

En Introduktion til Statistik

Bind 1A

Knut Conradsen

7. udgave

Lyngby 1999

IMM

Forord

Med denne (foreløbige) udgave af 'En Introduktion til Statistik' er der påbegyndt en mere gennemgribende revision af lærebogssystemet i statistik for civilingeniørstuderende ved DTU.

Udgaven er en bearbejdning af de tidligere udgaver, men nu skrevet i $\text{\TeX}$. Dette vil muliggøre en hurtigere opdatering af fremstillingen i kommende udgaver.

Der er udfoldet store bestræbelser for at sikre, at der ikke optræder for mange fejl; men det vil være naivt at tro, at der kan skrives mere end 500 siders formelfyldt tekst uden at der slipper en del fejl igennem korrekturlæsningen. Jeg vil derfor være taknemmelig for at blive gjort opmærksom (skriftligt) på tilbageværende fejl.

En række kolleger og studerende har bidraget meget under tilblivelsen og udviklingen, især nuværende og tidligere undervisere og studerende i faget Statistik 1. Omlægningen til $\text{\TeX}$ skyldes de (dengang) studerende Claus Ørum-Hansen, Ole Bøje Haagensen og Jan Nygaard Nielsen.

Endelig vil det være på sin plads her at fremhæve Poul Thyregod, som har ydet en stor indsats under udarbejdelsen af de tidligere udgaver, og Henrik Spliid, som ved omlægningen til $\text{\TeX}$ har været en uvurderlig støtte, ligesom en række forbedrede fremstillinger skyldes ham.

Knut Conradsen

Indhold

0	Forudsætninger og notation	9
0.1	Introduktion	9
0.2	Permutationer og kombinationer	26
0.3	Klassiske sandsynligheder	29
0.4	Sandsynlighedsfelter og stokastiske variable	31
0.4.1	Om sandsynligheder, hændelser og stokastiske variable	31
0.5	Betingede sandsynligheder	34
0.6	Fordelings- og frekvensfunktioner	40
0.7	Flerdimensionale stokastiske variable	43
0.8	Transformation af stokastiske variable	46
0.9	Momenter	53
0.10	Approksimative formler for middelværdi og varians	67
0.11	Konvergens	70
0.12	Notation	73
0.12.1	Γ -funktionen	74
0.13	Fortsættelse af tidligere eksempler	76
1	Sandsynlighedsteoretiske modeller	85
1.1	Lidt om stokastiske modellers verifikation	85
1.2	Modeller i forbindelse med Bernoulli forsøg	91
1.2.1	Bernoulli forsøg	91
1.2.2	Binomialfordelingen	92
1.2.3	Den negative binomialfordeling (Pascal's fordeling)	99
1.3	Nogle modeller om stikprøveudtagning	102
1.3.1	Den hypergeometriske fordeling	102
1.3.2	Polynomialfordelingen	104
1.4	Poisson modeller. Erlang- og Γ -fordelingen	108
1.4.1	Poisson fordelingen	108
1.4.2	Erlang- og Γ -fordelingen	118
1.5	Den normale fordeling	123
1.5.1	Analytiske egenskaber	123
1.5.2	Den centrale grænseværdisætning	125

1.5.3	Andre hypoteser, der fører til den normale fordeling	129
1.5.4	Den normale fordeling som tilnærmelse til andre fordelinger	130
1.6	Den logaritmiske normale fordeling	134
1.6.1	Analytiske egenskaber	134
1.6.2	Loven om proportional effekt	137
1.7	Ekstremværdiproblemer	142
1.7.1	Største og mindste observations fordeling	142
1.7.2	Asymptotiske ekstremværdifordelinger	147
1.7.3	Maximumsfordeling for eksponentiel type	148
1.7.4	Minimumsfordeling for eksponentiel type	156
1.7.5	Fordelinger af Cauchy type	161
1.7.6	Fordelinger af tredje type	169
1.7.7	Oversigter over asymptotiske ekstremværdifordelinger	177
1.8	Andre sandsynlighedsteoretiske modeller	180
1.8.1	Den rektangulære fordeling	180
1.8.2	Beta-fordelingen	181
1.8.3	Cauchy fordelingen	183
1.8.4	LaPlace fordelingen	184
1.8.5	Den logistiske fordeling	185
1.8.6	Pareto fordelingen	186
1.8.7	Ligefordelingen på $\{0, 1, \dots, n\}$	187
1.8.8	Den logaritmiske fordeling	187
1.9	Compound fordelinger	189
1.10	Fordelinger afledt af den normale fordeling	194
1.10.1	χ^2 -fordelingen	194
1.10.2	Rayleigh fordelingen	198
1.10.3	Student's t-fordeling	199
1.10.4	F-fordelingen	201
2	Estimationsteori	211
2.1	Generelt om estimationsteori	211
2.1.1	Statistisk inferens	211
2.1.2	Estimationsproblematikken	213
2.2	Estimatorers egenskaber	215
2.2.1	Centrale estimatorer	215
2.2.2	Konsistente estimatorer	218
2.2.3	Sufficiens	220
2.2.4	Efficiens	228
2.3	Estimationsmetoder	233
2.3.1	Maximum likelihood metoden	233
2.3.2	Mindste kvadraters metode	249
2.3.3	Momentmetoden	254
2.3.4	Intervalestimation (konfidensintervaller)	258
Bind 1B		296

3	Hypoteseprøvning	301
3.1	Generel problemstilling og metode	301
3.1.1	Indledning og definitioner	301
3.1.2	Testprincipper	314
3.2	Specielle tests	326
3.2.1	Tests i en binomialfordeling	326
3.2.2	Sammenligning af to binomialfordelinger	328
3.2.3	Tests i en Poissonfordeling	332
3.2.4	Sammenligning af to Poissonfordelinger	332
3.2.5	Tests i normalfordelingen	335
3.2.6	Test i Γ -fordelingen	354
3.2.7	Test i polynomialfordelingen	360
3.2.8	Test i kontingenstabel	363
3.2.9	Homogenitetstestet	367
3.3	Fordelingsfrie tests	370
3.3.1	Fortegnstestet og Wilcoxon-testet	370
3.3.2	Invers normalvægttest (van der Waerden-test)	379
3.3.3	Rangtest for skalaparametre (Siegel-Tukey)	381
4	Modelkontrol	385
4.1	Test for tilfældighed	385
4.1.1	Run test	386
4.1.2	Gennemsnittet af kvadrerede successive differenser	390
4.2	Kontrol af fordelingslov	394
4.2.1	Grafiske metoder	394
4.2.2	Tests for fordelingsstype	404
4.2.3	Beregning af empiriske momenter	410
5	Varians- og regressionsanalyser	413
5.1	Variansanalyser	413
5.1.1	Ensidet variansanalyse	413
5.1.2	Tosidet variansanalyse	423
5.1.3	Romersk kvadrat	441
5.1.4	Faktorforsøg	445
5.2	Regressionsanalyser	451
5.2.1	Regressionsanalyse med 1 uafhængig variabel	451
5.2.2	Sammenligning af 2 regressionslinier	466
5.2.3	Regressionsanalyse med 2 uafhængige variable	474
5.3	Tests for varianshomogenitet	482
5.3.1	Bartlett's test	482
5.3.2	Andre tests for varianshomogenitet	484
5.4	Fordelingsfrie tests	487
5.4.1	Måleskalaer	487
5.4.2	Invarians og rangtests	490
5.4.3	Kruskal-Wallis' test	491
5.4.4	Friedmans test	496

5.4.5	Rangkorrelationskoefficienter	500
5.4.6	Tabeller	507
6	Beslutningsteori	515
6.1	Generelt om beslutningsteori	515
6.1.1	Definitioner og metoder	516
6.1.2	Eksempel på analyse af et beslutningsproblem	522
6.2	Beslutningsteoriens anvendelse i statistikken	528
6.2.1	Beslutningsteoriens anvendelse i estimationsteorien	528
6.2.2	Beslutningsteoriens anvendelse i testteorien	535

Kapitel 0

Forudsætninger og notation

0.1 Introduktion

EKSEMPEL 0.1. I nedenstående tabel 0.1 er anført slagtevægten af 100 slagtegrise på et dansk andelssvineslagteri. Vi ser, at vægtene er ganske ens. De svinger mellem 53 og 71 kg, men langt de fleste ligger omkring 60-61 kg plus/minus et par kg. For at kunne udtrykke denne variabilitet lidt mere præcist, vil vi tælle op, hvor mange grise der har vægtene 53, 54, ..., 71, 71 kg. Hvis man laver optællingen manuelt kan man benytte 'havelåge'-princippet. Man skriver de mulige vægte op, og gennemgår derefter data. For hver vægtmåling slår man en streg (|), og når man har nået 5, sættes en skråstreg over de 4 foregående (||||). På denne måde fås let den søgte optælling. Resultatet er vist i nedenstående tabel. Her er også anført det **kumulerede** antal målinger, dvs. det antal målinger, der er mindre end lig den pågældende måling. Kaldes antallet af målinger med vægten $(52+i)$, $i = 1, \dots, 19$, for r_i , bliver det kumulerede antal altså

$$R_i = r_1 + \dots + r_i$$

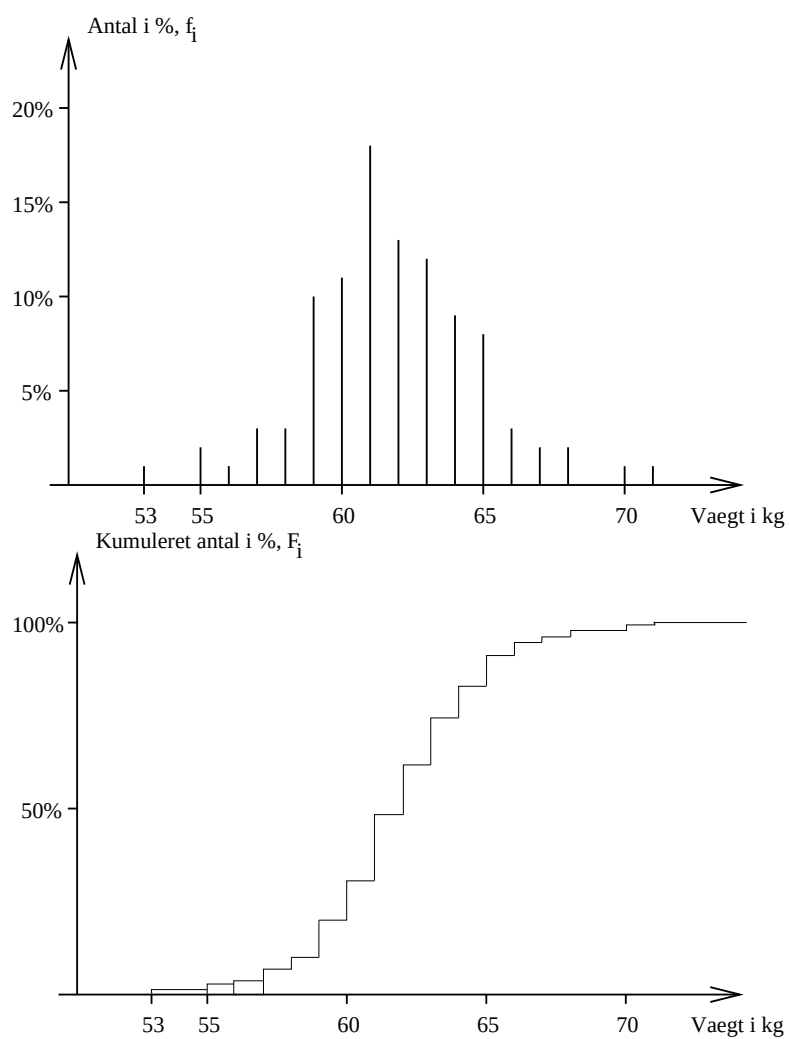
Værdierne af r_i og R_i er omsat til %-tal, nemlig den **relative hyppighed** eller **frekvens** $f_i = (r_i/100) \times 100\%$ og den summerede eller **kumulerede relative hyppighed (frekvens)** $F_i = (R_i/100) \times 100\%$. Værdierne af f_i og F_i er afbildet i Figur 0.1, henholdsvis som en 'pind' og som en trappekurve. Trappekurven er konstant mellem to måleværdier $(52+i-1)$ og $(52+i)$ og springet i $52+i$ er altså lig f_i . Funktionerne i Figur 0.1 kaldes også den **empiriske frekvensfunktion** og den **empiriske fordelingsfunktion**. ♦

63	62	57	60	65
65	61	63	61	59
62	61	64	60	60
60	59	64	62	63
65	62	61	64	61
65	61	63	63	63
61	68	61	61	62
59	67	59	59	60
66	61	60	59	64
61	58	59	61	63
63	63	59	56	67
55	61	62	64	62
71	65	60	63	60
64	61	70	64	66
65	62	62	60	60
63	64	61	62	55
62	65	59	66	62
58	61	57	59	57
62	64	61	68	60
63	58	53	65	61

Tabel 0.1: Slagtevægte i kg for 100 slagtegrise på et dansk andelssvineslagteri.

Vægt	'Havelåge'	Antal	Kumuleret antal
53		1	1
54		0	1
55		2	3
56		1	4
57		3	7
58		3	10
59		10	20
60		11	31
61		18	49
62		13	62
63		12	74
64		9	83
65		8	91
66		3	94
67		2	96
68		2	98
69		0	98
70		1	99
71		1	100

Tabel 0.2: Optælling af data om slagtevægte af grise.



Figur 0.1: Fordelingen af 100 målinger af slagtevægte og det kumulerede antal målinger i %.

I det foregående eksempel var der ikke flere forskellige måleværdier end at det var gørligt (og informativt) at tælle op, hvor mange målinger, der var af hver måleværdi. I det næste eksempel er målingerne givet med flere decimaler, og derfor vil vi **gruppere** dem.

EKSEMPEL 0.2. I forbindelse med miljømyndighedernes kontrol af forurenende aktiviteter opereres med 2 typer af kontroller, nemlig

1. mængdekontrol
2. tilstandskontrol

Ved mængdekontrol kontrolleres den totale mængde af forurenende stof, der udledes til det ydre miljø. (Man taler ofte om udledning til en **recipient**.) Ved tilstandskontrol forstås en kontrolform, hvor værdien af en given variabel ikke må overskride et givet niveau. Vi skal dvæle lidt ved, hvad dette sidste indebærer. I tabel 0.3 er anført

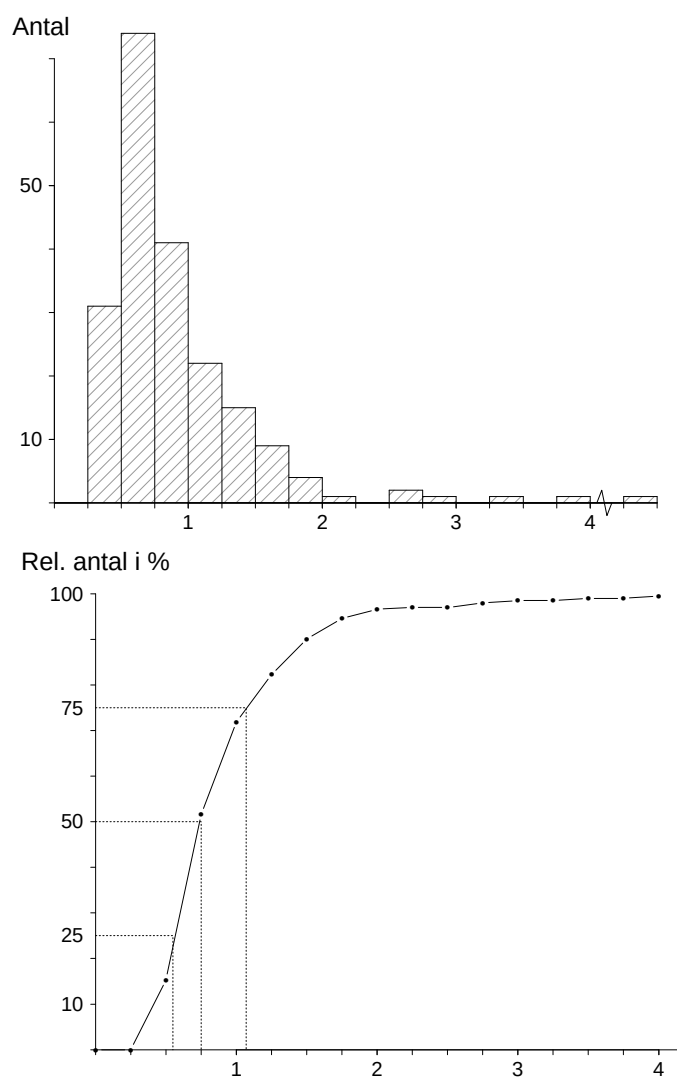
Klasse	Antal målinger	Kumuleret antal	Kumuleret antal i %
0.25-0.50	31	31	15.3
0.50-0.75	74	105	51.7
0.75-1.00	41	146	71.9
1.00-1.25	22	168	82.3
1.25-1.50	15	183	90.1
1.50-1.75	9	192	94.6
1.75-2.00	4	196	96.6
2.00-2.25	1	197	97.0
2.50-2.75	2	199	98.0
2.75-3.00	1	200	98.5
3.25-3.50	1	201	99.0
3.75-4.00	1	202	99.5
9.50-9.75	1	203	100

Tabel 0.3: Fordelingen af 203 målinger af en kontrolvariabel fordelt på 13 klasser. Højre endepunkt er regnet med til klassen. Miljømyndighedernes kravværdi er 1.

resultater af 203 målinger af en kontrolvariabel målt gennem et år på en dansk industrivirksomhed. Kontrolvariablen udtrykker et indhold (koncentration) af såkaldt bioaktivt organisk materiale. For at fremme overskueligheden er de observerede værdier grupperet i intervaller af længde 0.25, og det antal målinger, der ligger i hvert interval er angivet. Kaldes antallet af målinger i det i 'te interval a_i er størrelsen

$$h_i = \sum_{j \leq i} a_j = a_1 + \dots + a_i$$

ligeledes angivet. Størrelsen h_i kaldes analogt med det tidligere for det **kumulerede** antal og den angiver altså **det antal observationer**, der er **mindre end eller lig med højre klasseendepunkt i den i 'te klasse** (interval). Endelig fremkommer det kumulerede antal i % som $(h_i/203) \times 100\%$. Disse værdier er afbildet i figur 0.2. Øverst er a_i afbildet som funktion af 'intervallerne' og nederst er tegnet en funktion, der fremkommer ved lineær interpolation mellem værdierne $(h_i/203) \times 100\%$. Den øverste figur er et eksempel på det, vi kalder et **histogram**, og den nederste er en såkaldt **sumpolygon**.



Figur 0.2: Antal observationer pr klasse og det relative antal mindre end højre klasseendepunkt.

Det fremgår, at målingerne af kontrolvariablen ligger mellem 0.25 og 9.75, men ikke alle målinger optræder lige hyppigt! Vi ser, at 36.5% af målingerne ligger i intervallet $[0.50 - 0.75]$, men kun 0.5% i intervallet $[9.50 - 9.75]$. Af sumpolygonen fremgår, at 25% er mindre end 0.6 (ca.), 50% af målingerne er mindre end 0.75, og 25% er større end 1.1 (ca.). Disse 3 værdier

0.6
0.75
1.1

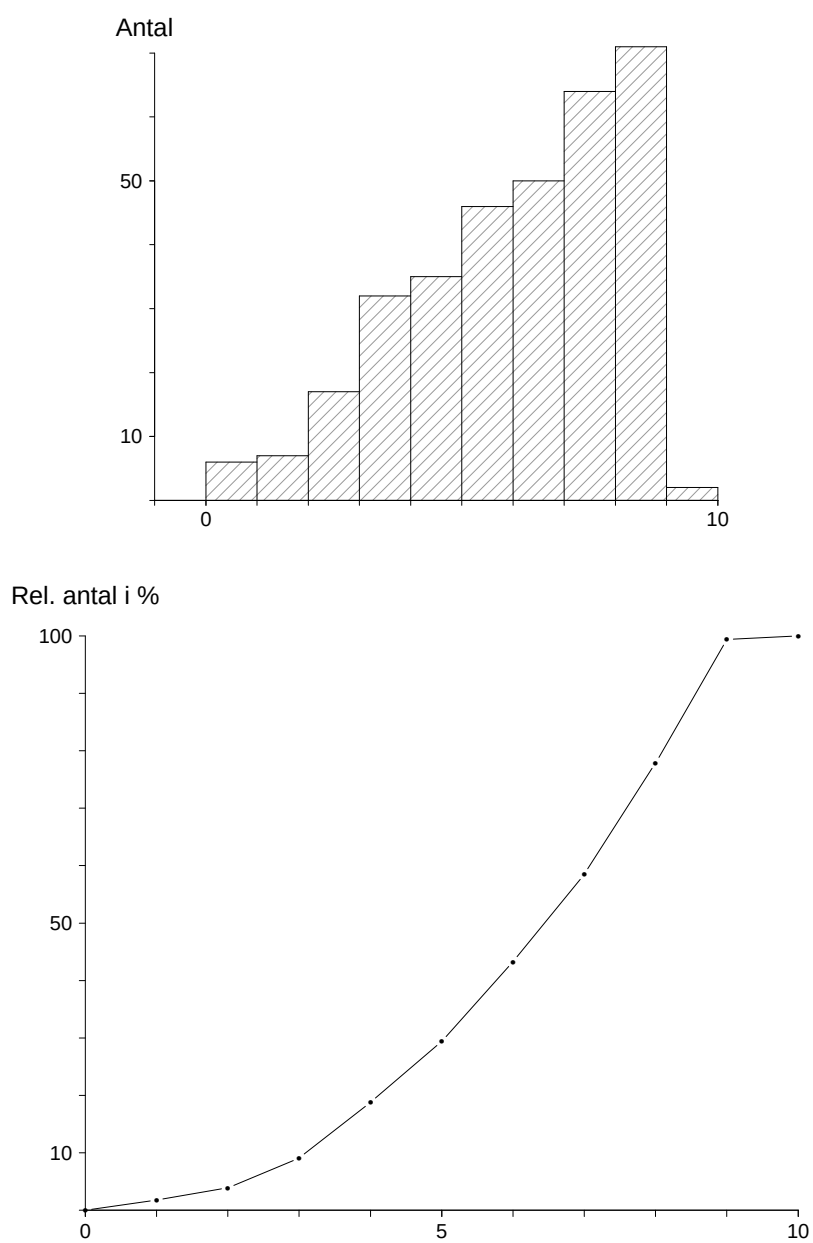
kaldes **nederste kvartil**, **medianen** og **øverste kvartil** for de observerede værdier af kontrolvariablen, og de bruges til at give en grov beskrivelse af, hvorledes målingerne fordeler sig på den reelle akse.

Miljømyndighedernes kravværdi er 1, dvs. at kontrolvariablen bør være mindre end 1, uden at det nærmere er specificeret, hvad der skal forstås herved. Det ses, at ca. 72% af målingerne er mindre end kravværdien, men omvendt altså er 28% større end denne. Er det acceptabelt? Den største måling er næsten 10 gange kravværdien! Når disse spørgsmål skal afgøres, må man tage højde for, at målingerne varierer meget bl.a. på grund af måleusikkerhed, på grund af tilfældige fluktuationer under produktionen, fx. grundet råvareinhomogeniteter, tilfældige, mindre driftsstop etc. Man kan ikke sætte kravværdien vilkårligt lav. Dette ville alene grundet de tilfældige fluktuationer kunne umuliggøre en fortsat produktion, dvs. **udlederrisikoen** bliver for stor. Omvendt vil en meget høj værdi medføre en for stor **recipientrisiko**. En rimelig afvejning af disse kræver, at der formuleres en passende **model** for målingernes variation, således at konsekvenserne af forskellige strategier kan beregnes. Herved kan der opnås en konsistent behandling af miljøsager. ♦

EKSEMPEL 0.3. Målingerne i dette eksempel svarer fuldstændigt til de eksempel 0.2 anførte. For den her anførte kontrolvariabel er miljømyndighedernes kravværdi 9. Vi ser, at kun 0.6% af målingerne overskrider denne værdi. Derimod ligger 21.5% af målingerne i klassen umiddelbart op til kravværdien! Vi ser her, at den tilfældige variation har en helt anden natur end det i eksempel 0.2 anførte. Det er bemærkelsesværdigt, at så få målinger overskrider kravværdien, når vi har så mange målinger lige op til denne. Det kunne her være af interesse at søge at beskrive fordelingen alene baseret på observationer, der e.g. er mindre end 5 kombineret med en passende model. Man kunne så sammenligne forudsigelser fra denne model med de faktisk konstaterede målinger. På denne måde kunne man måske få afdækket nogle relevante forhold omkring indsamlingen af data. ♦

Klasse	Antal målinger	Kumuleret antal	Kumuleret antal i %
0-1	6	6	1.8
1-2	7	13	3.9
2-3	17	30	9.1
3-4	32	62	18.8
4-5	35	97	29.4
5-6	46	143	43.3
6-7	50	193	58.5
7-8	64	257	77.9
8-9	71	328	99.4
9-10	2	330	100

Tabel 0.4: Fordelingen af 330 målinger af en kontrolvariabel fordelt på 10 klasser. Højreendepunkt er regnet med til klassen. Miljømyndighedernes kravværdi er 9.

