV

I Fase IV udføres undersøgelser i meget stor skala. Her er der tale om 2000+ patienter, og fasen har bl.a. til formål at afdække sjældne bivirkninger, fordele og ulemper ved hverdagsbrug af lægemidlet og i det hele taget give et billede af, hvilken indflydelse lægemidlet vil have på markedet og forbrugsmønstre.

Faserne i de kliniske undersøgelser er anskueliggjort nedenfor:

Prækliniske undersøgelser		Kliniske undersøgelser			
In vitro	In vivo	Fase I	Fase II	Fase III	Fase IV
Kemiske og biologiske undersøgelser	Forsøgsdyr	10-50 raske, yngre mænd	100-300 patienter med sygdommen	<2000 patienter	>2000 patienter, evt. globalt
Drug hunting	Farmakologi Toksikologi	Tolerance Bivirkninger	Optimal dosering Langtidsvirkning		Sjældne bivirkninger Farmakoøkonomi Forbrugsmønstre
Ca. 4 år		Ca. 2 år	Ca. 4-6 år		1-3 år
10000 stoffer	1000 stoffer	10 stoffer	4 lægemidler		1 lægemiddel

5.6 Hvis du vil vide mere om kliniske undersøgelser

Hvis du har lyst til at læse mere om kliniske undersøgelser og faserne i udvikling af ny medicin, er der to kilder til videre studier her:

- Brøsen, Simonsen, Dalhoff (2019): "Basal og klinisk farmakologi", FADL's Forlag
- Jensen, Jensen, Mogensen, Paulsen (1992): "Lægemiddelkemi", Systime

5.7 Case: Skørbug – første skridt mod kontrollerede forsøg

I slutningen af 1700 tallet var situationen i de store flådenationer, at sygdommen *skørbug* slog langt flere ihjel, end krigshandlinger. Den engelske flåde havde omkring 200.000 matroser indrulleret gennem de sidste 30 år af 1700 tallet. 130.000 af dem dør af skørbug. Det samme var tilfældet for fx den franske flåde. Så det siger sig selv, at man leder desperat efter årsagen til sygdommen og efter en kur.

Øvelse 5.10

Slå op i et leksikon eller på nettet og giv en kort beskrivelse af, hvad skørbug er, og hvad der forårsager sygdommen.

Denne case handler om denne bestræbelse. Det viste sig, at en engelsk søofficer, der desværre var langt nede i graderne og hierarkiet, ikke bare var tæt på at løse gåden, men at han i sin jagt efter at finde en løsning faktisk er tæt på at opstille en moderne undersøgelsesmetode, hvor et stof der af en eller anden grund forekommer interessant, testes op mod en række andre stoffer.

Fortællingen herom er gengivet i bogen *Randomistas. How radical researchers are changing our world* af Andrew Leigh. Den begynder som gengivet nedenfor, og du kan fortsætte læsningen ved at hente dokumentet her – [link til 26](#).

Læs afsnittet fra bogen og giv et kort referat.

As Commodore George Anson's six warships rounded the tip of South America, he knew his plans were unraveling. Commissioned by the British government to capture Spanish colonies in Panama and Peru, Anson had been at sea for half a year. The seas were rough and a Spanish squadron was chasing him, but the greatest threat to the men on board was a disease: scurvy.

Anson's expedition left England in 1740 with 1854 men. Four years later, just 188 of them returned. Of the eight ships that set out, only one made it back. Writer Stephen Bown describes the voyage as the worst medical disaster ever to take place at sea. Six months in, men were dying daily. At first, the corpses were sewn into their hammocks and thrown overboard. But after a time the living sailors became too weak to deal with the dead, and they were left below deck, where they had drawn their last breaths.

Sailing an eighteenth-century warship required many men. Anticipating that some would die of scurvy, captains often took along extra personnel. But Anson's losses exceeded expectations. With few men having the strength to work, some of the warships crashed into rocks off South America. One ship, the *Wager*, was ripped apart on the jagged rocks of southern Chile. Some of the sailors drowned because they were too weak to swim ashore.

Øvelse 5.10

- 1) Hvad var det for et forsøg, der blev foretaget? Beskriv de 6 kategorier, matroserne blev inddelt i.
- 2) Hvilke konklusioner kan man efter din mening drage af forsøget.
- 3) Sammenlign forsøgsopstillingen med moderne clinical trials. Hvad er anderledes, og hvad er fælles?

Et af problemerne med Linds opdagelse var, at man ikke anede, hvorfor citrusfrugter kan forebygge skørbug. Vi er over 100 år før, man opdager vitaminer, og finder ud af, hvor afgørende disse er for kroppens forskellige funktioner. Englænderne holdt opdagelsen hemmelig. De satte gang i en storproduktion af juice fra citrusfrugter, som blev en fast del af de engelske skibes forsyninger, når de var på langfart.

Derfor kunne det naturligvis ikke holdes totalt hemmeligt. Englænderne rådede over en eller anden hemmelig eliksir, der hjalp mod skørbug. Men da man ikke vidste, hvorfor det virkede, eller hvad det var i citrusfrugterne, der virkede, så drog man også af og til forkerte konklusioner.

Citrusfrugter er syrlige. Så det er måske det syrlige, der virker. Men der er jo meget andet, der er syrligt, fx eddiker. Det handler det næste kapitel i fortællingen om kampen mod skørbug om.

Konkurrencen om at nå først til Sydpolen

For godt 100 år siden var der en international konkurrence om at være den første, der nåede sydpolen. De to største navne og største konkurrenter var Scott fra Storbritannien og Amundsen fra Norge.

Den del af fortællingen, der interesserer os i forbindelse med historien om skørbug, er gengivet i Judea Pearl: *The Book of Why*. Du kan hente afsnittet her – [link til 27](#).

Øvelse 5.11

- 1) Læs beskrivelsen af Scotts forberedelser til og gennemførelsen af turen og giv et kort referat. Hvad er Scotts holdning til, hvordan man kan bekæmpe skørbug.?
- 2) Giv en vurdering ud fra datidens viden, hvor man ikke kendte vitaminer, af rationalet i Scotts overvejelser.

6.

Etiske problemer i clinical trials



I filmens afsluttende del spørger Casper, hvorfor man skal gennemføre et studie lige på denne måde, og hvorfor 26 uger. Hvis man "ved" allerede efter fx 10 uger, at A er bedre end B, hvorfor så fortsætte. Er det ikke uetisk at blive ved med at give halvdelen et dårligere præparat. Søren Andersen fra Novo Nordisk forklarer, at man kan designe sine studier, så man kan stoppe før tid, hvis nogle forudbestemte betingelser er opfyldte. Det skal dog være planlagt på forhånd. Men Søren tilføjer også, at der er situationer, hvor man er så presset, at man fra myndighedernes side vælger at fravige proceduren. I dette kapitel præsenteres tre cases om dette problemkompleks.

6. Etiske problemer i clinical trials

6.1. Case: Remdesivir og hydroxychloroquine

Under første del af corona-epidemien havde man ikke nogen medicin, der kunne hjælpe de hårdt ramte patienter. I den situation kom der fokus på især to præparater, dels hydroxychloroquine, et middel der anvendes mod visse malariasygdomme, og som ikke mindst præsident Trump advokerede for; og dels remdesivir, der var et kendt lægemiddel, som er virksomt mod visse virale sygdomme.

I en situation, hvor man ikke har noget i værktøjskassen griber man lettere efter de muligheder, der tilbydes, så i 2020 blev det anvendt i Kina og USA og en række andre lande. Der blev gennemført hurtige forsøg med gode resultater, men desværre var det ikke resultater, andre forskergrupper kunne genfinde. Så tvivlen opstod tidligt.

Øvelse 6.1 Hydroxychloroquine

Du kan hente en artikel fra Nature sammen med en lydfil, hvor der redegøres for resultaterne af forsøgene, samt et indlæg fra det danske Ugeskriftet (for læger) her – [link til 28](#).

- Læs artiklerne, lyt til podcasten og giv et referat heraf.
- Giv din vurdering af, om der er tale om lægefagligt begrundede forsøg eller i højere grad politisk dikterede.

Det er jo de valgte politikere, der har ansvaret for håndtering af krisen, og i sidste ende sidder med problemet.

- Hvorfor skulle det på den baggrund være odiøst, at politikerne blander sig og ønsker bestemte muligheder undersøgt?

Remdesivir er et lægemiddel, som i 2020 var kendt som et virksomt stof mod visse virale angreb. Især blev det brugt mod Hepatitis C, men det har også været i brug mod sygdomme, som man ikke har noget endegyldigt værn mod. Eksempelvis blev det brugt mod Ebola. Det var derfor naturligt at undersøge, om det kunne hjælpe patienter, der var hårdt angrebet på deres åndedrætsfunktioner.

Da midlet var afprøvet mht sikkerhed, bivirkninger mv, gik man direkte til storskalaforsøg. Her blev resultaterne vurderet af de amerikanske myndigheder som så gode – man havde jo heller ikke andet på dette tidspunkt – at der hurtigt blev givet autorisation til at bruge stoffet. Det medførte, at man i mange lande, herunder Danmark begyndte at bruge stoffet. Men allerede i starten af 2021 går WHO ud med stærke faglige argumenter og fraråder brugen af stoffet. Det ændrer ikke holdningen i USA. Og i Danmark kommer der heller ingen reel justering af anbefalingerne.

Øvelse 6.2 Remdesivir

Hent de tre artikler du kan finde her: en oversigtartikel fra Nature, en artikel fra det danske Ugeskriftet om WHO's stilling og en faglig artikel fra New England Journal of Medicine – [link til 29](#)

- Giv et referat af sagen ud fra artiklen i Nature.
- Læs summary fra den faglige artikel i New England Journal ..., find de statistiske fagudtryk, du møder her og giv en kort forklaring for hver.
- Giv en beskrivelse af, hvordan et stort *fase 3 studium* burde tilrettelægges.
- I artiklen fra Ugeskriftet refererer de WHO's indstilling til at stoppe med brugen af remdesivir. Hvad er who's begrundelser.
- Hvad bliver konsekvensen i Danmark?

6.2. Case: Theis Lange og Frank Shann, *Perils of stopping trials early*

Indholdet i denne case er følgende: Et nyt præparat afprøves klinisk på meget syge børn. Forsøget afbrydes lang tid før tid, fordi det ser ud til at præparatet har positiv virkning: Færre i gruppen der modtager aktiv medicin dør, end i gruppen, der modtager placebo. Materialet er egnet til at indgå i et studieretningsprojekt.

Øvelse 6.3

Hent artiklerne om og af lægen fra Bangla Desh her – [link til 30](#).

- Start med at læse artiklen fra Economist, der giver en overordnet præsentation af problemet.
- Skriv på grundlag af artiklen et kort referat af sagen: Hvad var baggrunden for at iværksætte forsøget? Hvordan blev forsøget tilrettelagt? Hvordan blev det udført?
- Hvornår i forløbet blev forsøget afbrudt? Hvad var den præcise begrundelse? Hvem er det der afbryder forsøget?
- Har de, som afbrød forsøget, et belæg fra andre lignende forsøg for, at man kan fremskrive resultaterne, der forelå, da man afbrød forsøget, og dermed forudsige med nogenlunde sikkerhed, hvad der ville ske, hvis man førte forsøget til ende?
- Hvad er konsekvenserne af, at et sådant forsøg afbrydes?

Øvelse 6.4

Orienter dig dernæst i henh. Mohammad Chisti et al.'s artikel *Bubble continuous positive airway pressure...an open, randomised controlled trial* samt i artiklen skrevet af Frank Shann og Theis Lange. Theis Lange er ansat på afdelingen for Biostatistik på Københavns Universitet. Begge artikler er bragt i *The Lancet*.

- Mohammad Chisti et al argumenterer for, at beslutningen om at stoppe forsøget var korrekt, mens Frank Shann og Theis Lange problematiserer denne beslutning. Hvor i artiklerne finder du disse argumenter.
- Giv en præsentation af, hvordan de to opponenter hver for sig håndterer relationen mellem de etiske problemstillinger på den ene side og den professionelle statistiske indsigt på den anden side.
- De to grupper er åbenlyst uenige, også om den statistiske metode. Giv en belysning af en af disse problemstillinger, der vedrører brugen af χ^2 -test contra brugen af Fishers eksakte test. Forklar, *hvornår* de to typer af test anvendes, *hvordan* de udføres og om der er "frit valg" mellem dem. Det berøres på side 2 af Shann og Lange og på s 5 af Mohammad Chisti et al.
- Det kan virke mærkeligt, at man ud fra de samme data kan nå til så forskellige konklusioner. Betyder det, at den ene fortolkning kan være lige så god som den anden? Eller hvad er din fortolkning af årsagen hertil?
- Læs det afsluttende afsnit af Mohammad Chisti et al.'s artikel med overskriften: Diskussion, og giv et referat heraf, idet du skal forholde dig til rimeligheden i, at der drages de konklusioner, som det sker i artiklen, i betragtning af at forsøget blev afbrudt

Bubble CPAP for pneumonia: perils of stopping trials early

Published Online
August 24, 2015
[http://dx.doi.org/10.1016/S0140-6736\(15\)60691-2](http://dx.doi.org/10.1016/S0140-6736(15)60691-2)

See [Articles](#) page 1057

"Avoid any analysis (or even brief inspection) of the data until some dozens of deaths have accumulated, for it is trials first looked at when very small that are most likely to be misleading."

Peto R et al¹

Experience with treating respiratory distress in neonates suggests that bubble continuous positive airway pressure (CPAP) might be a very effective treatment for severe pneumonia in children.² In *The Lancet*, Mohammad Chisti and colleagues³ report a randomised trial carried out in Bangladesh, in which 225 children aged under 5 years with severe pneumonia were enrolled, comparing bubble CPAP and high-flow oxygen therapy with standard low-flow oxygen therapy. After 21 treatment failures (the primary endpoint) and 13 deaths (a secondary endpoint) in two treatment groups combined, the bubble CPAP group had about a quarter the treatment failure rate and a quarter the mortality rate of the low-flow oxygen therapy group. Should such a trial be continued and risk excess deaths of children randomly assigned to low-flow oxygen therapy? Or should it be stopped early, despite the small sample size, and risk excess deaths in the future by providing misleading evidence, or by delaying widespread uptake of an effective treatment?

When trials begin, random variation often causes large treatment effects (random highs or lows) that diminish as the trial progresses (regression to the truth). Consequently, on average, randomised trials stopped

early for benefit overestimate treatment effects, especially when there are fewer than 500 events.⁴ Statistical techniques can adjust for multiple significance testing, but cannot prevent overestimation of treatment benefit from random highs that lead investigators to terminate a trial early.⁵ The bubble CPAP trial³ was stopped after reaching only 23% of the planned enrolment, primarily because the Data and Safety Monitoring Board deemed that low-flow oxygen therapy was associated with a higher mortality (10 [15%] of 67 children died) than bubble CPAP (3 [4%] of 79 children died). The Data and Safety Monitoring Board was in a very difficult ethical position; nevertheless, there are serious problems with the decision to stop the trial early.

First, after the first interim analysis, an independent statistician suggested that the Data and Safety Monitoring Board use the Haybittle-Peto stopping rule with p value less than 0.003. Peto and colleagues emphasise that their rule should not be applied "until some dozens of deaths have accumulated".¹ But this advice was not followed, and the trial was stopped after the second interim analysis with only 21 treatment failures and 13 deaths in the bubble CPAP and low-flow oxygen therapy groups combined (the high-flow oxygen therapy group was not considered in the decision to end the trial early). Many statisticians recommend using a p value less than 0.001 rather than 0.003 for the

Haybittle-Peto stopping rule,^{5,6} and even lower p values if less than half of the planned events have taken place (as in the bubble CPAP trial).^{6,7}

Second, rather than the χ^2 test, the authors should have used Fisher's exact test to determine whether the p value was less than 0.003 for the primary comparison of treatment failure with bubble CPAP (5 [6%] of 79 children) versus low-flow oxygen (16 [24%] of 67). The authors planned to use Fisher's exact test only when the number in any cell of a 2 × 2 table was less than five, but this rule of thumb depends on the level of significance being about 0.05; the lower the level of significance, the more data are needed to use the χ^2 test.⁸ Fisher's exact test does not have this problem and gives a p value of 0.0038, so the primary comparison did not have a p value of less than 0.003.

Third, the CI for the relative risks depends on the confidence level chosen (95%, 99.7%, or 99.9%) and the method of calculation. The authors present a 95% CI for death, with no correction for repeated testing or early stopping, and a 99.7% CI for treatment failure. A strong case can be made for using a CI of at least 99.9% for both treatment failure and death,^{5,6} but even with 99.7% CI as suggested by the statistician, Stata's cumulative incidence command with Fisher's exact test comparing bubble CPAP with low-flow oxygen therapy yields relative risks of 0.25 (99.7% CI 0.04–1.68) for death and 0.27 (99.7% CI 0.06–1.12) for treatment failure; neither comparison is significant even at the rather liberal criterion of a p value less than 0.003.

After only three (4%) deaths in the bubble CPAP group but 10 (15%) deaths in the low-flow oxygen therapy group (p=0.022), the investigators had an ethical dilemma: continue the trial and risk excess deaths in children randomly assigned to low-flow oxygen therapy or stop the trial early with weaker evidence than that gained from a full trial. Weak evidence can cause deaths if it leads to use of an ineffective treatment or delays widespread use of a beneficial one. As an interim analysis of a secondary comparison, a p value of 0.022 is potentially misleading; this p value should be compared with a p value of 0.003 (as recommended by the statistician) rather than a p value of 0.05, and it therefore provides only weak evidence that low-flow oxygen therapy actually has a higher mortality than bubble CPAP.^{1,4-6} Interim analyses of previous trials have found p values much lower than 0.022 for mortality, but no significant effects at completion (ie, regression to the truth); for example, p=0.0002 in a trial



David Silverman

of candesartan,⁹ p=0.002 in a trial of leukaemia drugs,¹⁰ and p=0.006 in a trial of tifacogin for sepsis⁷). To obtain evidence strong enough to change practice, investigators in Europe, North America, and Australia have considered it ethical to continue trials until the p value becomes much lower than 0.022 for mortality (eg, p=0.00006 in a trial of zidovudine,¹¹ p<0.0001 in a trial of bisoprolol,¹² and p<0.001 in trials of carvedilol and spironolactone¹³).

It is very difficult to do randomised trials in regions of high mortality, where limited resources and the high risk of death cause formidable logistical and ethical problems. The investigators in Bangladesh are to be congratulated for undertaking an ingenious and important trial that provides evidence suggesting that bubble CPAP might be a substantial improvement over the standard treatment of low-flow nasal-prong oxygen. Unfortunately, firm conclusions cannot be drawn because the trial was stopped early and incorrect statistical procedures were used. We hope it will be possible to perform another randomised trial (with rigorous group sequential analysis and perhaps mortality as the primary endpoint) or careful observational studies (if a randomised trial is not possible). Meanwhile, barriers to the use of bubble CPAP should be identified so this method of oxygen delivery can be made widely available if its benefits are confirmed.

*Frank Shann, Theis Lange

Intensive Care Unit, Royal Children's Hospital, Parkville, 3052 VIC, Australia (FS); Department of Paediatrics, University of Melbourne, Parkville, Australia (FS); and Section of Biostatistics, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark (TL)
frank.shann@rch.org.au

6.3. Case: Alzheimer – medicin godkendes, før vi kender sygdommens natur

I 1906 obducerer en tysk læge en netop afdød 51 årig kvinde for at prøve at forstå, hvad årsagen var til den galopperende demens, hun var blevet ramt af, og som gennem 5 år tydeligt for alle nedbrød ikke alene hendes kognitive evner, men hele hjernens funktion. Han fandt ved obduktionerne en række steder i hjernen, hvor der var ophobninger af bestemte proteiner, der klumpede sig sammen uden på nervecellerne i form af såkaldt beta-amyloid-plak.

Hans beskrivelse af fænomenet førte blandt andet til at sygdommen blev navngivet med hans navn, (Alois) Alzheimer. Og lige siden har der stort set været konsensus om, at han havde fundet årsagen til sygdommen. Men mærkeligt nok har der ikke været nogen som helst fremskridt i at forstå fænomenet. Man er stadig ikke kommet længere end det beskrivende niveau.

I mange år har en række af de store medicinalvirksomheder været på jagt efter et stof, der kan hæmme – og selvfølgelig i bedste fald helt stoppe – udviklingen af Alzheimer, men uden held. Det er så omkostningsfyldt, at virksomhederne naturligvis ikke afslører, hvad de har gang i. Men gennem de over 100 år har der tilsyneladende kun været ét overordnet mål for forskningen, nemlig at finde ud af, hvordan man kan hæmme opbygningen af den omtalte beta-amyloid-plak, og allerhelst direkte nedbryde det.

Og midt i 2021 lykkedes det for det amerikanske medicinfirma Biogen at få godkendt en nyudviklet medicin, der har fået navnet Aducanumab, og som ifølge firmaet netop leverer indenfor skiven: Det nedbryder det ophobede beta-amyloid-plak. Godkendelsen skete i et lyntempo, inspireret af godkendelsen af vaccinerne mod Covid – en procedure, der har mødt voldsom kritik. Men de amerikanske myndigheder har naturligvis været under et voldsomt pres fra både firma og pårørende.

Der er nu også sat beløb på, hvad medicinen kommer til koste: 56.000 \$ pr patient pr år. Man vurderer, at der er et potentiale af patienter i de velstående lande (som Danmark), der vil gøre dette præparat til det mest indtjenende nogensinde. I USA er der skønsmæssigt 6 millioner, der lider af Alzheimer i en eller anden grad. Kun de rigeste 5% regner man med vil få adgang til medicinen, dvs. 300.000. Det vil give en årlig indtjening alene fra USA på 17 milliarder dollars om året. Det er så mange penge, at det let kan komme til at overtrumfe faglige indvendinger.

Den mest basale indvending, som indtil for nylig *ikke* har været rejst, er spørgsmålet: Er der i hele forskningsindsatsen hidtil sket en forveksling af korrelation og kausalitet? Du kan finde en artikel, der adresserer dette problem, knyttet til øvelse 3.17 i kapitel 3. Ingen betvivler Alzheimers fund – det er bekræftet ved utallige andre obduktioner. Men at disse proteiner findes, er jo ikke et bevis på, at de udgør problemet. Et helt nyt dansk studie rejser dette spørgsmål, og mere end det: De har formuleret en hypotese om, at beta-amyloid-plak er udviklet som hjernens *forsvar* mod et angreb fra noget helt andet! Dvs de mener, at disse proteiner faktisk hæmmer udviklingen af Alzheimer, indtil kroppen må give op. Men hvad der så er årsagen, har de heller ikke et præcist bud på.

Hvis det viser sig, at de danske forskere har ret, så vil det faktisk accelerere Alzheimer at slippe medikamenter løs, der skal opløse beta-amyloid-plak! Når der således er åbnet op for en mulig ny Thalidomid-agtig skandale, så skyldes det altså, at myndighederne har accepteret en forcering af godkendelsesproceduren for dette stof.

Øvelse 6.5

Hent artiklerne fra Nature og Politiken, som du finder her – [link til 31](#).

- a) I artiklen kan vi læse, at firmaet havde opgivet præparatet ud fra et første fase 3 forsøg. Men så sker der noget: *But months later, the biotech firm brought the antibody back from the brink, after inspecting the data more closely*. Hvad er det, der gør, at de igen tager fat på det?
- b) Giv et referat af den efterfølgende diskussion. Hvad menes med formuleringen, der sammenligner måden data er analyseret på med følgende: "firing a shotgun at a barn and then painting a target around the bullet holes"?
- c) Det uafhængige ekspert-panel anbefaler med stemmerne 10-1, at den nye medicin *ikke* skal godkendes. Alligevel bliver den godkendt. Giv en vurdering af årsagerne hertil.
- d) I artiklen refereres også en diskussion blandt forskere om konsekvenserne af godkendelsen af Aducanumab. Nogle mener, det vil føre til en stagnation i forskningen vedr. Alzheimer-sygdommen, andre mener det vil booste forskningen. Præsenter og diskuter de to siders argumentation.
- e) Læs artiklen i Politiken om den nye danske forskning. Hvad er deres centrale pointe. Er det en opdagelse? Eller er det en begrundet hypotese? Eller er det blot en ny ide, de ønsker undersøgt?
- f) Antag at de danske forskere har helt eller delvist ret. Diskuter konsekvenserne heraf.