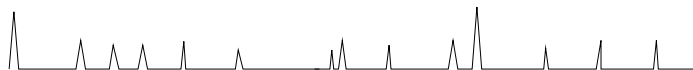


Figur 0.3: Antal observationer pr klasse og det relative antal mindre end højre klasseendepunkt.

Antal α -partikler i 2 sek.interval	Antal inter- valler
0	18
1	65
2	81
3	95
4	62
5	33
6	14
7	6
8	1

Tabel 0.5: Hyppigheden af antal α -partikler registreret i intervaller af 2 sekunders længde.

EKSEMPEL 0.4. I tabel 0.5 er anført resultatet af målinger af α -partikler ved hjælp af et Geiger-Müller rør. Dette er tilsluttet en x - t skriver, der bevæger sig med en hastighed på 1 cm/sekund. Resultatet af 15 sekunders forløb ses nedenfor.

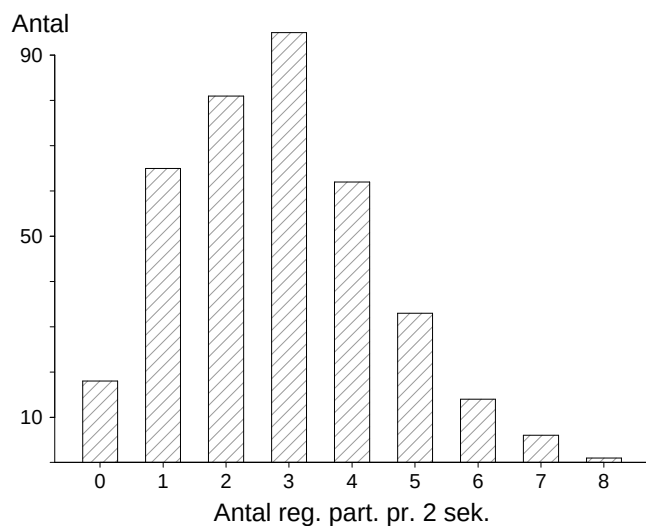


Strimlen er delt op i intervaller svarende til 2 sekunder, og antallet af spidser (svarende til en α -partikel) pr 2 sekunders interval er talt op. Der observeres i alt 375 intervaller. 18 af disse indeholdt ingen spidser, 65 en enkelt spids etc. I tabel 0.6 er angivet ventetiderne mellem α -partiklerne. Vi ser, at der er flest i det første interval fra 0-0.5. Derefter synes antallene at aftage mere eller mindre eksponentielt. Udsendelsen af

Ventetid	Hyppighed
0-0.5	539
0.5-1.0	289
1.0-1.5	115
1.5-2.0	71
2.0-2.5	36
2.5-3.0	12
3.0-3.5	2
3.5-4.0	2
4.0-4.5	2

Tabel 0.6: Ventetid mellem 2 på hinanden følgende α -partikler.

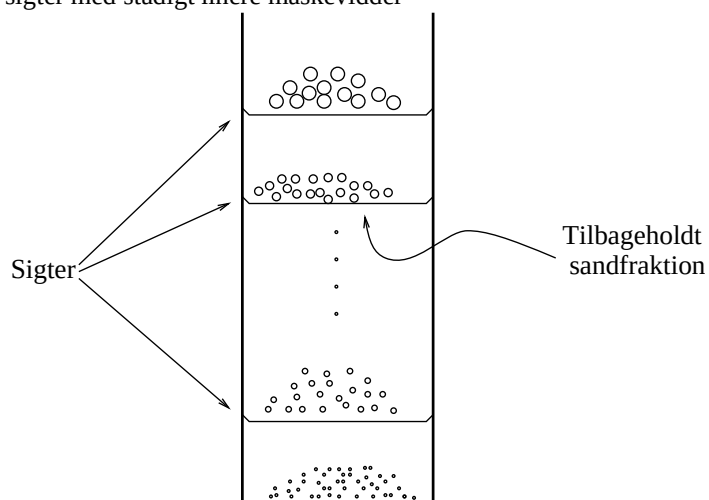
radioaktive partikler anses almindeligvis for at være en helt '**tilfældig**' proces. Vi har her karakteriseret denne ved dels at angive en såkaldt **diskret fordelt variabel**, nemlig det registrerede antal α -partikler/2 sek, og dels en såkaldt **kontinuert fordelt variabel**, nemlig ventetiden mellem α -partikler. Den diskret fordelte variabel kan antage værdierne 0, 1, 2, $\dots$, dvs. heltallige værdier, og den kontinuert fordelte variabel kan (i princippet) antage alle værdier større end 0. Vi bemærker, at den 'totale tilfældighed'



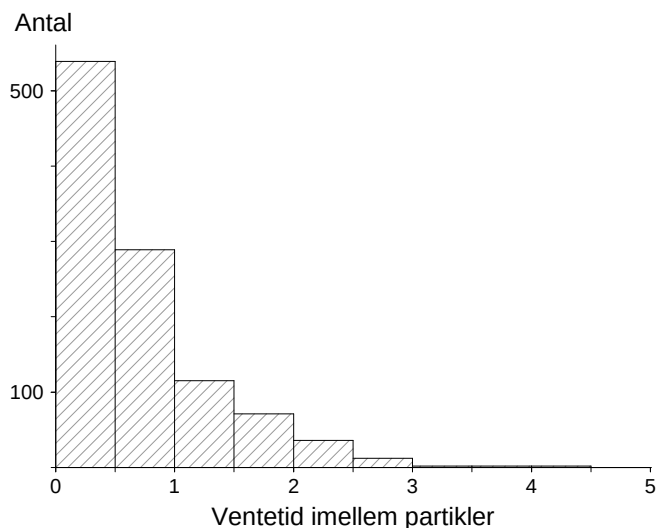
Figur 0.4: Fordeling af antal α -partikler/2 sekund.

giver sig udslag i en smuk regelmæssighed i fordelingen af de to fænomener. Efter en passende matematisk formulering kan dette omvendt udnyttes til at afgøre, om givne hændelser optræder helt 'tilfældigt'. ♦

EKSEMPEL 0.5. Størrelsesfordelinger af f.eks. støbesand bestemmes ved hjælp af et system af sigter med stadigt finere maskevidder



Der er efter sådan en sigtning intet principielt til hinder for at tælle antallet af sand-



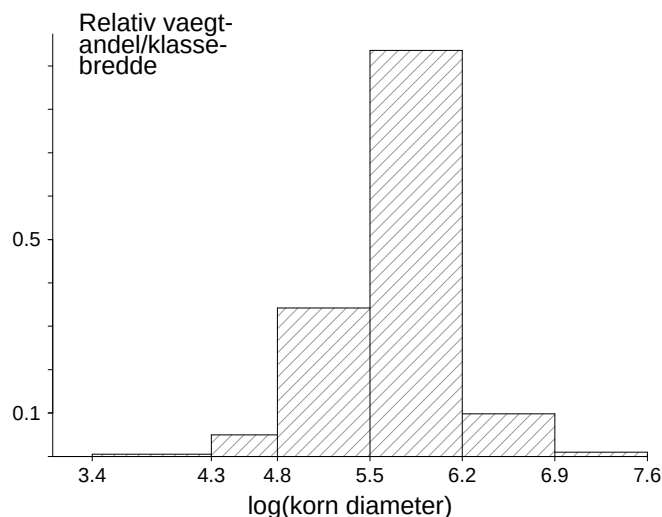
Figur 0.5: Fordeling af tidsafstanden mellem α -partikler.

skorn, der er tilbage i hver enkelt sigte. Selv for små prøver opstår dog et praktisk problem, nemlig det faktum, at der vil være et (for en manuel tælling) prohibitivt stort antal korn – i det mindste i de finere fraktioner.

Klassegrænser i μm	$\log_e$ (klas- segrænser)	Antal gram
31.25-75	3.4-4.3	3
75-125	4.3-4.8	18
125-250	4.8-5.5	168
250-500	5.5-6.2	460
500-1000	6.2-6.9	48
1000-2000	6.9-7.6	5

Tabel 0.7: Vægtfordeling for støbesand.

I stedet kan man veje mængden af sand i de forskellige fraktioner. Resultatet af et sådant forsøg er angivet i tabel 0.7. Da maskevidderne stort set er eksponentielt voksende er det nærliggende at gøre klasserne mere 'ensartede' ved at logaritmere værdierne. De herved frembragte data er afbildet i figur 0.6. Da klasserne (efter logaritmering) ikke er helt lige lange, er søjlernes højde ikke afbildet proportionalt med den fundne vægtfraktion. Højderne er divideret med de respektive klassebredder (efter logaritmeringen), således at **arealerne** af søjlerne er proportionale med klassens vægtandel. Det ses, at vi også her får et ganske 'regelmæssigt' billede af kornstørrelsefordelingen. Spørgsmålet er nu om, og i givet fald hvordan, man kan slutte sig til oplysninger om den fordeling, man ville have fået, hvis man havde 'talt' i stedet for 'vejet'. Disse spørgsmål kan

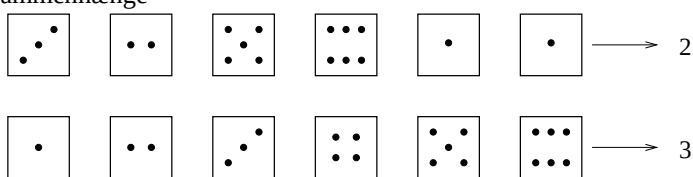


Figur 0.6: Relativ vægtandel/klassebredde, dvs. arealet af hver søjle er proportional med vægtandelen for klassen.

besvares bekræftende, og vi skal senere se, hvorledes vi også kan udlede resultater om overfladefordeling etc. ♦

Fælles for de foregående eksempler er, at den tilfældige variation, vi observerer, er resultatet af en række komplicerede mekanismer, som det ikke uden videre er muligt at beskrive. For at komme lidt tættere på nogle relevante modelklasser betragter vi nogle langt enklere forsøg i

EKSEMPEL 0.6. I tabel 0.8 er vist de enkelte udfald af forsøg, hvor der i hvert forsøg er lavet et kast med 6 terninger. Som resultat (udfald) af forsøget er så angivet antal terninger, der viser et lige antal øjne i det enkelte kast (med de 6 terninger). Eg. haves følgende sammenhænge



I det første tilfælde viser anden og fjerde terning et lige antal, dvs. udfaldet er 2. I det andet tilfælde er det anden, fjerde og sjette terning, dvs. udfaldet bliver 3. Det ses, at de mulige udfald er 0, 1, 2, 3, 4, 5 og 6.

I tabel 0.9 er anført udfaldet af 50 forsøg med kast med 1 terning, hvor man i det enkelte forsøg har angivet antal øjne. De mulige udfald er her 1, 2, 3, 4, 5 og 6.

Det fremgår, at vi i det første tilfælde har en præference for værdier omkring fordelings midte 3. I det andet tilfælde synes resultaterne at være mere ligeligt fordelte.

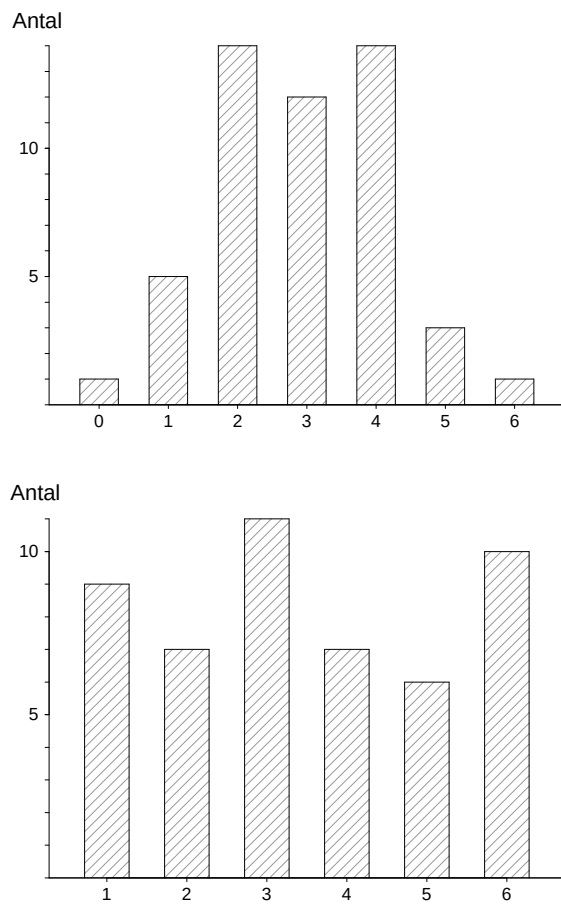
Disse resultater er i god overensstemmelse med den intuitive fornemmelse, de fleste har af sandsynlighedsbegrebet. ♦

Kast nr.	0	1	2	3	4	5	6	Kast nr.	0	1	2	3	4	5	6
1			x					26			x				
2					x			27			x				
3						x		28						x	
4				x				29					x		
5					x			30			x				
6					x			31				x			
7			x					32				x			
8				x				33					x		
9			x					34					x		
10						x		35					x		
11					x			36	x						
12			x					37				x			
13				x				38							x
14					x			39				x			
15		x						40			x				
16			x					41				x			
17		x						42					x		
18	x							43					x		
19			x					44				x			
20			x					45					x		
21		x						46	x						
22				x				47		x					
23			x					48				x			
24				x				49					x		
25			x					50				x			
I alt for de 50 forsøg									1	5	14	12	14	3	1

Tabel 0.8: Antal terninger med et lige antal øjne i 50 forsøg med kast med 6 terninger.

Kast nr.	1	2	3	4	5	6	Kast nr.	1	2	3	4	5	6
1						x	26				x		
2					x		27		x				
3						x	28	x					
4			x				29		x				
5		x					30						x
6				x			31						x
7						x	32					x	
8					x		33	x					
9	x						34						x
10						x	35	x					
11					x		36			x			
12				x			37	x					
13						x	38	x					
14		x					39			x			
15			x				40			x			
16				x			41				x		
17					x		42					x	
18						x	43				x		
19			x				44						x
20	x						45			x			
21	x						46	x					
22			x				47		x				
23			x				48		x				
24		x					49			x			
25			x				50				x		
I alt for de 50 forsøg								9	7	11	7	6	10

Tabel 0.9: Antal øjne i 50 forsøg med kast med 1 terning.



Figur 0.7: Øverst fordelingen af antal terninger med lige øjne på 6 terninger. Nederst fordelingen af antal øjne på 1 terning.

0.2 Permutationer og kombinationer

Vi indfører nogle funktioner, der er nyttige i forbindelse med optællinger mv.

1. Fakultetsfunktionen

$$n! = n(n-1)(n-2) \cdots 3 \cdot 2 \cdot 1, \quad n \in \mathbb{N} \quad (0.1)$$

2. Det r 'te nedstigende faktoriel

$$n^{(r)} = n(n-1) \cdots (n-r+1) = \frac{n!}{(n-r)!}, \quad n, r \in \mathbb{N} \quad (0.2)$$

3. Binomialkoefficienten

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n^{(r)}}{r!} \quad (0.3)$$

4. Polynomialkoefficienten

$$\binom{n}{r_1, \dots, r_k} = \frac{n!}{r_1! \cdots r_k!}, \quad r_1 + \dots + r_k = n \quad (0.4)$$

Bemærk, at binomialkoefficienten (0.3) er en polynomialkoefficient (0.4) svarende til $k = 2$, idet

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

Vi betragter nu en mængde $\mathcal{S}$ med n elementer. Et **ordnet udvalg** af elementer fra mængden kaldes en **permutation** og et ikke-ordnet kaldes en **kombination**. Betragter vi fx. heltallene fra 1 til 10 $\{1, 2, 3, \dots, 10\}$ vil talsættene

$$\begin{aligned} &\{1, 2, 5\} \\ &\{5, 1, 2\} \end{aligned}$$

blive opfattet som 2 forskellige permutationer, da rækkefølgen er forskellig, men som samme kombination.

Vi vil nu udtage r elementer fra mængden $\mathcal{S}$, dels **med tilbagelægning** og dels **uden tilbagelægning**. Ved udtagning med tilbagelægning udvælges et element, dets karakteristika noteres og det lægges tilbage. Derefter udtages igen et element. Det kan være

det, der lige var udtaget, og det kan være et andet. Ved udtagning uden tilbagelægning vil det samme element derimod ikke kunne udtages flere gange.

Vi ser først på udtagning uden tilbagelægning.

SÆTNING 0.1. Antallet af permutationer af r elementer taget uden tilbagelægning blandt de n elementer er

$$n^{(r)} \quad (0.5)$$

Hvis specielt $n = r$ fås

$$n! \quad (0.6)$$

Antallet af kombinationer af r elementer taget uden tilbagelægning blandt n er

$$\binom{n}{r} \quad (0.7)$$

Bevis. Ad permutationer:

Det første element kan vælges på n måder. For hver af disse kan det næste element vælges på $(n - 1)$ måder. I alt $n(n - 1)$ måder. For hver af disse kan det tredje element vælges på $n - 2$ måder, og der argumenteres som før etc.

Ad kombinationer:

Vi kalder antallet af kombinationer $K(n, r)$ og skal vise, at

$$r!K(n, r) = n^{(r)}$$

Her repræsenterer højresiden antallet af permutationer af r elementer blandt de n . Venstresiden repræsenterer antal måder vi kan udvælge r (uordnede) elementer ganget med antallet af måder sådanne r elementer kan ordnes på; men dette må jo netop være antal ordnede valg af r elementer. Derfor er

$$K(n, r) = \frac{n^{(r)}}{r!} = \binom{n}{r}$$

■

Dernæst vises

SÆTNING 0.2. Antallet af permutationer af r elementer med tilbagelægning fra en mængde med n elementer er

$$n^r \quad (0.8)$$

og antallet af kombinationer af r elementer taget med tilbagelægning fra n elementer er

$$\binom{n+r-1}{r} \quad (0.9)$$

Bevis. Ad permutationer:

Det første element kan vælges på n måder. Det næste ligeledes på n måder etc.

Ad kombinationer:

Vi symboliserer de (oprindelige) n elementer ved 1'er og tilføjer et $(n+1)$ 'te element, som også symboliseres ved 1. De r elementer symboliseres ved 0'er. Vi skriver disse op i en ordnet rækkefølge, hvor første og sidste nr. altid kan vælges at være et 1-tal. Vi får da en sekvens som

110010110111

Denne svarer så til en kombination, hvor element nr 2 er udtaget to gange (to 0'er efter 1-tal nr 2), element nr 3 en gang (et 0 efter 1-tal nr 3) og element nr 5 en gang (et 0 efter tal nr 5) blandt de i alt 7 elementer (der er otte 1-taller). Vi får en 'lovlig' kombination ved at placere de r 0'er på vilkårlige af de $n+r-1$ pladser, der er mellem de yderste 1-taller. Dette kan netop gøres på

$$\binom{n+r-1}{r}$$

måder. ■

Mange problemer i kombinatorikken kan føres tilbage til at se på fordelinger af kugler i kasser. Vi skal give en kort oversigt

1. r forskellige kugler i n forskellige kasser

Antal forskellige måder: n^r

2. r forskellige kugler i n forskellige kasser med højst 1 kugle i hver kasse.

Antal forskellige måder: $n^{(r)}$.

3. r ens kugler i n forskellige kasser.

Antal synligt forskellige måder: $\binom{n+r-1}{r}$

4. r ens kugler i n forskellige kasser med højst en kugle i hver kasse.

Antal synligt forskellige måder: $\binom{n}{r}$

5. r ens kugler i n forskellige kasser uden tomme kasser.

Antal synligt forskellige måder $\binom{r-1}{n-1}$

Bevis. Vi repræsenterer kassevæggene med 1-taller og kuglerne med 0'er. Vi anbringer først de r 0'er mellem 'yderkassernes ydervægge', dvs.

1000001

Ved at placere de $n - 1$ 1-taller imellem de to yderste får vi en fordeling af de r kugler i de n kasser. Hvis ingen kasser skal være tomme, skal 1-tallerne placeres i de $r - 1$ mellemrum mellem 0'erne med højst et 1-tal pr. mellemrum, og det kan netop gøres på $\binom{r-1}{n-1}$ måder. ■

0.3 Klassiske sandsynligheder

Sandsynlighedsregningen er oprindeligt udviklet for at løse spilteoretiske problemer. Det første solide fundament fik den omkring 1654 i en brevveksling mellem de franske matematikere Blaise Pascal og Pierre de Fermat. Den klassiske definition på en **sandsynlighed for et givet resultat af et eksperiment var forholdet mellem antallet af for resultatet gunstige udfald og det mulige antal udfald**. En sådan definition anvendtes af Laplace i hans fremstilling af sandsynlighedsregningen, og vi taler derfor også om **Laplaciske sandsynligheder**.

Lad os illustrere dette ved at se på et af de problemer, Fermat og Pascal beskæftigede sig med.

BEMÆRKNING 0.1 (CHEVALIER DE MÉRÉ'S PARADOKS). Chevalier de Méré var en spileglad person, der undrede sig over, at et spil, hvor man skulle slå mindst en sekser i 4 slag med en terning, gav andre odds end at slå mindst en dobbelt sekser i 24 slag med 2 terninger.

Lad os først se på, hvorfor de Méré mente, at spillene burde være ækvivalente. En dobbelt-sekser har sandsynligheden $\frac{1}{36}$ og der spilles 24 gange. Tilsvarende har en sekser sandsynligheden $\frac{1}{6}$, og der spilles 4 gange.

Da $\frac{24}{36} = \frac{4}{6}$ mente de Méré, at spillene burde være ækvivalente. Imidlertid er sandsynligheden for at få mindst 1 sekser i 4 forsøg lig

$$1 - \left(\frac{5}{6}\right)^4 = 0.518$$

og for at få en dobbelt-sekser i 24 forsøg lig

$$1 - \left(\frac{35}{36}\right)^{24} = 0.491$$

At de Méré har kunnet konstatere så små forskelle empirisk må betyde, at han har spillet dette spil tusindvis af gange. ▼

'Fysiske' sandsynligheder

Vi skal nu se på nogle modeller for partikelfordelinger, der har vundet indpas i fysikken. Vi betragter

- r partikler ($\simeq$ kugler)
- n celler i passende rum ($\simeq$ kasser)

Der er derfor

$$n^r$$

forskellige fordelinger af partiklerne i cellerne. Vi vil nu anføre nogle udtryk for, at cellerne $1, \dots, n$ indeholder $r_1, \dots, r_n$ partikler hvor $r_1 + \dots + r_n = r$.

I **Maxwell-Boltzmann modellen** opfattes partikelfordelingen som en fordeling af ens kugler i forskellige kasser. Hver fordeling har sandsynligheden n^{-r} og den samlede sandsynlighed bliver

$$\frac{\binom{r}{r_1, \dots, r_n}}{n^r} \quad (0.10)$$

I **Bose-Einstein modellen** skelnes mellem synligt forskellige arrangementer, og hver af disse tilordnes sandsynligheden

$$\frac{1}{\binom{n+r-1}{r}}$$

I **Fermi-Dirac modellen** antages, at ingen celler kan indeholde to partikler, og alle tilladelige fordelinger har samme sandsynlighed. Vi må her antage, at alle $k_i = 0$ eller 1 og får så sandsynligheden

$$\frac{1}{\binom{n}{r}}$$

Ifølge W. Feller (1957) er der ingen kendte fysiske partikler, der følger Maxwell-Boltzmann modellen. Bose-Einstein modellen dækker fotoner, atomkerner og atomer med et lige antal elementarpartikler. Fermi-Dirac modellen er velegnet til beskrivelse af elektroner, neutroner og protoner.