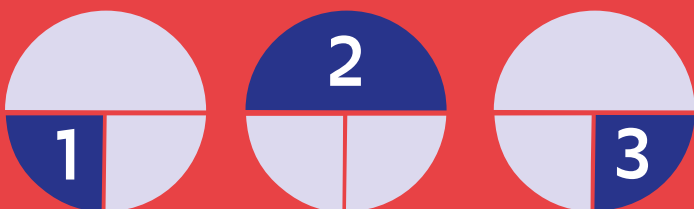
7.

Testteori 1.

Standardfordelingerne



Den matematiske statistik blev grundlagt i første del af 1900-tallet af originalt tænkende matematikere som Karl Pearson, William Gosset og Ronald Fisher. I et af sine hovedværker om, hvordan man skal designe eksperimenter, sammenfatter Ronald Fisher begreberne i hypotestest i en lille fortælling, der siden blev udgivet selvstændigt og har fået ikonisk status: A Lady Tasting Tea. Med denne fortælling introducerer vi begreberne i testteori, og gennemgår dernæst standardfordelingerne og testmetoderne for hver enkelt af disse.



7. Testteori 1. Standardfordelingerne

Når vi med **statistiske** metoder skal undersøge om en given påstand kan accepteres eller må forkastes, er det fordi vi ikke har muligheder for at give et **eksakt** svar. Om et bestemt stykke fyrværkeri virker, er der et eksakt svar på, men det er meningsløs at afprøve alle. Om et medicinsk præparat virker bedre end et andet, er der ikke et lige så præcist svar på – de enkelte tager forskelligt imod medicin, som Lea forklarer. Så vi må her inddrage en større gruppe, men det er klart ikke en mulighed at afprøve det på alle. Og der er ikke nødvendigvis - som med fyrværkeri - et *objektivt* kriterium for, hvornår man kan sige, at det ene virker bedre end det andet. Vi må *selv* opstille nogle mål for, hvornår vi vil drage den konklusion.

Det begrebsapparat og de metoder, der er udviklet inden for statistik, er kun omkring 100 år gammel. Selv om ny matematik bygger oven på, hvad andre før har skabt, så kan vi faktisk her datere, hvornår de metoder, der anvendes i filmen, blev præsenteret første gang i bogform.

7.1 Projekt om det statistiske begrebsapparat: A lady Tasting Tea

(Formen i dette afsnit lægger op til, at der arbejdes selvstændigt med det, individuelt eller i små grupper)

"A LADY declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup: We will consider the problem of designing an experiment by means of which this assertion can be tested"

Sådan indleder en af det 20. århundredes mest berømte statistikere, Ronald Fisher den artikel, hvor han populært præsenterede den moderne statistiks fundamentale begreber: Nulhypotese, randomisering, signifikansniveau mv.

Fisher lagde grundlaget for moderne statistik i en række værker i 1920'erne og 30'erne, først og fremmest *Statistical Methods for Research Workers* fra 1925, og *The Design of Experiments* fra 1935. Det er i den sidste han første gang fortæller historien.

Det er som indledning til afsnit II i bogen omhandlende "The principles of Experimentation....". Han var en fremragende formidler, og valget af denne case, har eftertiden vist, var et scoop. Med den historie formår han at præsentere de svære begreber, som man kan se her i indholdsfortegnelsen.

CONTENTS	
I. INTRODUCTION	
	PAGE
1. The Grounds on which Evidence is Disputed	1
2. The Mathematical Attitude towards Induction	3
3. The Rejection of Inverse Probability	6
4. The Logic of the Laboratory	8
II. THE PRINCIPLES OF EXPERIMENTATION, ILLUSTRATED BY A PSYCHO-PHYSICAL EXPERIMENT	
5. Statement of Experiment	13
6. Interpretation and its Reasoned Basis	14
7. The Test of Significance	15
8. The Null Hypothesis	18
9. Randomisation; the Physical Basis of the Validity of the Test	20
10. The Effectiveness of Randomisation	22
11. The Sensitiveness of an Experiment. Effects of Enlargement and Repetition	24
12. Qualitative Methods of increasing Sensitiveness	26
III. A HISTORICAL EXPERIMENT ON GROWTH RATE	
13.	30
14. Darwin's Discussion of the Data	31
15. Galton's Method of Interpretation	32
16. Pairing and Grouping	35
17. "Student's" t Test	38
18. Fallacious Use of Statistics	43
19. Manipulation of the Data	44
20. Validity and Randomisation	46
21. Test of a Wider Hypothesis	50

1. Summary af casen

Start med at lytte til følgende fremstilling af, hvad historien med *A lady Tasting Tea* handler om: <http://www.youtube.com/watch?v=lgs7d5saFFc>

En i gruppen giver en kort mundtlig fremstilling af, hvad indholdet i videoen var. De øvrige supplerer. I fællesskab noterer I 5 centrale stikord.

2. Eksperimentet

Fishers fortælling blev udgivet særskilt, og I kan hente den her – [link til 32](#). Læs den første side og giv et referat af Fishers diskussion af, hvordan et eksperiment skal designes.

3. Signifikans

På side 2 introducerer Fisher begrebet: Signifikans.

Gennemfør først en diskussion i gruppen om, hvordan I vil forklare begrebet, og hvordan I bruger begrebet i praksis. Er begrebet ligetil at anvende, eller er der her særlige fælder, man skal undgå? Læs så side 2 i artiklen. Hvordan stemmer Fishers beskrivelse overens med jeres?

Hvis I vil dykke dybere ned i spørgsmålet: Hvor kommer de 5% fra i signifikansniveauet, så kan I inddrage afsnit 7.6.

4. Nulhypotesen

På side 2-3 introducerer Fisher begrebet: Nulhypotese.

Gennemfør først en diskussion i gruppen om, hvordan I vil forklare begrebet, og hvordan I bruger begrebet i praksis. Er begrebet ligetil at anvende, eller er der her særlige fælder, man skal undgå? Læs så side 2-3 i artiklen. Hvordan stemmer Fishers beskrivelse overens med jeres?

5. Randomisering

På side 3-4 diskuterer Fisher begrebet: Randomisering

Læs afsnittene og forklar, hvad der menes hermed.

Overvej, om det er noget I er stødt på før i jeres eget arbejde med statistik – inddrag gerne andre fag.

Er det noget, I har hørt om, fx i forbindelse med udvikling af ny medicin / vacciner?

Opstil selv nogle vanskeligheder ved at gennemføre en fuldstændig randomisering.

6. Fishers metode til at udregne svaret på nulhypotesen

Fisher sætter forsøget i gang ved at stille de 8 kopper i en af ham valgt "tilfældig" rækkefølge. The Lady får oplyst, at der er 4 kopper, hvor der først er kommet mælk i og 4 kopper, hvor der først er hældt te i. Vis, at der er 70 forskellige muligheder for rækkefølgen af kopperne med 'først mælk' og 'først te'.

Nulhypotesen er, at hun ikke kan smage forskel.

Lad os fokusere på 'først mælk': Når hun har angivet de 4 kopper, hvor hun mener der er først hældt mælk i, så er de øvrige selvfølgelig også fastlagt. Hendes succes-muligheder er 0, 1, 2, 3 eller 4. Argumenter for, at vi under nulhypotesen får følgende skema:

Tea-Tasting Distribution Assuming the Null Hypothesis		
Success count	Combinations of selection	Number of Combinations
0	oooo	$1 \times 1 = 1$
1	ooox, ooxo, oxoo, xooo	$4 \times 4 = 16$
2	ooux, oxox, oxox, xoxo, xxoo, xoox	$6 \times 6 = 36$
3	oxxx, xoxx, xxox, xxxo	$4 \times 4 = 16$
4	xxxx	$1 \times 1 = 1$
	Total	70

som vi har lånt fra Wikipedias artikel om denne case.

Sandsynligheden for at få en succes på 4 ud af 4 ved et tilfældigt gæt er $\frac{1}{70} \approx 0.014$, dvs langt under 5%. Så nulhypotesen forkastes.

8 kopper er ikke meget til et forsøg, men der er jo grænser for hver meget te, vores Lady kan drikke. Hvordan ville I vurdere – uden at regne - hendes evner til at smage forskel, hvis hun gætter 3 ud af 4 rigtig? Foretag dernæst en udregning ud fra skemaet og kommenter.

7. Forbedringer af eksperimentet?

I den sidste del af artiklen diskuterer Fisher, hvordan eksperimentet kunne forbedres.

Har I selv nogle bud?

Hvis I har gennemgået teorien om *stikprøver*, kan I inddrage dette, herunder hvilken rolle *antallet* spiller i en stikprøve.

Hvis I selv har diskuteret, hvordan eksperimenter designes, kan I inddrage dette, fx begreberne *blindede* og *dobbeltblindede* forsøg. Er dette altid muligt at inddrage i eksperimenter?

8. Design selv et tilsvarende eksperiment med smageøvelser

Fx af forskellige typer af cola, og analysér resultaterne af dit eksperiment.

9. Hvad Fisher ikke omtaler – Pearsons teststørrelse

Har man lært om Chi-i-anden fordelingen, ville man opstille nulhypotesen: Der er ingen sammenhæng mellem rækkefølge og smag, og opstille disse skemaer:

Observerede (O)	Gætte: først mælk	Gætte: først te	I alt
Først mælk	4	0	4
Først te	0	4	4
I alt	4	4	8

Forventede (F)	Gætte: først mælk	Gætte: først te	I alt
Først mælk	2	2	4
Først te	2	2	4
I alt	4	4	8

Vi tester for uafhængighed, og udregner teststørrelsen til 8. Det giver en p-værdi på 0,0047, så nulhypotesen forkastes. Opstilling af en sådan teststørrelse var kendt af Fisher, den var introduceret af en anden stor statistiker Karl Pearson omkring år 1900.

- opstil formlen for teststørrelsen, generelt og med tabellens tal
- vis at værdien bliver 8
- hvad forstår vi ved *p-værdien*?

Den metode Fisher præsenterer er siden blevet betegnet *Fishers eksakte test*. Den er betydeligt mere besværlig at bruge, både mht. beregninger og mht. opstilling af antal muligheder. Selvom Fishers test som navnet siger er eksakt, så foretrækkes derfor normalt chi-i-anden test.

7.2 Introduktion af hypotesetest for folkeskolen

Hvis du i folkeskolen eller i starten af en 1. g klasse ønsker at arbejde med hypotesetest, så kan du hente et projektoplæg her – [link til 33](#). Projektet er skrevet, så begreber som fx boksploot anvendes.

7.3 Chi-i-anden test og Framingham Heart Study

Du kan i et projekt i HEM2, som kan hentes her: [Hvad er matematik? 2 - LRU.dk \(praxis.dk\)](#), finde en grundig og fyldestgørende introduktion til chi-i-anden testet.



Chi-i-anden testet anvendes til at afgøre spørgsmål om *kategoriske variable* som: Er der signifikant forskel på antal piger og drenge, der dumper ved en bestemt eksamen. Eller: Er der en sammenhæng mellem blodtyper og type 1 diabetes? Eller: Kan man smage forskel på Coca cola og Pepsi cola? Eller den case vi så på i 7.1.

Vi vil her give en kort introduktion båret af et datasæt fra **Framingham Heart Study**. Framingham Heart Study er et blandt flere store såkaldte kohorte studier, der blev sat i gang efter 2. verdenskrig, hvor man følger en bestemt befolkningsgruppe, her indbyggerne i den lille by Framingham i Massachusetts, USA, over en lang tidsperiode. Hensigten er at kunne studere langtidsvirkninger af en række variable, der muligvis påvirker sundhedstilstanden. Det kan være rygning, arbejdssituation, fysisk aktivitet, spisevaner, men det kan også være genetiske forhold og meget andet. Du kan finde en beskrivelse af studiet og nogle af de resultater, der er kommet ud af det, [her -link til 34](#)

I datasættet indgår oplysninger for 4240 personer. Hent det her - [link til 35](#)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
2	1	39.4		0.0	0		0	0	0.195	106	70.26.97	80.77				0
3	0	46.2		0.0	0		0	0	0.250	121	81.28.73	95.76				0
4	1	48.1		1.20	0		0	0	0.245	127.5	80.25.34	75.70				0
5	0	61.3		1.30	0		0	1	0.225	150	95.28.58	65.103				1
6	0	46.3		1.23	0		0	0	0.285	130	84.23.1	85.85				0
7	0	43.2		0.0	0		0	1	0.228	180	110.30.3	77.99				0
8	0	63.1		0.0	0		0	0	0.205	138	71.33.11	60.85				1
9	0	45.2		1.20	0		0	0	0.313	100	71.21.68	79.78				0
10	1	52.1		0.0	0		0	1	0.260	141.5	89.26.36	76.79				0
11	1	43.1		1.30	0		0	1	0.225	162	107.23.61	93.88				0
12	0	50.1		0.0	0		0	0	0.254	133	76.22.91	75.76				0
13	0	43.2		0.0	0		0	0	0.247	131	88.27.64	72.61				0
14	1	46.1		1.15	0		0	1	0.294	142	94.26.31	98.64				0
15	0	41.3		0.0	1		0	1	0.332	124	88.31.31	65.84				0

Vi er interesseret i at undersøge, om der er en sammenhæng mellem køn og diabetes.

Nulhypotesen er derfor: H_0 : Der er ingen sammenhæng mellem køn og diabetes.

Køn angives i kolonne A og diabetes i kolonne I. Sorter oplysningerne mht. køn, og derefter indenfor hver gruppe mht. diabetes, så du kan udfylde en tabel som denne (har du lært at lave pivottabeller, så gør gerne det):

Observerede	diabetes	ikke diabetes	I alt
mænd			1820
kvinder			2420
I alt			4240

a) Udregn kolonnesummer

b) Formuler nulhypotesen

Ideen i Chi-i-anden test er nu, at vi vil udregne de forventede værdier ud fra antagelsen af nulhypotesen. Hvis der ingen kønsmæssig forskel er, så ville vi forvente, at der var *samme procentandel* der havde diabetes blandt mænd og kvinder, og tilsvarende samme procentandel, der ikke havde det.

c) Udregn hvor stor en del af *hele populationen*, der har diabetes.

d) Udfyld nu denne tabel:

Forventede	diabetes	ikke diabetes	I alt
mænd			1820
kvinder			2420
I alt			4240

Hvis nulhypotesen holder, så skulle der ikke være særlig stor forskel på de to tabeller. Men selvfølgelig kan der være mindre tilfældige forskelle. Hvordan måler vi så forskellen på de to tabeller? Vi ønsker forskellen udtrykt i ét tal, som vi kalder *teststørrelsen*. Formlen for denne går tilbage til Karl Pearson, der sammen med Ronald Fisher grundlagde en væsentlig del af metoderne i den bekræftende statistik, som anvendes i dag.

Teststørrelsen, som Pearson introducerede i år 1900 er:

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - F_{ij})^2}{F_{ij}},$$

hvor vi summerer over alle tabellens celler (*i* og *j* angiver række og søjlenummer)

e) - Hvad ville χ^2 blive, hvis der ikke blev kvadreret i tælleren?

- Hvis der ikke blev divideret med F_{ij} , ville $(20 - 10)^2$ og $(1020 - 1010)^2$ give samme bidrag til teststørrelsen χ^2 . Diskuter det rimelige eller urimelig heri.

- Bestem teststørrelsen.

Et stort tal betyder, at tabellerne ligger "langt" fra hinanden, og at nulhypotesen derfor ikke holder. Men hvad er et *stort tal*? I nogle situationer har vi med større *tabeller* at gøre. Og det er klart, at større tabeller normalt vil give større tal som teststørrelse. Når vi vurderer, om det udregnede tal er så stort, at nulhypotesen må forkastes, så skal vi altså tage hensyn til, hvor mange søjler og rækker der er i tabellen. Dette gør man ved at tælle "*frihedsgrader*":

I en tabel som den ovenfor over *observerede* værdier, har vi bestemte *kolonnesummer* og *rækkesummer* – fx er de sidste henh. 1820 og 2420

Tabellen over de *forventede* værdier har *samme* kolonne- og rækkesummer.

Lad os se på et tænkt eksempel med samme kolonne- og rækkesummer, men andre observerede værdier. Tallene i de 4 celler kan ikke udfyldes frit! For har vi et bestemt antal mænd med diabetes, fx 210, så er der åbenbart 1610, der ikke har det. Og kolonnesummen giver samtidig, hvor mange kvinder der har diabetes – og dermed hvor mange der ikke har.

f) Gennemfør dette argument i detaljer.

Der var altså her ét tal vi kunne vælge frit: **Vi siger der er 1 frihedsgrad.**

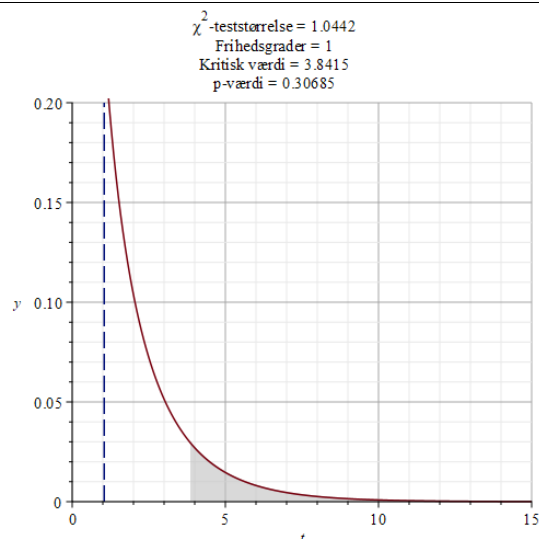
g) Argumenter for, at havde det være en tabel med 3 kolonner, så ville vi have 2 frihedsgrader.

Nulhypotesen forkastes, hvis udregningen af teststørrelsen giver et usædvanligt resultat. Hvad der er *usædvanligt* fastlægges af **signifikansniveauet**, som vi vælger til 5%. Dette betyder, at hvis vi gennemførte fx 10000 simuleringer af, hvordan data kan fordele sig i cellerne, når række- og kolonnesummerne er som angivet, og hver gang udregnede teststørrelsen, *så betegnes de 5% største tal som usædvanlige*.

Du kan finde et materiale om simulering af en chi-i-anden fordeling med anvendelse af dette eksempel fra Framingham Heart Study. Hent det [her – link til 36](#)

Antallet af frihedsgrader fastlægger, hvilken χ^2 -graf, vi skal anvende til at vurdere teststørrelsen. Værktøjsprogrammerne har normalt en indbygget kommando, der ud fra data både tegner den relevante graf og viser grænsen for, hvornår en teststørrelse betyder, at nulhypotesen forkastes eller accepteres.

h) Gennemfør et χ^2 -test på de observerede data. I Maple vil det se ud som tegningen her. Den stiplede linje angiver *teststørrelsen*, vi udregnede i punkt e). Den *kritiske værdi* angiver, at når teststørrelsen er herover forkastes nulhypotesen. *p-værdien* angiver sandsynligheden for at få en χ^2 -værdi på 1,0442 eller større. Nulhypotesen kan altså ikke forkastes.



k) Hent regnearket og formuler selv et spørgsmål du vil undersøge, som fx:

- Er der en sammenhæng mellem rygning og
- Er der en sammenhæng mellem BMI og ...

7.3.1 Sandsynligheden for at begå fejl

Formålet med et så imponerende stort studie som Framingham Heart Study er at samle viden om *hele populationens* sundhedstilstand målt på en lang række variable. Det ville være alt for kostbart og ressourcetrævendende at *foretage* denne undersøgelse på hele populationen. Derfor udføres studiet på en *stikprøve* med brug af en række statistiske metoder. Og derfor kan vi heller ikke få eksakte svar.

Der er altid en risiko for at begå fejl, ikke fordi vi ikke er dygtige nok, men fordi dette er indbygget i den statistiske metode. Vi fastlægger et signifikansniveau på fx 5%, og det betyder, at hvis vi gennemførte tusindvis af forsøg, og hver gang udregnede teststørrelsen, så ville de 5% største af disse værdier blive vurderet som så usædvanlige, at de ville medføre, at vi forkastede nulhypotesen. Men disse "usædvanlige" resultater indtræffer jo af og til – faktisk præcis i 5% af tilfældene!

Sagt på en anden måde betyder dette, at *i 5% af tilfældene forkaster vi en nulhypotese, der er sand*. Dette kaldes for en **fejl af type I**.

En anden situation, der også ville medføre en fejlslutning er følgende: Nulhypotesen er faktisk *falsk*, der er en sammenhæng mellem køn og diabetes, men en stikprøve, der er taget i Framingham viser det modsatte: Teststørrelsen ligger langt under den kritiske værdi, vi er tydeligt på "den sikre side", og vi konkluderer, at nulhypotesen ikke kan forkastes. Vi accepterer således den falske nulhypotese. Dette kaldes for en **fejl af type II**. Det er betydeligt sværere at beregne sandsynligheden for at begå denne fejl. Vi kan ikke bare bygge på de tidligere beregninger eller grafer, for de er jo beregnet og tegnet ud fra antagelsen, at nulhypotesen er sand.

Overvej selv: Når sandsynligheden for at begå en type 1 fejl er lig med signifikansniveauet på 5%, hvorfor sætter vi så ikke blot dette niveau meget lavt, fx til 1%, eller måske 0,1%?

7.4. Binomialmodel og binomialtest – beregning og simulering

Betragt følgende opgavetyper:

1. *En influenzaepidemi nærmer sig landet. Især ældre rammes hårdt og der har allerede været mange dødsfald i andre lande. I en by overtaler embedslægen alle 400 beboere på byens plejehjem til at lade sig vaccinere med en vaccine, der ifølge test slår an for 98,5% af alle vaccinerede. Hvad er sandsynligheden for, at der højst er 10 ældre, for hvem vaccinen ikke slår an?*
2. *Bestemte tulipanløg sælges med en spiringsgaranti på 90%. Der indkøbes og udplantes 50 løg. Hvad kan vi forvente til foråret? Vi ønsker svaret i form af en sandsynlighedstabel, der viser, hvad sandsynligheden er for at alle 50 spirer, og for at 49, 48, 47,..40, 39,... osv. løg spirer? Er det en usædvanlig hændelse, der bør give anledning til at klage, hvis kun 40 løg spirer?*

Selv om sproget og situationen er forskellig, er det alligevel ikke svært at se nogle fælles træk i de to eksempler. Det springer også i øjnene, at det minder om spilsituationer, som fx sandsynligheden for at få 10 rigtige ud af 13, når vi sætter krydserne tilfældigt på en tipskupon. Eller sandsynligheden for at få 4 seksere, når vi slår 12 gange med en terning. De matematiske modeller, vi opstiller til at beskrive sådanne situationer, kaldes for *binomialmodeller*.

At plante løg, at slå med en terning, at vaccinere en ældre eller at sætte kryds på en kupon kaldes alt sammen for *et eksperiment*. Den slags eksperimenter, vi her taler om, adskiller sig fra eksperimenter i fysik og kemi, hvor vi normalt forventer at kunne forudsige resultaterne. De omtalte eksperimenter har noget *tilfældigt* over sig. Ikke på den måde, at det er tilfældigt, om en vaccine slår an, men vi kan ikke vide det med sikkerhed. *Tilfældigt* er derfor ikke et godt ord, og i statistik og sandsynlighedsregning indfører vi en særlig betegnelse:

Praxis: Stokastisk eksperiment

Et eksperiment som et terningekast, hvor vi ikke kan forudsige udfaldet, kalder vi et *stokastisk eksperiment*.

Bemærk: Resultatet af et terningekast er principielt deterministisk, men vi kan ikke skaffe os viden til at kunne forudsige udfaldet.

Det helt særlige ved **binomialmodeller** er, at *det samme eksperiment gentages* en række gange. Antallet af gange forsøget gentages betegnes ofte med n . Vi antager at de enkelte eksperimenter ikke påvirker hinanden, så der er tale om samme eksperiment. Det er ikke altid let at sikre dette, men det er en vigtig faktor i designet af sådanne forsøg. Når det er tilfældet, bruger vi ofte ordet *basiseksperiment* eller *Bernoulliforsøg*.

Øvelse 7.1 Afhængige eksperimenter

Giv eksempler på gentagne eksperimenter, der ikke er uafhængige.

Resultatet af et sådant Bernoulliforsøg kalder vi for *et udfald*. Det er yderligere et særligt træk ved binomialmodellerne, at udfaldene kan deles op i to typer: Vaccinen slår an eller slår ikke an. Løget spirer eller spirer ikke. Vi får en sekser, eller vi får ikke en sekser. Verden er naturligvis meget mere nuanceret, vaccinen slog delvis an, løget spirede men gav en sølle plante osv. Men vi abstraherer fra alt det og deler op i to typer, som vi generelt kalder for *held* og *uheld*.

Til resultaterne held og uheld er der knyttet bestemte *sandsynligheder*. Sandsynligheden for held i et basiseksperiment kaldes for *basis-sandsynligheden*, og betegnes ofte p . Hvis sandsynligheden for held er p , så må sandsynligheden for uheld være $1 - p$, da der kun er de to muligheder.

Øvelse 7.2 Parametre i en binomial model

Angiv hvad n , p og $1 - p$ er i eksemplerne med:

- Influenzavaccinationen
- udplantningen af løg
- udfyldning af en tipskupon
- kast med en terning.

Vi er ofte interesseret i spørgsmål som: Hvor mange gange får vi held ud af de n mulige? Det kan vi jo ikke vide med sikkerhed, men for at beskrive det indfører vi en *stokastisk variabel* X , der angiver antallet af gange, vi får held i forsøget. Gentager vi forsøget, vil antallet kunne variere mellem 0 og n . Dvs. X antager værdier mellem 0 og n .

Vi sammenfatter nu:

Praxis: Binomialmodellen

Et stokastisk forsøg kan beskrives ved hjælp af en *binomialmodel*, når følgende er opfyldt:

- I forsøget *gentages* et bestemt *basiseksperiment*, eller Bernoulliforsøg i alt n gange.
- Resultaterne af de enkelte basiseksperimenter påvirker ikke hinanden. De siges at være *uafhængige*.
- Resultatet af basiseksperimentet kan opdeles i to typer, som vi betegner henholdsvis *held* og *uheld*.
- Sandsynligheden for held betegnes p og kaldes for *basissandsynligheden*. Sandsynligheden for uheld er $1 - p$.
- Den *stokastiske variabel* X angiver antallet af gange, vi i et forsøg får held.
- Sandsynligheden for at opnå held k gange betegnes med symbolet: $P(X = k)$.
- Tallene n og p kaldes også for henholdsvis *antalsparameteren* og *sandsynlighedsparameteren*.
- At et stokastisk forsøg kan beskrives med en binomialmodel, udtrykkes kort således: $X \sim b(n, p)$

Der findes en formel til udregning af $P(X = k)$, når X er binomialfordelt, $X \sim b(n, p)$. Formlen kan du finde i enhver matematikbog. Vi vil her udnytte, at ethvert matematisk værktøjsprogram har indbygget denne formel. Hvis $X \sim b(50, 0.90)$, så er notationen i eksempelvis Maple:

$$P(X = 45) = \text{binpdf}(50, 0.90, 45) = 0,1849$$

Dvs.: Sandsynligheden for at præcis 45 løg spirer med den givne spiringsgaranti, er 18,5%.

Det var ikke meget, kan man sige, men vi er jo heller ikke interesseret i om præcis 45 spirer, men derimod i at *mindst* 45 spirer. Dette udregnes i Maple med kommandoen:

$$P(X \geq 45) = \text{bincdf}(50, 0.90, 45, 50) = 0,6161$$

Øvelse 7.3. Terningkast

Vi kaster 12 gange med en terning. X angiver antal seksere. Angiv n og p i $X \sim b(n, p)$.

Hvad er sandsynligheden for at få præcis 2 seksere?

Hvad er sandsynligheden for at mindst 2 seksere?

Øvelse 7.4. Antal danskere med diabetes

Illustrationen og oplysningerne er hentet fra Diabetesforeningens web-site:

[Diabetes i Danmark | Viden og forskning | Diabetesforeningen](#)

- 252.000 danskere har type 2-diabetes (inklusive MODY og LADA)
- 28.000 danskere har type 1-diabetes

a) Slå op på nettet og find ud af, hvad de forskellige kategorier af diabetes betyder.

b) Forklar illustrationen med de blå, sorte osv mennesker.

c) Vi vælger en tilfældig dansker. Hvordan vil du estimere sandsynligheden for, at vi har trukket en med diabetes 1? Samme spørgsmål med diabetes 2?

d) I en forsamling på 500 tilfældigt udtrukne danskere, hvad er dér sandsynligheden for, at mindst 25 har diabetes?



- Type 2-diabetes ● Type 1-diabetes ● Type 1½-diabetes
- MODY (tælles ofte med under type 2-diabetes)

Fordeling på diabetestyper

- Type 1-diabetes udgør ca. 10 %
- Type 1½-diabetes udgør ca. 5-10 % (tælles ofte med under type 2-diabetes)
- Type 2-diabetes udgør ca. 80 %
- MODY udgør ca. 1-2 % (tælles ofte med under type 2-diabetes)

Middelværdien af en stokastisk variabel kaldes også forventningsværdien. Kaster vi en terning, er der hver gang $\frac{1}{6}$ chance for at få en sekser. Kaster vi 24 gange vil vores forventning være, at vi får omkring 4 seksere ($= \frac{1}{6} \cdot 24$), men vi kan selvfølgelig ikke være sikker.

Generelt gælder der: Hvis $X \sim b(n, p)$, så er *formlen for middelværdien*: $\mu = n \cdot p$

De enkelte sandsynligheder: $P(X = k)$, hvor k løber fra 0 til n , fordeler sig jævnt omkring middelværdien, og som i den deskriptive statistik kan vi udregne *spredningen* af disse dataværdier. For en binomialfordeling er der en særlig simpel formel:

Spredningen af en stokastisk variabel $X \sim b(n, p)$ kan udregnes med formelen: $\sigma = \sqrt{n \cdot p \cdot (1 - p)}$

(*Bemærkning*: Du kan finde et bevis for disse formler i projekt 9.9, i HEM2, se her: [Hvad er matematik? 2 - LRU.dk \(praxis.dk\)](#))

Øvelse 7.5. Grafen for en binomialfordeling

a) Undersøg, hvordan dit værktøjsprogram kan tegne grafer (pindediagrammer) over en binomialfordeling, $X \sim b(n, p)$, og fremstil dernæst en graf som den i billedet til højre.

b) Bestem middelværdi og spredning.

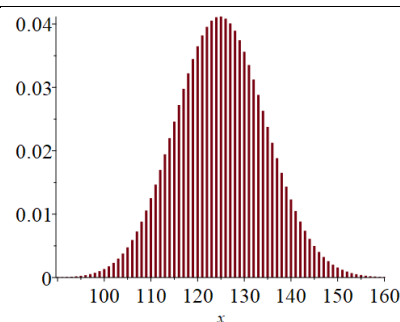
c) Fremstil grafer over $X_1 \sim b(500, 0.05)$ og $X_2 \sim b(500, 0.95)$

d) Udregn spredningen for de tre fordelinger.

e) Marker intervallerne $[\mu - 2 \cdot \sigma; \mu + 2 \cdot \sigma]$ på x-aksen for hver af de tre, fx med lodrette linjer.

f) Udregn $P(\mu - 2 \cdot \sigma \leq X \leq \mu + 2 \cdot \sigma)$ for hver af de tre.

g) Hvad iagttager du? Formuler en konklusion ud fra de tre



Pindediagram hørende til $b(500, 0.25)$

Grafen "ligner" grafen for en normalfordeling. Mon det altid gælder?

I filmen omtaler Lea kort opinionsmålinger, som en introduktion til en diskussion af usikkerhed og en beregning af konfidensinterval. Om et parti går frem eller tilbage bliver ikke altid præsenteret

med angivelse af den usikkerhed der må være. Vi giver et eksempel herpå, som også illustrerer brugen af binomialmodeller.

Praxis

Vi så i øvelse 7.5, at hvis empiriske data fordeler sig nogenlunde som en binomialfordeling, så vil ca 95% af alle ligge i intervallet $[\mu - 2 \cdot \sigma; \mu + 2 \cdot \sigma]$, hvor μ og σ er middelværdi og spredning.

Dette betyder også, at skal vi teste en *nulhypotese* om en binomialmodel, og har vi fastlagt et *signifikansniveau* på 5%, så vil vi anvende intervallet $[\mu - 2 \cdot \sigma; \mu + 2 \cdot \sigma]$ som vores *konfidensinterval*. Begreberne nulhypotese og signifikansniveau er forklaret i afsnit 7.1. Konfidensinterval forklares i eksemplet og øvelsen nedenfor

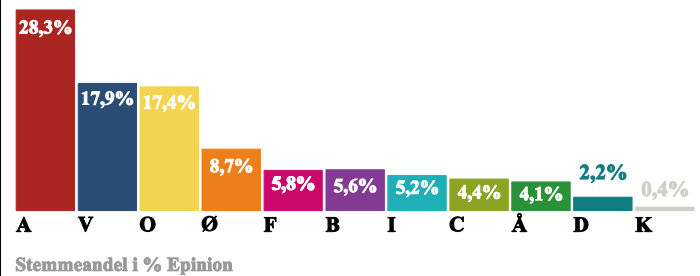
7.4.1 Eksempel og øvelse om opinionsmålinger

På figuren ses tallene fra en meningsmåling foretaget af Epinion d. 26.1.2018, hvor 1762 personer blev spurgt.

Ved folketingsvalget i 2015 fik liste Å (Alternativet) 4,8% procent af stemmerne.

a) Benyt et 95%-konfidensinterval til at afgøre, om andelen af personer, der stemmer på liste Å, har ændret sig signifikant.

MENINGSMÅLING D. 26. JANUAR 2018



Nulhypotese: Vi formulerer vores nulhypotese ud fra vores viden om populationen, dvs. alle der stemte på liste Å ved valget i 2015. Dermed har vi $p = 0,048$. I dette tilfælde er vores nulhypotese tosidet, da vi både kan få en tilbagegang eller en fremgang i andelen af stemmer. Vores **stikprøve** er de 1762 personer, som blev spurgt af Epinion den 26/1-2018.

Binomialfordeling: Vi har en stokastisk variabel X , der angiver antallet af personer, der stemmer på liste Å. Basissandsynligheden er $p = 0,048$. Den stokastiske variabel X er dermed binomialfordelt med $n = 1762$ og $p = 0,048$.

Konfidensintervallet for p er ved et signifikansniveau på 5%:

$$\left[0,048 \cdot 1762 - 2 \cdot \sqrt{0,048 \cdot (1 - 0,048) \cdot 1762} ; 0,048 + 2 \cdot \sqrt{0,048 \cdot (1 - 0,048) \cdot 1762} \right] = [66,6 ; 102,5]$$

Nu har vi opgivet procenttal og ikke absolutte tal, så vi omregner ovenstående til procenter og får, at konfidensintervallet er $[0,0378 ; 0,0582]$

Konklusion: Da stemmeandelen for liste Å i meningsmålingen er 0,041, så ligger værdien i konfidensintervallet, og dermed er der ikke belæg for, at stemmeandelen for liste Å har ændret sig signifikant.

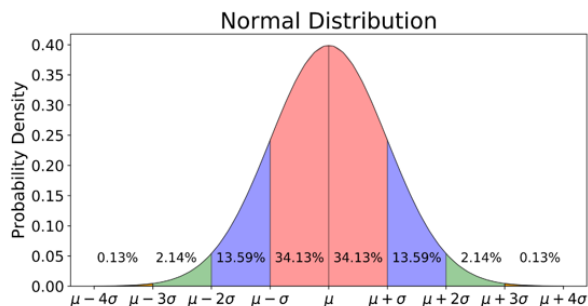
7.4.2 Simulering af en binomialmodel

Formlerne er lette at bruge i en binomialmodel, men det kan ofte være instruktivt at undersøge en nulhypotese eksperimentelt. Du kan her - [link til 37](#) - finde en opskrift på, hvordan man kan gennemføre dette i Maple.

7.5. Normalfordelingen og konfidensintervaller

Anvendelsen af normalfordelingen i hypotesetest bygger på denne grafiske fremstilling af, hvad der kendetegner normalfordelingen:

Grænserne, fx at 95% af de observerede i en normalfordeling ligger indenfor et område, der er bestemt af middelværdien ± 2 spredninger, er "tommefingerregler".



På illustrationen kan man se, at der ligger 4,28% udenfor området. Det præcise interval kan bestemmes i dit værktøjsprogram. I Maple opstilles fx ligningen:

$$\text{normalcdf}(0,1, -\infty, x) = 0,975, \text{ der giver: } x = 1,96,$$

så symmetrien giver, at for standardnormalfordelingen er 95%- intervallet: $[-1,96; 1,96]$

Øvelse 7.6

Overvej hvordan en ligning til at bestemme den øvre grænse for "det normale" område kan opstilles i dit værktøjsprogram, og løs ligningen.

Når vi opholder os ved dette, skyldes det naturligvis, at det *signifikansniveau* vi normalt vælger, når vi skal teste en nulhypotese, er 5%. Det betyder, at det *konfidensinterval*, vi udregner, for en *standardnormalfordeling* vil være lige præcis: $[-1,96; 1,96]$.

Har vi givet en stokastisk variabel $X \sim N(\mu, \sigma)$, som på illustrationen, så kan denne transformeres over i standardnormalfordelingen med transformationen: $z = \frac{X - \mu}{\sigma}$, der kaldes for en "z-score". Man kan opfatte z således, at den måler afstanden fra middelværdien i enheden σ , dvs: Hvor mange spredninger ligger vi væk fra middelværdien.

Normalfordelingstestet anvendes ofte til at svare på problemer, hvor vi har en *kendt normalfordeling* og spørger, om en bestemt værdi ligger indenfor normalområdet eller udenfor. Vi illustrerer det med et eksempel. I eksemplet får vi brug for følgende viden om stokastiske variable:

Hvis X_1, X_2, \dots, X_n er stokastiske variable med samme middelværdi μ og samme spredning σ , så vil **gennemsnittet** \bar{X} af de stokastiske variable have

middelværdi $\mu_{\bar{X}} = \mu$

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

Eksempel. Konfidensinterval og hypotesetest med normalfordeling

I et forsøg med gravide rotter undersøges, om dagligt indtag af alkohol i løbet af graviditeten påvirker den nyfødtes vægt. Man ved, at for normale rotter, der ikke udsættes for alkohol, er vægten af de nyfødte rotter normalfordelt med middelværdien $\mu = 18$ gram og en spredning på $\sigma = 4$. I forsøget indgår 16 gravide rotter, der i hele graviditeten har fået en daglig dosis alkohol. Der tages én nyfødt ud af hvert af de 16 kuld, så vi har en stikprøve på 16 nyfødte rotter.

Den gennemsnitlige vægt af de 16 nyfødte viser sig at være 15 gram.

Nulhypotesen er H_0 : Alkoholindtaget påvirker ikke ungernes vægt, dvs stikprøven på de 16 er fordelt som rottepopulationen generelt er: $\sim N(18, 4)$.

Det betyder, at *gennemsnittet* \bar{X} ifølge ovenstående er normalfordelt med middelværdi $\mu = 18$ og

spredning $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{16}} = \frac{4}{4} = 1$. Dvs: $\bar{X} \sim N(18, 1)$

Kan resultatet på $\bar{X} = 15$ nu forklares som en tilfældighed?

Vi undersøger om den transformerede z-score ligger i intervallet $[-1,96; 1,96]$:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{15 - 18}{1} = -3,0$$

Denne værdi er langt udenfor normalområdet. Værdien på -3 er usandsynlig under forudsætning af nulhypotesen. Derfor er konklusionen:

- Nulhypotesen forkastes. *Det kan ikke afvises, at der er en sammenhæng mellem alkoholindtag og den nyfødtes vægt*

Øvelse 7.7

I 1975 gennemførtes en landsdækkende undersøgelse i USA af, hvor meget tid skolebørn i afgangsklasser bruger på hjemmearbejde. Gennemsnitligt brugte de 8,4 timer ugentligt. Antallet af timer, eleverne brugte, var normalfordelt med en spredning på $\sigma = 3,2$.

25 år efter gennemførtes en helt tilsvarende undersøgelse af en stikprøve på 100 elever.

Det viste sig, at det gennemsnitlige timeforbrug for denne gruppe var $\bar{X} = 7,1$. Spørgsmålet er nu: Er der tale om et signifikant fald i det antal timer, eleverne bruger på hjemmearbejde.

Opstil en nulhypotese og undersøg om denne holder under et signifikansniveau på 5%.

(Bemærk: Der spørges om et *fald* så konfidensintervallet har alle 5% placeret til venstre!)

Anvendelse af et normalfordelingstest, der også kaldes for et **z-test**, kræver at vi kender populationen, herunder μ og σ . Hvis det ikke er tilfældet, må man estimere parametrene og bruge et såkaldt **t-test**, som omtales i kapitel 9.



7.6. Hvorfor er $p=0.05$?

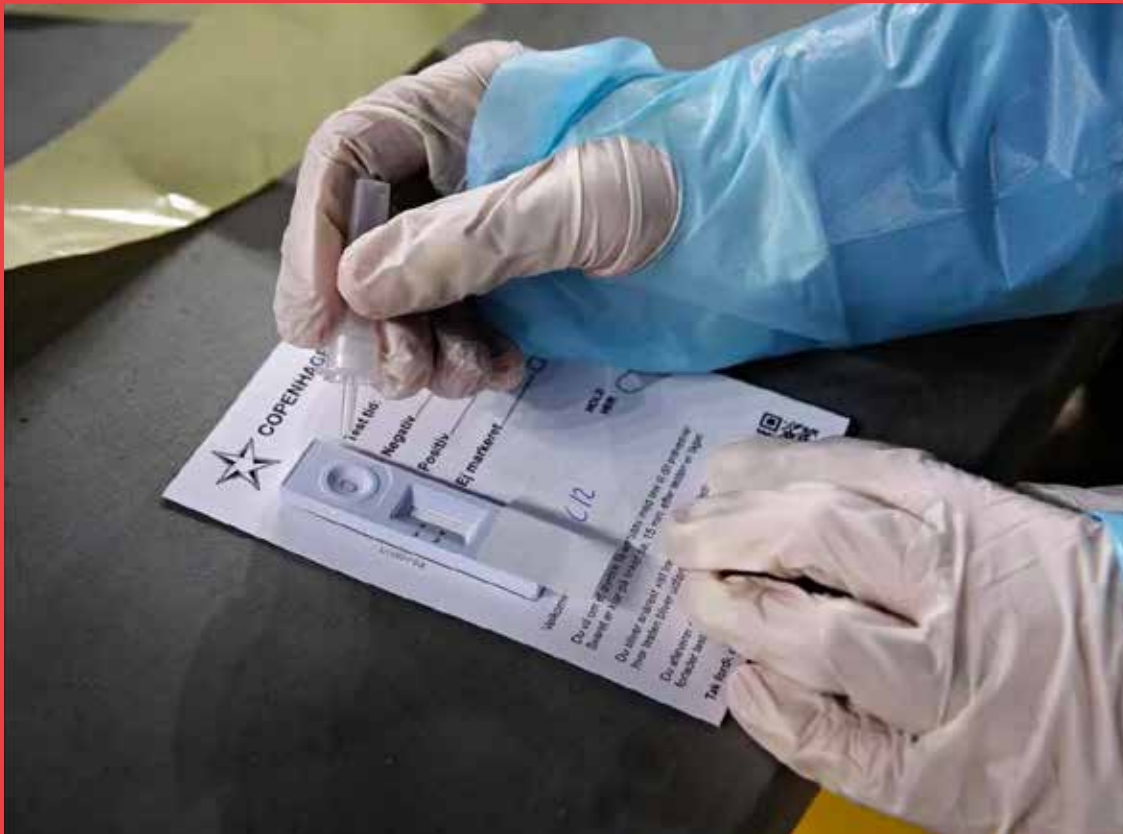
Du kan på hjemmesiden: [The Little Handbook of Statistical Practice \(jerrydallal.com\)](http://jerrydallal.com) finde et større materiale om statistik, hvor der er et opslag om *Why $p = 0,05$?* Det efterfølges af en række eksempler på, at man kan tage fejl, hvis man tror, at de 5% afgør sandheden.

Artiklen kan også hentes [her](#): [link til 38](#).

Du kan [her](#) [link til fil 39](#) finde en artikel, der redegør for den historiske oprindelse til signifikansniveauet, og som fører os langt tilbage i historien, før Fisher formulerede det i 1920'erne.

8.

Case: Testet positiv – men er man syg?



Coronaepidemien har gjort os alle bekendte med test. Undervejs hørte vi også en del om, hvor sikre de forskellige test var, og statistiske begreber som “falsk positiv” og “falsk negativ” blev anvendt i flæng uden seere og tilhørere nødvendigvis forstod, hvad begreberne egentlig dækker. Det er emnet for dette kapitel

8. Case: Testet positiv – men er man syg?

De første skridt ind i sandsynlighedsregningens verden er så lette, at det næsten er banalt: Vi kaster en terning – hvad er sandsynligheden for at få en sekser? Vi kaster en mønt, hvor vi kalder de to sider for plat og krone – hvad er sandsynligheden for at få krone? Vi trækker ét kort fra en godt blandet bunke med i alt 52 kort – hvad er sandsynligheden for trække et billedkort?

Det sidste var ikke så let som de første to. Vi opdager allerede ved sådanne indledende spørgsmål, at det er vigtigt at være omhyggelig og at kunne tælle mulighederne korrekt op: Regner vi kun knægt, dame og konge med som billedkort er der i alt 12 af disse ud af de 52 kort. Så sandsynligheden er $\frac{12}{52}$, der er det samme som $\frac{3}{13}$.

Øvelse 8.1 Kast med to terninger

Vi kaster to terninger. Hvad er sandsynligheden for at summen af øjnene er 5?

(Hint: Kald de to terninger for rød og grøn. Betegn fx et resultat af kastet (3,4). Hvor mange kombinationer af rød og grøn terning er der i alt? Hvor mange af disse giver den ønskede sum på 5?)

Ved kast med en terning er der 6 muligheder, og hvis terningen er retfærdig, er der lige stor sandsynlighed for hver af mulighederne. Derfor fastlægges sandsynligheden for at få en sekser til $\frac{1}{6}$.

Samme argument ved møntkast og ved trækning af ét kort giver sandsynligheder $\frac{1}{2}$ og $\frac{1}{52}$.

Men hvad er svaret, hvis vi spørger: Hvad er sandsynligheden for at trække en person med blodtype A, hvis vi helt tilfældigt udtrækker én dansker? For at svare på det, skal vi vide, hvor stor en andel af den danske befolkning, der har blodtype A. Dette er 44%. Det eneste rimelige svar er derfor, at trækker vi helt tilfældigt, så er sandsynligheden for at få en med blodtype A 44%. Dette kaldes også en *empirisk bestemt sandsynlighed*.

I stedet for at regne med *procenter*, er det som udgangspunkt ofte lettere at oversætte til *antal*: Vi har 100 personer ud af hvilke 44 har blodtype A. Var tallet 44,3% kunne det oversættes til: Vi har 1000 personer ud af hvilke 443 har blodtype A. Lad os nu se på et lidt mere kompliceret spørgsmål:

Øvelse 8.2 Klassikeren fra Harvard

Mange amerikanske lærebøger om statistik indeholder følgende eksempel på, hvor let det er at slutte forkert i sandsynlighedsregning:

Sandsynligheden for at udvikle en bestemt hjertesygdom er 1/1000. Der er udviklet en effektiv test, som fanger alle, der har sygdommen. Men denne test fanger yderligere 5% der faktisk ikke har sygdommen (i statistik kaldes dette: 5% falsk positive).

En person bliver testet positiv. Hvad er sandsynligheden for at vedkommende faktisk er syg?

Spørgsmålet blev stillet til 60 studenter og videnskabeligt ansatte på Harvard Medical School.

Næsten halvdelen svarede: 95%.

Det gennemsnitlige svar var: 56%.

Kun 11 af de 60 svarede korrekt.

Hvad er dit eget bud på et svar?

Denne øvelse handler helt grundlæggende om at være god til at tælle, og god til at opstille de forskellige tal logisk og overskueligt. Før vi giver det korrekte svar, så kan du hjælpes på vej ved at se en film, du kan hente [her](http://10.danske.matematikere-lru.dk/praxis.dk): [10 danske matematikere - LRU.dk \(praxis.dk\)](http://10.danske.matematikere-lru.dk/praxis.dk)
(Se den indledende del, gem forskningsdelen til senere).

Filmen giver flere eksempler på, hvor let det er at tage fejl i sandsynlighedsregning og statistik, bl.a. det vi har gengivet i afsnit 3.5: *Soldyrkere lever længere*. (Bemærk: Det er ikke nødvendigt for det følgende at have gennemgået afsnit 3.5).

Det centrale matematiske begreb i artiklen i afsnit 3.5 er *gennemsnitlig levealder*. Hvordan beregnes den?