0.4 Sandsynlighedsfelter og stokastiske variable

De definitioner og sætninger, der ikke findes i [10] og [32], kan man læse om i næsten enhver moderne bog i videregående sandsynlighedsregning. En rimelig letfattelig bog er [21].

0.4.1 Om sandsynligheder, hændelser og stokastiske variable

Vi indeleder med at definere et såkaldt **sandsynlighedsfelt** $(\Omega, \mathcal{A}, P)$. Her er Ω vort **univers**, $\mathcal{A}$ en **σ -algebra** af delmængder af Ω og P en **sandsynlighed**.

Universet eller **udfaldsrummet** Ω kan vi opfatte som foreningsmængden af alle mulige udfald for et eksperiment. Visse delmængder A af udfaldsrummet Ω kaldes **hændelser** (hændelserne er de mængder af udfald, vi er interesserede i at komme med udsagn om og bestemme sandsynligheder for). Af tekniske grunde er det hensigtsmæssigt, hvis samlingen $\mathcal{A}$, af alle hændelser er en **σ -algebra**, d.v.s. tilfredsstiller betingelserne

$$1. \quad \Omega \in \mathcal{A} \quad (0.11)$$

$$2. \quad A \in \mathcal{A} \Rightarrow \mathcal{C}A \in \mathcal{A} \quad (0.12)$$

$$3. \quad A_1, A_2, \dots \in \mathcal{A} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A} \quad (0.13)$$

Her betegner $\mathcal{C}A$ komplementærmængden til A .

Sandsynligheden P er en afbildning $P : \mathcal{A} \rightarrow \mathbb{R}$, der tilfredsstiller

$$1. \quad 0 \leq P(A) \leq 1 \quad \forall A \in \mathcal{A} \quad (0.14)$$

$$2. \quad P(\Omega) = 1 \quad (0.15)$$

$$3. \quad P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i), \quad (0.16)$$

hvor $A_1, A_2, \dots$ er en følge af disjunkte mængder fra $\mathcal{A}$.

Af (0.16) kan vi udlede nogle nyttige resultater om voksende og aftagende følger af mængder. Hvis $B_1, B_2, B_3, \dots$ er en voksende følge af mængder, dvs. $B_1 \subseteq B_2 \subseteq B_3 \subseteq \dots$, med grænseværdien

$$B = \bigcup_{i=1}^{\infty} B_i = \lim B_i,$$

fås

$$P(B) = \lim P(B_i).$$

Hvis tilsvarende $C_1, C_2, C_3, \dots$ er en aftagende følge af mængder, dvs. $C_1 \supseteq C_2 \supseteq C_3 \supseteq \dots$, med grænseværdien

$$C = \bigcap_{i=1}^{\infty} C_i = \lim C_i,$$

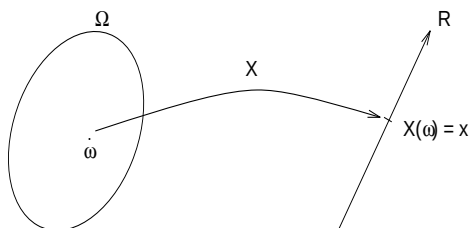
gælder

$$P(C) = \lim P(C_i).$$

En **stokastisk variabel** X er en afbildning fra udfaldsrummet Ω ind i de reelle tal, der tilfredsstiller

$$X^{-1}((-\infty, x]) = \{\omega | X(\omega) \leq x\} \in \mathcal{A}, \quad (0.17)$$

altså skal originalmængden (urbilledet) til et interval af formen $(-\infty, x]$ være indeholdt i hændelsesalgebraen $\mathcal{A}$:



For originalmængdeafbildningen bruges en speciel notation, nemlig

$$\{X = x\} \quad \text{i stedet for} \quad \{\omega | X(\omega) = x\} = X^{-1}(x)$$

$$\{X \leq x\} \quad \text{i stedet for} \quad \{\omega | X(\omega) \leq x\} = X^{-1}((-\infty, x]).$$

Bemærk, at $\{X = x\}$ og $\{X \leq x\}$ er hændelser.

Et simpelt tilfælde som støtte for forståelsen er følgende

EKSEMPEL 0.7. Vi betragter kast med en terning. Udfaldsrummet Ω er mængden af alle mulige udfald, d.v.s.

$$\Omega = \{\text{et øje}\} \cup \{\text{to øjne}\} \cup \dots \cup \{\text{seks øjne}\}.$$

En hændelse kan f.eks være ét øje og/eller fire øjne, d.v.s.

$$\{\text{et øje}\} \cup \{\text{fire øjne}\}$$

eller kun ét udfald

$$\{\text{fem øjne}\} \quad \text{etc.}$$

Vi går ud fra, at terningen er symmetrisk, d.v.s. vi har, at sandsynligheds målet P er bestemt ved

$$P\{\text{et øje}\} = \dots = P\{\text{seks øjne}\} = \frac{1}{6}.$$

Nu er det klart, at det, vi er interesseret i, er **antallet** af øjne. Man kunne derfor fristes til at definere en afbildning X ved

$$X(\{i \text{ øjne}\}) = i \quad i = 1, 2, 3, 4, 5, 6.$$

Dette X er en stokastisk variabel. Vi bruger X til at transformere vort problem over på den **reelle akse**. I stedet for at spørge om, hvad sandsynligheden er for at få hændelsen "et øje eller to øjne" ved terningkastet kan vi i stedet spørge om, hvad sandsynligheden er for, at $X \leq 2$, i.e.:

$$P(\{\text{et øje}\} \cup \{\text{to øjne}\}) = P\{X \leq 2\}.$$



0.5 Betingede sandsynligheder

EKSEMPEL 0.8. Lad os betragte et spil, hvor der kastes med 2 terninger, og hvor spillets udfald er summen af øjnene. Der er de mulige udfald

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

Sandsynligheden for hver af cellerne er $\frac{1}{36}$, hvis vi anvender en LaPlacisk sandsynlighedsdefinition.

Havde vi i stedet valgt at se bort fra alle spil, hvor anden terning viste 5 eller 6, dvs. kun ser på tilfælde, hvor anden terning viste ≤ 4 øjne, havde vi fået udfaldene

		2. terning			
		1	2	3	4
1. terning	1	2	3	4	5
	2	3	4	5	6
	3	4	5	6	7
	4	5	6	7	8
	5	6	7	8	9
	6	7	8	9	10

og sandsynligheden for hvert af disse udfald vil naturligvis sættes til $\frac{1}{24}$.

Endeligt kunne vi begrænse os til tilfældet, hvor de 2 terninger viser samme antal øjne, dvs. vi får udfaldene

		2. terning					
		1	2	3	4	5	6
1. terning	1	2					
	2		4				
	3			6			
	4				8		
	5					10	
	6						12

Vi indfører betegnelserne

- A : Summen af antal øjne i kastet med de to terninger er ≤ 6 .
 B : Antal øjne med anden terning er ≤ 4 .
 C : De to terninger viser det samme.

og finder, at disse har sandsynlighederne

$$\begin{aligned}
 P(A) &= \frac{15}{36} = \frac{5}{12} \\
 P(B) &= \frac{24}{36} = \frac{2}{3} \\
 P(C) &= \frac{6}{36} = \frac{1}{6}
 \end{aligned}$$

hvis vi betragter de oprindelige forsøg. Begrænser vi imidlertid til det andet forsøg, dvs. der hvor B gælder, dvs. hvor antallet af øjne for den anden terning er mindre end

eller lig med 4, da bliver sandsynligheden for at få en sum, der er mindre end eller lig med 6

$$Q(A) = \frac{14}{24} = \frac{7}{12}$$

hvor vi har brugt betegnelsen Q for sandsynligheden i dette eksperiment. Ser vi på det tilfælde, hvor C gælder, fås sandsynligheden

$$R(A) = \frac{3}{6} = \frac{1}{2}$$

hvor R tilsvarende er sandsynligheds målet i denne situation. Vi bemærker, at

$$Q(A) = \frac{14}{24} = \frac{14/36}{24/36} = \frac{P(A \cap B)}{P(B)}$$

og

$$R(A) = \frac{3}{6} = \frac{3/36}{6/36} = \frac{P(A \cap C)}{P(C)}$$



Inspireret af ovenstående eksempel kan vi introducere

DEFINITION 0.1 (BETINGET SANDSYNLIGHED). Den **betingede sandsynlighed** af en hændelse A **givet** en hændelse B er størrelsen

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (0.18)$$

forudsat at $P(B) > 0$.



En simpel anvendelse af definitionen anføres i

EKSEMPEL 0.9. Lad hændelsen A betegne udfaldet af et kast med en terning og hændelsen B , at udfaldet er et lige tal.

Såfremt det er givet, at udfaldet er et lige tal, er sandsynligheden for, at udfaldet er en sekser, lig

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

idet sandsynligheden for at få en sekser, som også er et lige tal, er $\frac{1}{6}$, og sandsynligheden for, at udfaldet er et lige tal er $\frac{1}{2}$. ♦

Hvis den betingede sandsynlighed $P(A|B)$ er lig $P(A)$ fås

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

I dette tilfælde får vi altså ikke noget at "vide" om sandsynligheden for hændelsen A ved at vide, at B er indtruffet. Derfor er det rimeligt at sige, at A og B er uafhængige. Dette er baggrunden for den efterfølgende definition.

DEFINITION 0.2. Vi siger, at et sæt hændelser $A_1, \dots, A_n$ er **stokastisk uafhængige**, hvis

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = P(A_{i_1}) \cdot P(A_{i_2}) \cdot \dots \cdot P(A_{i_k}) \quad (0.19)$$

for alle delmængder $\{i_1, \dots, i_k\}$ af $\{1, \dots, n\}$. Specielt er 2 hændelser A og B stokastisk uafhængige, hvis

$$P(A \cap B) = P(A) \cdot P(B) \quad (0.20)$$



Der gælder en lang række sætninger om sandsynligheder af forskellige kombinationer af hændelser. Vi nævner bl. a.

SÆTNING 0.3 (ADDITIONSSÆTNINGEN). For vilkårlige hændelser A og B gælder

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \quad (0.21)$$

Hvis A og B er disjunkte, reduceres dette til et specialtilfælde af (0.16).

Da $P(A|B) = P(A)$, såfremt A og B er uafhængige hvis $P(B) > 0$, fås umiddelbart den såkaldte **multiplikationssætning**

$$P(A \cap B) = P(A|B)P(B) \quad (0.22)$$

Endelig har vi

SÆTNING 0.4. Hvis $A_1, \dots, A_n$ er hinanden udelukkende hændelser (i.e. $A_i \cap A_j = \emptyset$ for $i \neq j$) med positive sandsynligheder, og hvis $\sum P(A_i) = 1$, da gælder for en vilkårlig hændelse B

$$1. \quad P(B) = \sum_i P(B \cap A_i) = \sum_i P(B|A_i)P(A_i), \quad (0.23)$$

hvis $P(B) > 0$

$$2. \quad P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum_i P(A_i)P(B|A_i)} \quad j = 1, \dots, n \quad (0.24)$$

Den sidste relation kaldes **Bayes regel**.

EKSEMPEL 0.10. Eksemplet er taget fra noter af Hofmann-Jøregensen. Betragt n forsikringstagere og lad os antage, at den i 'te forsikringstager har sandsynligheden p_i for at have et uheld i et enkelt år. Det antages desuden, at hændelserne held-uheld i de enkelte år er uafhængige for den enkelte forsikringstager, dvs. sandsynligheden for at en forsikringstager har uheld i to på hinanden følgende år er p_i^2 .

Vi udtrækker nu tilfældigt en forsikringstager og observerer denne i to på hinanden følgende år. Vi antager, at den i 'te forsikringstager har sandsynligheden λ_i ($\sum_i \lambda_i = 1$) for at blive udtrukket. Vi finder nu at

$$\begin{aligned} &P\{\text{forsikringstageren har uheld i det første år}\} \\ &= \sum P\{\text{uheld} | \text{nr } i\} P\{\text{nr } i\} = \sum_{i=1}^n p_i \cdot \lambda_i \end{aligned}$$

og

$$P\{\text{forsikringstageren har uheld i begge år}\} = \sum_{i=1}^n p_i^2 \cdot \lambda_i$$

Derfor bliver den betingede sandsynlighed

$$P\{\text{uheld i andet år} | \text{uheld i første år}\} = \frac{\sum_{i=1}^n p_i^2 \cdot \lambda_i}{\sum_{i=1}^n p_i \cdot \lambda_i}$$

Af Jensens ulighed følger umiddelbart, at

$$\left(\sum_{i=1}^n p_i \cdot \lambda_i \right)^2 \leq \sum_{i=1}^n p_i^2 \cdot \lambda_i,$$

hvorfor

$$\frac{\sum_{i=1}^n p_i^2 \cdot \lambda_i}{\sum_{i=1}^n p_i \cdot \lambda_i} \geq \sum_{i=1}^n p_i \cdot \lambda_i$$

eller med andre ord

$$\begin{aligned} P\{\text{uheld i andet år} | \text{uheld i første år}\} \\ \geq P\{\text{uheld i det første år}\} \end{aligned}$$

Dette tilsyneladende paradoks kaldes **falsk smitte**, og det indebærer altså, at hvis vi observerer et uheld for en tilfældigt udvalgt person vil dette øge sandsynligheden for, at denne samme person har et uheld året efter.

Vi bemærker, at hvis alle $p_i = p$, bliver de involverede sandsynligheder ens. ♦

EKSEMPEL 0.11. Vi betragter en familie, der får n børn, og vi antager, at der er sandsynligheden $\frac{1}{2}$ for at få et drengebarn og derfor også $\frac{1}{2}$ for at få en pige. Sandsynligheden for at få netop x drenge og $n - x$ piger ses at blive

$$\binom{n}{x} \left(\frac{1}{2}\right)^x \left(1 - \frac{1}{2}\right)^{n-x} = \binom{n}{x} 2^{-n}$$

Vi betragter nu hændelserne

- A : alle børn er af samme køn
- B : højst et af børnene er en pige

Da bliver

$$\begin{aligned} P(A \cap B) &= P\{\text{alle er drenge}\} = 2^{-n} \\ P(A) &= 2^{-n} + 2^{-n} = 2^{-n+1} \\ P(B) &= 2^{-n} + n2^{-n} = (n+1)2^{-n} \end{aligned}$$

Vi har da

$$P(A)P(B) = P(A \cap B) \Leftrightarrow 2^{-n+1}(n+1)2^{-n} = 2^{-n}$$

eller

$$n+1 = 2^{n-1} \Leftrightarrow n = 3,$$

dvs de to hændelser er uafhængige netop, hvis familien har 3 børn. $\blacklozenge$

0.6 Fordelings- og frekvensfunktioner

DEFINITION 0.3 (FORDELINGSFUNKTIONEN). F for en stokastisk variabel X er defineret ved

$$F(x) = P\{X \leq x\}.$$

$\blacktriangle$

Fordelingsfunktionen "indeholder" alle oplysninger om de sandsynligheder, hvormed hændelser af formen $\{a \leq X \leq b\}$ optræder. I stedet for at undersøge et sandsynlighedsmål P direkte kan vi efter transformation med den stokastiske variabel X i stedet for beskæftige os med en reel funktion af en reel variabel, nemlig fordelingsfunktionen F . Dette betyder selvfølgelig en stor teknisk lettelse (teorien for funktioner $F : \mathbb{R} \rightarrow \mathbb{R}$ er velkendt, hvorimod teorien for afbildninger $P : \mathcal{A} \rightarrow \mathbb{R}$ er noget vanskeligere tilgængelig).

En fordelingsfunktion er **voksende**. Dette følger af, at der for $g > 0$ gælder

$$\begin{aligned} F(x+h) &= P\{X \leq x+h\} \\ &= P\{X \leq x\} + P\{x < X \leq x+h\} \\ &\geq F(x) \end{aligned}$$

Definerer vi hændelserne

$$A_n = \{X \leq x + \frac{1}{n}\}$$

er

$$\lim A_n = \bigcap_{i=1}^{\infty} A_i = \{X \leq x\}$$

og vi får

$$\lim_{n \rightarrow \infty} F\left(x + \frac{1}{n}\right) = F(x)$$

dvs. F er kontinuert fra højre. Endelig gælder at

$$F(x) \rightarrow 1 \text{ for } x \rightarrow \infty \quad \text{og} \quad F(x) \rightarrow 0 \text{ for } x \rightarrow -\infty.$$

Ved **diskret** fordelte stokastiske variable forstås variable, der alene kan antage heltallige værdier (eller repræsentere forskellige kvalitetstyper etc), medens **kontinuert** fordelte stokastiske variable benyttes til at repræsentere kontinuerte størrelser (målelige, f.eks. vægt, længde, etc.).

DEFINITION 0.4 (FREKVENSFUNKTIONEN). eller **tætheden** f for X forstås funktionen givet ved

$$\begin{aligned} 1. \quad f(x) &= P\{X = x\} & X \text{ diskret fordelt} \\ 2. \quad f(x) &= F'(x) & X \text{ kontinuert fordelt} \end{aligned} \tag{0.25}$$

▲

I (0.25) forudsættes betingelse 2 blot gyldig i de punkter, hvor F er differentiabel. I det kontinuerte tilfælde taler vi også om **tætheden** f .

Fordelingsfunktionen bestemmes ud fra frekvensfunktionen ved

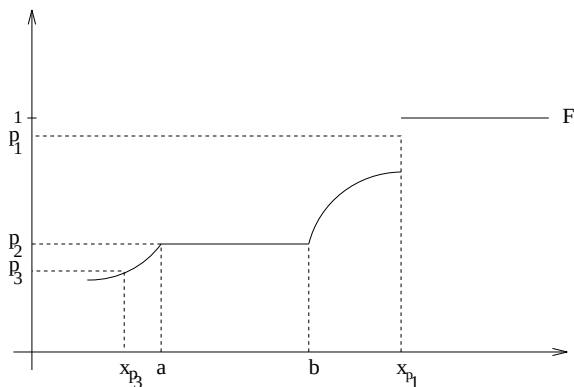
$$\begin{aligned} 1. \quad F(x) &= \sum_{i \leq x} f(i) & X \text{ diskret fordelt} \\ 2. \quad F(x) &= \int_{-\infty}^x f(t) dt & X \text{ kontinuert fordelt} \end{aligned} \tag{0.26}$$

Frekvensfunktioner er ikke-negative, og der gælder, at $\sum_x f(x) = 1$ respektive $\int_{-\infty}^{\infty} f(x) dx = 1$.

Til karakterisering af fordelinger anvendes hyppigt de såkaldte **fraktiler** (engelske navne: fractiles, quantiles, percentiles etc.). En p -fraktil x_p for en fordelingsfunktion F er groft taget defineret ved, at $F(x_p) = p$ eller $x_p = F^{-1}(p)$. Denne definition holder imidlertid kun for strengt monotone, kontinuerte F . Den stringente definition er, at x_p er en **p-fraktil**, netop hvis

$$F(x_p - 0) \leq p \leq F(x_p + 0). \quad (0.27)$$

Dette illustreres på nedenstående figur.



p_1 -fraktilen x_{p_1} og p_3 -fraktilen x_{p_3} er entydigt bestemte. Derimod er ethvert punkt i intervallet $[a, b]$ en p_2 -fraktil.

50%-fraktilen i en fordeling kaldes dennes **median**.

Uden bevis nævner vi

SÆTNING 0.5. En fordeling er entydigt bestemt af dens fraktiler. Anderledes udtrykt: Hvis 2 fordelinger har de samme p -fraktiler for ethvert p , da er fordelingerne identiske.

Vi slutter dette afsnit med at omtale nogle i statistikken yderst vigtige begreber, nemlig **positions-** og **skalaparametre**. Lad X være en stokastisk variabel med fordelingsfunktion F . Lad α og β være reelle tal med $\beta > 0$. Vi søger da fordelingen af

$$Y = \alpha + \beta X \quad (0.28)$$

Vi har

$$\begin{aligned} G(y) &= P\{Y \leq y\} \\ &= P\{\alpha + \beta X \leq y\} \\ &= P\left\{X \leq \frac{y - \alpha}{\beta}\right\} \end{aligned}$$

eller

$$G(y) = F\left\{\frac{y - \alpha}{\beta}\right\} \quad (0.29)$$

Frekvensfunktionen bliver

$$g(y) = \begin{cases} f\left(\frac{y - \alpha}{\beta}\right) & X \text{ diskret} \\ \frac{1}{\beta} f\left(\frac{y - \alpha}{\beta}\right) & X \text{ kontinuert} \end{cases} \quad (0.30)$$

DEFINITION 0.5. En fordeling G , der fremkommer ved en affin transformation af en fordeling F ((0.28)-(0.30)), siges at være af samme **type** som F , men med "**positionsparameter** α og **skalaparameter** β ". ▲

En fornemmelse af disse parameters betydning får man bedst ved at se på tætheden for kontinuerte stokastiske variable. Vi skitserer nogle grafer for tætheder af samme type, men med forskellige positions- og skalaparametre.

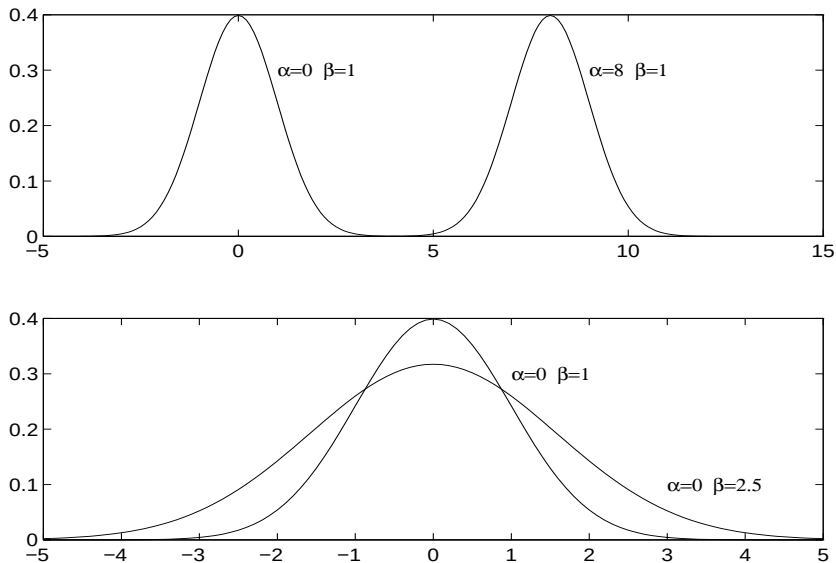
Vi ser, at positionsparameteren blot "flytter" fordelingen på akse, hvorimod en større skalaparameter "tværer" fordelingen ud over akse.

0.7 Flerdimensionale stokastiske variable

En n -dimensional stokastisk variabel er en vektor $(X_1, \dots, X_n)$, hvor de enkelte komponenter er endimensionale stokastiske variable.

Den **simultane fordelingsfunktion** F for $(X_1, \dots, X_n)$ er givet ved

$$F(x_1, \dots, x_n) = P\{X_1 \leq x_1 \wedge \dots \wedge X_n \leq x_n\} \quad (0.31)$$



Den **simultane frekvensfunktion** for $(X_1, \dots, X_n)$ er en afbildning f med egenskaben

$$F(x_1, \dots, x_n) = \begin{cases} \sum_{t_1 \leq x_1} \cdots \sum_{t_n \leq x_n} f(t_1, \dots, t_n) \\ \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t_1, \dots, t_n) dt_n \cdots dt_1 \end{cases} \quad (0.32)$$

i henholdsvis det diskrete og det kontinuerte tilfælde. Hvis $(X_1, \dots, X_n)$ er af kontinuert type, taler vi også om **tætheden** f , og vi har

$$f(x_1, \dots, x_n) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F(x_1, \dots, x_n) \quad (0.33)$$

i punkter, hvor F har kontinuerte partielle afledede.

Vi betragter igen en n -dimensional stokastisk variabel $(X_1, \dots, X_n)$. Ved den **marginale fordelingsfunktion** for X_i ($1 \leq i \leq n$) forstås simpelt hen X_i 's fordelingsfunktion F_i givet ved $F_i(x) = P\{X_i \leq x\}$. Analogt selvfølgelig for **marginale frekvensfunktioner** og **tætheder**.