Øvelse 8.3 Gennemsnitlig levealder

Antag vi har et land, hvor den demografiske struktur er meget stabil over tid, og fordeler sig således, hvis vi fokuserer på levealder:

Andel af befolkningen	Alder når de dør
15%	under 45 år
10%	45-55
20%	55-65
30%	65-75
20%	75-85
5%	85-95

For nyfødte børn gælder altså, at 20% af dem bliver mellem 75 og 85, mens 15% vil dø inden de bliver 45.

- Hvad skal vi forstå ved gennemsnitlig levealder? Udregn denne! (*Hint*: Tænk på middeltal af grupperede talsæt)
- Hvad er den gennemsnitlige levealder for dem, der er blevet 55? Og for dem der er blevet 75?

Resultatet af øvelse 8.3 giver den grundlæggende forklaring på den fejl, der blev begået af de læger, som anbefalede solarier.

Hvis en sygdom normalt først udvikles sent i livet, så tages stikprøven af syge blandt en gruppe af befolkningen, der i gennemsnit bliver betydeligt ældre end den normale levealder. En af statistikens fædre, William Farr, som grundlagde den engelske pendant til Danmarks Statistik, udtrykte det i en artikel så tidligt som i 1840'erne således: *Når generaler, biskopper og dommere med stor sandsynlighed bliver ældre end sergenter, præster og notarer, er det hverken på grund af helbred eller arbejdsforhold - dem der udnævnes til sådanne titler er allerede oppe i årene ved udnævnelsen.*

Situationen i sandsynlighedsregning kan sammenlignes med situationen i procentregning:

- Når man skal beregne, eller man får oplyst nogle procenttal, skal man altid spørge, hvad det er *procent af*? Ellers kan man komme til at sammenligne procenttal, der ikke kan sammenlignes. En nyfødt har 5% chance for at blive over 85 år gammel. Men er man nået forbi 65, så er chancen for at blive over 85 nu $\frac{5}{55} = \frac{1}{11} = 9,1\%$.
- Når man skal beregne sandsynligheden for en bestemt hændelse, skal man altid overveje, hvad det er for en afgrænset del af verden vi undersøger. Hvad er sandsynligheden for, at man bliver ramt af en influenzaepidemi den kommende vinter? Det afhænger af hvilken aldersgruppe vi ser på.

Vi vender nu tilbage til "Harvard-klassikeren". Sygdommen rammer en ud af 1000. Derfor opstiller vi en *matrix*, eller en *antalstabel* baseret på en population på 1000.

Der er én syg, som testes positiv – det er markeret i det grønne felt. Der er ingen syge, der testes negativ, for testen fanger alle syge. Derfor 0 i det gule felt.

Da der er en syg, er der 999 ikke-syge. Antal falsk positive er så 5% af 999. Dette er lig med 50, hvilket er skrevet i det orange felt – vi regner ikke med decimaler af mennesker. Så i alt testes 51 positivt. Tilbage er der 949, som ikke er syge, og som ikke testes positiv – det blå felt

test \ tilstand	syg	ikke syg	I alt
positivt udslag	1	50	51
ikke positivt udslag	0	949	949
I alt	1	999	1000

Betragt nu tabellen. En person er testet positiv og er altså blandt de 51. I denne gruppe er der 1 syg.

Så sandsynligheden for at vedkommende har sygdommen er $\frac{1}{51}$. Vi skriver dette således:

$$P(\text{syg, når man er testet positiv}) = \frac{1}{51} = 1,96\% \approx 2\%$$

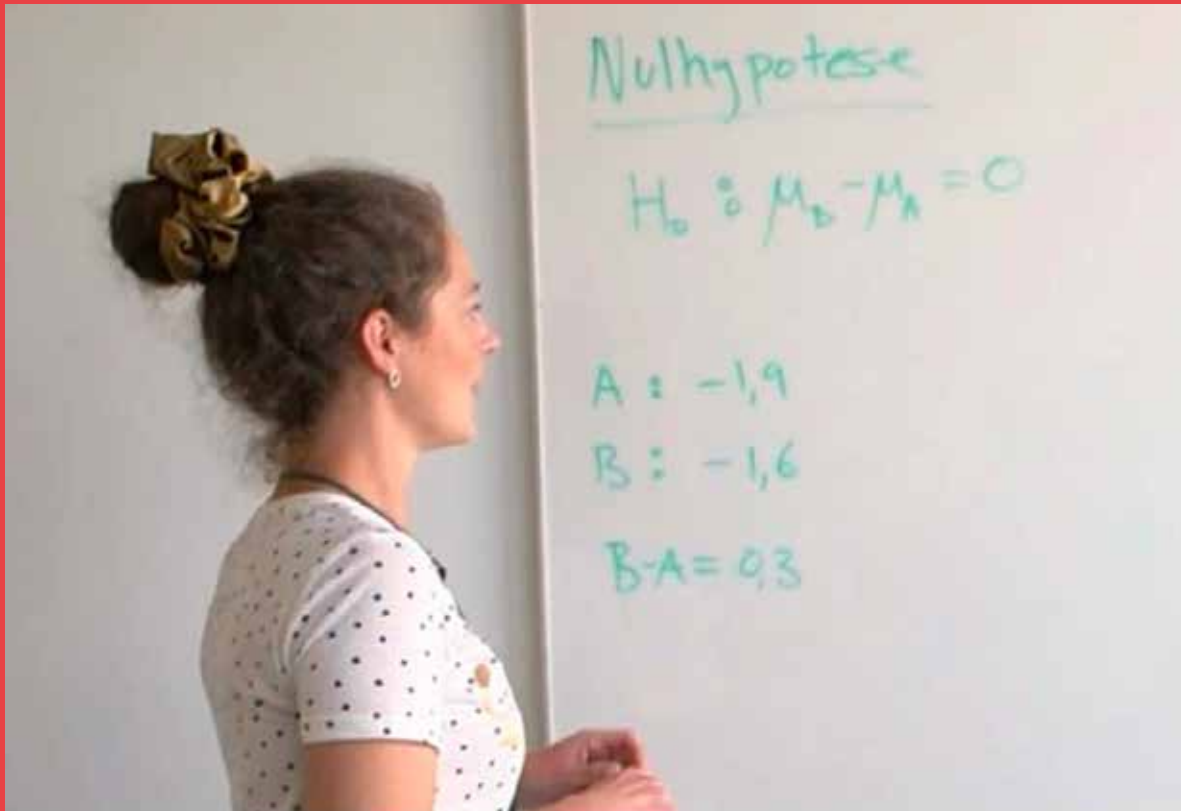
Var det dit svar? Se evt. filmklippet igen for at høre Susanne Ditlevsens forklaring.

Øvelse 8.4 Yderligere materialer om falsk positive

På hjemmesiden *sandsynligvis.dk* kan du finde film og materialer indenfor samme problemstilling, som hos Susanne Ditlevsen: [sandsynligvis.dk - Videoer](https://sandsynligvis.dk)

9.

Testteori 2. t-test og sammenligning af to dataserier



Sammenligningen af de to præparater A og B med grafiske metoder gav en indikation på, at A performede bedst. Men stadig kun en indikation. Med statistiske metoder kan vi ikke bevise, at A er bedst. Men vi kan kontrollere den sikkerhed, hvormed vi formulerer den slags konklusioner. Nu er vi nødt til at regne på data, siger Lea, og det gør vi i dette kapitel.

9. Testteori 2. t-test og sammenligning af to dataserier

Et normalfordelingstest kaldes også et **z-test**. Navnet stammer fra den notation vi normalt anvender, når vi transformerer en given normalfordeling $X \sim N(\mu, \sigma)$ over i standardnormalfordelingen ved

hjælp af transformationen:
$$z = \frac{X - \mu}{\sigma}.$$

Denne transformation bygger åbenlyst på, at vi kender middelværdi μ og spredningen σ (som også kaldes standardafvigelsen, engelsk: **Standard Deviation**). Men situationen er meget ofte, at det gør vi ikke. Derfor må vi estimere disse parametre.

Vi har altså et datasæt, der er en stikprøve af en større population, som vi har grund til at formode, er normalfordelt. Begrundelsen kan være rent grafisk ud fra et histogram, eller lidt mere avanceret ud fra et qq-plot (behandlet i afsnit 4.6). Men vi kender ikke parametrene. De eneste talværdier, vi har, er datasættets værdier. Derfor må vi bruge dem til at estimere parametrene.

At *estimere en parameter* bygger grundlæggende på "sund fornuft", men dog tilsat en analyse fra den mere avancerede statistik, der skal sikre, at estimerterne har nogle "fornuftige" egenskaber. Det kan fx være, at der ikke er bias mellem estimerterne, at de er uafhængige og andet. Det vil vi ikke komme yderligere ind på her.

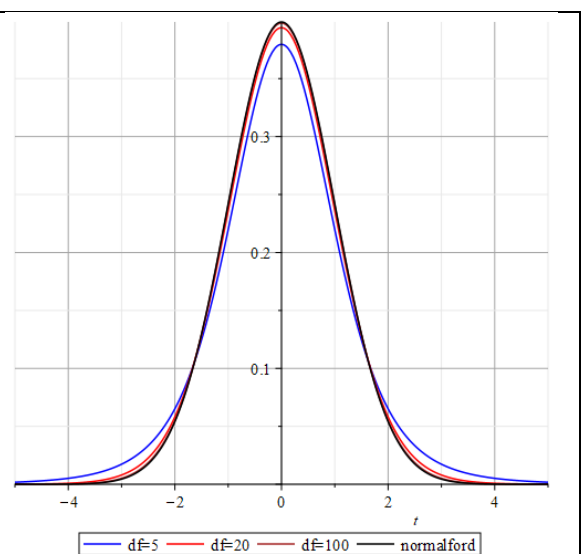
- **Middelværdien** μ estimeres med *stikprøvens gennemsnit*, \bar{x} .
- **Variansen** estimeres med *stikprøvens varians*, $Var(X)$, der beregnes som stikprøvevariansen i den deskriptive statistik:

$$Var(X) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}$$

- **Standardafvigelsen** estimeres så som kvadratroden af variansen: $s = \sqrt{Var(X)}$

Når vi estimerer parametrene ud fra dataværdierne, så er der naturligvis en grad af *usikkerhed* i dette: I et andet tilsvarende eksperiment ville vi have lidt andre dataværdier. Og da vi anvender spredningen til at bestemme konfidensintervallet, bliver der også en usikkerhed på dette, dvs vi kan ikke være så skarpe i afgrænsningen af, hvornår en teststørrelse falder indeni eller udenfor intervallet.

Man kan vise, at sandsynlighedsfordelingen dermed også ændrer sig, så vi ikke uden videre kan anvende normalfordelingen, men skal anvende en såkaldt **t-fordeling**. Denne kan betragtes som en modificeret eller justeret normalfordeling. Er der mange dataværdier, fx 200 eller 1000, så er sandsynligheden for at vores estimerter er tæt ved de sande værdier stor. Så i det tilfælde skal normalfordelingen ikke justeres ret meget. Men er der kun 10 dataværdier, så er usikkerheden betydeligt større, og normalfordelingskurven skal justeres mere, så 95% konfidensintervallerne bliver bredere. Der er således en *t-fordeling* knyttet til hvert antal af dataværdier i stikprøven. Er der 10 dataværdier, taler vi om, at der er 9 *frihedsgrader* og skriver "df=9". Det grafiske billede viser normalfordelingen sammen med tre t-fordelinger



Når antal frihedsgrader bliver meget stort, vil t-fordelingerne nærme sig normalfordelingen

Antallet af frihedsgrader for en t-fordeling er altid $n - 1$, hvor n er antal i stikprøven. Det kan vi forstå således: Se på formlen for variansen. Vi har udregnet en værdi ud fra de 10 dataværdier. *Samme værdi* kunne komme fra et andet datasæt. Vi kan skrue på de 9 ud af de 10 dataværdier – men dermed vil den 10. værdi være fastlagt, da vi skal have samme værdi for variansen.

Øvelse 9.1

Følgende 10 dataværdier er endpoints for 10 af A-patienterne i regnearket der behandles i filmen: {9, 6.2, 6.8, 5.6, 7.3, 6.4, 7.1, 5.8, 5.8, 6}.

- Udregn den estimerede middelværdi af denne stikprøve.
- Udregn den estimerede varians, ud fra formlen ovenfor.
- Hvad er antallet af frihedsgrader?
- Opstil et antal andre datasæt, der giver samme middelværdi og varians.

Øvelse 9.2

t-fordelingen findes i dit værktøjsprogram. I Maple hedder *t-sandsynlighedsfunktionen* med 9 frihedsgrader: $tpdf(9, x)$.

- Tegn grafen af $tpdf(9, x)$ sammen med grafen for normalfordelingens sandsynlighedsfunktion (tæthedsfunktion)
- Bestem værdierne af $tpdf(9, 1.7)$ og $normalpdf(0, 1, 1.7)$ og kontroller på din graf

Den tilsvarende fordelingsfunktion hedder $tcdf(9, x)$.

- Tegn grafen for $tcdf(9, x)$ og $tpdf(9, x)$ i samme koordinatsystem.

$tcdf(9, 1.5)$ angiver arealet under grafen for $tcdf(9, x)$ i området fra $-\infty$ til 1.5.

- Bestem værdien af $tcdf(9, 1.5)$, og kontroller på grafen.
- Udregn arealet under grafen for $tcdf(9, x)$ i området fra $-\infty$ til 1.5 som et integral.
- Giv en fortolkning af værdien $tcdf(9, 1.5)$ med brug af sandsynligheder.

Et **t-test** af en hypotese om en middelværdi gennemføres efter samme opskrift som et **z-test**, men her skal vi anvende de estimerede værdier og udregne konfidensintervallet under hensyn til antal frihedsgrader. Vi demonstrerer dette i et eksempel og øvelse. I eksemplet får vi brug for følgende viden om stokastiske variable:

Hvis Y_1, Y_2, \dots, Y_n er stokastiske variable med samme middelværdi μ og samme spredning σ , så vil **gennemsnittet** \bar{Y} af de stokastiske variable have

$$\text{middelværdi } \mu_{\bar{Y}} = \mu$$

$$\text{spredning } \sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

Eksempel og øvelse 9.3. Nulhypotese om middelværdi undersøgt med t-test

Eksemplet er hentet fra en amerikansk lærebog, og derfor er målene i ounces. Men det er ligegyldigt, vi regner uden enheder.

Vi har 15 dataværdier, der er kontrolmålinger på vægten af sukkerpakker.

{16.1, 16.2, 15.7, 15.8, 16.0, 15.8, 15.8, 15.9, 16.0, 15.9, 16.0, 16.0, 16.1, 15.7, 15.8}

Sukkerpakkerne skal indeholde (mindst) 16 ounces. Vi vil derfor teste nulhypotesen:

$$H_0 : \mu = 16 \text{ på et 95\% signifikansniveau}$$

Vi antager, at de 15 værdier er værdier af 15 uafhængige, normalfordelte stokastiske variable:

$$Y_i \sim N(\mu, \sigma) \quad i = 1, 2, \dots, 15$$

- Hvad er antallet af frihedsgrader
- Beregn estimater for middelværdi og varians. Du skal få 15.92 og 0.023
- Beregn estimatet for standardafvigelsen. Du skal få $s = 0.15$.

Når vi skal udregne teststørrelsen, så anvender vi formlerne anført ovenfor: $t = \frac{\bar{y} - 16}{\sigma_{\bar{y}}} = \frac{\bar{y} - 16}{\frac{s}{\sqrt{15}}}$

(Bemærk: Kender du z-testet, vil du genkende formlen!)

- Indsæt i udtrykket og beregn teststørrelsen t . Du skal få -2.1 .
- Bestem 95% konfidensintervallet i t -fordelingen med 14 frihedsgrader. Du skal få:
 $[-2.145; 2.145]$
- Konklusionen formuleres noget i retning af: *På et 95% signifikansniveau kan vi ikke forkaste nulhypotesen og påstå, at firmaet ikke lever op til deklarationen.* Men vi vil normalt tilføje, at når det er så tæt ved den kritiske værdi, vil vi foreslå yderligere undersøgelse

I filmen gennemfører Lea et test for at undersøge, om middelværdierne for A og B er signifikant forskellige. Dvs. her skal middelværdierne af to stikprøver sammenlignes. Det siges ikke udtrykkeligt i filmen, men det som er i spil her, er et t-test, da vi ikke kender de sande værdier af parametrene.

Eksempel og øvelse 9.4. t-test af om der er forskel på to middelværdier.

(Eksemplet er lånt fra Thomas Bendsens noter om statistik)

For at vurdere forskellen i kalcium-koncentrationen hos raske mænd og kvinder, har man målt P-Kalcium (Total) (i enheden mmol / liter) hos 10 mænd og kvinder og fået følgende data:

Mænd	2.45	2.42	2.33	2.40	2.46	2.24	2.29	2.26	2.33	2.24
Kvinder	2.34	2.52	2.32	2.31	2.31	2.39	2.25	2.27	2.43	2.41

Vi antager, at de 10 værdier for mændene og de 10 værdier for kvinderne begge er 10 uafhængige, normalfordelte stokastiske variable:

$$Y_{m,i} \sim N(\mu_m, \sigma_m) \quad i = 1, 2, \dots, 10 \quad Y_{k,i} \sim N(\mu_k, \sigma_k) \quad i = 1, 2, \dots, 10$$

Vi vil teste nulhypotesen:

H_0 : Der er ingen forskel på kalcium-koncentrationen hos mænd og kvinder, dvs:

H_0 : $diff = \mu_m - \mu_k = 0$ på et 95% signifikansniveau

- Udregn middelværdierne. Du skal få: $\bar{y}_{mænd} = 2.342$, $\bar{y}_{kvinder} = 2.355$, dvs $d = -0.013$

Standardafvigelsen på en differens kan med god tilnærmelse beregnes med formlen:

$$\sigma_{m,k} = \sqrt{\frac{\sigma_m^2}{n_m} + \frac{\sigma_k^2}{n_k}},$$

hvor σ_m og σ_k er standardafvigelserne på data for mænd og kvinder, og n angiver antal.

- Beregn værdien af standardafvigelsen på differensen. Du skal få 0.0375.
- Antal frihedsgrader er summen af antal for mænd og antal for kvinder. Bestem tallet.

Konfidensintervallet omkring $d = -0.013$ bestemmes nu til:

$$[d - t_{0.975}(f) \cdot \sigma_{m,k}; d + t_{0.975}(f) \cdot \sigma_{m,k}], \text{ med } 2,5\% \text{ til hver side.}$$

f er antallet af frihedsgrader, og $t_{0.975}(f)$ er denne t -fordelings kritiske værdi.

- Bestem konfidensintervallet. Du skal få: $[-0,0918 ; 0,0658]$
- Hvor ligger $diff = \mu_m - \mu_k = 0$ i forhold til intervallet? Formuler en konklusion.

Øvelse 9.5. Test af nulhypotesen om middelværdierne for A og B i filmen

Overvej selv, hvordan dette eksempel hænger sammen med det, som Lea gennemgik i filmen. Med de givne tal i filmen for middelværdi, standardafvigelse og antal, udregn da selv et 95% konfidensinterval for nulhypotesen.

Bemærk: I t-fordelings tabeller er der ikke særskilte opslag for ethvert antal af frihedsgrader, fx ikke for $df=198$, så her blev anvendt $t_{0,975}(f) = 1.96$. Det korrekte er $t_{0,975}(198) = 1.97$. Det giver ingen forskel i resultaterne.

Øvelse 9.6 t-test-calculator

Der findes en række websites, hvor man kan indtaste oplysninger om data og få gennemført et test i en slags black-box. Nogle er mere oplysende end andre, se fx siden, der linkes til nedenfor, og som giver dette skærmbillede:

Efter indtastning vil du få et svar af typen:

Result

$$\mu_1 - \mu_2 = (M_1 - M_2) = 0.3, 95\% \text{ CI } [0.011, 0.589].$$

You can be 95% confident that the difference between your two population means ($\mu_1 - \mu_2$) lies between 0.011 and 0.589.

Men denne side giver også formler og oplysende mellemregninger

Gå ind på siden, indtast de ønskede værdier, og tryk på 'calculate'

Independent Samples Confidence Interval Calculator

This simple confidence interval calculator uses a t statistic and two sample means (M_1 and M_2) to generate an interval estimate of the difference between two population means (μ_1 and μ_2).

The formula for estimation is:

$$\mu_1 - \mu_2 = (M_1 - M_2) \pm t_{\alpha/2}(M_1 - M_2)$$

where:

M_1 & M_2 = sample means

t = t statistic determined by confidence level

$s_{(M_1 - M_2)}$ = standard error = $\sqrt{(s_p^2/n_1) + (s_p^2/n_2)}$

To perform this calculation you need to know your two sample means, the number of items in your samples, and the standard deviations for your two samples. (If you need to calculate means and standard deviations from sets of raw scores, you can do so using our [descriptive statistics tools](#).) The calculation works on the assumption that the two population variances are equal (i.e., it uses a pooled standard deviation in order to calculate the standard error portion of the confidence interval calculation).

The Calculation

Please enter your data into the fields below, with your sample 1 mean being the higher of your two means, select a confidence level (the calculator defaults to 95%), and then hit Calculate. Your result will appear at the bottom of the page.

Sample 1 Mean (M_1):	<input type="text"/>
Sample 1 Size (n_1):	<input type="text"/>
Standard Deviation 1 (s_1):	<input type="text"/>
Sample 2 Mean (M_2):	<input type="text"/>
Sample 2 Size (n_2):	<input type="text"/>
Standard Deviation 2 (s_2):	<input type="text"/>
Confidence Level:	95% <input type="button" value="v"/>

Please enter your values above, and then hit the calculate button.

[Confidence Interval Calculator: Independent Samples T-Test \(socsistatistics.com\)](http://socsistatistics.com)

Projekt 9.1. Eksperimentel undersøgelse af formlen for varians

I formlen for variansen af en stikprøve, som anvendes til at estimere standardafvigelsen, indgår at vi dividerer med $n - 1$, og ikke med n . Hvorfor egentlig? Dette spørgsmål undersøges eksperimentelt i et projekt, som du kan finde her:

[kap9 Projekt 9 2 Den bedste estimator for populationens sande varians.pdf \(praxis.dk\)](#)

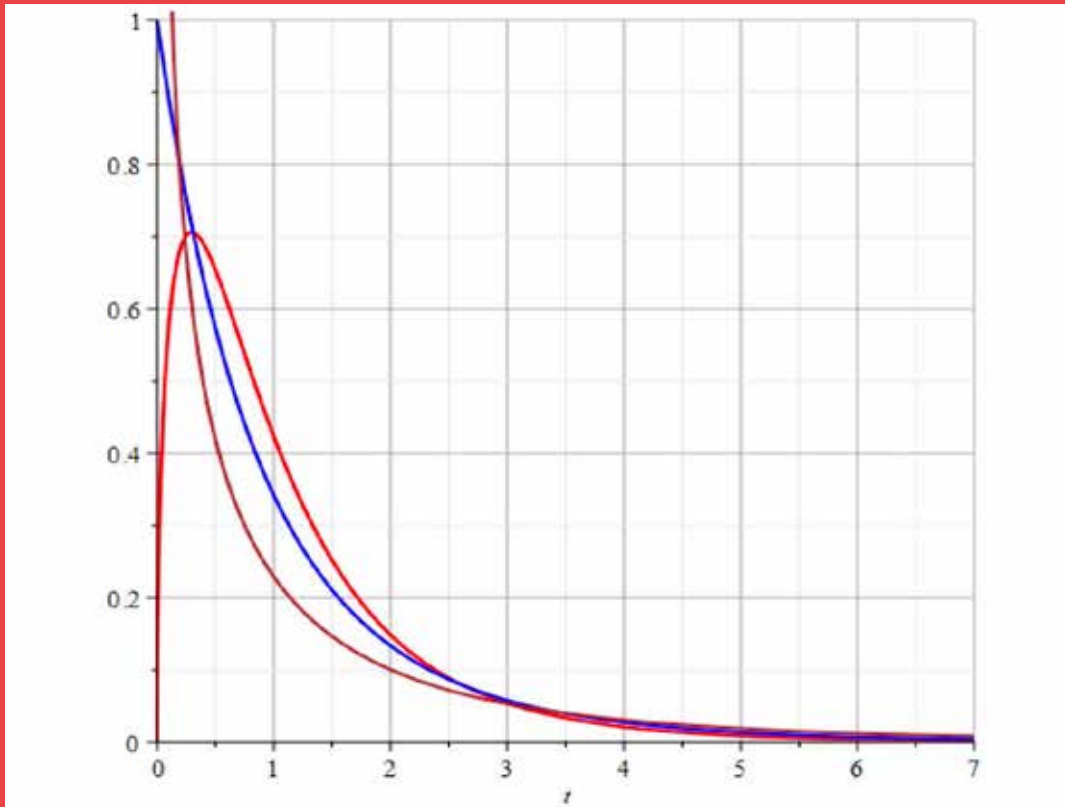
Projekt 9.2. Analyse af unges adfærd med brug af chi-i-anden-test og t-test

Hvis du vil arbejde videre med hypotesetest, så kan det være en god ide at kombinere anvendelsen chi-i-anden testet, der undersøger spørgsmål om kategoriske variable, og t-testet, der undersøger spørgsmål om numeriske variable. Projekt 8.10 fra HEM3 rummer dette og arbejder med et stort datasæt om unge tyskeres adfærd:

[Hvad er matematik? 3 - LRU.dk \(praxis.dk\)](#)

10.

Testteori 3. SRP-inspiration: ANOVA og F-fordelingen



I kapitel 9 sammenlignes to sæt af dataværdier, men ofte vil vi have en situation, hvor vi ønsker at teste flere end to mod hinanden. Måske endda sammenligne datasæt, der er grupperet efter flere kriterier. Det kan gøres ved at gennemføre en række test to og to, men dels er det besværligt og dels er der så større risiko for at begå den slags "fejl", der er indbygget i de statistiske metoder. I dette kapitel præsenteres en metode, hvormed vi på en gang kan sammenligne data grupperet efter flere kriterier. Sammen med dette nye test, der hedder ANOVA følger også en ny fordeling, der er beslægtet med chi-i-anden fordelingen, og kaldes for F-fordelingen

10. Testteori 3. SRP-inspiration: ANOVA og F-fordelingen

(Eksemplet, der anvendes i præsentationen af ANOVA er lånt fra Per Vejrup-Hansens bog 'Statistik med Excel'. Den rummer generelt en glimrende indføring i Excels mange værktøjer til statistik)

I disse materialer præsenteres en række forskellige metoder til at teste nulhypoteser. Men emnet er slet ikke udtømt med dette. I filmen spørger Casper, om det, de har talt om, så også er de metoder, hun og Novos statistikere bruger i praksis. De bruger naturligvis hele værktøjskassen, men Lea svarer også, at de ikke mindst anvender ANOVA og ANCOVA. Vi vil her åbne vinduet til denne større verden af statistiske fordelinger, men emnet er for omfattende til en grundigere gennemgang, og det teoretiske grundlag vil vi lade ligge.

Den tabel af dataværdier, som Novo har stillet til rådighed, forklarede Lea var et lille udsnit af et meget større materiale med måske op mod 100 variable og flere måleresultater som kontrol på, hvordan patienterne tager imod medicinen. Til en sådan stor database kan man stille mange forskellige interessante spørgsmål. Hvis vi fx brugte flere inddelingskriterier, såsom alder, vægt / BMI og etnicitet, så kunne vi spørge om A performer bedre end B i alle grupper. Der kunne fx være 4 grupper: (A, under 50 år), (A, over 50 år), (B, under 50 år), (B, over 50 år). Inddelingen kunne tilsvarende være (A, BMI>30), (A, 30>BMI>20), osv.

Hvordan skulle man teste, om A performer bedre end B i alle grupper? Med computerkraft til rådighed er besværet ved at teste på kryds og tværs ikke det største problem i det her. Det er derimod *risikoen for at begå fejl*. Vi ved fra introduktionen til hypotesetest, at fastlæggelsen af et signifikansniveau på fx 95% betyder, at der er 5% risiko for at begå en fejl af type 1, hvor vi forkaster en nulhypotese der er sand. Lad os sige, vi havde opdelt dataværdierne i 4 grupper, som vi ville teste op mod hinanden 2 og 2. Det ville give 6 testsituationer – overvej selv dette!

Øvelse 10.1. Sandsynligheden for fejl af type 1 ved mange test.

Hvis signifikansniveauet er 95%, så er sandsynligheden for en fejl af type 1 lig med 5%.

- Hvad er sandsynligheden for *ikke* at begå fejl af type 1
- Hvad er sandsynligheden for *ikke* at begå en type 1 fejl i nogen af de 6 testsituationer? Du skal få 73,5%
- Hvad er sandsynligheden for, at vi begår en type 1-fejl i *mindst en* af de 6 situationer?

ANOVA er svaret på det problem, øvelse 10.1 peger på. **ANOVA** betyder **A**nalysis of **V**ariance, og er altså et test, der bygger på variansanalyse i stedet for vurderinger på middelværdierne.

Vi vil arbejde med et eksempel, og så undervejs forklare og demonstrere, hvad metoden går ud på, og hvad den svarer på. I eksemplet er data indhentet fra indkøb af den *samme* kurv af varer fra 5 forskellige butikskæder, der alle findes i 4 forskellige byer. Resultatet var dette:

Butik	By 1	By 2	By 3	By 4	Gennemsnit
A	750	800	810	680	
B	780	790	790	740	
C	820	830	840	750	
D	790	770	800	730	
E	740	770	770	750	
Gennemsnit					*)

*) I dette felt skrives det totale gennemsnit, dvs: gennemsnittet af gennemsnittene!

Formålet er at undersøge, om der er *signifikant* forskel på priserne i de forskellige byer og de forskellige kæder. Strategien i denne undersøgelse er dels at se på, hvordan priserne varierer **mellem**

byerne og dels hvordan priserne varierer **indenfor de enkelte byer**, dvs. mellem butikkerne. Undersøgelsen skal ende ud med et samlet svar på nulhypotesen:

H_0 : Der er hverken signifikant forskel på priserne mellem byerne eller mellem butikskæderne.

Der er rigtig mange beregninger i ANOVA, og det er ikke mindst beregninger af varianser, som er ret tunge at udføre manuelt. Derfor vil vi trække på Excels indbyggede analyseredskaber¹.

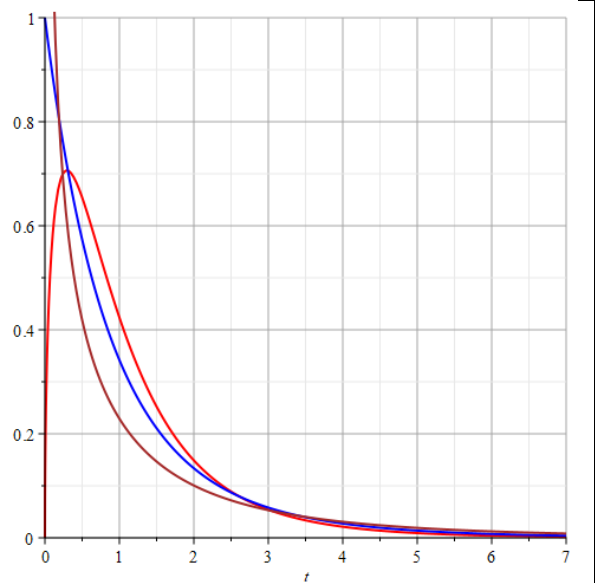
10.1 One way ANOVA – (Enkeltfaktor-ANOVA)

Vi starter med at undersøge om der er forskel på priserne *mellem* byerne. Her får vi brug for *gennemsnitspriserne* for de 4 byer. *Variansen* af disse 4 tal er et samlet mål for hvor meget tallene afviger fra *deres* gennemsnit. En stor varians må være udtryk for, at der er stor forskel på priserne mellem byerne – men hvordan vurderer vi, om der er "stor forskel". Det gør vi ved at sammenligne med variationen *indenfor* byerne. Disse udregnes hver for sig, hvorefter vi tager gennemsnittet af dem. Dette tal er vores "målestok" for om prisforskellen mellem byerne er stor eller om vi må konkludere, at den ikke er *signifikant*.

At det er en målestok betyder, at vi dividerer variansen mellem byerne med denne gennemsnitsvarians indenfor byerne:
$$F = \frac{\text{varians}_{\text{mellem}}}{\text{varians}_{\text{indenf}}}$$

Dette tal er stort, hvis variansen mellem byerne er stor *i forhold til* den gennemsnitlige varians indenfor byerne. Hvis omvendt der ikke er stor forskel på variansen mellem byerne og indenfor de enkelte byer, så er F af mere beskedene størrelse.

Formlen for F giver anledning til en ny sandsynlighedsfordeling, den såkaldte *F-fordeling*. Man kan vise, at både tæller og nævner i formelen følger chi-i-anden fordelinger, og ud af det dels udlede formelen for sandsynlighedsfunktionen, og dels få tegnet grafen for den. De fleste matematiske værktøjsprogrammer har F -fordelingen indbygget, i Maples notation betegnes sandsynlighedsfunktionen (tæthedsfunktionen) $Fpdf(df_1, df_2, t)$, hvor df står for antal frihedsgrader i henholdsvis tæller og nævner. Her er gengivet graferne af $Fpdf(3, 16, t)$, $Fpdf(2, 13, t)$ og $Fpdf(1, 9, t)$. Testes en hypotese på et 95% signifikansniveau, er det interessant at kende den kritiske værdi.



Ovenstående bliver mere klart når vi regner på eksemplet

Øvelse 10.1 ANOVA – enkeltfaktor

- Læg tabellen ovenfor ind i et Excelark.
- Udregn selv gennemsnittene både lodret og vandret (gerne med brug af *middel*)

¹ Funktionen *Dataanalyse* skal ofte hentes særskilt: I et åbent Excelark går man ind i *filer / indstillinger* og vælger *Tilføjelsesprogrammer*. Her klikkes på *Analysis Toolpack* eller hvis det er på dansk: *Dataanalyse*. Det vil så blive gjort aktivt og bliver lagt under fanen *Data*. Excel kalder ANOVA for ANOVA (!).

- c) Hent i *Data / Dataanalyse* programmet ANAVA- Enkeltfaktor. I den dialogboks, der kommer frem, angives området, hvor data er indtastet og hvor overskrifterne: By1, By2 osv medtages (gør det evt ved at markere området). Vælg gruppering efter kolonner, alpha (signifikansniveau) på 0,05 og hak af, at etiketterne er i første række. Vælg selv, hvor du vil have resultatet skrevet.
- d) Kontroller, at du får et svar som dette, og at gennemsnittene er som dine fra b):

Anava: Enkelt faktor						
RESUME						
Grupper	Antal	Sum	Gennemsnit	Varsians		
By 1	5	3880	776	1030		
By 2	5	3960	792	620		
By 3	5	4010	802	670		
By 4	5	3650	730	850		

ANAVA						
Variationskilde	SK	fg	MK	F	P-værdi	F krit
Mellem grupper	15220	3	5073,33333	6,401682	0,004688	3,238872
Inden for grupper	12680	16	792,5			
I alt	27900	19				

Forklaring på forkortelserne:

Mellem grupper: Her regnes på gennemsnittene af kolonnerne.

Inden for grupper: Her regnes på hver kolonne for sig.

SK: Sum af kvadratafvigelse fra gennemsnit. Dette er *tælleren* i formelen for varians.

fg: Dette er antal frihedsgrader (se evt kapitel 9). Dette er *nævneren* i formelen for varians.

MK: Dette står for 'Middel Kvadratafvigelse'. Det er **SK/ fg**, altså *variansen*.

F: Dette er teststørrelsen, der anvendes i *F-fordelingen*. Den udregnes som $\frac{MK_{mellem}}{MK_{indenf}} = \frac{5073,3}{792,5}$

- e) Betragt formelen for F , og redegør med dine egne ord for, hvad en stor F -værdi betyder mht variansen mellem grupperne og variansen indenfor gruppen. Fx i stil med: *En stor F-værdi betyder, at der er ... på priserne mellem byerne, mens der er ... indenfor byerne.*
- f) Tegn selv grafen for denne F -fordeling, $Fpdf(3,16,t)$.
- g) Bestem den kritiske værdi ved at løse en ligning som $Fcdf(3,16,x) = 0.95$, kontroller, du får det samme som **F-krit** i skemaet ovenfor, og afsæt værdien som en lodret linje i det grafiske billede.
- h) Kontroller, at $F = 6.40164$, og afsæt tilsvarende denne værdi i det grafiske billede.
- i) P-værdien** er sandsynligheden for at få en værdi der er 6,40164 eller større. Denne kan vi beregne på to måder – gør det og kontroller tabellens tal på 0,004688:
- Som en funktionsværdi af fordelingsfunktionen $Fcdf(3,16,t)$,
 - Som et areal under grafen for sandsynlighedsfunktionen $Fpdf(3,16,t)$
- j) Formuler en konklusion.

Analysen giver ikke en forklaring på, *hvad årsagen er* til den signifikante forskel mellem byerne. Det kunne være, at det bare var én af byerne, hvor gennemsnittet afveg væsentligt fra de øvrige, og som derved gav det udslag, at H_0 blev forkastet. Det kræver en videre analyse at undersøge det.

Før vi går videre, opfordres du til selv at gennemføre en af Excels beregninger. Det kan give større indsigt i, hvad det er der sker i ANOVA-beregningerne. Lad os se på SK-søjlen i ANOVA delen af tabellen. Vi udregner først kvadratsummen *mellem grupperne*, dvs mellem byerne.

Øvelse 10.2. Manuel udregning af tal fra ANOVA-tabellen - 1

Variationen mellem grupperne er en beregning af afvigelserne mellem *gruppe-gennemsnittene* og det samlede gennemsnit ("grand-gennemsnit"). Det er altså følgende tal:

Gennemsnit	776	792	802	730
-------------------	------------	------------	------------	------------

De kvadrerede afvigelser multipliceres hver med antallet af observationer, dvs her 5, for at kunne sammenligne tallet med det, vi udregner i næste øvelse, hvor der indgår dette antal observationer i hver gruppe. Sammenligningen sker i den udregning af teststørrelsen F , som vi foretog i øvelse 10.1.

- Vis, at *grand-gennemsnit* = 775
- Vis, at variationen mellem grupperne bliver:

$$5 \cdot (776 - 775)^2 + 5 \cdot (792 - 775)^2 + 5 \cdot (802 - 775)^2 + 5 \cdot (730 - 775)^2 = 15220$$

Dernæst udregnes kvadratsummen *indenfor grupperne*, dvs indenfor byerne:

Øvelse 10.3. Manuel udregning af tal fra ANOVA-tabellen - 2

Se på den oprindelige tabel:

Butik	By 1	By 2	By 3	By 4
A	750	800	810	680
B	780	790	790	740
C	820	830	840	750
D	790	770	800	730
E	740	770	770	750
Gennemsnit	776	792	802	730

- Opret en tabel i regnearket med samme antal rækker og søjler, men uden tal. Lad os sige, at tallene i kolonnen med By 1 står i cellerne B7 ... B11. I det nye ark er de tilsvarende celler B18 ... B22. Vi udfylder nu cellerne i det nye ark med formler i cellerne. Formlerne skal i hver celle give os *kvadratet på afstanden til gennemsnittet*. I celle B18 skriver vi derfor: $= (B7 - 776)^2$.

Når vi taster 'enter' udregnes tallet og vi får det første tal, 676 i nedenstående:

Kvadratsummen indenfor grupper:

Butik	By 1	By 2	By 3	By 4
A	676			
B	16			
C	1936			
D	196			
E	1296			
SUM	4120			

- Med hjælp fra Excels kommandoer kan vi nu trække formlen ned og få udfyldt B18 .. B22. Dernæst anvendes Excels kommando til udregning af summer, og vi får 4120.
- Udfyld nu selv hele tabellen, idet du i hver søjle indsætter byens gennemsnitstal.
- Udregn den samlede sum og kontroller, at du får det samme som i den automatiske udregning, nemlig 12680.

Øvelse 10.4. Kontroller tabellens tal

Kontroller endelig alle de øvrige tal i tabellen for One Way ANOVA

10.2 Two way ANOVA – (To faktor Anova)

Vi inddrager nu den anden faktor i den oprindelige tabel, nemlig *butikskæderne*. Derved undersøges om forskellen, vi fandt i 'One Way ANOVA', kan forklares med bestemte forskelle mellem butikkerne. I et excelark med samme data og samme søjler og rækker gennemføres denne undersøgelse med brug af *Data / Dataanalyse*, hvor vi vælger: *Anava – to faktor uden gentagelse*.

I skemaet angives hele tabellen som input, dvs inklusive række med byer og søjle med butikker (men uden at medtage gennemsnit). Vi hakker etiketter af, sikrer os der står 5%, og vælger et passende outputsted i arket. Du skal få et resultat som dette:

Anava: To-faktor uden gentagelse					
RESUME	Antal	Sum	Gennemsnit	Varians	
A	4	3040	760	3533,333	
B	4	3100	775	566,6667	
C	4	3240	810	1666,667	
D	4	3090	772,5	958,3333	
E	4	3030	757,5	225	
By 1	5	3880	776	1030	
By 2	5	3960	792	620	
By 3	5	4010	802	670	
By 4	5	3650	730	850	

ANOVA						
Variationskilde	SK	fg	MK	F	P-værdi	F krit
Rækker	7050	4	1762,5	3,756661	0,033222	3,259167
Kolonner	15220	3	5073,33333	10,8135	0,000996	3,490295
Fejl	5630	12	469,166667			
I alt	27900	19				

Fra resumeet springer det i øjnene, at der er større varians pr butik end der er pr by. Det fortæller, at de største forskelle findes *mellem* byerne. Det indtryk bekræftes af de udregnede *p*-værdier i ANOVA-tabellen: *p*-værdien for byerne (kolonner) er meget beskeden, og fortæller, at det er helt usandsynligt den forskel skulle optræde tilfældigt hvis nulhypotesen var sand. For butikkerne (rækker) indbyrdes er der også signifikant forskel på priserne, idet $p=3,3\%$, men altså mindre markant.

I øvelse 10.2 udregnede vi *variationen mellem byerne* (kolonnerne) til 15220. Det tal genfinder vi ovenfor.

Vi udregner nu tilsvarende *variationen mellem butikskæderne* (rækkerne) ved at se på hvor meget rækkegennemsnittene varierer fra "gran gennemsnit" på 775. De kvadrerede afvigelse multipliceres som i 10.2 hver med antallet af observationer, dvs her 4, for at kunne sammenligne. Vi får i overensstemmelse med ANOVA- tabellen:

$$4 \cdot (760 - 775)^2 + 4 \cdot (775 - 775)^2 + 4 \cdot (810 - 775)^2 + 4 \cdot (772,5 - 775)^2 + 4 \cdot (757,5 - 775)^2 = 7050$$

F-størrelsen udregnes som forholdet mellem variationen *mellem* grupperne og variationen *indenfor* grupperne. Men en del af den sidste har vi konkret fået tilskrevet variationen mellem butikskæderne. Den del af variationen, der *ikke "fanges"* af variationen mellem butikskæderne, kalder Excels ANOVA-ark for *Fejl*: Læg mærke til, at: $5630 = 12680 - 7050$.

Det er en lidt sær betegnelse, men det korte af det lange er, at vi nu skal sammenligne variationen mellem grupperne med denne resterende variation indenfor grupperne.

Øvelse 10.5. Beregning af teststørrelserne F

- Argumenter for antallet af frihedsgrader. (*Hint*: Kategorien 'Fejl' repræsenterer cellerne i arket, hvor vi nu har lagt to bånd på: lodret gennemsnit og vandret gennemsnit)
- Udregn de gennemsnitlige kvadratafvigelse (MK).
- Udregn geststørrelserne F ved at sammenholde variationen *mellem* rækkerne (eller mellem kolonnerne) med variationen *indenfor* rækkerne / kolonnerne. Det sidste er for begge situationer det samme tal, nemlig MK-tallet i 'Fejl'.
- Hvilke F -fordelinger er det, vi skal anvende her?
- Anvend de relevante F -fordelinger og bestem de kritiske værdier. Kontroller tabellen.
- Formuler nu en samlet konklusion i relation til nulhypotesen, i relation til nulhypotesen.

Øvelse 10.6. Anvend ANOVA på filmens datasæt

Vælg nu selv nogle inddelingskriterier for det datasæt, vi har fået fra Novo, hvor A og B klart er det ene kriterium. Gennemfør dernæst en undersøgelse med brug af ANOVA, hvor du kan lade din undersøgelse følge gangen ovenfor.

10.3 ANCOVA

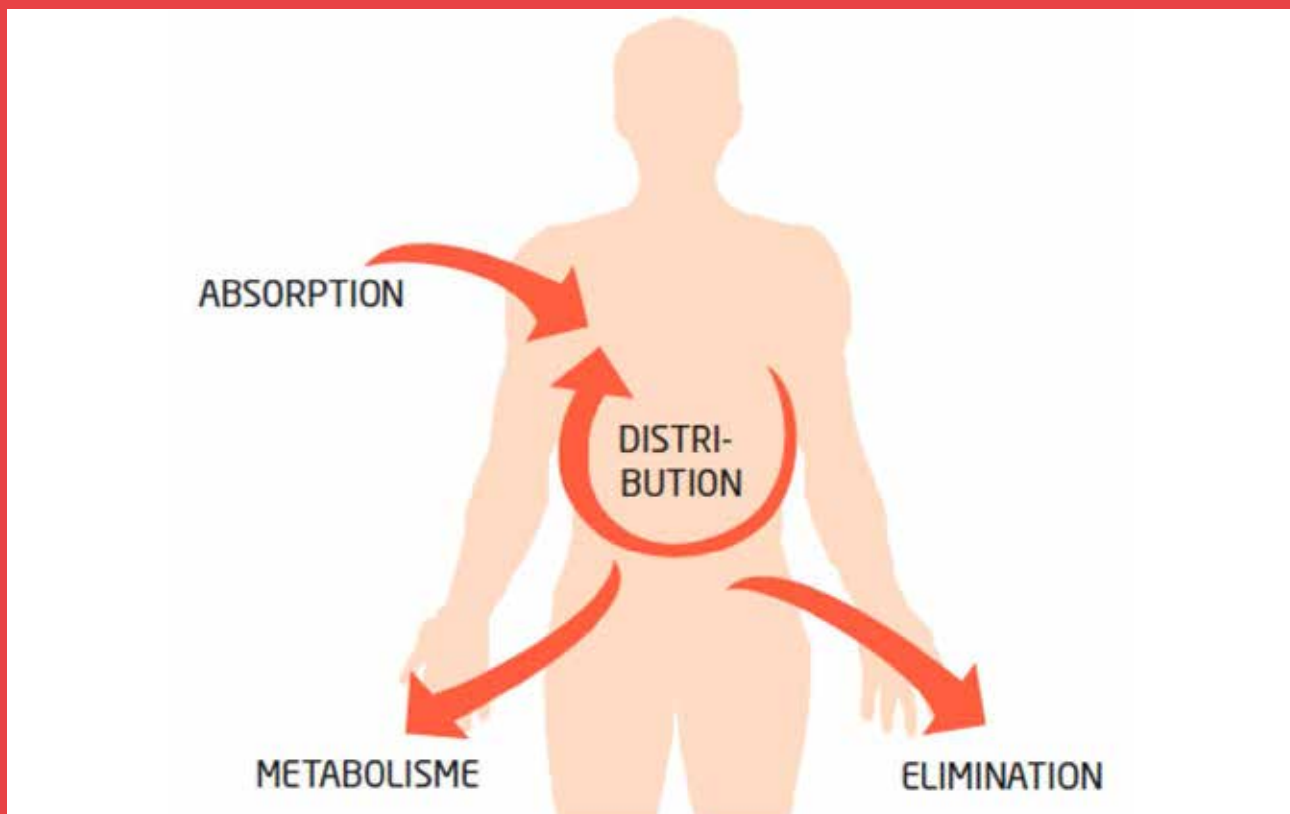
I Novo-eksemplet, hvor vi sammenlignede to behandlinger, A og B, undersøgte vi *ændringen* ift baseline. Dvs. vi så at sige nulstillede værdien ved baseline. Men kunne det ikke tænkes, at denne spillede ind?

I en såkaldt ANCOVA-undersøgelse inddrages dette aspekt: Vi undersøger stadig *ændringerne* ift baseline, men nu under hensyntagen til værdien ved baseline. Datasættet med værdier fra baseline kaldes et *covariat*, og ANCOVA-undersøgelsen sker ved at udbygge ANOVA med en slags regressionsanalyse, idet vi spørger: Kunne forskellen på A og B forklares ved, at de to grupper havde lidt forskellige udgangspunkter?

Desværre er ANCOVA ikke en del af hverken Excels eller Maples statistikpakker. Man er nødt til at downloade dette særskilt. Du kan se nærmere herom her – [link til 40](#) – hvor vi gennemgår metoden via nogle eksempler.

11.