Til bestemmelse af de marginale fordelinger ud fra den simultane fordeling har vi følgende

SÆTNING 0.6. Den marginale frekvensfunktion f_i , for $i = 1, \dots, n$, er givet ved

$$f_i(x_i) = \begin{cases} \sum_{x_1=-\infty}^{\infty} \cdots \sum_{x_{i-1}=-\infty}^{\infty} \sum_{x_{i+1}=-\infty}^{\infty} \cdots \sum_{x_n=-\infty}^{\infty} f(x_1, \dots, x_n) \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) dx_n \cdots dx_{i+1} dx_{i-1} \cdots dx_1 \end{cases} \quad (0.34)$$

i henholdsvis det diskrete og det kontinuerte tilfælde.

Til sidst indfører vi begrebet stokastisk uafhængige variable.

DEFINITION 0.6. De stokastiske variable $X_1, \dots, X_n$ er (**stokastisk**) **uafhængige**, netop hvis den simultane fordeling for $(X_1, \dots, X_n)$ er produktet af de marginale fordelinger for X_i 'erne, i. e. hvis

$$F(x_1, \dots, x_n) = F_1(x_1) \cdots F_n(x_n)$$

eller tilsvarende

$$f(x_1, \dots, x_n) = f_1(x_1) \cdots f_n(x_n).$$

▲

Ved den **betingede fordeling** af en (eventuelt n -dimensional) stokastisk variabel X **givet** den (eventuelt k -dimensionale) stokastisk variabel Y forstår vi i det tilfælde, hvor Y følger en diskret fordeling, simpelthen sandsynligheds målet Q_y givet ved

$$Q_y(B) = P\{X \in B | Y = y\}. \quad (0.35)$$

Hvis X også er diskret fordelt, bliver den **betingede frekvensfunktion** af X **givet** $Y = y$

$$\begin{aligned} g(x|y) &= P\{X = x | Y = y\} \\ &= \frac{P(\{X = x\} \cap \{Y = y\})}{P\{Y = y\}} = \frac{f(x, y)}{h(y)}, \end{aligned} \quad (0.36)$$

hvor f er den simultane frekvensfunktion af X og Y , og h er Y 's marginale frekvensfunktion. Det er forudsat, at $h(y) > 0$.

Hvis fordelingerne ikke er diskrete, er begrebet en betinget fordeling kompliceret; men det kan vises, at man også i det kontinuerte tilfælde får, at den **betingede fordeling** af X givet Y er givet ved tætheden

$$g(x|y) = \frac{f(x, y)}{h(y)}, \quad (0.37)$$

såfremt $h(y) > 0$.

0.8 Transformation af stokastiske variable

Vi anfører hovedresultatet, nemlig

SÆTNING 0.7. Lad $\mathbf{X}$ være en n -dimensional stokastisk variabel med tæthed

$$f(x_1, \dots, x_n) \neq 0 \quad \text{for} \quad \mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbf{A} \subseteq \mathbb{R}^n.$$

Vi betragter en injektiv afbildning $h : A \rightarrow B \subseteq \mathbb{R}^n$, der har kontinuerte partielle afledede. Da har

$$\mathbf{Y} = h(\mathbf{X}) = h(\mathbf{X}_1, \dots, \mathbf{X}_n)$$

tætheden

$$g(y_1, \dots, y_n) = f(h^{-1}(y_1, \dots, y_n)) \cdot \left| \det \frac{\partial(x_1, \dots, x_n)}{\partial(y_1, \dots, y_n)} \right|,$$

for $\mathbf{y} \in \mathbf{B}$ og 0 ellers.

Bevis. Forbigås. Sætningen er ensbetydende med den fra analysen kendte integral-transformationssætning. ■

Det kan ved beregninger være nyttigt at erindre, at **Jacobi-determinanten** ofte lettest beregnes ved hjælp af relationen

$$\det \frac{\partial(x_1, \dots, x_n)}{\partial(y_1, \dots, y_n)} = \frac{1}{\det \frac{\partial(y_1, \dots, y_n)}{\partial(x_1, \dots, x_n)}}.$$

KOROLLAR 0.1. Lad X være en endimensional kontinuert fordelt stokastisk variabel med tæthed $f(x) \neq 0$ for $x \in A \subseteq \mathbb{R}$ og lad h være en monoton afbildning. Da er tætheden for $Y = h(X)$ givet ved

$$g(y) = f(h^{-1}(y)) \left| \frac{dh^{-1}(y)}{dy} \right|$$

for $y \in h(A)$ og 0 ellers.

Bevis. Vi indskrænker os til det tilfælde, hvor h er voksende. Vi kalder fordelingsfunktionerne for X og Y for $F(x)$ henholdsvis $G(y)$. Vi finder da

$$\begin{aligned} G(y) &= P\{Y \leq y\} \\ &= P\{h(X) \leq y\} \\ &= P\{X \leq h^{-1}(y)\} \\ &= F(h^{-1}(y)) \end{aligned}$$

Frekvensfunktionen for Y findes nu ved differentiation

$$\begin{aligned} g(y) &= G'(y) \\ &= F'(h^{-1}(y)) \cdot \frac{dh^{-1}(y)}{dy} \end{aligned}$$

Da $F' = f$ følger corollaret umiddelbart. ■

Da $x = h^{-1}(y)$ kan corollaret formuleres - med en lidt upræcis men mnemoteknisk set udmærket notation -

$$f(x) = g(y) = f(x) \left| \frac{dx}{dy} \right|$$

hvilket i det **monotont voksende** tilfælde kan skrives

$$g(y)dy = f(x)dx$$

EKSEMPEL 0.12. Lad

$$\mathbf{y} = h(\mathbf{x}) = \mathbf{A} \mathbf{x} + \mathbf{b},$$

hvor $\mathbf{A}$ er en regulær $n \times n$ -matrix. Da er

$$\mathbf{x} = \mathbf{A}^{-1} \mathbf{y} - \mathbf{A}^{-1} \mathbf{b},$$

d.v.s.

$$\det \frac{\partial (x_1, \dots, x_n)}{\partial (y_1, \dots, y_n)} = \det \mathbf{A}^{-1},$$

hvorfor Jacobi-determinanten er

$$\det \mathbf{A}^{-1} = [\det(\mathbf{A})]^{-1}.$$

Tætheden for $\mathbf{Y} = h(\mathbf{X})$ bliver derfor

$$g(\mathbf{y}) = \frac{1}{|\det(\mathbf{A})|} f(\mathbf{A}^{-1} \mathbf{y} - \mathbf{A}^{-1} \mathbf{b}).$$



Vi illustrerer transformationssætningen med yderligere et

EKSEMPEL 0.13. Lad X_1 og X_2 være uafhængige med samme frekvensfunktion

$$f(x) = \frac{1}{\beta} \exp(-x/\beta) \quad \text{for } x > 0$$

og 0 ellers. Vi vil bestemme den simultane frekvensfunktion for $X_1 + X_2$ og X_1/X_2 .

Vi har transformationen $(y_1, y_2) = h(x_1, x_2)$ med

$$\begin{aligned} y_1 &= x_1 + x_2 \\ y_2 &= x_1/x_2 \end{aligned}$$

Den inverse transformation er

$$\begin{aligned}x_1 &= \frac{y_1 + y_2}{1 + y_2} \\x_2 &= \frac{y_1}{1 + y_2}\end{aligned}$$

For illustrationens skyld bestemmer vi funktionalmatricerne for såvel transformationen som for den inverse transformation. Vi finder

$$\frac{\partial(y_1, y_2)}{\partial(x_1, x_2)} = \begin{Bmatrix} \frac{dy_1}{dx_1} & \frac{dy_1}{dx_2} \\ \frac{dy_2}{dx_1} & \frac{dy_2}{dx_2} \end{Bmatrix} = \begin{Bmatrix} 1 & 1 \\ \frac{1}{x_2} & -\frac{x_1}{x_2^2} \end{Bmatrix}$$

Denne har determinanten

$$-\frac{x_1}{x_2^2} - \frac{1}{x_2} = -\frac{x_1 + x_2}{x_2^2},$$

der skrevet op som funktion af y_1 og y_2 bliver

$$-\frac{y_1(1+y_2)^2}{y_1^2} = -\frac{(1+y_2)^2}{y_1}$$

Analogt finder vi

$$\frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} = \begin{Bmatrix} \frac{dx_1}{dy_1} & \frac{dx_1}{dy_2} \\ \frac{dx_2}{dy_1} & \frac{dx_2}{dy_2} \end{Bmatrix} = \begin{Bmatrix} \frac{y_2}{1+y_2} & \frac{y_1}{(1+y_2)^2} \\ \frac{1}{1+y_2} & -\frac{y_1}{(1+y_2)^2} \end{Bmatrix}$$

Denne har determinanten

$$\frac{1}{(1+y_2)^3} \{-y_2 y_1 - y_1\} = -\frac{y_1}{(1+y_2)^2}$$

og vi ser, at denne størrelse som forventet er den reciprokke til det tidligere fundne udtryk.

Vi ser, at transformationen afbilder hele det område, hvor $f_1(x_1)f_2(x_2) \neq 0$, dvs. første kvadrant i (x_1, x_2) planen over i første kvadrant i (y_1, y_2)

Den samlede frekvensfunktion for (X_1, X_2) er

$$f_{X_1, X_2}(x_1, x_2) = \begin{cases} \frac{1}{\beta^2} \exp(-(x_1 + x_2)/\beta) & \text{for } x_1, x_2 > 0 \\ 0 & \text{ellers} \end{cases}$$

og derfor bliver den simultane frekvensfunktion for (Y_1, Y_2) lig

$$f_{Y_1, Y_2}(y_1, y_2) = \begin{cases} \frac{1}{\beta^2} y_1 \exp(-y_1/\beta) \frac{1}{(1+y_2)^2} & \text{for } y_1, y_2 > 0 \\ 0 & \text{ellers} \end{cases}$$

Vi ser, at $f_{Y_1, Y_2}(y_1, y_2)$ er produktet af

$$g_1(y_1) = \frac{1}{\beta^2} y_1 \exp(-y_1/\beta)$$

og

$$g_2(y_2) = \frac{1}{(1+y_2)^2}.$$

Derfor er Y_1 og Y_2 uafhængige og har frekvensfunktionerne g_1 og g_2 (ved integration checkes, at hvert af udtrykkene er multipliceret med konstanter, så den totale sandsynlighedsmasse bliver 1). ♦

Vi skal nu benytte den almindelige transformationssætning til at bestemme fordelingen af summen af to (kontinuert fordelte) stokastiske variable. Vi anfører resultatet i

SÆTNING 0.8. Lad de stokastiske variable X og Y have den simultane frekvensfunktion $f(x, y)$. Da har summen $X + Y$ frekvensfunktionen

$$g(s) = \int_{-\infty}^{\infty} f(s-t, t) dt \quad (0.38)$$

Hvis X og Y er uafhængige med frekvensfunktioner f_X og f_Y fås

$$g(s) = \int_{-\infty}^{\infty} f_X(s-t) f_Y(t) dt \quad (0.39)$$

Integralet betegnes også **foldningen** af f_X og f_Y .

Bevis. Vi betragter transformationen

$$\begin{aligned} u &= x + y \\ v &= y \end{aligned}$$

Den inverse afbildning er givet ved

$$\begin{aligned} x &= u - v \\ y &= v \end{aligned}$$

Funktionaldeterminanten er

$$\det \begin{pmatrix} \frac{dx}{du} & \frac{dx}{dv} \\ \frac{dy}{du} & \frac{dy}{dv} \end{pmatrix} = 1$$

og den simultane frekvensfunktion for $U, V) = (X + Y, Y)$ bliver derfor

$$h(u, v) = f(u - v, v)$$

og den marginale frekvensfunktion for U fremkommer ved at integrere v ud, dvs

$$g(u) = \int_{-\infty}^{\infty} f(u - v, v) dv$$

Resultatet for uafhængige X og Y følger trivielt. ■

Vi betragter et

EKSEMPEL 0.14. Lad X_1 og X_2 være uafhængige med frekvensfunktioner

$$f_i(x) = \begin{cases} \frac{1}{\beta} \exp(-x/\beta) & x \geq 0 \\ 0 & x < 0 \end{cases} \quad i = 1, 2$$

Da bliver frekvensfunktionen for $X_1 + X_2$

$$g(s) = \int_{-\infty}^{\infty} f_1(s - t) f_2(t) dt$$

Nu er

$$\begin{aligned} f_1(s-t) &= 0 & \text{for } t > s \\ f_2(t) &= 0 & \text{for } t < 0 \end{aligned}$$

Derfor bliver

$$\begin{aligned} g(s) &= \int_{-\infty}^{\infty} f_1(s-t) f_2(t) dt \\ &= \int_0^s \frac{1}{\beta^2} \exp(-(s-t+t)/\beta) dt \\ &= \frac{1}{\beta^2} \exp(-s/\beta) \int_0^s dt \\ &= \frac{1}{\beta^2} s \exp(-s/\beta) \end{aligned}$$

for $s > 0$ og 0 ellers. ♦

Man kan på samme vise resultaterne for $X - Y$, $X \cdot Y$ og X/Y . Vi samler for overskuelighedens skyld resultaterne i tabel 0.10.

Funktion	Frekvensfunktion generelle tilfælde	Frekvensfunktion X, Y uafhængige
$X + Y$	$\int_{-\infty}^{\infty} f(s-t, t) dt$	$\int_{-\infty}^{\infty} f_X(s-t) f_Y(t) dt$
$X - Y$	$\int_{-\infty}^{\infty} f(t, t-s) dt$	$\int_{-\infty}^{\infty} f_X(t) f_Y(t-s) dt$
$X \cdot Y$	$\int_{-\infty}^{\infty} f(t, \frac{s}{t}) \frac{1}{ t } dt$	$\int_{-\infty}^{\infty} f_X(t) f_Y(\frac{s}{t}) \frac{1}{ t } dt$
X/Y	$\int_{-\infty}^{\infty} f(st, t) t dt$	$\int_{-\infty}^{\infty} f_X(st) f_Y(t) t dt$

Tabel 0.10: Frekvensfunktioner for simple funktioner af to kontinuert fordelte stokastiske variable X og Y med simultan frekvensfunktion $f(x, y)$ såvel i det generelle tilfælde som i tilfældet med uafhængige X og Y . I det sidste tilfælde er frekvensfunktionerne for X og Y $f_X(x)$ henholdsvis $f_Y(y)$.

0.9 Momenter

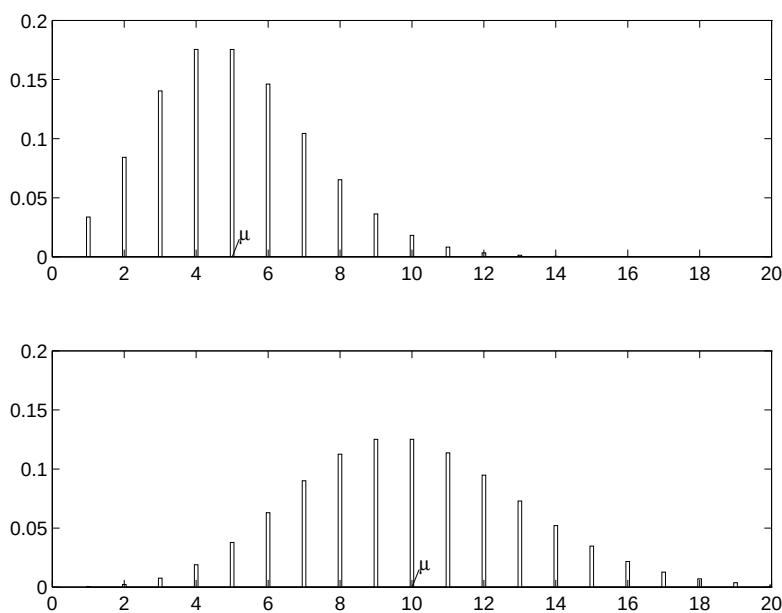
DEFINITION 0.7. Middelværdien eller **forventningsværdien** af en stokastisk variabel X med frekvensfunktionen f er tallet

$$E(X) = \begin{cases} \sum x f(x) & X \text{ diskret} \\ \int_{-\infty}^{\infty} x f(x) dx & X \text{ kontinuert} \end{cases}$$

▲

Det kræves, at $\int |x|f(x) dx$ respektive $\sum |x|f(x)$ eksisterer. $E()$ står for expectation = forventning. Det ses, at middelværdien eller forventningsværdien netop svarer til tyngdepunktet i en (masse-)fordeling. Som standardbetegnelse for middelværdien anvendes ofte μ .

Forventningsværdiens betydning som mål for en fordelings placering fremgår af nedenstående figur, hvor der er anført nogle frekvensfunktioner med forskellige middelværdier.



SÆTNING 0.9. Lad g være en reel funktion af n variable. Hvis den (flerdimensionale)

stokastiske variable $(X_1, \dots, X_n)$ har frekvensfunktionen f , da er

$$\begin{aligned} E(g(X_1, \dots, X_n)) &= \\ &\begin{cases} \sum_{x_1=-\infty}^{\infty} \cdots \sum_{x_n=-\infty}^{\infty} g(x_1, \dots, x_n) f(x_1, \dots, x_n) \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \dots, x_n) f(x_1, \dots, x_n) dx_n \cdots dx_1 \end{cases} \end{aligned}$$

i henholdsvis det diskrete og det kontinuerte tilfælde.

Bevis. Vi indskrænker os til det endimensionale tilfælde med en kontinuert fordelt stokastisk variabel X , der transformeres af en monotont voksende funktion g . Vi har, at frekvensfunktionen for $Y = g(X)$ er

$$h(y) = f(g^{-1}(y))(g^{-1})'(y)$$

Ved anvendelse af substitutionen $x = g^{-1}(y)$ fås

$$\begin{aligned} \int_{-\infty}^{\infty} g(x) f(x) dx &= \int_{-\infty}^{\infty} g(g^{-1}(y)) f(g^{-1}(y)) (g^{-1})'(y) dy \\ &= \int_{-\infty}^{\infty} y h(y) dy \\ &= E(Y) \end{aligned}$$

hvilket netop er det søgte resultat. ■

BEMÆRKNING 0.2. Vi ser, at e.g. X_1 's middelværdi naturligvis er den samme, hvad enten vi finder den i den marginale fordeling for X_1 , eller vi beregner den ved hjælp af ovenstående sætning. ▼

Ved hjælp af Sætning 0.9 fås umiddelbart følgende vigtige regneregler for forventningsværdier

$$E(a + bX) = a + bE(X) \quad (0.40)$$

$$E(X + Y) = E(X) + E(Y) \quad (0.41)$$

$$E(XY) = E(X)E(Y), \text{ } X \text{ og } Y \text{ uafhængige} \quad (0.42)$$

Vi skal nu indføre begrebet momenter i en fordeling.

DEFINITION 0.8. Den stokastiske variabel X 's k 'te moment er

$$\mu'_k = E(X^k)$$

Det k 'te centrale moment er størrelsen

$$\mu_k = E((X - E(X))^k) = E((X - \mu)^k)$$

▲

Det andet centrale moment kaldes også **variansen** og er altså

$$V(X) = E((X - E(X))^2) = E(X^2) - (E(X))^2 \quad (0.43)$$

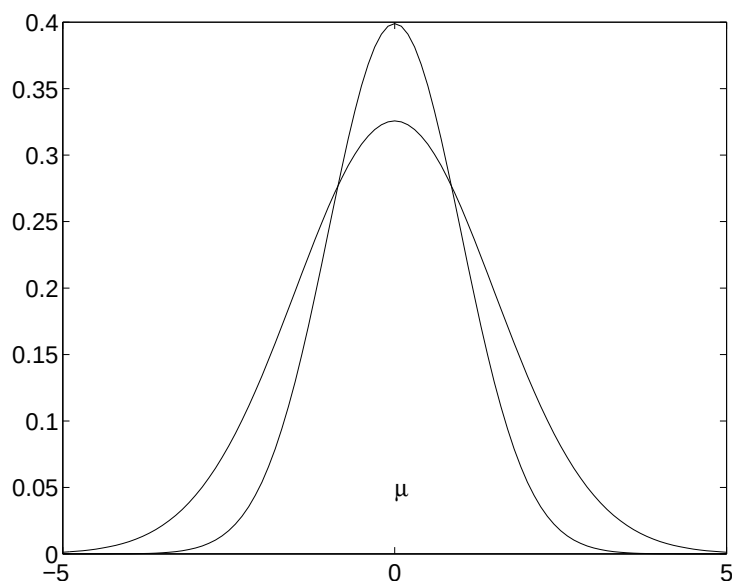
(Det sidste lighedstegn ses ved direkte regning). Man bemærker, at variansen svarer til inertimomentet i en massefordeling. Som standardbetegnelse for variansen anvendes ofte σ^2 .

Variansen er et mål for, hvor meget sandsynlighedsmassen er spredt ud over akse. Dette fremgår måske klarere, hvis man anvender skrivemåden

$$V(X) = \begin{cases} \sum_{i=1}^{\infty} (x_i - \mu)^2 f(x_i) \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \end{cases} ,$$

idet variansens betydning som mål for den gennemsnitlige kvadrerede afvigelse fra forventningsværdien nu klarere træder frem.

I nedenstående figur er anført to frekvensfunktioner med samme forventningsværdi, men med forskellige varianser.



Kvadratroden af variansen kaldes **spredningen** eller **standardafvigelsen**.

Som et relativt mål for "bredden" af en sandsynlighedsfordeling anvendes ofte forholdet mellem spredningen og middelværdien

$$\frac{\sigma}{\mu} = \frac{\sqrt{V(X)}}{E(X)}.$$

Denne størrelse kaldes **variationskoefficienten**. Den er dog kun et fornuftigt mål, såfremt der findes et "naturligt" nulpunkt (forventningsværdien ændres jo ved en nulpunktsforskydning).

Sandsynlighedsmål på den positive halvakse kan foruden ved de sædvanlige beliggenhedsmål og spredningsmål også beskrives ved de såkaldte geometriske mål, nemlig **geometrisk middelværdi** og **geometrisk standardafvigelse**. For en stokastisk variabel X , der kun antager positive værdier, defineres disse ved

$$\begin{aligned}\mu_g &= \exp(E(\log_e X)) \\ \sigma_g &= \exp(\sqrt{V(\log_e X)}),\end{aligned}$$

naturligvis forudsat, at forventningsværdierne eksisterer.

Endelig defineres **kovariansen** mellem 2 stokastiske variable ved

$$\text{Cov}(X, Y) = E(X - E(X)) E(Y - E(Y)) \quad (0.44)$$

og vi har

$$\text{Cov}(X, Y) = E(X \cdot Y) - E(X) E(Y) \quad (0.45)$$

Hvis X og Y er uafhængige ses, at $\text{Cov}(X, Y) = 0$.

Der gælder de vigtige regneregler

$$V(aX) = a^2 V(X) \quad (0.46)$$

$$V(X + b) = V(X) \quad (0.47)$$

$$V(X \pm Y) = \begin{cases} V(X) + V(Y) \pm 2\text{Cov}(X, Y) \\ V(X) + V(Y), X \text{ og } Y \text{ uafhængige} \end{cases} \quad (0.48)$$

$$\text{Cov}((a_1 X + b_1), (a_2 Y + b_2)) = a_1 a_2 \text{Cov}(X, Y) \quad (0.49)$$

Ved kombination af (0.48) og (0.49) fås

$$\begin{aligned} V(a_1 X_1 + \dots + a_n X_n) &= a_1^2 V(X_1) + \dots + a_n^2 V(X_n) \\ &\quad + 2a_1 a_2 \text{Cov}(X_1, X_2) + \dots \\ &\quad + 2a_{n-1} a_n \text{Cov}(X_{n-1}, X_n) \end{aligned}$$

Endelig gælder, at **korrelationskoefficienten** ρ defineres ved

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{V(X) V(Y)}} \quad (0.50)$$

Vi bemærker, at $|\rho| \leq 1$.

Korrelationen er et mål for den lineære sammenhæng mellem to (ikke-udartede) stokastiske variable. Kun hvis der med sandsynlighed 1 gælder en relation af formen

$$\alpha X + \beta Y + \gamma = 0,$$

hvor $(\alpha, \beta) \neq (0, 0)$, gælder det, at $|\rho| = 1$. Hvis α og β har modsat fortegn, er $\rho = 1$, og hvis de har samme fortegn, er $\rho = -1$. I alle andre tilfælde er $|\rho| < 1$. Hvis $\rho = 0$, kaldes de variable **ukorrelerede**. Det skal bemærkes, at uafhængige variable altid er ukorrelerede, hvorimod det modsatte ikke nødvendigvis er tilfældet.

For en flerdimensional stokastisk variabel (skrives af notationsmæssige grunde som en søjlevektor)

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

definerer vi

$$\mathbf{E}(\mathbf{X}) = \begin{pmatrix} \mathbf{E}(X_1) \\ \vdots \\ \mathbf{E}(X_n) \end{pmatrix}$$

$\mathbf{D}(\mathbf{X}) =$

$$\begin{pmatrix} \mathbf{V}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \mathbf{V}(X_2) & & \\ \vdots & \vdots & & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \cdots & \mathbf{V}(X_n) \end{pmatrix}$$

Denne matrix kaldes også for **kovariansmatricen** eller **dispersionsmatricen**. Vi bemærker, at den er symmetrisk.

Betragtes en linearkombination $\mathbf{a}^T \mathbf{X}$ af $\mathbf{X}$'s komponenter, kan det vises, at

$$\mathbf{V}(\mathbf{a}^T \mathbf{X}) = \mathbf{a}^T \mathbf{D}(\mathbf{X}) \mathbf{a}$$

Da variansen på venstresiden altid er ≥ 0 må den kvadratiske form på højresiden også være ≥ 0 . Heraf følger umiddelbart, at $\mathbf{D}(\mathbf{X})$ må være positiv semidefinit.

Betingede middelværdier og **varianser** af X for givet Y defineres som middelværdier henholdsvis varianser i de betingede fordelinger. Man anvender symbolerne

$$E(X|Y)$$

henholdsvis

$$\begin{aligned} V(X|Y) &= E((X - E(X|Y))^2|Y) \\ &= E(X^2|Y) - (E(X|Y))^2. \end{aligned}$$

For hver realisation af Y fås en realisation af $E(X|Y)$ henholdsvis $V(X|Y)$, og disse størrelser er således stokastiske variable, og der gælder følgende relationer

$$E(E(X|Y)) = E(X)$$

$$V(X) = V(E(X|Y)) + E(V(X|Y))$$

$$\text{Cov}(X, Z) = E(\text{Cov}(X, Z|Y)) + \text{Cov}(E(X|Y), E(Z|Y))$$

EKSEMPEL 0.15. Lad X og Y være indbyrdes uafhængige stokastiske variable med sammen forventningsværdi μ og varianser lig σ^2 henholdsvis δ^2 . Vi betragter nu den stokastiske variabel

$$Z = wX + (1 - w)Y, \quad 0 \leq w \leq 1$$

Vi finder umiddelbart

$$\begin{aligned} E(Z) &= w\mu + (1 - w)\mu = \mu \\ V(Z) &= w^2\sigma^2 + (1 - w)^2\delta^2 = g(w) \end{aligned}$$

Vi ser altså, at forventningsværdien af Z er lig μ uafhængigt af værdien af w . Vi vil nu bestemme den værdi af w , der gør variansen af Z så lille som mulig. Vi differentierer udtrykket med hensyn til w og finder

$$\begin{aligned} \frac{dg}{dw} &= 2w\sigma^2 - 2(1 - w)\delta^2 \\ &= 2w(\sigma^2 + \delta^2) - 2\delta^2 \end{aligned}$$

Derfor er

$$\frac{dg}{dw} = 0 \quad \Leftrightarrow \quad w = \frac{\delta^2}{\sigma^2 + \delta^2}$$

De den anden afledede af g er negativ, er dette et minimum. Vi omskriver resultatet let og får, at variansen af Z er minimal for vægtene

$$w = \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma^2} + \frac{1}{\delta^2}}$$

og

$$1 - w = \frac{\frac{1}{\delta^2}}{\frac{1}{\sigma^2} + \frac{1}{\delta^2}}$$

dvs. hvis vægtene er proportionale med de reciprokke varianser. $\blacklozenge$

Vi generaliserer resultatet i ovenstående eksempel til summer af flere stokastiske variable i

EKSEMPEL 0.16. Lad $X_1, \dots, X_n$ være indbyrdes uafhængige stokastiske variable med

$$\begin{aligned} E(X_i) &= \mu \\ V(X_i) &= \sigma_i^2 \end{aligned}$$

og sæt

$$Z = w_1 X_1 + \dots + w_n X_n, \quad w_1 + \dots + w_n = 1$$

Da er

$$V(Z) = w_1^2 \sigma_1^2 + \dots + w_n^2 \sigma_n^2 = g(w_1, \dots, w_n)$$

For at minimalisere denne under bibetingelsen $w_1 + \dots + w_n = 1$ indføres en Lagrange multiplikator λ , og vi får, at vi skal minimalisere

$$h(w_1, \dots, w_n, \lambda) = g(w_1, \dots, w_n) + \lambda \left(\sum_j w_j - 1 \right)$$

Vi differentierer og finder

$$\begin{aligned}\frac{dh}{dw_i} &= 2w_i\sigma_i^2 + \lambda, \quad i = 1, \dots, n \\ \frac{dh}{d\lambda} &= \sum_j w_j - 1\end{aligned}$$

Sættes $\frac{dh}{dw_i} = 0$ fås

$$w_i = -\frac{\lambda}{2} \frac{1}{\sigma_i^2}$$

Indsættes dette i $\frac{dh}{d\lambda} = 0$, dvs. $\sum_j w_j = 1$, fås

$$-\frac{\lambda}{2} \sum_j \frac{1}{\sigma_j^2} = 1$$

eller

$$\lambda = -\left(\frac{1}{2} \sum_j \frac{1}{\sigma_j^2}\right)^{-1}$$

Dette giver følgende udtryk for w_i

$$w_i = \frac{\frac{1}{\sigma_i^2}}{\sum_j \frac{1}{\sigma_j^2}}$$

dvs., at vægtene w_i er proportionale med de reciprokke varianser. ♦

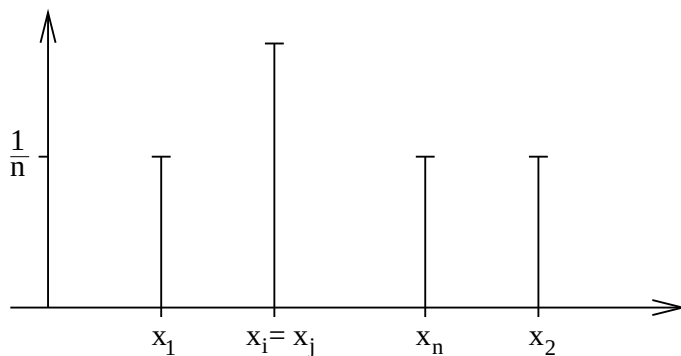
EKSEMPEL 0.17 (DEN DISKRETE FORDELING PÅ n TAL.). Vi definerer tallene

$$x_1, \dots, x_n$$

og definerer frekvensfunktionen givet ved

$$f(x_i) = \frac{1}{n}$$

Grafen er



I denne fordeling er middelværdien eller forventningsværdien

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

dvs gennemsnittet af tallene. Betegnelsen $\bar{x}$ (læs: x streg) er en standardnotation for denne størrelse.

Tilsvarende bliver variansen lig

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

Formlen

$$V(X) = E(X^2) - [E(X)]^2$$

bliver til

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \left[\frac{1}{n} \sum_{i=1}^n x_i \right]^2$$

eller

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

en formel, der kan være hensigtsmæssigt ved numeriske beregninger. ♦

EKSEMPEL 0.18. Lad $X \in U(0, 1)$, dvs uniformt fordelt over intervallet $[0, 1]$. Da finder vi

$$E(X^n) = \int_0^1 x^n dx = \frac{1}{n+1}$$

Derfor er

$$\begin{aligned} E(X) &= \frac{1}{2} \\ V(X) &= \frac{1}{3} - \frac{1}{4} = \frac{1}{12} \end{aligned}$$

Desuden finder vi, som et eksempel,

$$\begin{aligned} \text{Cov}(X^2, X) &= E([X^2 - E(X^2)][X - E(X)]) \\ &= E(X^3) - E(X^2)E(X) \\ &= \frac{1}{4} - \frac{1}{3} \cdot \frac{1}{2} \\ &= \frac{1}{12} \end{aligned}$$

Nu er

$$V(X^2) = E(X^4) - [E(X^2)]^2 = \frac{1}{5} - \frac{1}{4} = \frac{4}{45}$$

hvorfor korrelationskoefficienten mellem X^2 og X bliver

$$\rho = \frac{1/12}{\sqrt{4/45}\sqrt{1/12}} = \frac{\sqrt{15}}{4} \simeq 0.9682$$

ikke overraskende et højt tal. ♦

EKSEMPEL 0.19. Lad de stokastiske variable X og Y have den simultane frekvens-

funktion

$$f(x, y) = \frac{1}{2\pi} \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}[x^2 - 2\rho xy + y^2]\right)$$

Vi bestemmer først den marginale fordeling for X ved at integrere Y ud. Vi finder

$$\begin{aligned} h(x) &= \int_{-\infty}^{\infty} f(x, y) dy \\ &= c \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2(1-\rho^2)}[x^2(1-\rho^2) + (y-\rho x)^2]\right) dy \\ &= c \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}x^2\right) \exp\left(-\frac{1}{2(1-\rho^2)}(y-\rho x)^2\right) dy \\ &= c \exp\left(-\frac{1}{2}x^2\right) \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2(1-\rho^2)}(y-\rho x)^2\right) dy \\ &= c_1 \exp\left(-\frac{1}{2}x^2\right) \end{aligned}$$

idet det sidste integral bliver en konstant, der er uafhængig af x .

Derfor finder vi, at den marginale fordeling for X må have frekvensfunktionen

$$h(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right),$$

Denne fordeling vil vi senere kalde en normalfordeling med middelværdi 0 og varians 1.

Den betingede fordeling af Y for givet X bliver

$$\begin{aligned} g(y|x) &= \frac{f(x, y)}{h(x)} \\ &= d \cdot \exp\left(-\frac{1}{2(1-\rho^2)}(y-\rho x)^2\right) \end{aligned}$$

Dette svarer til en normalfordeling med middelværdi

$$E(Y|X = x) = \rho x$$

og varians

$$V(Y|X = x) = 1 - \rho^2$$

Havde vi i stedet betragtet X og Y med samme korrelation ρ , men med $E(X) = \mu_x$ og $V(X) = \sigma_x^2$, og $E(Y) = \mu_y$ og $V(Y) = \sigma_y^2$ havde vi fået

$$E(Y|X = x) = \mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x) \quad (0.51)$$

$$V(Y|X = x) = \sigma_y^2 (1 - \rho^2) \quad (0.52)$$

For hvert udfald af X har vi en værdi af $E(Y|X = x)$. Det vil derfor være naturligt at skrive

$$E(Y|X) = \mu_y + \rho \frac{\sigma_y}{\sigma_x} (X - \mu_x) \quad (0.53)$$

Vi finder da

$$\begin{aligned} E(E(Y|X)) &= \mu_y + \rho \frac{\sigma_y}{\sigma_x} (E(X) - \mu_x) \\ &= \mu_y \\ &= E(Y) \end{aligned}$$

Tilsvarende bliver

$$\begin{aligned} V(E(Y|X)) &= \rho^2 \frac{\sigma_y^2}{\sigma_x^2} V(X) \\ &= \rho^2 \sigma_y^2 \end{aligned}$$

Da

$$E(V(Y|X)) = \sigma_y^2 (1 - \rho^2)$$

genfinder vi relationen

$$V(Y) = E(V(Y|X)) + V(E(Y|X))$$



Som et mål for **skævheden** af en fordeling anvendes ofte størrelsen (med $E(X) = \mu$)

$$\gamma_1 = \frac{E((X - \mu)^3)}{(V(X))^{3/2}},$$

d.v.s. det tredje centrale moment divideret med spredningen i tredje. Det må straks bemærkes, at man ofte ser γ_1^2 anført som et skævhedsmål (og da ofte under betegnelsen β_1). Hvis $\gamma_1 > 0$, er den højre hale i fordelingen større end den venstre. Omvendt, hvis $\gamma_1 < 0$. Hvis fordelingen er symmetrisk, er $\gamma_1 = 0$.

Som et mål for **topstejlheden (kurtosis)** af en fordeling anvendes

$$\gamma_2 = \frac{E((X - \mu)^4)}{(V(X))^2} - 3,$$

d.v.s. det fjerde centrale moment divideret med variansens kvadrat minus 3. Grunden til forekomsten af leddet -3 er, at γ_2 derved bliver 0 for den såkaldte normale fordeling. Størrelsen γ_2 bliver dermed et mål relativt til den normale fordeling.

Hvis γ_2 er positiv, har fordelingen relativt tykke haler og dermed en spidsere top end den normale fordeling. Vi taler om en **leptokurtisk** fordeling. Hvis γ_2 er negativ, har fordelingen tyndere haler og en fladere top end den normale fordeling. Vi taler om en **platykurtisk** fordeling. Hvis $\gamma_2 = 0$, siges fordelingen at være **mesokurtisk**.

Vi slutter med at nævne et sidste mål for en fordelings beliggenhed, nemlig **modus**. Modus er defineret som et punkt, hvor frekvensfunktionen har et lokalt maksimum. Der er altså tale om "mest sandsynlige" værdier.

Hvis en fordeling har et modus, siges den at være **unimodal**. Hvis der er to, kaldes den **bimodal** etc.

DEFINITION 0.9 (KARAKTERISTISK FUNKTION). Den karakteristiske funktion for en stokastisk variabel X er defineret ved

$$\phi(t) = E(e^{itX}) \quad (0.54)$$

Ved trivial anvendelse af sætning 0.9 fås

$$\phi(t) = \begin{cases} \sum e^{itx} f(x) & X \text{ diskret} \\ \int_{-\infty}^{\infty} e^{itx} f(x) dx & X \text{ kontinuert} \end{cases} \quad (0.55)$$

▲

Den karakteristiske funktion kan bl.a. benyttes til at bestemme fordelingen af transformationer af stokastiske variable og til at udregne momenter for stokastiske variable.

Til enhver bestemt sandsynlighedsfordeling svarer én og kun én karakteristisk funktion og omvendt. Man kan altså med kendskab til den karakteristiske funktion afgøre, hvilken fordeling en given stokastisk variabel følger. Dette udnyttes flere steder i nærværende note.

Endvidere gælder følgende meget vigtige sætninger:

SÆTNING 0.10. Lad den til den stokastiske variabel X svarende karakteristiske funktion være $\phi(t)$. Da er X 's middelværdi og varians henholdsvis

$$E(X) = -i\phi'(0) \quad (0.56)$$

$$V(X) = -\phi''(0) - (E(X))^2 \quad (0.57)$$

Bevis. Trivielt, ses ved at finde $\phi'(0)$ henholdsvis $\phi''(0)$ ud fra definitionen på $\phi(t)$. ■

SÆTNING 0.11. Lad X og Y være uafhængige stokastiske variable med karakteristiske funktioner $\phi_x(t)$ og $\phi_y(t)$ henholdsvis. Da er den karakteristiske funktion for $Z = X + Y$ givet ved:

$$\phi_z(t) = \phi_x(t)\phi_y(t) \quad (0.58)$$

Bevis. Forbigås. Kan vises ved direkte regning. ■

0.10 Approksimative formler for middelværdi og varians

Ofte er det vanskeligt at finde et eksakt udtryk (f. eks. ved hjælp af sætning 0.9) for forventningsværdien $E(h(X))$ af en funktion h af en stokastisk variabel. Hvis h er lineær (affin), kan man dog ifølge resultaterne i forrige afsnit uden videre udtrykke $E(h(X))$ og $V(h(X))$ ved $E(X)$ og $V(X)$. I det generelle tilfælde vil det derfor være nærliggende at forsøge at approksimere h med en funktion, der er affin i det område, hvor X varierer, d.v.s. vi vil foretage en Taylor-approksimation omkring $\mu = E(X)$. Vi får da

$$Y = h(X) \simeq h(\mu) + (X - \mu)h'(\mu),$$

og dette giver de approksimative formler

$$\begin{aligned} E(Y) &= E(h(X)) \simeq h(\mu) \\ V(Y) &= V(h(X)) \simeq (h'(\mu))^2 V(X). \end{aligned}$$

Nøjagtigheden af disse formler er selvsagt størst for stokastiske variable med et meget snævert variationsområde (det vil ca. sige med en lille varians).

Vi kan nu finde approksimative udtryk for variansen af f. eks. $\log_e X$ og $\sqrt{X}$. Vi finder

$$V(\log_e X) \simeq \frac{V(X)}{(E(X))^2}$$

d.v.s.

$$\sqrt{V(\log_e X)} \simeq \frac{\sqrt{V(X)}}{E(X)},$$

hvilket netop er varianskoefficienten for X .

Helt tilsvarende bliver

$$V(\sqrt{X}) \simeq \frac{\frac{1}{4}V(X)}{E(X)}.$$

I mange situationer vil man betragte en variabel Y , der er en funktion af flere stokastiske variable $X_1, \dots, X_n$, i. e.

$$Y = h(X_1, \dots, X_n).$$

Vi foretager igen en Taylor-approksimation omkring forventningsværdierne, i.e.

$$\begin{aligned} Y &= h(X_1, \dots, X_n) \\ &\simeq h(\mu_1, \dots, \mu_n) + (X_1 - \mu_1) \frac{\partial h}{\partial x_1} + \dots + (X_n - \mu_n) \frac{\partial h}{\partial x_n}, \end{aligned}$$

hvor de partielle afledede skal tages i punktet $(\mu_1, \dots, \mu_n)$. Vi må igen antage, at X 'ernes variationsområde ligger så snævert omkring forventningsværdierne, at Taylor-approksimationen er anvendelig. Vi finder da

$$\begin{aligned}
E(Y) &\simeq h(\mu_1, \dots, \mu_n) \\
V(Y) &\simeq \left(\frac{\partial h}{\partial x_1}\right)^2 V(X_1) + \dots + \left(\frac{\partial h}{\partial x_n}\right)^2 V(X_n) \\
&\quad + 2\frac{\partial h}{\partial x_1}\frac{\partial h}{\partial x_2}\text{Cov}(X_1, X_2) \\
&\quad + \dots + \frac{\partial h}{\partial x_{n-1}}\frac{\partial h}{\partial x_n}\text{Cov}(X_{n-1}, X_n).
\end{aligned}$$

Hvis X_i 'erne er uafhængige (eller blot ukorrelerede), fås specielt

$$V(Y) \simeq \left(\frac{\partial h}{\partial x_1}\right)^2 V(X_1) + \dots + \left(\frac{\partial h}{\partial x_n}\right)^2 V(X_n).$$

Denne relation kaldes ofte **fejlophobningsloven**.

Vi skal nu give et lille eksempel på anvendelse af fejlophobningsloven.

EKSEMPEL 0.20. Vi vil finde et approksimativt udtryk for

$$V(Y) = V(X_1 \cdot X_2),$$

hvor X_1 og X_2 er uafhængige stokastiske variable. Vi har Taylor-approksimationen

$$h(x_1, x_2) = x_1 x_2 = \mu_1 \mu_2 + (x_1 - \mu_1) \frac{\partial h}{\partial x_1} + (x_2 - \mu_2) \frac{\partial h}{\partial x_2}.$$

Da

$$\left. \frac{\partial h}{\partial x_1} \right|_{\substack{x_1 = \mu_1 \\ x_2 = \mu_2}} = \mu_2$$

$$\left. \frac{\partial h}{\partial x_2} \right|_{\substack{x_1 = \mu_1 \\ x_2 = \mu_2}} = \mu_1$$

fås

$$h(x_1, x_2) \simeq \mu_1 \mu_2 + \mu_2 (x_1 - \mu_1) + \mu_1 (x_2 - \mu_2).$$

Dette giver udtrykket

$$V(Y) = V(X_1 \cdot X_2) \simeq \mu_2^2 V(X_1) + \mu_1^2 V(X_2).$$



0.11 Konvergens

SÆTNING 0.12 (CHEBYCHEVS ULIGHED). Lad X være en stokastisk variabel med middelværdi μ og varians σ^2 . Da er

$$P\{|X - \mu| > \epsilon\} \leq \frac{\sigma^2}{\epsilon^2} \quad (0.59)$$

Specielt haves

$$P\{|X - \mu| > k\sigma\} \leq \frac{1}{k^2} \quad (0.60)$$

Bevis. Vi betragter indikatorfunktionen

$$I_{\{|X - \mu| > \epsilon\}} = \begin{cases} 0 & \text{hvis } |X - \mu| \leq \epsilon \\ 1 & \text{hvis } |X - \mu| > \epsilon \end{cases}$$

Der gælder trivielt, at

$$\begin{aligned} (X - \mu)^2 &\geq (X - \mu)^2 I_{\{|X - \mu| > \epsilon\}} \\ &\geq \epsilon^2 I_{\{|X - \mu| > \epsilon\}} \end{aligned}$$

Derfor er

$$E((X - \mu)^2) \geq \epsilon^2 E(I_{\{|X - \mu| > \epsilon\}})$$

eller

$$\sigma^2 \geq \epsilon^2 P\{|X - \mu| > \epsilon\}$$

da forventningsværdien af en vilkårlig indikatorfunktion I_A er sandsynligheden $P(A)$ for hændelsen A . ■

DEFINITION 0.10 (SVAG KONVERGENS). Vi betragter en følge af kumulerede fordelingsfunktioner F_n , $n = 1, 2, \dots$. Lad endvidere F være en fordelingsfunktion. Vi siger da, at følgen F_n **konvergerer svagt mod** F , hvis

$$F_n(x) \rightarrow F(x) \quad (0.61)$$

for alle de x , hvor $F(x)$ er kontinuert. Vi skriver $F_n \xrightarrow{w} F$, hvor w kommer fra det engelske weakly. ▲

DEFINITION 0.11 (KONVERGENS I FORDELING). Hvis $X_1, X_2, \dots$ er en følge af stokastiske variable, hvis fordelingsfunktioner $F_1, F_2, \dots$ konvergerer svagt mod F siger vi, at $X_1, X_2, \dots$ **konvergerer i fordeling** med F som grænsefordeling. ▲

EKSEMPEL 0.21. Lad $X_1, X_2, \dots$ være en følge af stokastiske variable med

$$P\{X_n = x\} = \binom{n}{x} p_n^x (1 - p_n)^{n-x} = f_n(x)$$

for $x = 0, 1, \dots, n$.

Lad $p_n \rightarrow 0$ for $n \rightarrow \infty$, så

$$np_n \rightarrow \lambda$$

Da vil som bekendt

$$f_n(x) \rightarrow \frac{\lambda^x}{x!} e^{-\lambda}$$

Derfor vil

$$P\{X_n \leq z\} = \sum_{i=0}^{[z]} f_n(i),$$

hvor $[z]$ er heltalsdelen af z . Da alle $f_n(i) \rightarrow f(i)$, og da der er et fast antal, nemlig $[z] + 1$ led i summen, vil

$$P\{X_n \leq z\} \rightarrow F(z)$$

hvor

$$F(z) = \sum_{i=0}^{[z]} f_n(i)$$

Med andre ord konvergerer X_n i fordeling med F som grænsefordelingsfunktion. $\blacklozenge$

DEFINITION 0.12 (KONVERGENS I SANDSYNLIGHED). Lad $X_1, X_2, \dots$ være en følge af stokastiske variable og lad X være en stokastisk variabel, hvorom der gælder

$$P\{|X_n - X| > \epsilon\} \rightarrow 0 \quad n \rightarrow \infty \quad (0.62)$$

Vi siger da, at X_n **konvergerer i sandsynlighed** mod X , og vi skriver

$$X_n \xrightarrow{P} X$$

$\blacktriangle$

Konvergens i sandsynlighed er et stærkere begreb end konvergens i fordeling, idet vi har

SÆTNING 0.13. Hvis X_n konvergerer i sandsynlighed mod X , og hvis

$$\begin{aligned} F_n(x) &= P\{X_n \leq x\} \\ F(x) &= P\{X \leq x\} \end{aligned}$$

er de respektive kumulerede fordelingsfunktioner, vil F_n konvergere i fordeling mod F , dvs

$$F_n \xrightarrow{w} F$$

Bevis. Forbigås. $\blacksquare$

Vi skal nu formulere den sætning, der går under navnet

SÆTNING 0.14 (DE STORE TALS LOV). Lad $X_1, X_2, \dots$ være en følge af uafhængige, identisk fordelte stokastiske variable med forventningsværdi μ .

Da vil

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

konvergere i sandsynlighed mod μ for $n \rightarrow \infty$.