Matematisk modellering af medicins optag og forbrænding i kroppen



Filmen handler om en sammenligning af to insulinpræparater, og i en animation illustreres insulins betydning for kroppens optag af glukose. Men hvad sker der egentlig, når et lægemiddel tilføres kroppen? Det er en selvstændig forskningsgren indenfor farmakologi. Indtages det fx i tabletform, så skal det eksempelvis først passere et meget barsk miljø i maven. Lægemidlets rejse gennem kroppen analyseres i dette kapitel vha. compartment-modeller, hvilket indebærer anvendelse af differentiaalligninger.

11. Medicins optag og forbrænding i kroppen undersøgt vha. compartment modeller

I filmen forklares, hvordan den mad, vi indtager, omdannes til glukose, som via blodet transporteres til kroppens celler, og hvordan hormonet insulin spiller en afgørende rolle i den proces. I det følgende afsnit vil vi gå i dybden med matematiske aspekter vedrørende medicins optagelse, fordeling, nedbrydning og udskillelse fra kroppen. Afsnittet forudsætter, du har kendskab til løsning af differentiaalligninger. Fra og med afsnit 11.2.7 går vi over til at se på anden ordens differentiaalligning. Dette emneområde er behandlet med udledning af løsningsformlerne og gennemregnede eksempler i HEM 3, kap 6. Dette kapitel kan hentes her: [link til 41](#)

Læren om lægemidlers optagelse, fordeling og nedbrydning i kroppen kaldes ofte under et for *farmakokinetik* og *farmakodynamik*. Ordene er sammensat af *farmakologi* – dvs. læren om lægemidler, samt *kinetik* og *dynamik* – dvs. læren om kraftpåvirkninger og bevægelse. Man skelner faktisk ofte mellem disciplinerne farmakokinetik og farmakodynamik ved at definere, at farmakokinetik vedrører, hvad "kroppen gør ved medicinen": hvordan optages, fordeles og udskilles et lægemiddel, og hvor hurtigt eller langsomt går det? Farmakodynamikken derimod beskæftiger sig med, hvad "medicinen gør ved kroppen": hvordan og via hvilke kemiske reaktioner og biologiske mekanismer virker lægemidlet?

Emnefeltet er derfor et helt oplagt tværfagligt samarbejdsområde mellem matematik og kemi, biologi eller bioteknologi. Matematikken fokuserer mest på farmakokinetikken - ofte centreret omkring modellering i form af opstilling af såkaldte compartmentmodeller og løsning af tilhørende differentiaalligninger. Kemi/biologi zoomer derimod ofte mere ind på farmakodynamikken i form af lægemidlers konkrete virkning på kroppen. Kort sagt kan man sige, at biologien og kemien beskriver, *hvordan* lægemidlet rejser gennem kroppen, hvor matematikken tager sig af, hvor *hurtigt eller langsomt* det går.

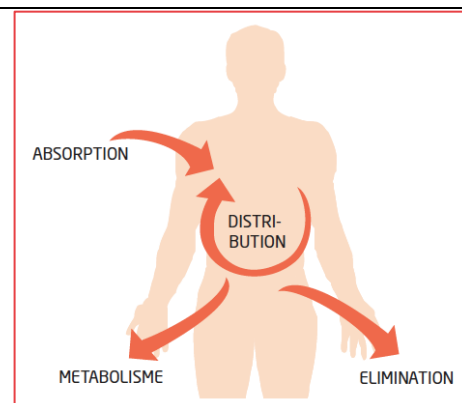
11.1 Et lægemiddels rejse gennem kroppen – lidt baggrundsteori

Når et lægemiddel indføres i kroppen, kommer virkningen først til udtryk, idet lægemidlet når frem til "målet". Målet kan være en del af et organ, en biologisk celle, en receptor eller noget helt tredje, og vejen til målet kan ofte være "lang" for lægemidlet. I langt de fleste tilfælde transporteres lægemidlet via blodbanen. Det hele foregår altså i væskefaser, og lægemidlet skal derfor typisk være til stede i målorganet i en passende koncentration, førend virkningen indtræffer, hvilket også er årsagen til, at både dosering og tid spiller afgørende roller i farmakokinetikken.

Lægemidlets rejse gennem kroppen er karakteriseret ved fire faser:

- absorption (optagelse),
- distribution (fordeling),
- metabolisme (nedbrydning) og til slut
- elimination (udskillelse).

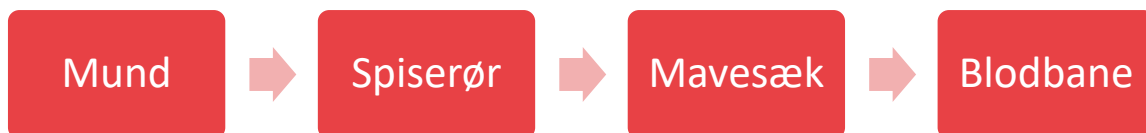
I det følgende gennemgås de fire faser kort.



"Medicin og matematik", Matematiske horisonter 2009

11.1.1 Absorption

Et lægemiddel kan indføres i kroppen på en lang række forskellige måder. Nogle af de mest almindelige metoder er: intravenøs dosering, hvor medicinen vha. en kanyle eller et drop sprøjtes direkte i blodbanen, subkutan dosering, hvor medicinen først sprøjtes ind i fedtvævet for derefter langsomt at fordele sig ud i blodbanen, intramuskulær dosering, hvor medicinen med kanyle indføres i muskelvævet, oral dosering (f.eks. piller eller miksturer), nasal dosering (næsesprays) eller rektal dosering (stikpiller). Når lægemidlet er indført, skal det gennemtrænge en eller flere af kroppens barrierer for at nå frem til blodbanen, hvor det absorberes i blodet. Hvis man eksempelvis indtager en pille, vil pillen først blive opløst i mavesækken for derefter at passere mave/tarmvæggen og passere ud i blodet:



Hvis man derimod indfører medicinen intravenøst er vejen til blodbanen naturligvis meget kortere. Det er for eksempel tilfældet, når det gælder insulinindsprøjtninger til mennesker med sukkersyge:



Øvelse 11.1

Hvilke fordele og ulemper er der ved de forskellige doseringsmetoder? Hvad er fordelene ved at dosere medicin intravenøst? Hvad er ulemperne? Hvilke fordele er der ved oral dosering? Hvilke ulemper?

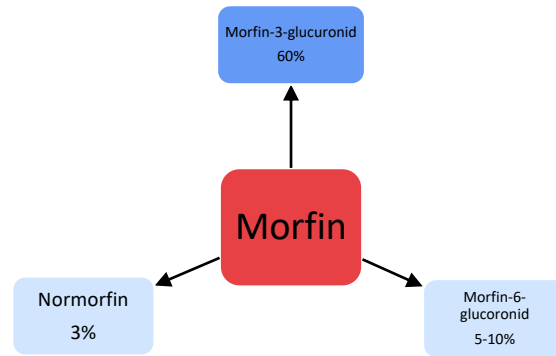
11.1.2 Distribution

Når lægemidlet er absorberet i blodbanen, transporteres det via blodets kredsløb til målstedet i kroppen. Her er det afgørende for tidsforløbet og virkningsgraden, om medicinen skal have effekt i et organ med stor blodgennemstrømning som f.eks. hjernen, hjertet eller et mål, hvor blodgennemstrømningen er mindre som f.eks. ekstremiteterne eller slet ikke eksisterende som f.eks. i knoglemarven.

11.1.3 Metabolisme

Metabolismen – eller stofskiftet – er betegnelsen for de processer i kroppen, hvor lægemidlet omdannes til andre kemiske forbindelser. Ofte forløber metabolismeprocesserne i leveren. For nogle medikintyper er det eksempelvis i leveren, det aktive stof omdannes til inaktivt stof for senere at blive udskilt fra kroppen. Nogle lægemidler metaboliseres faktisk slet ikke men udskilles derimod fra kroppen i sin oprindelige kemiske form, når virkningen er ophørt – det gælder for eksempel mange typer antibiotika. Atter andre medikintyper udnytter metabolismen til at danne en række andre stoffer, som så har en virkning.

Eksempelvis skyldes morfins smertestillende virkning hovedsageligt nogle af de stoffer, som via enzymer dannes gennem metabolismen i leveren og faktisk ikke morfinen selv. Undersøgelser har vist, at ca. 75% af en morfindosis hovedsageligt omdannes til tre forskellige kemiske forbindelser efter følgende fordeling:



Og her er det faktisk kun morfin-6-glucoronid og normorfin, der har den ønskede smertestillende virkning. Eksemplet med morfin illustrerer på glimrende vis, at når man skal undersøge virkningen af et stof, er det ikke nok at holde styr på stoffet, men også nødvendigt at kende produkterne af de kemiske processer, som medicinen gennemgår. Dette er ikke nogen simpel opgave, for ganske ofte forløber metabolismen af et lægemiddel samtidigt flere forskellige steder i kroppen. Hvordan og hvor meget, den enkelte proces fjerner, kan ikke altid bestemmes. Derfor bliver den matematiske modellering ofte en meget simplificeret udgave af virkeligheden.

11.1.4 Elimination

Eliminationen beskriver, hvordan medicinen til slut udskilles fra kroppen. Her er der ikke som sådan tale om en kemisk omdannelse, men om en direkte fjernelse fra kroppen. Mange stoffer ender i urinen efter at være blevet filtreret ud af blodet i nyrene. Kort sagt fungerer nyrene som en si, hvor store molekyler bliver siet fra og små molekyler slipper igennem.

11.2 Compartmentmodellering

For at danne sig et overblik over modelleringen af de forholdsvis komplicerede processer på lægemidlets rejse gennem kroppen benytter man ofte såkaldte compartmentmodeller. Et *compartment* i denne sammenhæng er en enhed eller afdeling af kroppen, hvor man antager, at lægemidlet opfører sig nogenlunde ensartet i forhold til eksempelvis distribution og metabolisme. Ideen er nu, at man forsøger at modellere koncentrationsflowet mellem de enkelte compartments. Koncentrationen angiver man typisk som masse pr. volumen med enheden g/L eller stofmængde pr. volumen med enheden for stofmængdekonzentration mol/L eller M (*molær*). Volumen for et compartment er distributionsvolumen. Hvis vi for eksempel regner blodet for et compartment vil distributionsvolumen således typisk være i størrelsesorden 5 L. Man skelner mellem 1-, 2- eller endda 3-compartmentmodeller alt efter hvor mange afdelinger, der inkluderes i modelleringen. I det følgende vil vi begrænse os til at se på 1- og 2-compartmentmodeller, da disse vil være dækkende i de fleste modelleringssituationer.

11.3 1-compartmentmodeller

I en 1-compartmentmodel interesserer man sig kun for koncentrationen af medicin i et enkelt compartment. Det kunne for eksempel være relevant i en situation, hvor et lægemiddel indføres intravenøst for derefter af virke direkte i blodbanen, hvor det efterfølgende udskilles eller elimineres fra blodbanen i nyrene. Compartmentmodellen kan illustreres grafisk således:



Hvis vi betegner den aktuelle stofmængdekonzentration i mol/L af lægemidlet i blodet til tiden t med $c_b(t)$, kan vi nu opstille forskellige scenarier, for hvordan $c_b(t)$ aftager som funktion af tiden.

11.3.1 0. ordens elimination

Hvis eliminationen af lægemidlet fra blodbanen forløber med konstant hastighed, kan vi opstille differentialligningen:

$$c_b'(t) = -K \quad (1)$$

Eliminationskonstanten K er positiv, hvilket forklarer minustegnet på differentialligningen højre side. Jo større K er, des hurtigere aftager koncentrationen af lægemidlet i blodbanen. Differentialligningen svarer til det, der inden for reaktionskinetik populært kaldes for en *0. ordens reaktion*.

Øvelse 11.2

Vis, at løsningen til differentialligningen (1) er det følgende lineære udtryk:

$$c_b(t) = -K \cdot t + c_0 \quad (2)$$

Løsningskurven er naturligvis en ret linje med hældningskoefficienten $-K$. Skæringen med andenaksen c_0 svarer til startkoncentrationen af lægemidlet i blodbanen.

Øvelse 11.3

Eftersom, der er tale om lineær vækst, er halveringstiden for et 0. ordens forløb ikke konstant. Den afhænger både af koncentrationen c og af eliminationskonstanten K . Vis, at halveringstiden er givet ved formlen:

$$T_{1/2} = \frac{c}{2 \cdot K} \quad (3)$$

På grafen her ses et eksempel på et 0. ordens eliminationsforløb for et lægemiddel, der er indført intravenøst i blodbanen.

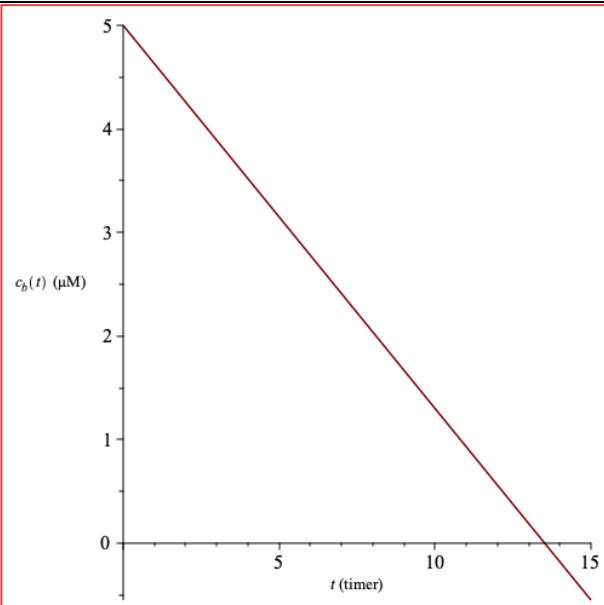
Startkoncentrationen er på $c_0 = 5,0mM$ og eliminationskonstanten $K = 0,37mM/time$.

Hældningskoefficienten er dermed $-0,37mM/time$.

Jo større K -værdi des hurtigere forløber eliminationen.

Øvelse 11.4

Giv en fortolkning af grafens skæringer med koordinatsystemet akser.



En 0. ordens elimination af medicin i kroppen

Der er kun få eksempler på lægemidler, der elimineres vha. 0. ordens kinetik. Det mest kendte er nok kroppen enzymatiske nedbrydning af "lægemidlet" alkohol (ethanol). Uanset hvor meget alkohol, man har drukket, eller hvor fysisk aktiv man er, nedbrydes der nemlig kun den samme mængde alkohol pr. tidsenhed. Mængden, som kan nedbrydes, afhænger af leverens størrelse, da det er her nedbrydningen foregår. Den viser sig at være omtrent proportional med kropsvægten, så en gennemsnitlig voksen person kan nedbryde cirka 0,1 g ethanol pr. kg kropsvægt pr. time.

Øvelse 11.5

En bestemt flaske vin indeholder 66,28g ethanol. Opstil et funktionsudtryk for massen af ethanol i din krop som funktion af tiden i timer ud fra din kropsvægt. Tegn grafen for sammenhængen og beregn, hvornår al ethanol er nedbrudt.

11.3.2. 1. ordens elimination

0. ordens elimination som ved nedbrydning af alkohol er dog en sjældenhed i virkelighedens verden. Den mængde lægemiddel, som kroppen er i stand til at fjerne pr. tid, afhænger ofte af den samlede mængde medicin i det pågældende compartment. Det kan sammenlignes med vand, der løber ud af hullet i en spand. Hvis der er meget vand i spanden, vil trykket være stort, og strålen af vand, der løber ud af spanden, vil være kraftig, men efterhånden som vandstanden falder, bliver strålen mindre og mindre kraftig. Dette svarer til en situation, hvor eliminationen af lægemidlet fra blodbanen er proportional med den samlede koncentration, og vi kan derfor opstille differentialligningen:

$$c_b'(t) = -K \cdot c_b(t) \quad (4)$$

Denne differentialligning genkender vi som en "eksponentiel type".

Øvelse 11.6

Vis, at løsningen til differentialligningen (4) er givet ved formlen: $c_b(t) = c_0 \cdot e^{-K \cdot t}$ (5)

Her er løsningen altså en eksponentielt aftagende funktion, hvor halveringstiden udelukkende afhænger af eliminationskonstanten K og derfor er konstant: $T_{1/2} = \frac{\ln(2)}{K}$ (6)

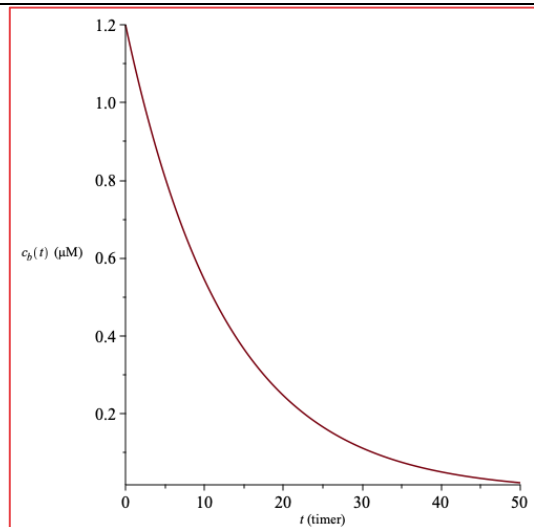
På denne graf ses et eksempel på et 1. ordens eliminationsforløb for et lægemiddel, der er indført intravenøst i blodbanen med en startkoncentration på: $c_0 = 1,2 \text{ mmol/L}$ og eliminationskonstant på:

$$K = 0,079 \text{ timer}^{-1}.$$

Vi genkender grafens eksponentielle forløb.

Øvelse 11.7

Beregn halveringskonstanten i eksemplet ovenfor. Hvor skærer grafen koordinatsystemets akser?



En 1. ordens elimination af medicin i kroppen

11.3.3. 2. ordens elimination

I princippet kan man også forestille sig mange andre muligheder for eliminationsmekanismer – f.eks. at lægemidlet udskilles fra blodbanen via 2. ordens kinetik. I så fald vil følgende differentialligning være opfyldt: $c_b'(t) = -K \cdot (c_b(t))^2$ (7)

Øvelse 11.8

Vis, at denne differentialligning har løsningsformlen: $c_b(t) = \frac{1}{K \cdot t + \frac{1}{c_0}}$ (8)

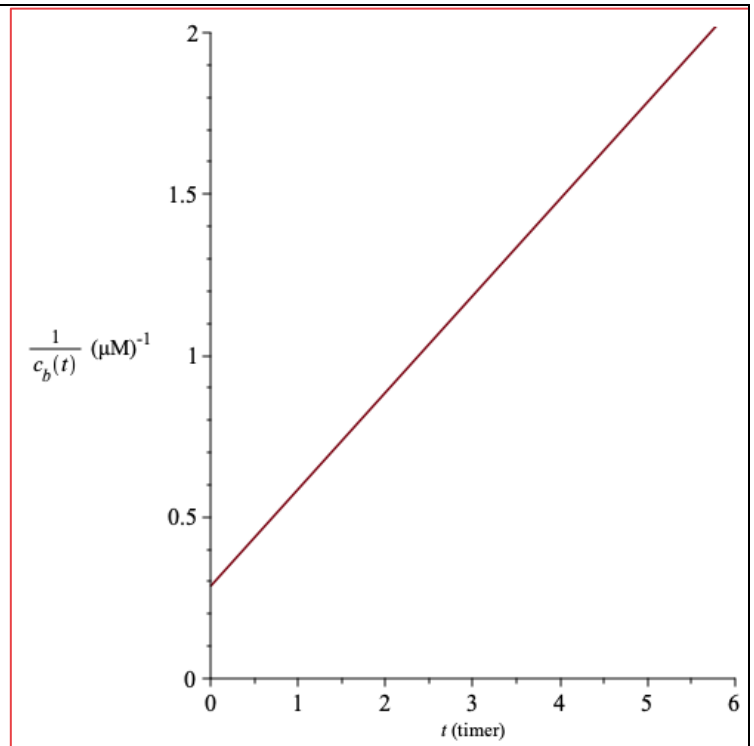
Oftentimes one writes the solution to the 2. order case in reciprocal form:

$$\frac{1}{c_b(t)} = K \cdot t + \frac{1}{c_0} \quad (9)$$

The above mentioned writing mode has the advantage, that a plot of the reciprocal concentration of the drug as a function of time will be linear with the elimination coefficient

K and the intercept with the y-axis $\frac{1}{c_0}$.

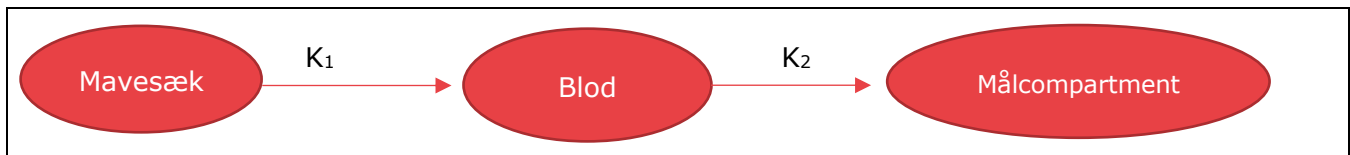
In the graph an example of a 2. order elimination process for a drug, which is administered intravenously in the bloodstream with an initial concentration $c_0 = 3,5 \text{ mM}$, and an elimination constant $K = 0,3 (\text{mM} \cdot \text{timer})^{-1}$



En 2. ordens elimination af medicin i kroppen

11.4. 2-compartmentmodeller

If the drug is taken orally, a 1-compartment model typically will be poor at reflecting reality. The model should at least include a stomach compartment, a compartment for the blood and possibly a target compartment for the area of effect in the body. This makes the mathematical modeling, although it is still possible to give good advice on solutions to the problem.



Let us look at an example corresponding to the above figure. Assume that the overall concentration change of the drug in the stomach will be determined by the differential equation corresponding to 1. order elimination:

$$c_m'(t) = -K_1 \cdot c_m(t) \quad (10)$$

The drug concentration in the blood will be determined by a positive contribution corresponding to diffusion from the stomach: $F \cdot K_1 \cdot c_m(t)$ and a negative contribution caused by the amount of drug, which is eliminated or sent further to the target compartment: $-K_2 \cdot c_b(t)$

K_1 and K_2 are proportionality constants and F is a so-called bioavailability of the drug. The above mentioned should be taken into account, as some drug will always be "lost", when it is administered without the bloodstream. This can for example be in the form of chemical transformation in the stomach or breakdown into inactive products. The value for F lies in the interval from 0 to 1. Overall, the situation corresponds to a bathtub, where water flows in from the tap with a certain speed and water flows out of the bottom of the tub. We also have the differential equation:

$$c_b'(t) = F \cdot K_1 \cdot c_m(t) - K_2 \cdot c_b(t) \quad (11)$$

Tilsammen kaldes differentialligningerne (10) og (11) for et koblet system af differentialligninger:

$$c_m'(t) = -K_1 \cdot c_m(t) \quad (10)$$

$$c_b'(t) = F \cdot K_1 \cdot c_m(t) - K_2 \cdot c_b(t) \quad (11)$$

Der er flere forskellige måder at løse systemet på. Vi vil her gennemgå to forskellige tilgange:

11.4.1. Løsning vha. lineær 1. ordens differentialligning ('Panserformlen')

Man kan f.eks. starte med at løse differentialligningen (10), som jo er en 1. ordens elimination, hvilket vil sige, at:

$$c_m(t) = c_0 \cdot e^{-K_1 t} \quad (12)$$

Hvor c_0 er startkoncentrationen af medicin i mavecompartmentet.

Øvelse 11.9

Vis, at man ved at indsætte (12) i differentialligningen (11) kan komme frem til:

$$c_b'(t) + K_2 \cdot c_b(t) = F \cdot K_1 \cdot c_0 \cdot e^{-K_1 t} \quad (13)$$

Differentialligningen (13) genkender vi som en lineær 1. ordens differentialligning, der kan løses vha. "Panserformlen". I formelsamlingen til MAT-A STX er løsningsformlen angivet i formel (180) således:

$$(180) \quad y' + a(x) \cdot y = b(x) \quad y = e^{-A(x)} \int b(x) \cdot e^{A(x)} dx + c \cdot e^{-A(x)},$$

hvor $A(x)$ er stamfunktion til $a(x)$

Øvelse 11.10

Benyt formel (180) fra Formelsamlingen til at vise, at løsningen til (13) bliver:

$$c_b(t) = F \cdot K_1 \cdot c_0 \cdot e^{-K_2 t} \cdot \int e^{(K_2 - K_1)t} dt + c \cdot e^{-K_2 t} \quad (14)$$

Hvilke regneregler har du benyttet i omskrivningen?