Bevis. Vi nøjes med at se på det tilfælde, hvor X_i 'erne også har en varians. I dette tilfælde finder vi

$$\begin{aligned} E(\bar{X}_n) &= \mu \\ V(\bar{X}_n) &= \sigma^2/n \end{aligned}$$

og Chebychevs ulighed giver umiddelbart

$$P\{|\bar{X}_n - \mu| > \epsilon\} \leq \frac{\sigma^2}{n\epsilon^2} \rightarrow 0, \quad n \rightarrow \infty$$

dvs $\bar{X}_n \xrightarrow{P} \mu$. ■

EKSEMPEL 0.22. Hvis X_i fx. viser udfaldet i et møntkast ($1 \simeq$ krone, $0 \simeq$ plat) er $E(X_i) = p$, sandsynligheden for kr., og vi finder, at det gennemsnitlige antal kr. konvergerer i sandsynlighed mod p . ◆

0.12 Notation

Vi nævner til sidst, at hvis en stokastisk variabel følger en vilkårlig fordeling, som vi kalder ABC-fordelingen, vil vi kort udtrykke dette ved at skrive $X \in \text{ABC}$, d.v.s.

$$X \in \text{ABC} \stackrel{\text{d}}{\Leftrightarrow} X \text{ er ABC-fordelt.}$$

Hvis X er ABC-fordelt, i.e. $X \in \text{ABC}$, vil vi endvidere ofte skrive

$$P\{\text{ABC} \leq x\} \quad \text{i stedet for } P\{X \leq x\}.$$

Hvis k er et reelt tal, vil vi – stadig for $x \in ABC$ – skrive, at

$$k \cdot X \in k \text{ } ABC,$$

d.v.s. vi siger, at en konstant k gange en ABC -fordelt stokastisk variabel er $k \text{ } ABC$ -fordelt. Endelig vil det være hensigtsmæssigt at benævne **p-fraktilerne** i en ABC -fordeling med

$$ABC_p.$$

Disse konventioner der bl. a. går ud på, at vi anvender betegnelser som ABC både som navn for en fordeling og som betegnelse for en stokastisk variabel, der følger den pågældende fordeling, kan nok virke lidt forvirrende her, men det viser sig senere at være meget tidsbesparende vedtagelser.

Undertiden er det også hensigtsmæssigt at have en kort betegnelse for fordelingen af en stokastisk variabel X . Vi vil anvende symbolet $\mathbb{L}(X)$ ($\mathbb{L}$ for law).

For at lette senere referencer starter vi med at anføre nogle resultater om de fra analysen kendte Γ - og β -funktioner (gamma- og beta-funktioner). Vi vil endvidere give nogle enkelte nye og repetere andre definitioner og sætninger om stokastiske variables fordelinger.

Vi får kun brug for at kende disse funktioner for reelle argumenter. En mere fuldstændig fremstilling af teorien kan findes i [1] eller [5].

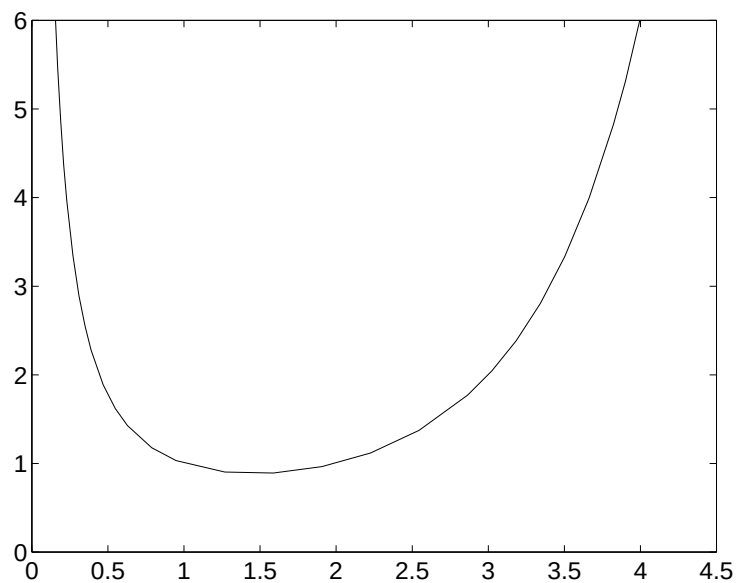
0.12.1 Γ -funktionen

DEFINITION 0.13. For reelle, positive argumenter defineres Γ -funktionen ved Eulers integral

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt, \quad x > 0 \quad (0.63)$$

▲

For positivt argument er grafen som vist på nedenstående figur.



Vi opremsrer nogle nyttige egenskaber ved Γ -funktionen i

SÆTNING 0.15. Der gælder, at

$$x\Gamma(x) = \Gamma(x+1) \quad (0.64)$$

$$\Gamma(1) = 1 \quad \wedge \quad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \quad (0.65)$$

Specielt haves for naturlige tal n

$$\Gamma(n) = 1 \cdot 2 \cdot 3 \cdots (n-1) = (n-1)! \quad (0.66)$$

$$\Gamma\left(n + \frac{1}{2}\right) = \frac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{2^n} \sqrt{\pi} \quad (0.67)$$

Bevis. Relationen (0.64) følger umiddelbart ved partiel integration i definitions ligningen (0.63),

$$\begin{aligned} \int_0^\infty t^x e^{-t} dt &= \left[-t^x e^{-t} \right]_0^\infty + x \int_0^\infty t^{x-1} e^{-t} dt \\ &= x \int_0^\infty t^{x-1} e^{-t} dt, \end{aligned}$$

hvilket er ensbetydende med (0.64).

Den første relation i (0.65) er let, hvorimod den anden kræver lidt længere regninger, som vi udelader (se f. eks [5][p. 169]). Relationerne (0.66) og (0.67) følger let ved rekursiv anvendelse af (0.64). Vi viser e.g. (0.67)

$$\begin{aligned}\Gamma\left(n + \frac{1}{2}\right) &= \left(n - \frac{1}{2}\right) \Gamma\left(n - \frac{1}{2}\right) \\ &= \cdots = \left(n - \frac{1}{2}\right) \left(n - \frac{2}{3}\right) \cdots \frac{3}{2} \cdot \frac{1}{2} \Gamma\left(\frac{1}{2}\right)\end{aligned}$$

■

Vi anfører endvidere en yderst nyttig approksimation til Γ -funktionen, nemlig

SÆTNING 0.16 (STIRLINGS FORMEL). For $x \rightarrow \infty$ er

$$\Gamma(x) \approx \sqrt{2\pi} x^{x-\frac{1}{2}} e^{-x} \quad (0.68)$$

Bevis. Forbigås (se [5][p. 180]).

■

0.13 Fortsættelse af tidligere eksempler

EKSEMPEL 0.23. Vi betragter de tidligere givne data (Eksempel 0.2, side 13) over indhold af bioaktivt-organisk materiale. Det fremgår, at miljømyndighedernes kontrolværdi er 1, og vi ønsker at undersøge, om data lever op til dette krav.

En måde at gøre dette på ville være, at bestemme et passende konfidensinterval for fx. middelværdien og dernæst se, om dette indeholder værdier ≥ 1 . Hvis dette er tilfældet kan man sige, at miljømyndighedernes krav ikke er overholdt.

Hyppigt anvende værdier for konfidensgraden er i sådanne tilfælde 70% og 95%.

Det fremgår af histogrammet (Figur 0.2), at fordelingen er skæv, og at en normalfordelingsantagelse ikke vil være rimelig. Som et bedre alternativ ligger den logaritmiske normalfordeling lige for. Vi har angivet fordelingen af $\log_e$ (kontrolvariablen) i Tabel 0.8. I Figur 0.8 er histogrammet sammenlignet med den normale fordeling med parametre $(-0.24, 0.50^2)$ svarende til værdierne

$$\begin{aligned}\bar{z} &= -0.24 \\ s_z &= 0.50\end{aligned}$$

for de transformerede data $Z_i = \log_e X_i$. Det ses, at tilpasningen ikke er helt perfekt. Den estimerede frekvensfunktion ligger 'for langt' til højre. Dette skyldes, at der er nogle store positive, ekstreme observationer. Ser vi på intervallet $[\bar{Z} - s_z, \bar{Z} + s_z]$, ligger der 1 observation under nederste intervalgrænse, men 6 observationer over øverste grænse. Den datagenererende proces giver altså lidt hyppigere meget store værdier end meget små værdier. Genberegnes gennemsnit for de 196 observationer, der ligger i ovennævnte interval fås værdien -0.28, og det ses, at normalfordelingen med denne middelværdi passer 'bedre' til data.



Figur 0.8: Fordeling af $\log_e$ (kontrolvariabel) samt 203×0.2 frekvensfunktionen for en normalfordeling med middelværdi -0.25 og spredning 0.50.

Da udelukkelsen af de ekstreme observationer vil være i virksomhedens favør og recipientens disfavør vil vi fortsætte beregningerne på det fuldstændige materiale. Vi finder

$$\begin{aligned} t(202)_{85\%} &\simeq 1.04 \\ t(202)_{97.5\%} &\simeq 1.97 \end{aligned}$$

hvorfor 70% og 95% konfidensintervallerne er givet ved

$$\begin{aligned} \left[-0.24 - 1.04 \frac{0.50}{\sqrt{203}}, -0.24 + 1.04 \frac{0.50}{\sqrt{203}} \right] &= [-0.28, -0.20] \\ \left[-0.24 - 1.97 \frac{0.50}{\sqrt{203}}, -0.24 + 1.97 \frac{0.50}{\sqrt{203}} \right] &= [-0.31, -0.17] \end{aligned}$$

Det fremgår, at begge intervaller ligger langt under 0, og det må derfor være rimeligt at fastslå, at processens middelniveau $\mu_{\log}$ (for de logaritmerede data) overholder kravspecifikationen $\mu_{\log} \leq 0$.

Konfidensintervallet for fordelings geometriske middelværdi

$$\mu_g = e^{\mu_{\log}}$$

bliver uden videre (konfidensgrad 95%)

$$[e^{-0.31}, e^{-0.17}] = [0.73, 0.84],$$

idet eksponentialfunktionen er voksende, og derfor er

$$P\{U \leq \mu \leq V\} = P\{e^U \leq e^\mu \leq e^V\}$$

Følgelig kan et konfidensinterval for den transformerede parameter fås ved at transformere intervalgrænserne.

Den geometriske middelværdi er angivet i den oprindelige måleskala, og vi kan derfor sammenligne kravværdien fra den oprindelige skala, som var 1, med ovenstående interval og igen konstatere, at intervallet ligger til venstre for kravværdien.

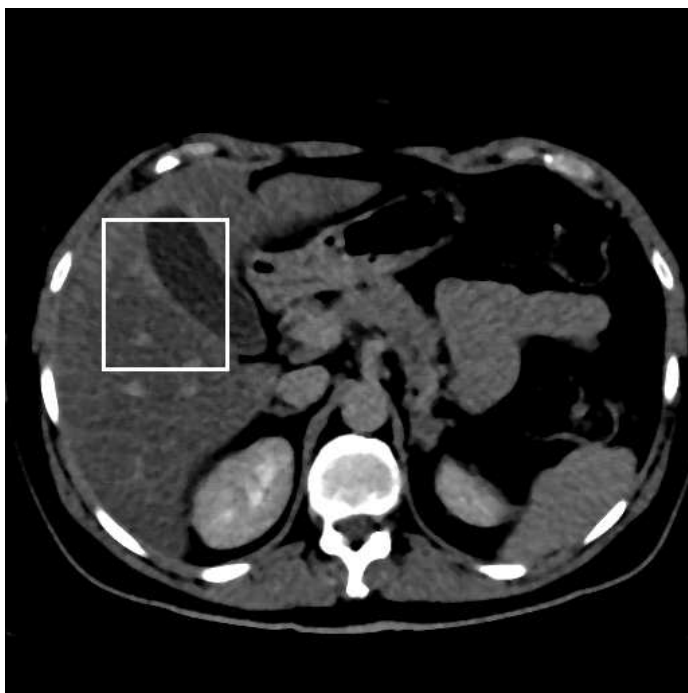
Det er ikke helt så enkelt at finde et konfidensinterval for middelværdien

$$E(X) = e^{\mu_{\log} + \frac{1}{2}\sigma_{\log}^2}$$

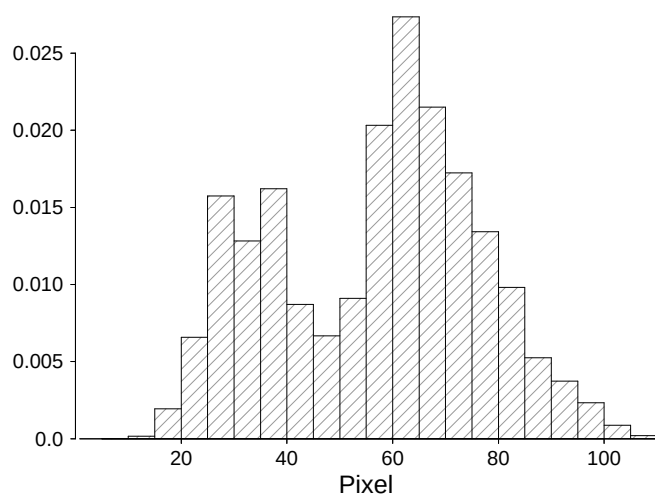
Maximum likelihood estimatet for denne er

$$\exp(-0.23 + \frac{1}{2}0.50^2) = 0.90$$

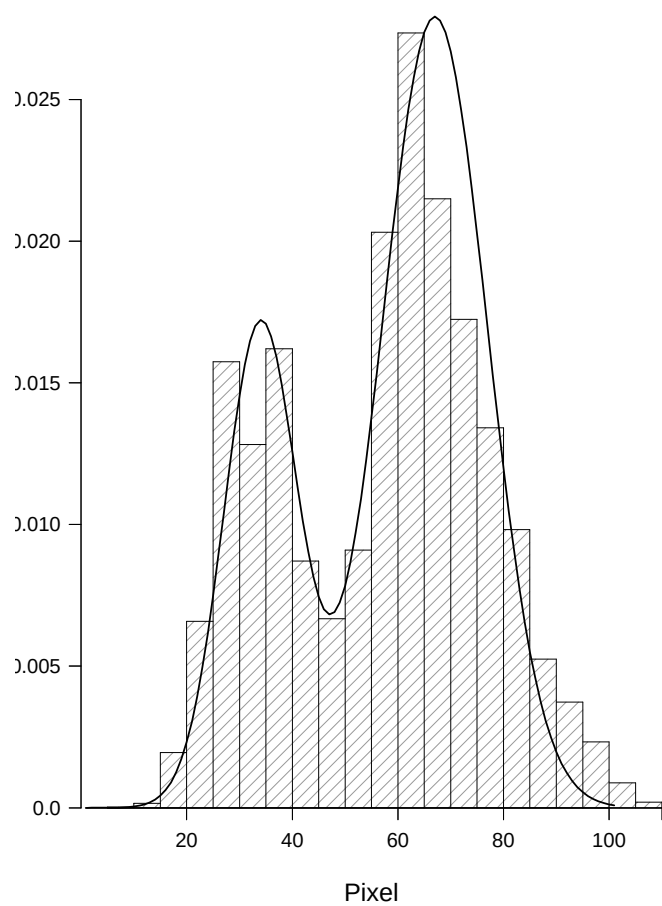
men da $\sigma_{\log}^2$ også er ukendt kan vi ikke blot gå frem som før. Vi skal ikke komme nærmere ind på disse forhold, men blot konstatere, at det faktum, at den statistiske analyse er enkel for de logaritmerede data (under forudsætning af, at observationerne er logaritmisk normalfordelte) er en primær begrundelse for at vælge at betragte de logaritmerede data i stedet for de oprindelige. ♦



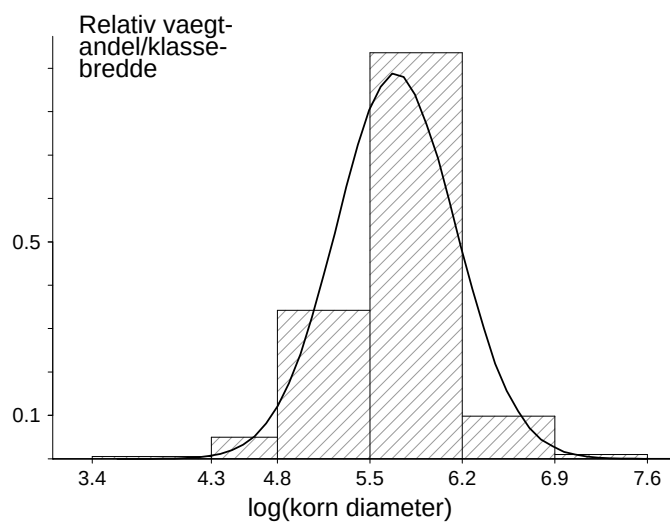
Figur 0.9: Magnetresonansbillede (MR-billede) af et snit gennem en menneskehjerne.



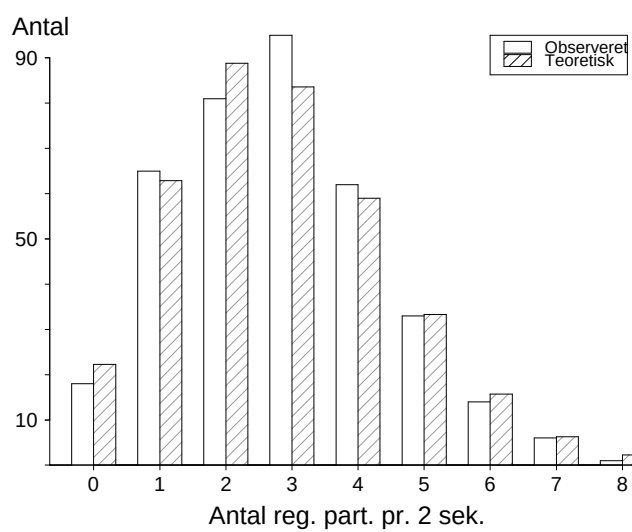
Figur 0.10: Histogram over pixelværdierne for området i det hvide rektangel i figur 0.9.



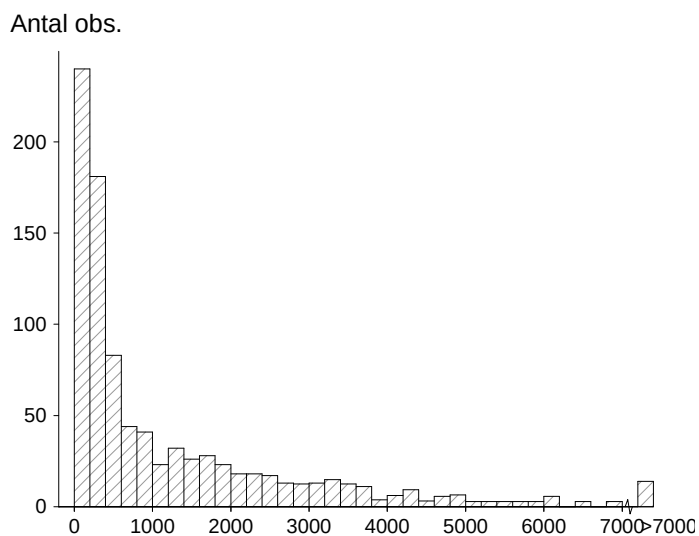
Figur 0.11: Compound-fordeling bestående af 2 normale fordelinger fittet til histogrammet i figur 0.10.



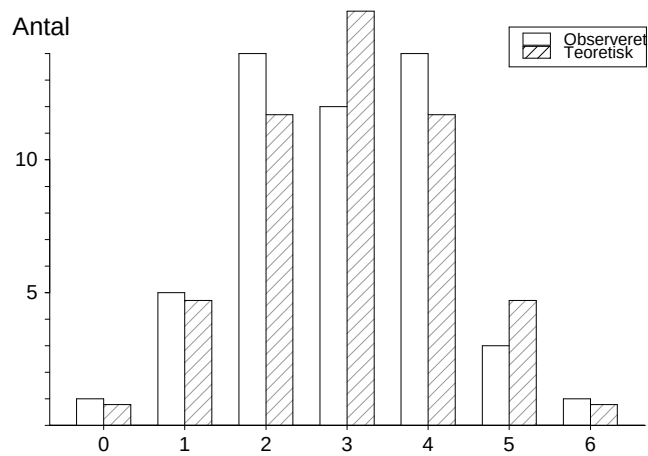
Figur 0.12: En normal fordeling fittet til log(korn-diameter) fordelingen fra figur 0.6.



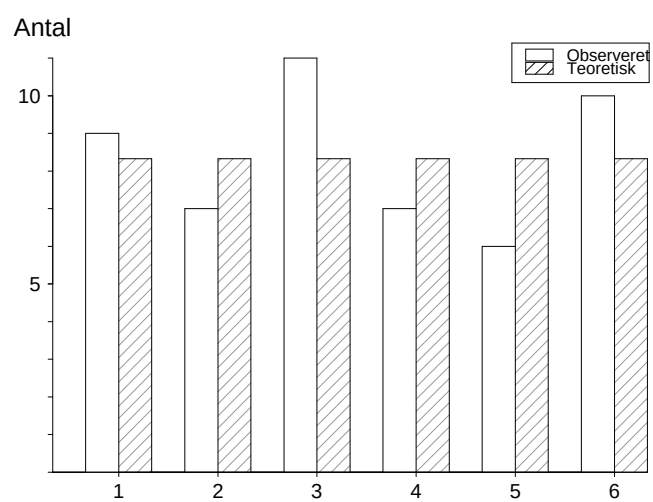
Figur 0.13: En Poisson-fordeling fittet til data fra figur 0.4.



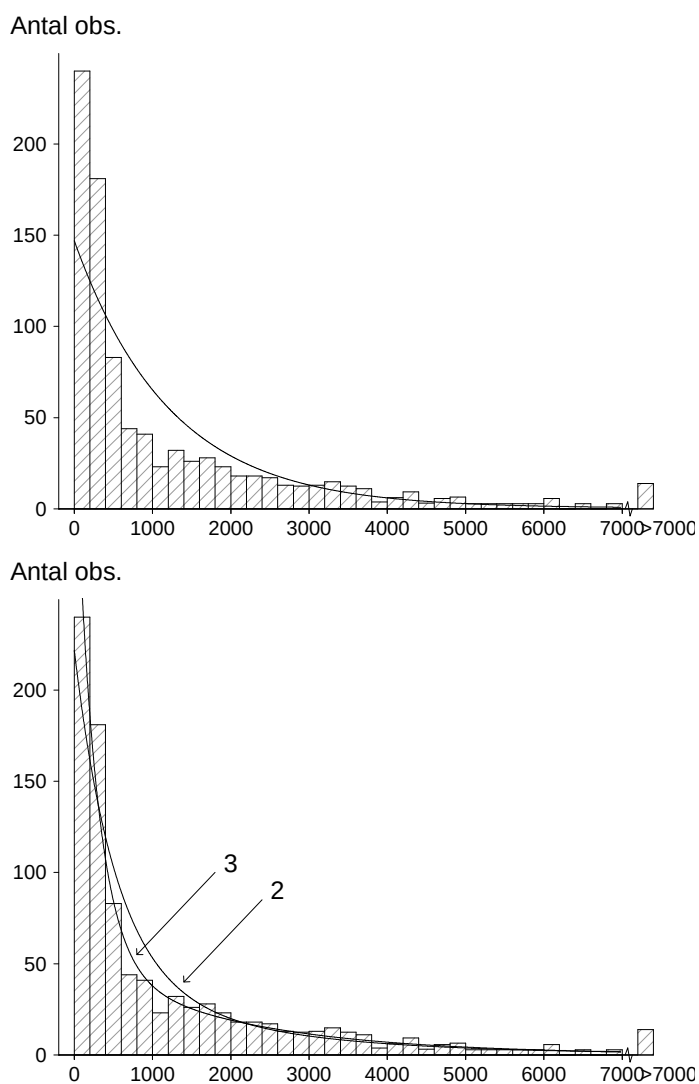
Figur 0.14: Fordelingen af 903 samtaletider fra Ordrup central.



Figur 0.15: $B(6, \frac{1}{2})$ -fordeling fittet til data fra figur 0.7.



Figur 0.16: Ligefordelingen på tallene fittet til data fra figur 0.7.



Figur 0.17: En tilpasning med en eksponentialfordeling (øverst) og med compoundfordelinger (2 eksponentialfordelinger). "2" er estimeret med momentmetoden og "3" med maksimum-likelihood metoden.