Hvis nu de to eliminationskonstanter K_1 og K_2 er forskellige, fås ud fra (14):

$$c_b(t) = F \cdot K_1 \cdot c_0 \cdot e^{-K_2 t} \cdot \frac{1}{K_2 - K_1} \cdot e^{(K_2 - K_1)t} + c \cdot e^{-K_2 t}, \quad K_1 \neq K_2 \quad (15)$$

Og:

$$c_b(t) = \frac{F \cdot K_1 \cdot c_0}{K_2 - K_1} \cdot e^{-K_1 t} + c \cdot e^{-K_2 t}, \quad K_1 \neq K_2 \quad (16)$$

Lad os nu yderligere antage, at koncentrationen af medicin i blodbanen er 0 ved start: $c_b(0) = 0$

Øvelse 11.11

Vis at man ved at indsætte den betingelse i (16) kan komme frem til: $c = \frac{F \cdot K_1 \cdot c_0}{K_1 - K_2}$

Dette giver os endeligt følgende udtryk for lægemiddelkoncentrationen i blodbanen:

$$c_b(t) = \frac{F \cdot K_1 \cdot c_0}{K_2 - K_1} \cdot e^{-K_1 t} + \frac{F \cdot K_1 \cdot c_0}{K_1 - K_2} \cdot e^{-K_2 t}, \quad K_1 \neq K_2 \quad (17)$$

Øvelse 11.12

Tegn i dit CAS-værktøj grafen for $c_b(t)$ fra ligning (17) med nogle passende parameterverdier – f.eks. $c_0=1.8$, $K_1=0.3$, $K_2=0.2$, $F=0.5$

Undersøg, hvordan grafen ændrer sig, når du ændrer på parameterverdierne

Alternativt kan du eksperimentere med parameterverdierne på følgende GeoGebra-graf:

<https://www.geogebra.org/classic/jbmkavsa>

Prøv at beskrive, hvilken betydning parametrene har for grafens udseende.

Brug så præcise fagord som muligt.

Hvis i (14) K_1 og K_2 er ens fås det noget simple udtryk:

$$c_b(t) = F \cdot K_1 \cdot c_0 \cdot e^{-K_1 t} \cdot t + c \cdot e^{-K_1 t} = (F \cdot K_1 \cdot c_0 \cdot t + c) \cdot e^{-K_1 t} \quad (18)$$

Hvis vi endnu engang antager, at koncentrationen af medicin i blodbanen er 0 ved start: $c_b(0) = 0$, da fås ud fra (18):

$$0 = (F \cdot K_1 \cdot c_0 \cdot 0 + c) \cdot e^{-K_1 \cdot 0} \Rightarrow c = 0$$

Hvilket giver dette udtryk for medicinkoncentrationen i blodet:

$$c_b(t) = F \cdot K_1 \cdot c_0 \cdot t \cdot e^{-K_1 t}, \quad K_1 = K_2 \quad (19)$$

Øvelse 11.13

Overvej, om det mon er rimeligt at antage, at $K_1 = K_2$. Hvad taler for? Hvad taler imod?

Øvelse 11.14

Tegn i dit CAS-værktøj grafen for $c_b(t)$ fra ligning (19) med nogle passende parameterverdier – f.eks. $c_0 = 1,8$, $K_1 = 0,3$, $F = 0,5$

Undersøg, hvordan grafen ændrer sig, når du ændrer på parameterverdierne

Alternativt kan du eksperimentere med parameterverdierne på følgende GeoGebra-graf:

<https://www.geogebra.org/classic/wrdk9vqj>

Prøv at beskrive, hvilken betydning parametrene har for grafens udseende.

Brug så præcise fagord som muligt.

11.4.2. Løsning ved omskrivning til en 2. ordens differentialligning

I dette afsnit trækker vi kraftigt på det omtalte afsnit fra HEM3.

Det koblede differentialligningssystem kan også løses ved at omskrive til en 2. ordens differentialligning. Vi tager igen udgangspunkt i det koblede system (10) og (11):

$$c_m'(t) = -K_1 \cdot c_m(t) \quad (10)$$

$$c_b'(t) = F \cdot K_1 \cdot c_m(t) - K_2 \cdot c_b(t) \quad (11)$$

Vi starter med at differentiere mht. t i (11) på begge sider af lighedstegnet:

$$c_b''(t) = F \cdot K_1 \cdot c_m'(t) - K_2 \cdot c_b'(t) \quad (20)$$

I ligning (20) indsætter vi nu udtrykket for $c_m'(t)$ fra (10):

$$c_b''(t) = -F \cdot K_1^2 \cdot c_m(t) - K_2 \cdot c_b'(t) \quad (21)$$

I ligning (21) vil vi gerne af med variabelen $c_m(t)$, så derfor isoleres denne først i ligning (11):

$$\begin{aligned}c_b'(t) &= F \cdot K_1 \cdot c_m(t) - K_2 \cdot c_b(t) \\c_b'(t) + K_2 \cdot c_b(t) &= F \cdot K_1 \cdot c_m(t) \\c_m(t) &= \frac{c_b'(t) + K_2 \cdot c_b(t)}{F \cdot K_1}\end{aligned}$$

Dette udtryk for $c_m(t)$ indsættes nu i (21), hvorved vi får en 2. ordens differentialligning:

$$c_b''(t) = -F \cdot K_1^2 \cdot \frac{c_b'(t) + K_2 \cdot c_b(t)}{F \cdot K_1} - K_2 \cdot c_b'(t)$$

Som kan omskrives til:

$$c_b''(t) = -K_1 \cdot (c_b'(t) + K_2 \cdot c_b(t)) - K_2 \cdot c_b'(t)$$

Og videre til:

$$c_b''(t) + (K_1 + K_2) \cdot c_b'(t) + K_1 \cdot K_2 \cdot c_b(t) = 0 \quad (22)$$

Denne 2. ordens differentialligning har det karakteristiske polynomium:

$$p(x) = x^2 + (K_1 + K_2) \cdot x + K_1 \cdot K_2$$

Øvelse 11.15

Vis at diskriminanten for det karakteristiske andengradspolynomium er lig med:

$$d = (K_1 - K_2)^2$$

Her skal vi atter dele op i to tilfælde alt efter om eliminationskonstanterne K_1 og K_2 er forskellige. Hvis K_1 og K_2 er forskellige, er diskriminanten positiv, og så fås rødderne for det karakteristiske polynomium til:

$$x = \frac{-b \pm \sqrt{d}}{2 \cdot a} = \frac{-K_1 - K_2 \pm \sqrt{(K_1 - K_2)^2}}{2 \cdot 1} = \frac{-K_1 - K_2 \pm (K_1 - K_2)}{2} = \begin{cases} K_1 \\ K_2 \end{cases}$$

Løsningen til differentialligningen (22) bliver af typen:

$$c_b(t) = a \cdot e^{-K_1 t} + b \cdot e^{-K_2 t}, \text{ hvor } a \text{ og } b \text{ er vilkårlige konstanter.}$$

Lad os nu endnu engang antage, at koncentrationen af medicin i blodbanen er 0 ved start: $c_b(0) = 0$. Herved fås:

$$\begin{aligned}c_b(0) &= a \cdot e^{-K_1 \cdot 0} + b \cdot e^{-K_2 \cdot 0} \\0 &= a + b \\b &= -a\end{aligned}$$

Løsningen kan så omskrives til:

$$c_b(t) = a \cdot e^{-K_1 t} - a \cdot e^{-K_2 t} \quad (23)$$

Koefficienten a afhænger naturligvis af parametrene F , c_0 , K_1 og K_2 , og vi ser, at vi heldigvis er nået frem til et funktionsudtryk, der svarer til (17):

$$c_b(t) = \frac{F \cdot K_1 \cdot c_0}{K_2 - K_1} \cdot e^{-K_1 t} + \frac{F \cdot K_1 \cdot c_0}{K_1 - K_2} \cdot e^{-K_2 t}, \quad K_1 \neq K_2 \quad (17)$$

Hvis eliminationskonstanterne K_1 og K_2 er ens, bliver diskriminanten i det karakteristiske polynomium 0, hvorved vi får dobbeltroden:

$$x = \frac{-b \pm \sqrt{d}}{2 \cdot a} = \frac{-K_1 - K_1}{2 \cdot 1} = K_1$$

Herved bliver løsningen til differentialligningen (22) af typen:

$$c_b(t) = a \cdot t \cdot e^{-K_1 t} + b \cdot e^{-K_1 t}, \text{ hvor } a \text{ og } b \text{ igen er vilkårlige konstanter.}$$

Lad os nu endnu engang antage, at koncentrationen af medicin i blodbanen er 0 ved start: $c_b(0) = 0$. Herved fås:

$$\begin{aligned} c_b(0) &= a \cdot 0 \cdot e^{-K_1 \cdot 0} + b \cdot e^{-K_1 \cdot 0} \\ 0 &= 0 + b \cdot 1 \\ b &= 0 \end{aligned}$$

hvorved løsningen omskrives til: $c_b(t) = a \cdot t \cdot e^{-K_1 t}$ (24)

Og som før, så afhænger konstanten a af parametrene F , c_0 , K_1 og K_2 , så vi er atter engang nået frem til et funktionsudtryk, der svarer til (19), som vi fandt frem til vha. "Panserformlen":

$$c_b(t) = F \cdot K_1 \cdot c_0 \cdot t \cdot e^{-K_1 t}, \quad K_1 = K_2 \quad (19)$$

Øvelse 11.16 Koncentrationen i blodet af et bestemt lægemiddel er blevet undersøgt over tid, og man kom frem til følgende data: a) Lav et punktplot af data i dit CAS-værktøj b) Benyt CAS-værktøjets faciliteter til at fitte data til en funktion med forskriften som i ligning (19): $c_b(t) = F \cdot K_1 \cdot c_0 \cdot t \cdot e^{-K_1 t}, \quad K_1 = K_2$ c) Hvad bliver parameterværdierne F , c_0 , K_1 og K_2 ? d) Tegn grafen af $c_b(t)$	Tid (timer)	Koncentration (nmol/L)
	0,08	0,519
	0,17	1,04
	0,25	1,22
	0,33	1,53
	0,50	1,63
	0,67	1,59
	0,83	1,57
	1,00	1,50
	1,25	1,43
	1,83	1,01
	2,33	0,745
	3,00	0,399

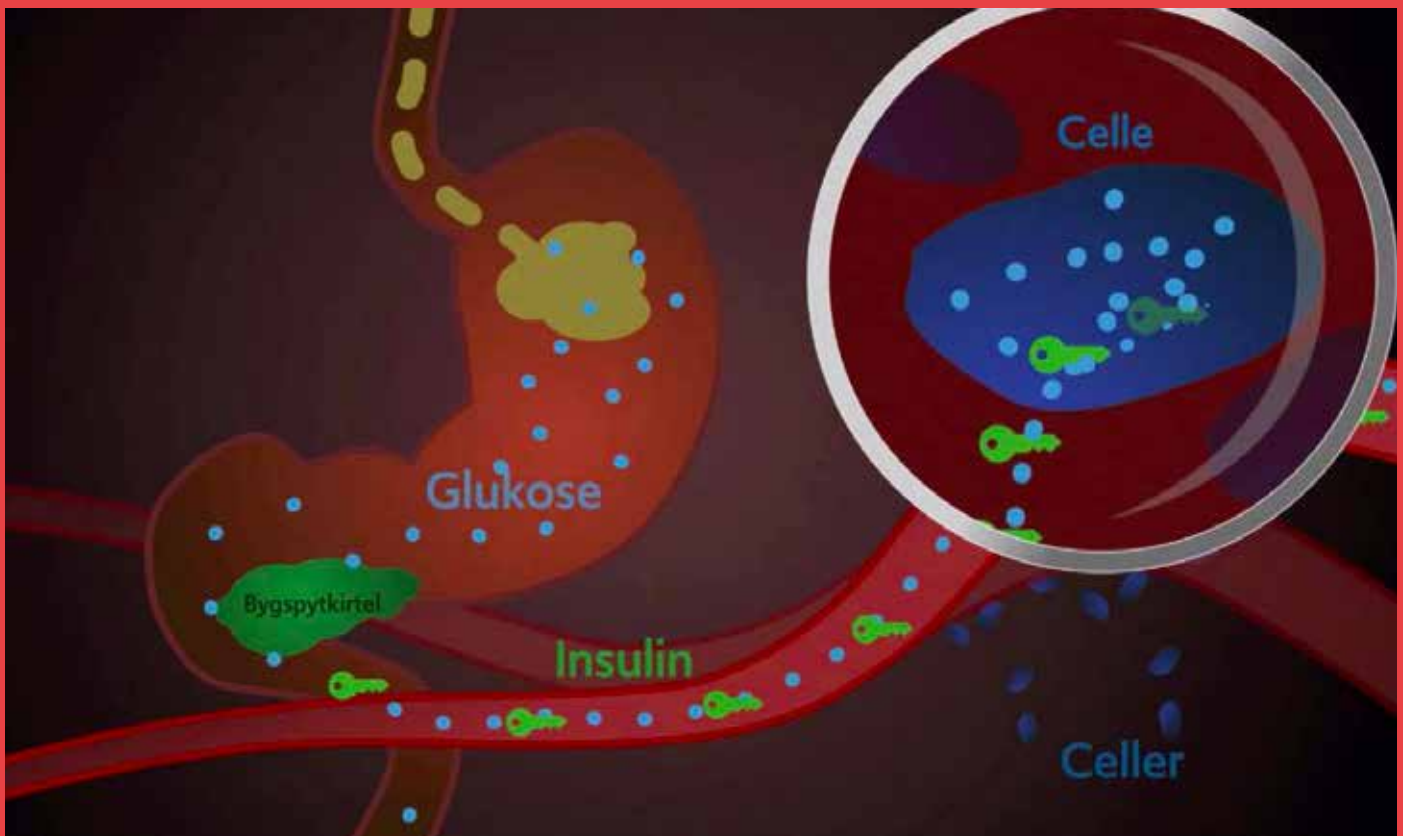
11.5 Mere om compartmentmodeller

Hvis du har lyst til at læse mere om compartmentmodellering og den matematiske side af medicins rejse gennem kroppen er der nogle kilder til videre studier her:

- Baktoft, Allan (2017): "Matematik i virkeligheden, Bind 4", Forlaget Natskyggen
- Brydesholt, Ebbesen (2012): "Lærebog i matematik – Bind 4", Systeme
- Brøsen, Simonsen, Dalhoff (2019): "Basal og klinisk farmakologi", FADL's Forlag
- Jensen, Jensen, Mogensen, Paulsen (1992): "Lægemiddelkemi", Systeme
- Jónsdóttir, Klim, Mortensen og Madsen (2009): "Matematik i medicinudvikling", Matematiske Horisonter, DTU, <https://www.compute.dtu.dk/om-os/matematiske-horisonter>

12.

Matematisk modellering af diabetes



I filmen fortælles, hvordan næringsstoffer omdannes til glukose, der for raske mennesker transporteres ud til cellerne. I dette kapitel vil vi opstille en matematisk model for dette, hvor vi både ser på mængden af glukose i fordøjelsessystemet og i blodet. Målet med modelleringen er at få en matematisk beskrivelse af glukoseniveauet i blodet, og en model der gælder både for diabetikere og for raske mennesker.

12. Matematisk modellering af diabetes

(Kapitel 12 bygger i vid udstrækning på artiklen: *Mathematical Model of ingested glucose in Glucose-Insulin Regulation*. Artiklen kan hentes [her](#) [link til 42](#)

Modellen oversættes til en inhomogen anden ordens differentiaalligning. Dette emneområde er behandlet med udledning af løsningsformlerne og gennemregnede eksempler i HEM 3, kap 6. Kapitlet kan hentes her: [link til 41](#). Løsningsmetoderne kan findes i projekt 6.5 i HEM3 – [link til 43](#)

12.1 Introduktion til modellen

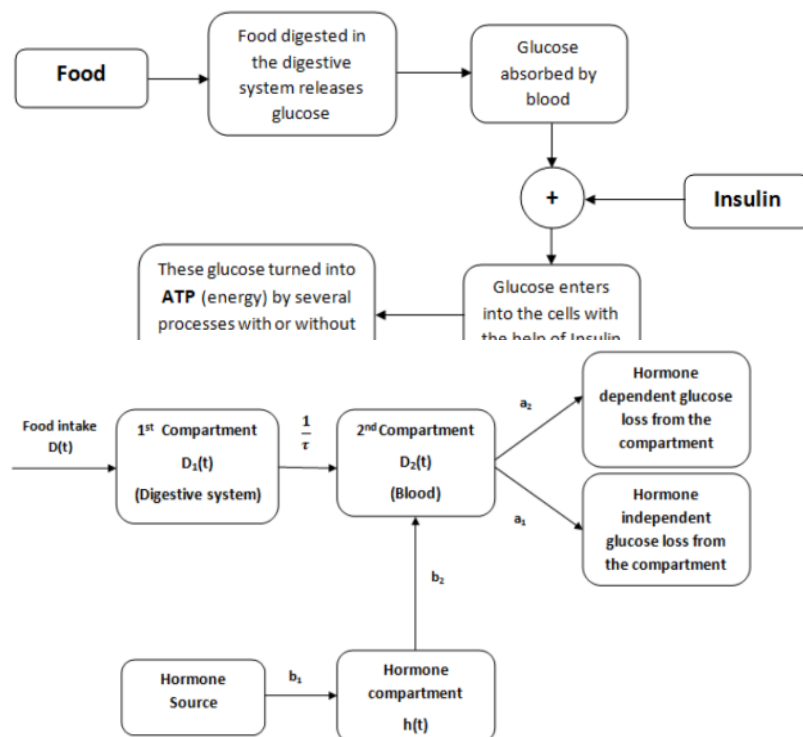
I filmen fortælles, hvordan næringsstoffer omdannes til glukose, der for raske mennesker transporteres ud til cellerne. Det kan illustreres med en compartment-model således:

Vi vil her modellere situationen, idet vi både ser på mængden af glukose i fordøjelsessystemet og i blodet.

De variable, der indgår i vores model, er her angivet på lidt udbygget compartment-model.

I det følgende kan du relatere de variable og nogle af parametrene til denne illustration.

Målet med modelleringen er at få en matematisk beskrivelse af glukoseniveauet i blodet



12.2 Opstilling af modellen

I modellen haves et system af 3 differentiaalligninger

$$\begin{aligned} \frac{dD_1}{dt} &= -\frac{D_1}{\tau} \\ \frac{dD_2}{dt} &= \frac{D_1}{\tau} - a_1 \cdot D_2 - a_2 \cdot h \\ \frac{dh}{dt} &= b_1 \cdot D_2 - b_2 \cdot h \end{aligned} \quad (*)$$

med begyndelsesbetingelserne

$$D_1(0) = D \cdot A_G \quad D_2(0) = 0 \quad h(0) = 0.$$

De tre afhængige variable D_1 , D_2 og h afhænger alle af tiden t .

- $D_1(t)$ beskriver glukose **ændringen** i fordøjelsessystemet, og da det antages, at glukose niveauet ved tom mave (faste) er 0, så svarer $D_1(t)$ til glukose niveauet i fordøjelsessystemet.
- $D_2(t)$ beskriver glukose **ændringen** i blodet ud fra $D_2(t) = G(t) - G_b$, hvor $G(t)$ er glukose niveauet i blodet og G_b er glukose niveauet i blodet ved tom mave (faste).
- $h(t)$ er **ændringen** i det effektive hormon niveau ud fra $h(t) = H(t) - H_b$, hvor $H(t)$ er det effektive hormon niveau, og H_b er det effektive hormon niveau ved tom mave (faste).

Alle enheder måles mg/dl og timer.

Konstanterne τ , a_1 , a_2 , b_1 og b_2 i modellen kan beskrives som

- τ positiv konstant, der beskriver den reciprokke rate, hvormed glukose niveauet i fordøjelsessystemet ændres.
- a_1 positiv konstant, der beskriver raten, hvormed glukose niveauet i blodet, uafhængigt af hormoner, ændres.
- a_2 positiv konstant, der beskriver raten, hvormed glukose niveauet i blodet, afhængigt af hormoner, ændres.
- b_1 positiv konstant, der beskriver raten, hvormed hormon niveauet, afhængigt af ændringen i glukose niveauet i blodet, ændres.
- b_2 positiv konstant, der beskriver raten, hvormed hormon niveauet, uafhængigt af ændringen i glukose niveauet i blodet, ændres.

12.3 Løsning af differentialligningssystemet

Trin 1

Vi lægger mærke til, at D_2 og h ikke indgår i differentialligningen $\frac{dD_1}{dt} = -\frac{D_1}{\tau}$. Så denne differentialligning kan vi genkende som en førsteordens differentialligning i D_1 af typen $\frac{dy}{dt} = k \cdot y$, hvor den fuldstændige løsning er $y = c \cdot e^{k \cdot x}$.

Dvs. den fuldstændige løsning til $\frac{dD_1}{dt} = -\frac{D_1}{\tau}$ er: $D_1(t) = k \cdot e^{-\frac{t}{\tau}}$.

Øvelse 12.1

- Argumenter for dette.
- Bestem konstanten k , ud fra begyndelsesbetingelsen $D_1(0) = D \cdot A_G$
- Vis, at den partikulære løsning er: $D_1(t) = D \cdot A_G \cdot e^{-\frac{t}{\tau}}$.

Trin 2

En strategi for løsning af de to andre differentialligninger, som er koblede, kan være at omskrive dem til **én** andenordensdifferentialligning.

Vi differentierer differentialligningen $\frac{dD_2}{dt} = \frac{D_1}{\tau} - a_1 \cdot D_2 - a_2 \cdot h$ mht. t :

$$\frac{d^2 D_2}{dt^2} = \frac{1}{\tau} \frac{dD_1}{dt} - a_1 \cdot \frac{dD_2}{dt} - a_2 \cdot \frac{dh}{dt}.$$

Øvelse 12.2

Vis at ovenstående omskrivning er korrekt.

I denne andenordens differentialligning kan vi indsætte den tredje differentialligning

$$\frac{dh}{dt} = b_1 \cdot D_2 - b_2 \cdot h.$$

Vi får dermed

$$\begin{aligned} \frac{d^2 D_2}{dt^2} &= \frac{1}{\tau} \frac{dD_1}{dt} - a_1 \cdot \frac{dD_2}{dt} - a_2 \cdot (b_1 \cdot D_2 - b_2 \cdot h) \\ \frac{d^2 D_2}{dt^2} &= \frac{1}{\tau} \frac{dD_1}{dt} - a_1 \cdot \frac{dD_2}{dt} - a_2 \cdot b_1 \cdot D_2 + a_2 \cdot b_2 \cdot h \end{aligned} \quad (**)$$

Hvor vi i den anden ligning blot har ganget ind i parentesen.

Vores mål er at opstille en andenordensdifferentialligning i D_2 . Vi ønsker derfor at eliminere D_1 og h fra ligningen.

Først h : Vi isolerer h i den anden af differentialligningerne i systemet (*):

$$\frac{dD_2}{dt} = \frac{D_1}{\tau} - a_1 \cdot D_2 - a_2 \cdot h \Leftrightarrow a_2 \cdot h = -\frac{dD_2}{dt} + \frac{D_1}{\tau} - a_1 \cdot D_2.$$

Indsættes dette udtryk i (**) får vi:

$$\frac{d^2 D_2}{dt^2} = \frac{1}{\tau} \frac{dD_1}{dt} - a_1 \cdot \frac{dD_2}{dt} - a_2 \cdot b_1 \cdot D_2 + b_2 \cdot \left(-\frac{dD_2}{dt} + \frac{D_1}{\tau} - a_1 \cdot D_2 \right)$$

Øvelse 12.3

Vis at ligningen kan omskrives til:

$$\frac{d^2 D_2}{dt^2} + (a_1 + b_2) \cdot \frac{dD_2}{dt} + (a_2 \cdot b_1 + a_1 \cdot b_2) \cdot D_2 = \frac{1}{\tau} \frac{dD_1}{dt} + b_2 \cdot \frac{D_1}{\tau}$$

Vi kan nu udnytte, at vi har et løsningsudtryk for $D_1(t) = D \cdot A_G \cdot e^{-\frac{t}{\tau}}$.