Kapitel 1

Sandsynlighedsteoretiske modeller

I dette kapitel vil vi give en oversigt over de vigtigste af de fordelinger fra sandsynlighedsregningen, vi kommer til at beskæftige os med i statistikken.

Der er tale om visse repetitioner fra sandsynlighedsregningen. Disses berettigelse er først og fremmest, at der derved vil skabes et (forhåbentlig) anvendeligt og overskueligt overblik over de i bogen anvendte fordelinger.

Efter en summarisk opremsning af egenskaber ved de betragtede fordelinger er der givet nogle eksempler på, under hvilke forhold man kan forvente, at de respektive fordelinger optræder i ingeniørvidenskaberne. Disse eksempler er vigtige, idet man ved anvendelser af statistikken altid må foretrække modeller, der bygger på generelle love vedrørende den proces, som frembringer data. Hvis man nemlig blot empirisk har fastlagt en model, vil man formentlig nok kunne **interpolere** sig frem til anvendelige resultater, hvorimod man bør være noget varsom med at anvende resultater, der er fremkommet ved **ekstrapolation** i en sådan empirisk model.

1.1 Lidt om stokastiske modellers verifikation

Blandt statistikere (og erkendelsesteoretikere) foregår en livfuld diskussion om, hvilke af en række sandsynlighedsbegreber (logiske, subjektive, frekventielle o.a.), der skal danne grundlaget for den praktiske anvendelse af stokastiske modeller. Det er ikke meningen her at referere, endsige bidrage til denne diskussion. En rimelig tilgængelig

oversigt findes i [2]. Se også [26]. Vi vil her gå ud fra eksistensen af - og, at læseren er indforstået med - begrebet **sandsynligheden** for givne udfald af visse såkaldte **tilfældige** fysiske eksperimenter, og at vi kan udnytte **stokastiske variable** til at angive **målinger** af disse udfald.

I dette kapitel skal vi som nævnt beskæftige os dels med de analytiske egenskaber ved sandsynlighedsfordelinger og dels med disse fordelingers genesis (oprindelse).

For at lette læsningen vil det være af værdi, at der knyttes konkrete eksempler til fremstillingen. Vi vil derfor først give en beskrivelse af de matematiske resultater, som gør det muligt at vurdere en models rimelighed ud fra en række **gentagne** observationer af et fænomen. En nærmere diskussion af de problemer, der er forbundet med vendingen "en gentagelse af et eksperiment" gives i indledningen til kapitel 2. I nærværende afsnit kan vi nøjes med det almindeligt accepterede indhold, ordet umiddelbart tillægges.

Vi omtaler først begrebet **indikatorfunktionen** I_A for en vilkårlig mængde A . Den er defineret ved

$$I_A(x) = \begin{cases} 1, & \text{hvis } x \in A \\ 0, & \text{hvis } x \notin A \end{cases}.$$

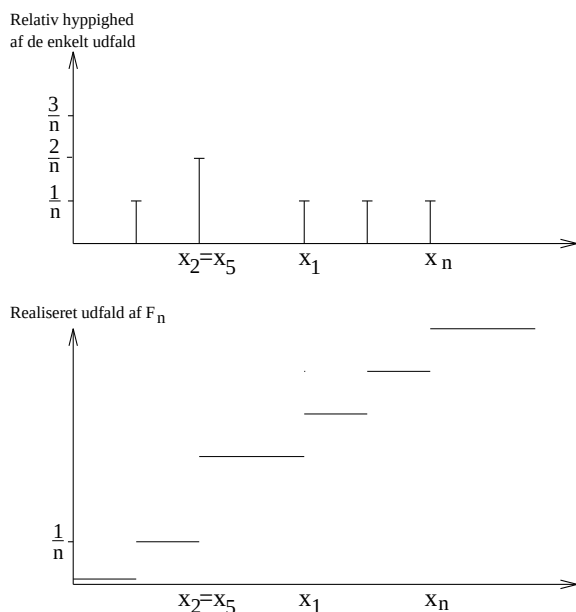
Vi er nu i stand til at indføre et meget vigtigt begreb, nemlig den empiriske fordelingsfunktion for en række stokastiske variable $X_1, \dots, X_n$. Vi har

DEFINITION 1.1 (DEN EMPIRISKE FORDELINGSFUNKTION). F_n for $X_1, \dots, X_n$ er givet ved

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{\{X_i \leq x\}},$$

hvor I betegner indikatorfunktionen. Verbalt kan vi udtrykke det, at den empiriske fordelingsfunktion er den relative hyppighed af de X_i 'er, der er mindre end eller lig med x . ▲

BEMÆRKNING 1.1. Vi ser, at den empiriske fordelingsfunktion taget i et vilkårligt punkt er en stokastisk variabel. Har vi givet realiserede udfald $x_1, \dots, x_n$, da er det realiserede udfald af $F_n(x)$ lig værdien af fordelingsfunktionen for ligefordelingen over $x_1, \dots, x_n$ (den fordeling, der tilordner hvert af punkterne $x_1, \dots, x_n$ sandsynligheden $\frac{1}{n}$). Dette er skitseret i nedenstående graf. ▼



Hvis $X_1, \dots, X_n$ er uafhængige og identisk fordelte med fordelingsfunktionen F , kunne man håbe, at et realiseret udfald af F_n kan anvendes som skøn over F . Det forholder sig virkelig således, idet der gælder følgende meget vigtige sætning.

SÆTNING 1.1 (DEN MATEMATISKE STATISTIKS HOVEDSÆTNING). Lad $X_1, \dots, X_n$, være indbyrdes uafhængige og identisk fordelte med fordelingsfunktionen F . Da er sandsynligheden for, at den empiriske fordelingsfunktion F_n konvergerer ligeligt mod den teoretiske fordelingsfunktion F , lig 1. Udtrykt formelt har vi

$$P \left\{ \lim_{n \rightarrow \infty} \left(\sup_{-\infty < x < \infty} |F_n(x) - F(x)| \right) = 0 \right\} = 1.$$

Bevis. Forbigås. Se f.eks. [47, p. 335]. ■

På basis af denne sætning kan man ud fra gentagne observationer af et stokastisk fænomen vurdere, om en postuleret fordeling er rimelig ved at sammenligne det realiserede udfald af den empiriske fordelingsfunktion med den postulerede fordelingsfunktion.

Hvis afvigelserne mellem de to funktioner skønnes store, har man en indikation af, at modellen ikke er fyldestgørende. En nærmere præcisering af, hvad der menes med "store", vil blive gennemgået i afsnittet om testteori og uddybet i kapitel 4.

I stedet for at sammenligne fordelingsfunktionerne vil man ofte se på frekvensfunk-

tioner. Hvis der er tale om diskrete fordelinger, er dette ganske ligefremt. Hvis der derimod foreligger observationer fra en kontinuert fordeling, vil den realiserede empiriske fordelingsfunktion ikke svare til en kontinuert fordeling, men til en diskret fordeling med støtte i observationsværdierne.

Dette problem omgår vi ved at foretage en såkaldt **gruppering** af materialet, d.v.s. vi inddeler et interval, der indeholder alle observationer i et antal delintervaller, såkaldte **klasser**. Vi tænker os nu, at alle observationer, der ligger i et delinterval, er jævnt fordelt over intervallet, og inddelingen giver således anledning til

DEFINITION 1.2 (HISTOGRAM). Den tæthedsfunktion, som er konstant i ethvert delinterval, og hvis integral over ethvert interval er proportionalt med antallet af observationer i intervallet, kaldes det til inddelingen svarende **histogram**. ▲

DEFINITION 1.3 (SUMPOLYGON). Den til et histogram svarende fordelingsfunktion kaldes en **sumpolygon**. ▲

Problemer vedrørende valg af delintervaller (antal, størrelse, etc.) behandles i kapitel 4.

Vi illustrerer nu begreberne i følgende eksempel.

EKSEMPEL 1.1. I nedenstående tabel er anført målingen af hovedets diameter på 36 nittehoveder. (Uddrag af større datamateriale i [25]).

Diameter i mm					
13.62	13.26	13.24	13.39	13.51	13.52
13.44	13.33	13.25	13.26	13.13	13.56
13.20	13.35	13.28	13.55	13.47	13.38
13.29	13.44	13.49	13.54	13.40	13.20
13.26	13.34	13.33	13.37	13.56	13.37
13.63	13.33	13.26	13.31	13.35	13.33

Vi vil nu konstruere et histogram og en sumpolygon svarende til disse målinger. Vi har, at den mindste og den største observation er 13.13, henholdsvis 13.63.

Vi vælger derfor at inddele intervallet $[13.105, 13.705]$, der omslutter alle observationer, i 6 lige store dele. Vi bemærker, at **intervalendepunkterne har en decimal mere end målingerne** på nittehovedet. Dette sikrer, at der ikke opstår tvivl om, hvilket interval en måling skal placeres i. Vi bestemmer nu antallet af observationer i hver klasse og får følgende skema. I skemaet er endvidere anført de **kumulerede** antal, d.v.s. de successive summer af antallene i de enkelte klasser.

Klasse	Antal	Kumuleret antal	Kum. ant./36
13.105–13.205	111 = 3	3	0.08
13.205–13.305	11111 111 = 8	11	0.31
13.305–13.405	11111 11111 111=13	24	0.67
13.405–13.505	1111 = 4	28	0.78
13.505–13.605	11111 1 = 6	34	0.94
13.605–13.705	11 = 2	36	1.00

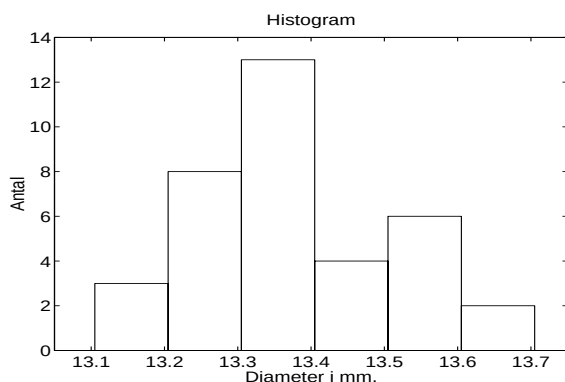
Histogrammets værdi i f. eks. klassen $[13.205, 13.305]$ skal åbenbart være

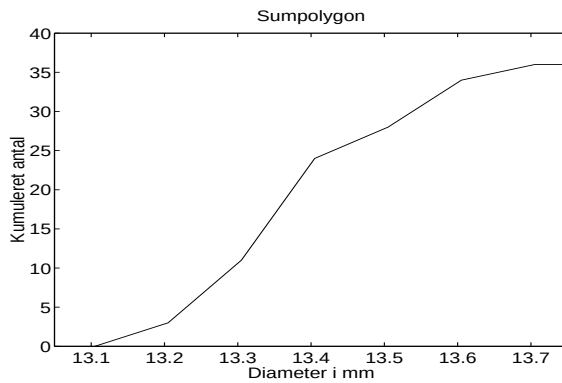
$$\begin{aligned} \frac{\text{antal i klassen}}{\text{totale antal} \times \text{klassebredden}} &= \frac{\text{relative antal i klassen}}{\text{klassebredden}} \\ &= \frac{8}{36 \cdot 0.100} = 2.222. \end{aligned}$$

Hvis alle klasser har samme bredde (som det er tilfældet her), bliver histogrammet proportionalt med den funktion, der i det i 'te interval antager værdien a_i , hvor a_i er antallet af observationer i den i 'te klasse. Derfor anføres ofte to inddelinger på ordinataksen svarende til hver af de ovenstående funktioner.

De samme betragtninger gør sig selvfølgelig gældende for sumpolygonens vedkommende.

Vi anfører nu histogram og sumpolygon for det ovenfor anførte datamateriale i nedenstående to figurer. ♦





Oftentimes one will not be interested in comparing an empirical distribution with a specific distribution, but rather in finding out, whether the empirical distribution can be assumed to stem from a **family** of distributions, e.g. given by the distribution functions

$$F(\cdot, \theta), \quad \theta \in \Omega,$$

where θ is a parameter.

More concretely one can imagine, that one wishes to decide, whether an empirical distribution can be assumed to be of the form

$$F(x, \theta) = \begin{cases} 1 - e^{-x/\theta} & , \quad x \geq 0 \\ 0 & , \quad x < 0 \end{cases},$$

but one is not interested in the specific value of θ .

In this situation it is natural to choose the θ , which gives the "best" agreement, and then evaluate the deviations between the empirical distribution and the distribution, which corresponds to the chosen θ . This way of proceeding will be followed in this chapter, but the methods and principles, by which such a "best" θ is chosen, will be discussed first in chapter 2. The chosen value of θ will only be postulated in the next chapter.

After these introductory considerations over verification of stochastic models we will go over to the concrete cases.

1.2 Modeller i forbindelse med Bernoulli forsøg

Vi indleder med at beskrive de simple forsøg, der er basis for hele afsnittet.

1.2.1 Bernoulli forsøg

Vi siger, at en stokastisk variabel er en **Bernoulli variabel**, hvis den kun antager 2 værdier, f.eks. 0 og 1. Valget af 0 og 1 er egentlig arbitrært, men medmindre andet nævnes, vil vi overalt i det følgende antage, at der er tale om 0 og 1.

Hvis vi sætter $P\{X = 1\} = p$, fås $P\{X = 0\} = 1 - p$. Frekvensfunktionen for X er altså

$$f(x) = P\{X = x\} = p^x(1-p)^{1-x} \quad \text{for } x = 0, 1. \quad (1.1)$$

Heraf fås umiddelbart

$$E(X) = p \quad \text{og} \quad V(X) = p(1-p) \quad (1.2)$$

En sådan 2-punktsfordeling kaldes en **Bernoulli fordeling**.

Et eksperiment med 2 gensidigt udelukkende og udtømmende mulige udfald kaldes et **Bernoulli eksperiment**. Man benævner i mange sammenhænge udfaldene i et sådant eksperiment for **succes** og **fiasko**.

DEFINITION 1.4. En følge af simple Bernoulli eksperimenter, hvor udfaldene af eksperimenterne er uafhængige, og hvor sandsynligheden for succes forbliver uændret, kaldes en følge af **Bernoulli forsøg**. ▲

EKSEMPEL 1.2. Hvis man ved en løbende fabrikation af en bestemt vare kan gå ud fra, at sandsynligheden for, at en enhed af varen ikke er defekt, er konstant, og at en enheds tilstand ikke påvirker de øvrige enheders tilstande, da har man i den løbende konstatering af defekt/ikke defekt et typisk eksempel på Bernoulli forsøg. ◆

Vi vil nu bestemme nogle sandsynlighedsfordelinger i forbindelse med Bernoulli forsøg.

1.2.2 Binomialfordelingen

Lad der være givet n Bernoulli forsøg, hvor sandsynligheden for at få succes (i.e. 1) i et enkelt eksperiment er p . Antallet af succeser X er da en stokastisk variabel med frekvensfunktion

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \dots, n \quad (1.3)$$

Leddene $p^x (1-p)^{n-x}$ er sandsynligheden for en given sekvens af x succeser og $n-x$ fiaskoer. Leddet $\binom{n}{x}$ angiver, på hvor mange måder en sådan sekvens kan forekomme. Heraf følger resultatet umiddelbart.

DEFINITION 1.5. En stokastisk variabel X med frekvensfunktionen (1.3) siges at være **binomialt fordelt** med parametre n og p . Kort skrevet

$$X \in B(n, p) \stackrel{\text{d}}{\Leftrightarrow} f(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \dots, n.$$

▲

BEMÆRKNING 1.2. X kan opfattes som summen af n uafhængige Bernoulli fordelte variable, idet det totale antal succeser i de n forsøg trivielt er lig summen af udfaldene i de enkelte forsøg (da udfaldet 1 $\Leftrightarrow$ succes og udfaldet 0 $\Leftrightarrow$ fiasko). ▼

Vi samler nogle resultater om binomialt fordelte variable i

SÆTNING 1.2. Lad $X \in B(n, p)$. Da er den karakteristiske funktion (jfr. p. 66)

$$\phi(t) = (1 + p(e^{it} - 1))^n.$$

Middelværdi og varians er

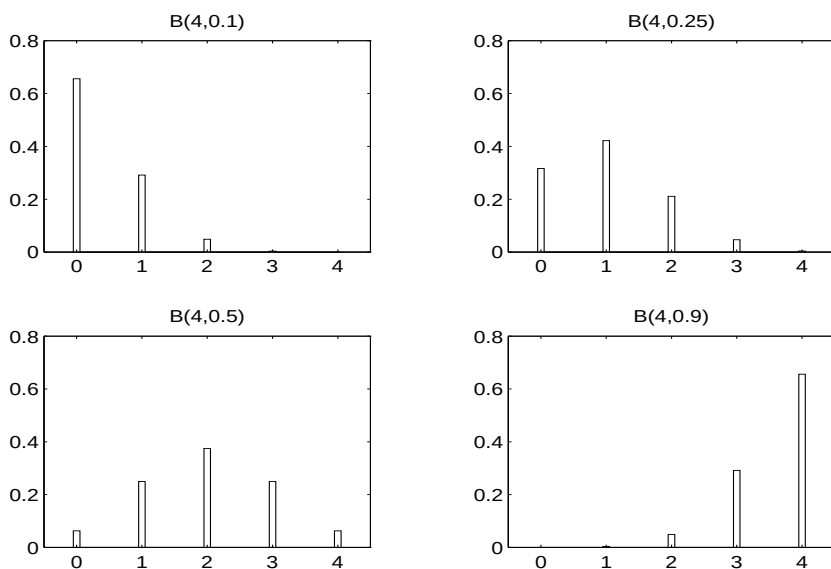
$$\begin{aligned} E(X) &= np \\ V(X) &= np(1-p). \end{aligned}$$

Bevis. Resultatet om middelværdi og varians følger umiddelbart af fortolkningen af X som en sum $Y_1 + \cdots + Y_n$ af n uafhængige Bernoulli fordelte variable, idet vi e.g. har

$$\begin{aligned} V(X) &= V(Y_1 + \cdots + Y_n) \\ &= V(Y_1) + \cdots + V(Y_n) \\ &= np(1-p), \end{aligned}$$

idet vi benytter (0.48) i kapitel 0 og (1.2) i dette kapitel. Tilsvarende vises resultatet om middelværdien. ■

Frekvenskurvens udseende afhænger meget af parametrene n og p . Vi angiver grafer for visse værdier af n og p .



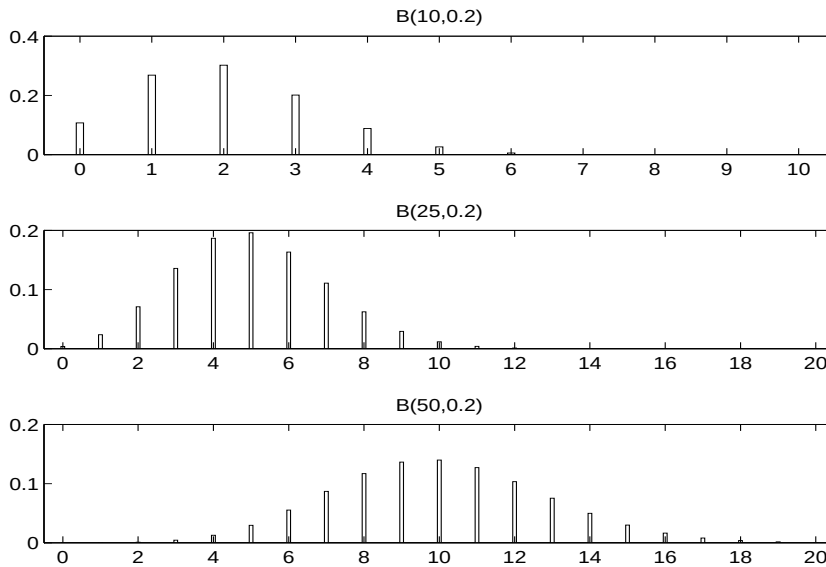
Ved tabelopslag får man ofte brug for følgende

SÆTNING 1.3. Lad $X \in B(n, p)$. Da er $n - X \in B(n, 1 - p)$.

Bevis. X kan opfattes som en sum af Bernoulli variable, d.v.s.

$X = Y_1 + \cdots + Y_n$. Dette giver

$$n - X = (1 - Y_1) + \cdots + (1 - Y_n).$$



Nu er $1 - Y_i$ igen Bernoulli fordelt, men med parameter $1 - p$. Heraf følger, at $n - X \in B(n, 1 - p)$.

■

Vi har endvidere følgende

SÆTNING 1.4 (REPRODUKTIVITETSSÆTNINGEN). Lad X og Y være uafhængige og stokastiske variable. Da gælder

$$X \in B(n_1, p) \wedge Y \in B(n_2, p) \Rightarrow X + Y \in B(n_1 + n_2, p).$$

Bevis. Umiddelbart, når man erindrer, at $X + Y$ kan tolkes som det totale antal succeser i $n_1 + n_2$ Bernoulli forsøg.

■

EKSEMPEL 1.3. Vi minder om, at binomialfordelingen har forbindelse med **stikprøveudtagning med tilbagelægning**. Lad der være givet en samling objekter, der hver er forsynet med netop et af to karakteristika A og B (e.g. defekte varer >< varer, der er i orden, røde kugler >< hvide kugler, emner over en vis længde >< emner under denne længde, etc). Lad brøkdelen af objekter med karakteristikum A være p . Vi udvælger tilfældigt et objekt og noterer, hvilket karakteristikon objektet har. Vi lægger det tilbage og gentager proceduren. Hvis vi efter n udvælgelser kalder antallet af objekter med karakteristikum A for X , er det klart, at $X \in B(n, p)$.

◆

I det følgende eksempel anfører vi nogle data, som man efter ovenstående eksempel kunne vente ville følge en binomialfordeling.

EKSEMPEL 1.4. Ved en stikprøvekontrol i en produktionsproces har man fundet følgende antal defekte emner i 50 stikprøver af hver 100 emner (data stammer fra [25]).

Stikprøve nr.	Andet ciffer									
	0	1	2	3	4	5	6	7	8	9
0	-	2	0	3	2	2	0	4	3	2
1	3	1	6	1	1	2	0	4	0	1
Første	1	2	2	4	2	3	0	1	3	2
ciffer	3	4	4	0	2	3	1	3	0	1
4	4	1	1	3	1	3	4	2	2	0
5	1	-	-	-	-	-	-	-	-	-

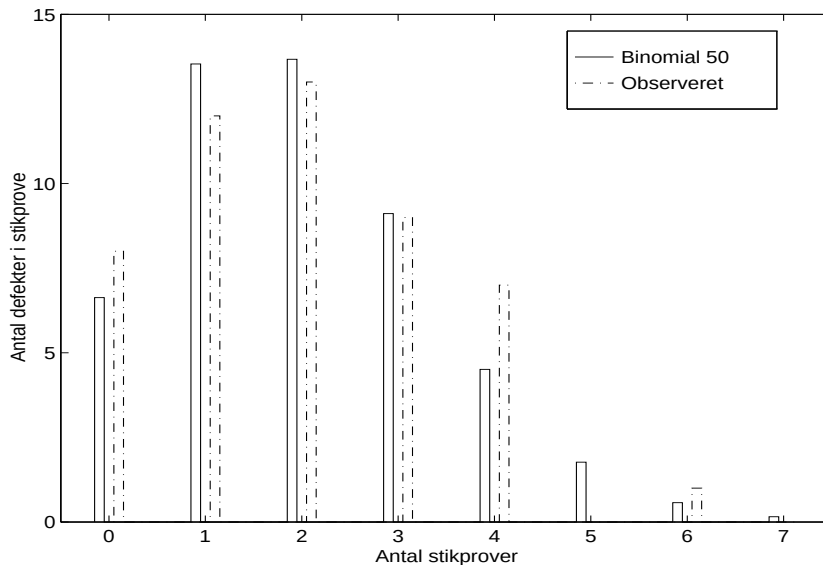
Såfremt processen kan antages at tilfredsstille de almindelige krav til en følge af Bernoulli forsøg, kan ovenstående data opfattes som realiserede udfald $x_1, \dots, x_{50}$ af indbyrdes uafhængige $B(100, p)$ -fordelte variable.

Vi vil derfor sammenligne de relative hyppigheder for $0, 1, \dots$ defekte med en passende valgt binomialfordeling. Vi finder

Antal defekte emner	Antal stikprøver	Relative antal stikpr.
0	8	0.16
1	12	0.24
2	13	0.23
3	9	0.18
4	7	0.14
5	0	0.00
6	1	0.02
7	0	0.00

På nedenstående figur har vi sammenlignet den empiriske fordeling med frekvensfunktionen for en $B(100, 0.02)$ -fordeling.