Øvelse 12.4

Indsæt udtrykket for $D_1(t)$ i andenordensdifferentialligningen, og vis, at vi derved får den første af differentialligningerne nedenfor. Og vis dernæst, at denne kan reduceres til den anden af differentialligningerne:

$$\begin{aligned} \frac{d^2 D_2}{dt^2} + (a_1 + b_2) \cdot \frac{dD_2}{dt} + (a_2 \cdot b_1 + a_1 \cdot b_2) \cdot D_2 &= \frac{1}{\tau} \cdot \left(-\frac{1}{\tau} \right) \cdot D \cdot A_G \cdot e^{-\frac{t}{\tau}} + \frac{b_2}{\tau} \cdot D \cdot A_G \cdot e^{-\frac{t}{\tau}} \\ \frac{d^2 D_2}{dt^2} + (a_1 + b_2) \cdot \frac{dD_2}{dt} + (a_2 \cdot b_1 + a_1 \cdot b_2) \cdot D_2 &= \left(\frac{b_2}{\tau} - \frac{1}{\tau^2} \right) \cdot D \cdot A_G \cdot e^{-\frac{t}{\tau}} \end{aligned}$$

Vi har nu fået en inhomogen lineær andenordensdifferentialligning i D_2

Vi omdøber konstanterne således:

$$a_1 + b_2 = 2a, \quad a_2 \cdot b_1 + a_1 \cdot b_2 = \omega_0^2 \quad \text{og} \quad \left(\frac{b_2}{\tau} - \frac{1}{\tau^2} \right) \cdot D \cdot A_G = C.$$

Med denne omdøbning får vi en inhomogen lineær andenordensdifferentialligning på formen

$$\frac{d^2 D_2}{dt^2} + 2a \cdot \frac{dD_2}{dt} + \omega_0^2 \cdot D_2 = C \cdot e^{-\frac{t}{\tau}}.$$

Trin 3

For at løse denne *inhomogene* lineære andenordensdifferentialligning, løser vi først den *homogene* andenordens differentialligning:

$$\frac{d^2 D_2}{dt^2} + 2a \cdot \frac{dD_2}{dt} + \omega_0^2 \cdot D_2 = 0$$

Vi betragter derfor det karakteristiske polynomium.

$$x^2 + 2a \cdot x + \omega_0^2 = 0, \text{ hvor } a = 1, b = 2a \text{ og } c = \omega_0^2.$$

Øvelse 12.5

Vis, at diskriminanten er $d = 4 \cdot (a^2 - \omega_0^2)$

Vi leder efter en løsning, der grafisk er en dæmpet svingning, fordi $D_2(t)$ skal klinge af.

Men så skal diskriminanten opfylde $d < 0$, jfr HEM3 s. 293-98. Dette betyder:

$$a^2 - \omega_0^2 < 0 \Leftrightarrow a^2 < \omega_0^2 .$$

Øvelse 12.6

Hent løsningsformlen fra HEM3 s. 294, og opskriv løsningen til den homogene andenordens differentiaalligning. Du skal få:

$$D_2(t) = e^{-\frac{2a}{2}t} \cdot (A_1 \cdot \cos(\omega \cdot t) + A_2 \cdot \sin(\omega \cdot t)) , \text{ hvor}$$

$$\omega = \frac{\sqrt{4 \cdot 1 \cdot \omega_0^2 - (2a)^2}}{2} = \frac{\sqrt{4 \cdot (\omega_0^2 - a^2)}}{2} = \frac{\sqrt{4} \cdot \sqrt{(\omega_0^2 - a^2)}}{2} = \sqrt{(\omega_0^2 - a^2)} .$$

Øvelse 12.7

Gennemfør ovenstående omskrivninger.

Dette må betyde, at $\omega^2 = \omega_0^2 - a^2$.

Øvelse 12.8

Forklar valget af $a_1 + b_2 = 2a$ og $a_2 \cdot b_1 + a_1 \cdot b_2 = \omega_0^2$.

Løsningen til den homogene andenordens differentiaalligning kan omskrives til

$$D_2(t) = e^{-at} \cdot (A_1 \cdot \cos(\omega \cdot t) + A_2 \cdot \sin(\omega \cdot t)).$$

Trin 4 (Gættemetoden til løsning af den inhomogene differentiaalligning er behandlet i HEM3, projekt 6.5, som du kan hente [her](#) - link til 43)

Vi skal nu gætte på en løsning til den inhomogene andenordensdifferentiaalligning, og da højresiden i denne er en eksponentialfunktion, gætter vi på $f(t) = c_0 \cdot e^{-\frac{t}{\tau}}$.

Øvelse 12.9

Forklar hvorfor gættet på eksponentialfunktionen skal have ovenstående form.

Vi skal nu undersøge ved indsættelse, hvilket udtryk konstanten c_0 skal have.

Øvelse 12.10

- a) Givet $f(t) = c_0 \cdot e^{-\frac{t}{\tau}}$ bestem da $f'(t)$ og $f''(t)$.
 b) Vi indsætter vores gæt på en løsning til den inhomogene andenordens differentialligning.

$$\left(c_0 \cdot e^{-\frac{t}{\tau}}\right)'' + 2a \cdot \left(c_0 \cdot e^{-\frac{t}{\tau}}\right)' + \omega_0^2 \cdot c_0 \cdot e^{-\frac{t}{\tau}} = C \cdot e^{-\frac{t}{\tau}}$$

$$\frac{c_0}{\tau^2} \cdot e^{-\frac{t}{\tau}} - \frac{2a}{\tau} \cdot c_0 \cdot e^{-\frac{t}{\tau}} + \omega_0^2 \cdot c_0 \cdot e^{-\frac{t}{\tau}} = C \cdot e^{-\frac{t}{\tau}}$$

$$\left(\frac{c_0}{\tau^2} - \frac{2a}{\tau} \cdot c_0 + \omega_0^2 \cdot c_0\right) \cdot e^{-\frac{t}{\tau}} = C \cdot e^{-\frac{t}{\tau}}$$

$$\left(\frac{c_0}{\tau^2} - \frac{2a}{\tau} \cdot c_0 + \omega_0^2 \cdot c_0\right) = C$$

Øvelse 12.11

Forklar i alle detaljer udregningerne og reduktionerne i ovenstående. Forklar specielt, hvad det er for en egenskab ved eksponentialfunktionen, som gør, at vi kan forkorte den væk i den sidste ligning.

Da vi er på jagt efter et udtryk for c_0 , så løser vi $\frac{c_0}{\tau^2} - \frac{2a}{\tau} \cdot c_0 + \omega_0^2 \cdot c_0 = C$ mht. c_0 .

$$\frac{c_0}{\tau^2} - \frac{2a}{\tau} \cdot c_0 + \omega_0^2 \cdot c_0 = C$$

$$c_0 \cdot \left(\frac{1}{\tau^2} + \omega_0^2 - \frac{2a}{\tau}\right) = C$$

For at undgå brøker, så multiplicerer vi med τ^2 i ovenstående ligning.

$$c_0 \cdot (1 + \omega_0^2 \cdot \tau^2 - 2a \cdot \tau) = C \cdot \tau^2 \quad (***)$$

Øvelse 12.12

Forklar udregningerne og reduktionerne i ovenstående.

Da $\omega^2 = \omega_0^2 - a^2$ er $\omega_0^2 = \omega^2 + a^2$.

Og C blev indført som en ny variabel i stedet for udtrykket: $\left(\frac{b_2}{\tau} - \frac{1}{\tau^2}\right) \cdot D \cdot A_G$

Indsættes disse to udtryk for henh. ω_0^2 og C i (***), får vi:

$$c_0 \cdot (1 + \omega^2 \cdot \tau^2 + a^2 \cdot \tau^2 - 2a \cdot \tau) = \left(\frac{b_2}{\tau} - \frac{1}{\tau^2}\right) \cdot D \cdot A_G \cdot \tau^2$$

$$c_0 \cdot ((1 - a \cdot \tau)^2 + \omega^2 \cdot \tau^2) = (b_2 \cdot \tau - 1) \cdot D \cdot A_G$$

$$c_0 = \frac{(b_2 \cdot \tau - 1) \cdot D \cdot A_G}{(1 - a \cdot \tau)^2 + \omega^2 \cdot \tau^2}$$

Øvelse 12.13

Forklar i alle detaljer udregningerne og reduktionerne i ovenstående.

Konklusion: Hvis $c_0 = \frac{(b_2 \cdot \tau - 1) \cdot D \cdot A_G}{(1 - a \cdot \tau)^2 + \omega^2 \cdot \tau^2}$, så er $f(t) = c_0 \cdot e^{-\frac{t}{\tau}}$ en løsning til den inhomogene andenordens differentialligning.

Vi har dermed, at den fuldstændige løsning til den inhomogene differentialligning kan skrives på formen:

$$D_2(t) = e^{-at} \cdot (A_1 \cdot \cos(\omega \cdot t) + A_2 \cdot \sin(\omega \cdot t)) + c_0 \cdot e^{-\frac{t}{\tau}}$$

Trin 5

Vi kan nu anvende begyndelsesbetingelserne $D_2(0) = 0$ og $D_2'(0) = \frac{A_G \cdot D}{\tau}$.

Øvelse 12.14

Forklar $D_2'(0) = \frac{A_G \cdot D}{\tau}$ ud fra den oprindelige differentialligning i D_2 og de oprindelige begyndelsesbetingelser.

Indsættes $t = 0$ fås:

$$D_2(0) = e^{-a \cdot 0} \cdot (A_1 \cdot \cos(\omega \cdot 0) + A_2 \cdot \sin(\omega \cdot 0)) + c_0 \cdot e^{-\frac{0}{\tau}} = A_1 + c_0$$

Og da $D_2(0) = 0$, så får vi $A_1 = -c_0$.

Øvelse 12.15

Vis, at differentiation af $D_2(t)$ giver:

$$D_2'(t) = -a \cdot e^{-at} \cdot (A_1 \cdot \cos(\omega \cdot t) + A_2 \cdot \sin(\omega \cdot t)) + e^{-at} \cdot (-A_1 \cdot \omega \cdot \sin(\omega \cdot t) + A_2 \cdot \omega \cdot \cos(\omega \cdot t)) - \frac{c_0}{\tau} \cdot e^{-\frac{t}{\tau}}$$

Indsættes $t = 0$ heri fås:

$$D_2'(0) = -a \cdot e^{-a \cdot 0} \cdot (A_1 \cdot \cos(\omega \cdot 0) + A_2 \cdot \sin(\omega \cdot 0)) + e^{-a \cdot 0} \cdot (-A_1 \cdot \omega \cdot \sin(\omega \cdot 0) + A_2 \cdot \omega \cdot \cos(\omega \cdot 0)) - \frac{c_0}{\tau} \cdot e^{-\frac{0}{\tau}}$$

$$D_2'(0) = -a \cdot A_1 + \omega \cdot A_2 - \frac{c_0}{\tau}$$

Indsæt nu heri udtrykket for $D_2'(0)$ fra begyndelsesbetingelsen, samt udnyt din viden om at $A_1 = -c_0$:

$$\frac{A_G \cdot D}{\tau} = -a \cdot A_1 + \omega \cdot A_2 - \frac{c_0}{\tau}$$

$$\frac{A_G \cdot D}{\tau} - a \cdot c_0 + \frac{c_0}{\tau} = \omega \cdot A_2 \quad \text{indsæt og ryk over}$$

$$A_G \cdot D - a \cdot c_0 \cdot \tau + c_0 = \tau \cdot \omega \cdot A_2 \quad \text{gang på begge sider med } \tau$$

$$A_2 = \frac{A_G \cdot D - a \cdot c_0 \cdot \tau + c_0}{\tau \cdot \omega}$$

Hermed er konstanterne A_1 og A_2 bestemt.

Den partikulære løsning til den inhomogene andenordens differentialligning er således:

$$D_2(t) = e^{-at} \cdot \left(-c_0 \cdot \cos(\omega \cdot t) + \left(\frac{A_G \cdot D - a \cdot c_0 \cdot \tau + c_0}{\tau \cdot \omega} \right) \cdot \sin(\omega \cdot t) \right) + c_0 \cdot e^{-\frac{t}{\tau}},$$

hvor
$$c_0 = \frac{(b_2 \cdot \tau - 1) \cdot D \cdot A_G}{(1 - a \cdot \tau)^2 + \omega^2 \cdot \tau^2}, \quad \omega^2 = \omega_0^2 - a^2 \text{ og } a^2 < \omega_0^2.$$

Og dermed kan udviklingen i glukose niveauet i blodet som funktion af tiden udtrykkes ved funktionen:

$$G(t) = G_b + D_2(t)$$

$$G(t) = G_b + e^{-at} \cdot \left(-c_0 \cdot \cos(\omega \cdot t) + \left(\frac{A_G \cdot D - a \cdot c_0 \cdot \tau + c_0}{\tau \cdot \omega} \right) \cdot \sin(\omega \cdot t) \right) + c_0 \cdot e^{-\frac{t}{\tau}}$$

Øvelse 12.16

Forklar ovenstående udtryk for $G(t)$.

12.4 Estimering af parametrene og tegning af grafiske billeder

Vi har nu et formeludtryk, og ville kunne tegne et grafisk billede, hvis vi kendte størrelserne af de parametre, der indgår. Men dem kan vi estimere, ud fra et sæt af datapunkter.

Eksempel: Parameterestimering

En person måler sit blodsukker efter 12 timers faste og indtag af 70 g glukose.

Tabellen viser en række sammenhørende data for personens målinger.

Tid (målt i timer)	0	$\frac{1}{3}$	$\frac{2}{3}$	1	$\frac{4}{3}$	$\frac{5}{3}$	2	$\frac{7}{3}$
Blodsukker (målt i mmol/l)	4,4	4,8	7,2	7,5	4,3	4,7	4,5	4,1

Ved hjælp af NonLinearFit i Maple kan parametrene i modellen bestemmes med regression

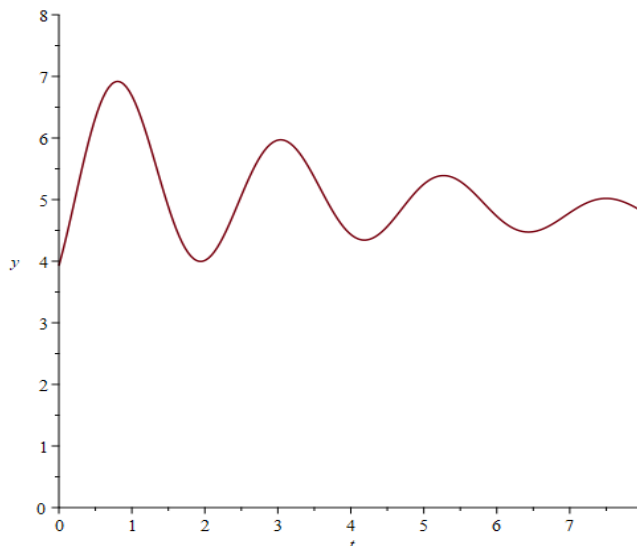
$$G(t) = G_b + e^{-at} \cdot \left(-c_0 \cdot \cos(\omega \cdot t) + \left(\frac{A_G \cdot D - a \cdot c_0 \cdot \tau + c_0}{\tau \cdot \omega} \right) \cdot \sin(\omega \cdot t) \right) + c_0 \cdot e^{-\frac{t}{\tau}}$$

hvor vi antager, at personen optager 80% af de 70 g glukose dvs. $A_G \cdot D = 0,8 \cdot 70 = 56$.

I Maplearket ser det således ud:

```
with(Gym) :
with(Statistics) :
datatid := [0, 1/3, 2/3, 1, 4/3, 5/3, 2, 7/3]
datatid := [0, 1/3, 2/3, 1, 4/3, 5/3, 2, 7/3]
dataglukoseperson := [4.4, 4.8, 7.2, 7.5, 4.3, 4.7, 4.5, 4.1]
dataglukoseperson := [4.4, 4.8, 7.2, 7.5, 4.3, 4.7, 4.5, 4.1]
G(t) := NonlinearFit(G_b + e^{-alfa*t} * (-c_0*cos(omega*t) + (56 - alfa*c_0*tau + c_0)/(tau*omega)*sin(omega*t)) + c_0*e^{-t/tau}, datatid, dataglukoseperson, t)
G := t -> NonlinearFit(G_b + e^{-alfa*t} * (-c_0*cos(omega*t) + (56 - alfa*c_0*tau + c_0)/(tau*omega)*sin(omega*t)) + c_0*e^{-t/tau}, datatid, dataglukoseperson, t)
G(t)
3.92569148747282 + e^{-0.270743482638729 t} (-1.46265709477275 cos(2.80692382981074 t) + 1.39812021472327 sin(2.80692382981074 t))
+ 1.46265709477275 e^{-0.0751866001648030 t}
```

Grafen kan nu tegnes:



Øvelse 12.18

En person måler sit blodsukker efter 12 timers faste og indtag af 50 g glukose. Tabellen viser en række sammenhørende data for personens målinger.

Tid (målt i timer)	0	$\frac{1}{3}$	$\frac{2}{3}$	1	$\frac{4}{3}$	$\frac{5}{3}$	2	$\frac{7}{3}$
Blodsukker (målt i mmol/l)	4,6	4,7	6,9	7,2	4,4	4,8	4,6	4,2

a) Bestem ved regression parametrene i modellen

$$G(t) = G_b + e^{-at} \cdot \left(-c_0 \cdot \cos(\omega \cdot t) + \left(\frac{A_G \cdot D - a \cdot c_0 \cdot \tau + c_0}{\tau \cdot \omega} \right) \cdot \sin(\omega \cdot t) \right) + c_0 \cdot e^{-\frac{t}{\tau}}$$

hvor vi antager, at personen optager 80% af de 50 g glukose dvs. $A_G \cdot D = 0,8 \cdot 50 = 40$.

b) Plot grafen for G , når tiden er fra 0 timer til 10 timer.

c) Bestem personens blodsukker niveau efter 3 timer.

d) Bestem de tidsintervaller, hvor personens blodsukker ligger indenfor normalområdet jf.

<https://diabetes.dk/diabetes-1/fakta-om-type-1/blodsukker-og-maling>, hvis vi antager, at personen ikke er diabetiker.

e) Bestem størrelsen $T_0 = \frac{2 \cdot \pi}{\omega}$. Størrelsen indikerer, hvor lang tid, der går førend personen er sulten igen.

Øvelse 12.19

En anden person måler sit blodsukker efter 12 timers faste og indtag af 50 g glukose. Tabellen viser en række sammenhørende data for personens målinger.

Tid (målt i timer)	0	$\frac{1}{3}$	$\frac{2}{3}$	1	$\frac{4}{3}$	$\frac{5}{3}$	2	$\frac{7}{3}$
Blodsukker (målt i mmol/l)	6,1	5,9	10,1	10,3	5,9	6,1	5,8	6,1

a) Bestem ved regression parametrene i modellen

$$G(t) = G_b + e^{-at} \cdot \left(-c_0 \cdot \cos(\omega \cdot t) + \left(\frac{A_G \cdot D - a \cdot c_0 \cdot \tau + c_0}{\tau \cdot \omega} \right) \cdot \sin(\omega \cdot t) \right) + c_0 \cdot e^{-\frac{t}{\tau}}$$

hvor vi antager, at personen optager 80% af de 50 g glukose dvs. $A_G \cdot D = 0,8 \cdot 50 = 40$.

- b) Plot grafen for G , når tiden er fra 0 timer til 10 timer.
- c) Bestem personens blodsukker niveau efter 3 timer.
- d) Bestem de tidsintervaller, hvor personens blodsukker ligger indenfor normalområde, jf. <https://diabetes.dk/diabetes-2/forebyggelse>, hvis vi antager, at personen prediabetiker.
- e) Bestem størrelsen $T_0 = \frac{2 \cdot \pi}{\omega}$. Størrelsen indikerer, hvor lang tid, der går førend personen er sulten igen.

Øvelse 12.20

Bestem størrelsen G_b i modellen

$$G(t) = G_b + e^{-at} \cdot \left(-c_0 \cdot \cos(\omega \cdot t) + \left(\frac{A_G \cdot D - a \cdot c_0 \cdot \tau + c_0}{\tau \cdot \omega} \right) \cdot \sin(\omega \cdot t) \right) + c_0 \cdot e^{-\frac{t}{\tau}}$$

for en person med diabetes eller prediabetes ud fra

<https://diabetes.dk/diabetes-2/fakta-om-type-2/blodsukker-og-maling>

og <https://diabetes.dk/diabetes-2/forebyggelse>

Litteratur

Lærebøger, fysisk og digitalt

- Gravetter & Wallnau, *Statistics for the Behavioral Sciences*, Wadsworth, 2000
- Susanne Ditlevsen og Helle Sørensen, *Introduktion til Statistik*, Institut for matematiske fag, 2015
- Jørgen Larsen: *Et antal statistiske standardmodeller*, IMFUFA 120, 1986
- Jørgen Larsen, *Sandsynlighedsregning og statistik for matematikstuderende*, IMFUFA 452, 2006
- Thomas Bendsen, *Noter i statistik for bioanalytikere*, 2009-2022, [Noter i statistik \(statnoter.dk\)](http://statnoter.dk)
- Karian & Tannis, *Probability and Statistics, Explorations with Maple*, 2008
- Per Brockhoff et al, *Introduction to Statistics at DTU*, 2018 (*Bygger på brug af R*)
- *Hvad er matematik?* 1, 2, 3 (af Bjørn Grøn, Bodil Bruun og Olav Lyndrup) indeholder en grundig indføring i alle sider af den deskriptiv og bekræftende statistik, vi møder i disse undervisningsmaterialer, bortset fra emnerne ANOVO og ANCOVA.

Specielle emner indenfor statistik

- Per Vejrup-Hansen, *Statistik med Excel*, forlaget Samfundslitteratur, 2002
- Ronald Fisher, *A Lady Tasting Tea*, indgår som kapitel 2 i *The Design of Experiment*, Oliver and Boyd, 1935. Kapitlet kan hentes via projekt 9.12 i HEM3: [kap9 Projekt 9 12 A lady Tasting Tea.pdf \(praxis.dk\)](http://praxis.dk/kap9_Projekt_9_12_A_lady_Tasting_Tea.pdf)
- Susanne Ditlevsen: *Statistiske metoder i hverdagsliv og neurovidenskab*, film af Bjørn Grøn og David Binzer, [10 danske matematikere - LRU.dk \(praxis.dk\)](http://praxis.dk/10_danske_matematikere_-_LRU.dk)
- Bjørn Grøn, *Projektmateriale 1* i tilknytning til Susanne Ditlevsens film: *Statistiske metoder i hverdagsliv og i neurovidenskab*, hentes samme sted.
- Inge Henningsen, *Matematiske og statistiske modeller i undervisningen*, Dansk Center for Naturvidenskabsdidaktik, 2001
- Hesterberg et al, *Bootstrap Methods and Permutation tests*, Freeman and co. 2003
- Hesterberg, *What Teachers Should Know about the Bootstrap*, The American Statistician, 2015
- Jørgen Hoffmann-Jørgensen, *Sandsynligheder*, Mathilde nr 35, 2008
- Jónsdóttir, Klim, Mortensen og Madsen, *Matematik i medicinudvikling*, i *Matematiske Horisonter*, DTU, 2009, [Matematik i medicinudviklingen - DTU Compute](http://DTU Compute)
- Trine Saugstrup, *Statistik og lægemidler*, Aktuel Naturvidenskab 1, 2010

Specielle kapitler i øvrige bøger:

- Brøsen, Simonsen, Dalhoff, *Basal og klinisk farmakologi*, FADL's Forlag, 2019
- Andrew Leigh, *Randomistas*, Yale University Press, 2018
- Pearl & Mackenzie, *The Book of Why*, Basic Books, 2018
- Jordan Ellenberg, *How not to be Wrong, The Power of Mathematical Thinking*, The Penguin Press, 1971 / 2014

Opslagsværk

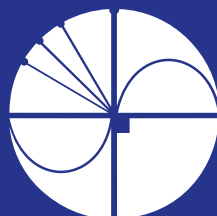
- [The Little Handbook of Statistical Practice \(jerrydallal.com\)](http://jerrydallal.com)

Projektet **Træk virksomhederne ind i undervisningen** fortæller, hvordan matematik er uundværlig for avancerede virksomheder i et moderne samfund. I filmene besøger Casper og Nicoline 12 forskellige virksomheder og lærer noget om den matematik, de anvender.

Casper og Nicoline møder ansatte, der er med helt i front af det matematiske arbejde: statistikere, ingeniører, matematik-økonomer eller epidemiologer m.fl. Alle har uddannelser med et betydelig matematisk indhold.

Til hver film kan du downloade undervisningsmaterialer. Her finder du øvelser, opgaveforløb, projekter og oplæg til studieretningsprojekter. Materialerne er opdelt i tre niveauer af sværhedsgrad fra 9 kl til 3 g. Emnerne er inspireret af filmen om virksomheden og den anvendte matematik, men man kan arbejde med materialet uden at have filmen kørende.

Hæftet om Matematikken bag ny medicin hos NOVO knytter sig til filmen af samme navn.



**matematik
i arbejde**