Parameterværdien $p = 0.02$ er valgt, fordi der er konstateret ca. 2% defekte blandt alle de inspicerede emner.



Der ses at være en god overenstemmelse mellem de to frekvensfunktioner, og Bernoulli modellen og binomialfordelingen synes derfor at være rimeligt egnede til at beskrive stokastiske fænomener af omtalte slags, jfr. sætning 1.1, side 87. ♦

EKSEMPEL 1.5. Vi søger sandsynligheden for, at en stokastisk variabel $X \in B(19, 0.65)$ er større end 7. Nu indeholder tabeller over binomialfordelingen som hovedregel kun indgange for $p \leq 0.5$. Vi anvender derfor sætning 1.3 til at foretage en omskrivning. Vi har

$$\begin{aligned}
 P\{X > 7\} &= P\{19 - X < 12\} \\
 &= P\{B(19, 0.35) < 12\} \\
 &= P\{B(19, 0.35) \leq 11\} \\
 &= 0.9886
 \end{aligned}$$

♦

EKSEMPEL 1.6 (GROUP TESTING). Der er givet et stort antal - lad os sige N - elektroniske komponenter, der hver er indkapslet i et heliumfyldt hylster.

Man ønsker at undersøge, om nogle af disse hylstre lækker, således at heliumet siver ud. Til rådighed haves et apparat, der kan afgøre, om der befinder sig helium i et lukket rum, men vanskeligt, hvor meget der er af luftarten.

Undersøgelsen kan naturligvis nu foregå ved, at man undersøger hver komponent for sig. Dette kræver N forsøg.

Vi kan imidlertid også anbringe k komponenter i måleudstyret og undersøge, om der er helium til stede. Hvis dette **ikke** er tilfældet, er der ingen af de k komponenter, der lækker, således at dette ene test har været tilstrækkeligt for de k komponenter. Er udfaldet positivt, i.e. er der helium til stede, testes hver af de k komponenter separat, og vi har anvendt $k + 1$ målinger til de k komponenter.

Man er nu selvfølgelig interesseret i, hvad det forventede antal målinger er ved anvendelse af ovenstående procedure.

Vi antager, at sandsynligheden p for, at en komponent er defekt, er den samme for alle komponenter, og at komponenter er stokastisk uafhængige.

Vi kalder det nødvendige antal målinger X , og vi vil bestemme fordelingen af X . Vi finder først sandsynligheden for, at en måling på k komponenter giver positivt udfald. Den er

$$\begin{aligned} & P\{\text{mindst 1 ud af } k \text{ er defekt}\} \\ &= 1 - P\{\text{ingen ud af } k \text{ er defekt}\} \\ &= 1 - (1 - p)^k \\ &= \rho. \end{aligned}$$

Her har vi anvendt uafhængighedsforudsætningen ved det 2. lighedstegn.

Der er $n = \frac{N}{k}$ grupper (vi antager for simpeltheds skyld, at n er heltallig). Hvis vi observerer et positivt udfald i netop ν grupper, skal der udføres $n + \nu k$ målinger i alt. Sandsynligheden for denne hændelse er

$$\binom{n}{\nu} \rho^\nu (1 - \rho)^{n - \nu},$$

idet udfaldet af 1. måling i hver gruppe kun har 2 mulige udfald, nemlig positivt (helium til stede) eller negativt (helium ej til stede). Sandsynligheden for positivt udfald er den samme for alle grupper, nemlig ρ , og endelig er grupperne uafhængige. Vi kan derfor anvende (1.3), side 92, og resultatet følger. Vi har altså

$$P\{X = n + \nu k\} = \binom{n}{\nu} \rho^\nu (1 - \rho)^{n - \nu} \quad \text{for } \nu = 0, 1, \dots, n.$$

Vi ser, at $\frac{X-n}{k} \in B(n, \rho)$, d.v.s., at

$$E\left(\frac{X-n}{k}\right) = n\rho$$

eller

$$E(X) = n + k \cdot n\rho = N \left(1 - (1-p)^k + \frac{1}{k}\right), \quad (1.4)$$

hvilket altså er det forventede antal målinger ved den anvendte teknik. For kendt p kan det optimale k bestemmes ved differentiation af (1.4) med hensyn til k og derefter sætte differentialkoefficienten lig 0. Den fremkomne ligning må løses numerisk. Vi indsætter nogle værdier for p og k .

For $p = 0.1$ og $k = 5$ fås

$$E(X) = N \left(1 - 0.9^5 + \frac{1}{5}\right) = 0.6N,$$

og for $p = 0.01$ og $k = 10$ er

$$E(X) = N \left(1 - 0.99^{10} + \frac{1}{10}\right) = 0.16N.$$

Det er helt klart, at man kan reducere antallet af målinger endog særdeles kraftigt ved at anvende en teknik som ovenfor beskrevet. Den kan selvfølgelig forfines meget. I anden omgang behøver man naturligvis ikke at måle på samtlige k elementer, men man kan igen foretage en opdeling i undergrupper etc. ♦

Det anførte eksempel, der her er givet i en lettere simplificeret form, stammer fra Bell Laboratorierne, hvor metoden anvendes med stor succes (d.v.s. man opnår store besparelser i antallet af målinger).

Teknikken kaldes **group testing**. En fremstilling af de mere specielle metoder findes i [50] og [51].

EKSEMPEL 1.7. Under 2. verdenskrig anvendtes den i eksempel 1.6 beskrevne teknik til at undersøge blodprøver hos amerikanske soldater. Fremgangsmåden var helt analog til den ovenfor omtalte. Man blandede blodprøver fra k soldater og undersøgte den blandede prøve for eventuelle bakterier. Man opnåede besparelser i antal prøver på op til 80%. En nøjere beskrivelse findes i [16]. ♦

Vi går nu over til at behandle den negative binomialfordeling.

1.2.3 Den negative binomialfordeling (Pascal's fordeling)

Det næste problem, der naturligvis rejser sig i forbindelse med en række Bernoulli forsøg, er spørgsmålet om, hvornår vi får den 1. og hvornår vi (generelt) får den k 'te succes.

Sandsynligheden for at skulle vente $k + x$ forsøg for at få k succeser er (jfr. [32][p. 23])

$$\binom{k+x-1}{x} p^k (1-p)^x \quad \text{for } x = 0, 1, 2, \dots, \quad (1.5)$$

hvilket med konventionen

$$\binom{-k}{x} = (-1)^x \binom{k+x-1}{x} \quad (1.6)$$

kan skrives

$$\binom{-k}{x} p^k (p-1)^x \quad \text{for } x = 0, 1, 2, \dots \quad (1.7)$$

Resultatet (1.5) er åbenbart. I løbet af de første $k + x - 1$ forsøg skal der opnås $k - 1$ succeser og x fiaskoer. På det $k + x$ 'te forsøg skal der opnås en succes. Faktoren $\binom{k+x-1}{x}$ angiver, på hvor mange måder vi kan placere de x fiaskoer blandt de $k + x - 1$ første forsøg. Sandsynligheden for en af disse er $(1-p)^x$, og p^k er sandsynligheden for at få "placeret" en succes på hver af de resterende k pladser, herunder én på den sidste.

DEFINITION 1.6. En stokastisk variabel X , der har en frekvensfunktion givet ved (1.5) (eller (1.7)), siges at være **negativt binomialt fordelt** eller **Pascal fordelt** med parametrene (k, p) . I tilfældet $k = 1$ taler vi også om den **geometriske fordeling**. Kort skrevet

$$X \in \text{NB}(k, p) \stackrel{\text{d}}{\Leftrightarrow} f(x) = \binom{k+x-1}{x} p^k (1-p)^x \quad x = 0, 1, 2, \dots$$

▲

Vi samler nogle resultater om den negative binomialfordeling i

SÆTNING 1.5. Lad $X \in \text{NB}(k, p)$. Da er den karakteristiske funktion lig

$$\phi(t) = \frac{p^k}{(1 - (1-p)e^{it})^k}.$$

Middelværdi og varians er

$$\begin{aligned} E(X) &= \frac{k(1-p)}{p} \\ V(X) &= \frac{k(1-p)}{p^2} \end{aligned}$$

Bevis. Forbigås. Benyt f.eks. sætning 0.10, side 67. ■

Ligesom for binomialfordelingen har vi en reproduktivitetssætning.

SÆTNING 1.6 (REPRODUKTIVITETSSÆTNING). Lad X og Y være uafhængige stokastiske variable. Da gælder

$$X \in \text{NB}(k_1, p) \wedge Y \in \text{NB}(k_2, p) \Rightarrow X + Y \in \text{NB}(k_1 + k_2, p).$$

Bevis. Resultatet er næsten trivielt, idet $X + Y$ kan opfattes som det antal forsøg, der skal benyttes ud over $k_1 + k_2$ forsøg, før vi har fået $k_1 + k_2$ succeser.

Man kan også blot gange de karakteristiske funktioner for X og Y sammen og konstatere, at det er den karakteristiske funktion for en $\text{NB}(k_1 + k_2, p)$ -fordeling. ■

EKSEMPEL 1.8 (KVALITETSKONTROL). En mulig måde at undersøge kvaliteten af et parti varer er følgende: Man vælger tilfældigt et emne og undersøger, om det er defekt. Det lægges tilbage, og proceduren gentages, indtil man har fået 5 defekte emner. Hvis man når at udtage 50 emner uden at have fundet 5 defekte emner, godkendes partiet.

For at vurdere, om en sådan metode (**stikprøveplan**) er velegnet, kan man undersøge klassifikationssandsynlighederne. Eksempelvis kan vi finde sandsynligheden for at acceptere et parti, der indeholder 10% defekte emner.

Med X benævner vi den stokastiske variable, der angiver, hvor mange forsøg ud over 5, der er nødvendige for at finde 5 defekte. Da er $X \in \text{NB}(5, 0.10)$. Vi accepterer

et parti, hvis vi skal bruge mere end 50 forsøg i alt, d.v.s. hvis $X > 45$. Den søgte sandsynlighed er derfor

$$\begin{aligned} P\{X > 45\} &= 1 - P\{X \leq 45\} \\ &= 1 - \sum_{x=0}^{45} \binom{5+x-1}{x} 0.10^5 0.90^x. \end{aligned}$$

Med et lille program beregnes

$$P\{X > 45\} = 0.43.$$



1.3 Nogle modeller om stikprøveudtagning

Vi har hidtil kun beskæftiget os med stikprøveudtagning med tilbagelægning og det kun i det tilfælde, hvor der blot er 2 muligheder for udfald. Vi vil nu først betragte problemet uden tilbagelægning.

1.3.1 Den hypergeometriske fordeling

Vi betragter igen problemstillingen i eksempel 1.3. Der var givet N objekter hver forsynet med netop 1 af 2 karakteristika, f.eks. defekt og ikke-defekt. Lad der være M defekte. Hvis vi udtager n elementer **med** tilbagelægning som i eksempel 1.3, fås, at det totale antal defekte følger en $B(n, \frac{M}{N})$ -fordeling.

Hvis vi nu **ikke** lægger emnerne tilbage efter at have undersøgt dem, ændrer brøkdelen af defekte emner sig fra træk til træk. Antallet af defekte emner følger derfor ikke længere en binomialfordeling. Sandsynligheden for, at der i stikprøven på n elementer er x defekte bliver

$$f(x) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} \quad (1.8)$$

Argumentet herfor er kombinatorisk. $\binom{M}{x}$ angiver, på hvor mange måder vi kan udtage de x defekte elementer blandt de M , og $\binom{N-M}{n-x}$ angiver på hvor mange måder de ikke-defekte i stikprøven kan udtages. Tælleren angiver altså antallet af måder, hvorpå vor stikprøve kan dannes. Nævneren angiver det totale antal måder, hvorpå en stikprøve på n elementer kan dannes, og resultatet følger heraf.

Udtrykket (1.8) er formelt defineret for $x = 0, 1, \dots, n$, som det er tilfældet for binomialfordelingen. $f(x)$ er imidlertid kun forskellig fra 0 for

$$\max(0, n + M - N) \leq x \leq \min(n, M).$$

DEFINITION 1.7. En stokastisk variabel X , hvis frekvensfunktion er givet ved (1.8), siges at være **hypergeometrisk fordelt** med parametre (n, M, N) . Kort skrevet

$$X \in H(n, M, N) \stackrel{\text{d}}{\Leftrightarrow} f(x) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} \quad x = \max(0, n + M - N), \dots, \min(n, M).$$

▲

Vi anfører middelværdi og varians i

SÆTNING 1.7. Lad $X \in H(n, M, N)$. Da er

$$\begin{aligned} E(X) &= n \cdot \frac{M}{N} \\ V(X) &= n \cdot \frac{M}{N} \left(1 - \frac{M}{N}\right) \frac{N-n}{N-1}. \end{aligned}$$

Bevis. Forbigås. Resultatet fås ved en direkte anvendelse af definitionerne. ■

Idet $p = \frac{M}{N}$ er brøkdelen af defekte emner, ser vi, at middelværdien i den hypergeometriske fordeling er den samme som i den tilsvarende binomialfordeling. Derimod er variansen mindre i den hypergeometriske fordeling end i binomialfordelingen.

Hvis N og M er store, kunne man forvente, at den hypergeometriske fordeling ville ligge tæt ved binomialfordelingen, idet der så ikke kan være stor forskel på stikprøveudtagning med tilbagelægning og uden tilbagelægning. Sådan er det også, og vi præciserer resultatet i

SÆTNING 1.8. Lad N og M gå mod uendelig på en sådan måde, at

$$\frac{M}{N} \rightarrow p \in (0, 1).$$

Da vil for fast n og x

$$\frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} \rightarrow \binom{n}{x} p^x (1-p)^{n-x}.$$

Bevis. Beviset anføres, da det er relativt simpelt og illustrerer, hvordan man under grænseovergange på hensigtsmæssig vis omformer binomialkoefficienter. Vi omskriver venstresiden og får

$$\begin{aligned} \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} &= \binom{n}{x} \frac{M!}{N^x (M-x)!} \\ &\times \frac{(N-M)!}{N^{n-x} (N-M-(n-x))!} \frac{(N-n)! N^n}{N!} \end{aligned}$$

Nu er f.eks.

$$\frac{M!}{N^x(M-x)!} = \frac{M}{N} \left(\frac{M}{N} - \frac{1}{N} \right) \cdots \left(\frac{M}{N} - \frac{x-1}{N} \right).$$

Der er x led, og hvert led går mod p for $N, M \rightarrow \infty$, d.v.s.

$$\frac{M!}{N^x(M-x)!} \rightarrow p^x.$$

Tilsvarende med de øvrige led, og sætningen følger. ■

Sætningen retfærdiggør, at man ved stikprøveundersøgelser i store partier (næsten) altid anvender binomialsandsynligheder, skønt der oftest er tale om stikprøveudtagning uden tilbagelægning.

Vi vil senere se eksempler på den hypergeometriske fordelings anvendelse i praktiske problemer, især i forbindelse med kvalitetskontrol.

Vi går nu over til stikprøveudtagning, hvor der er tale om mere end 2 klassifikationsmuligheder.

1.3.2 Polynomialfordelingen

Lad der til et eksperiment være knyttet k hinanden udelukkende hændelser $A_1, \dots, A_k$, som bestemmer alle muligheder (en kugle har en af farverne rød, grøn, blå eller gul; et emne er for kort, har en passende længde eller er for langt etc.). Lad $A_1, \dots, A_k$ have sandsynlighederne $p_1, \dots, p_k$ henholdsvis. Vi spørger da om sandsynligheden for, at A_i indtræffer x_i gange ved n indbyrdes uafhængige gentagelser af eksperimentet ($x_1 + \dots + x_k = n$). Sandsynligheden for, at A_1 indtræffer x_1 gange, $\dots$, A_k indtræffer x_k gange i en bestemt rækkefølge, er

$$p_1^{x_1} \cdots p_k^{x_k},$$

den samme uanset rækkefølgen. Den søgte sandsynlighed er derfor lig ovenstående ganget med antallet af måder, hvorpå de x_1 pladser for $A_1, \dots$ de x_k pladser for A_k kan udvælges blandt de n . Dette antal benævnes

$$\binom{n}{x_1 \cdots x_k}$$

og man kan ved et forholdsvis enkelt kombinatorisk argument vise, at

$$\binom{n}{x_1 \cdots x_k} = \frac{n!}{x_1! \cdots x_k!}, \quad x_1 + \cdots + x_k = n \quad (1.9)$$

Disse koefficienter kaldes **polynomialkoefficienter**. For $k = 2$ har vi specielt

$$\binom{n}{x_1 x_2} = \frac{n!}{x_1! x_2!} = \binom{n}{x_1},$$

d.v.s. en binomialkoefficient. Den søgte sandsynlighed bliver altså

$$\binom{n}{x_1 \cdots x_k} p_1^{x_1} \cdots p_k^{x_k} = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}, \quad (1.10)$$

hvor altså $p_1 + \cdots + p_k = 1$ og $x_1 + \cdots + x_k = n$. Der er altså kun $k - 1$ parametre foruden n , idet e.g. $p_k = 1 - p_1 - \cdots - p_{k-1}$.

DEFINITION 1.8. En flerdimensional stokastisk variabel $(X_1, \cdots, X_k)$ siges at være **polynomialfordelt** $(n, p_1, \cdots, p_k)$ hvis dens frekvensfunktion har formen (1.10), d.v.s.

$$\begin{aligned} (X_1, \cdots, X_k) &\in \text{Pol}(n, p_1, \cdots, p_k) \\ \stackrel{d}{\Leftrightarrow} f(x_1, \cdots, x_k) &= \binom{n}{x_1, \cdots, x_k} p_1^{x_1} \cdots p_k^{x_k}, \\ x_1 \geq 0, \cdots, x_k \geq 0, &\quad x_1 + \cdots + x_k = n. \end{aligned}$$

▲

Ved lidt regning bringes man let til at indse

SÆTNING 1.9. Lad $\underline{X} = (X_1, \cdots, X_k)' \in \text{Pol}(n, p_1, \cdots, p_k)$. Da er

$$E(\underline{X}) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_k) \end{pmatrix} = \begin{pmatrix} np_1 \\ \vdots \\ np_k \end{pmatrix}$$

og

$$\begin{aligned} V(\underline{X}) &= \begin{pmatrix} V(X_1) & \cdots & \text{Cov}(X_1, X_k) \\ \vdots & & \vdots \\ \text{Cov}(X_k, X_1) & \cdots & V(X_k) \end{pmatrix} \\ &= \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_k \\ -np_2p_1 & np_2(1-p_2) & \cdots & -np_2p_k \\ \vdots & \vdots & & \vdots \\ -np_kp_1 & -np_kp_2 & \cdots & np_k(1-p_k) \end{pmatrix} \end{aligned}$$

BEMÆRKNING 1.3. I overensstemmelse med definitionen i afsnit 0.9 skriver vi her vektorer som søjlevektorer. ▼

Da polynomialfordelingen er en flerdimensional generalisering af binomialfordelingen, kunne man forvente, at der også ville findes en reproduktivitetssætning for den. Sætningen eksisterer, og den lyder som følger.

SÆTNING 1.10 (REPRODUKTIVITETSSÆTNING). Lad $\underline{X} = (X_1, \dots, X_k)$ og $\underline{Y} = (Y_1, \dots, Y_k)$ være stokastisk uafhængige. Da gælder

$$\begin{aligned} \underline{X} \in \text{Pol}(n_1, p_1, \dots, p_k) \quad \wedge \quad \underline{Y} \in \text{Pol}(n_2, p_1, \dots, p_k) &\Rightarrow \\ \underline{X} + \underline{Y} &\in \text{Pol}(n_1 + n_2, p_1, \dots, p_k). \end{aligned}$$

Bevis. Beviset er ganske analogt til det tilsvarende for binomialfordelingen. Komponenten $X_i + Y_i$ er det totale antal udfald af hændelse A_i i n forsøg, og resultatet følger nu direkte af definitionen af polynomialfordelingen. ■

EKSEMPEL 1.9. For en given type glødelamper gælder, at brændetiden for en tilfældig udvalgt glødelampe falder i intervallerne 0-700 timer, 700-800 timer og over 800 timer med sandsynligheder på henholdsvis 0.2, 0.7 og 0.1. Glødelamperne sælges i pakninger på 5. Vi vil finde sandsynligheden for at få 4 glødelamper med en brændetid på mere en 700 timer og 3 med en brændetid på mere en 800 timer i en enkelt pakning.

Antallene (X_1, X_2, X_3) af glødelamper i de 3 klasser $\in \text{Pol}(5, 0.2, 0.7, 0.1)$. Den søgte sandsynlighed er $P\{X_2 + X_3 \geq 4 \wedge X_3 \geq 3\}$. De værdier af (x_1, x_2, x_3) , der tilfredsstiller disse uligheder, er

$$\{(0, 0, 5), (0, 1, 4), (0, 2, 3), (1, 0, 4), (1, 1, 3)\} = A.$$

Den søgte sandsynlighed er nu

$$\sum_{(x_1, x_2, x_3) \in A} \binom{5}{x_1 \ x_2 \ x_3} 0.2^{x_1} 0.7^{x_2} 0.1^{x_3} = 0.00816,$$

d.v.s. mindre end 1%.



Polynomialfordelingen spiller en vigtig rolle f.eks. ved tests i antalstabeller, kontrol af fordelingslov m.v., som vi senere skal se.

1.4 Poisson modeller. Erlang- og Γ -fordelingen

Poisson-fordelingen er en yderst nyttig fordeling til at beskrive mange former for tælleprocesser med. Vi vil i de kommende afsnit anføre nogle eksempler på "fysiske" forhold, der fører til Poisson modeller. Vi indleder med

1.4.1 Poisson fordelingen

Vi anfører først definitionen på Poisson fordelingen.

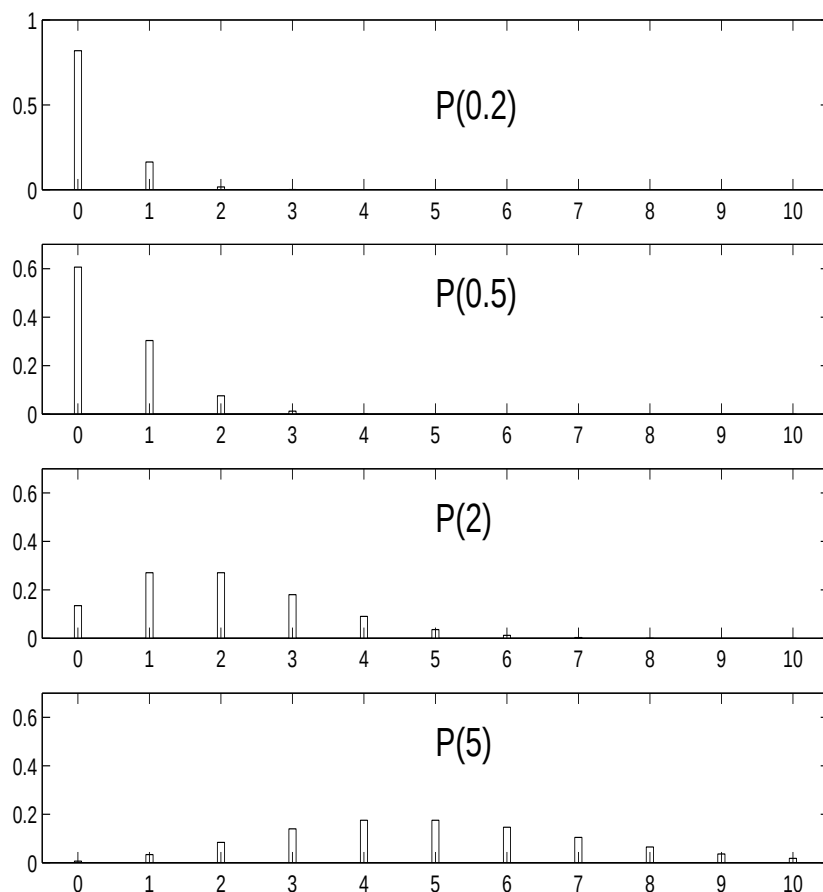
DEFINITION 1.9. Vi siger, at en stokastisk variabel X er **Poisson fordelt** med parameter λ , hvis den har frekvensfunktionen

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

Vi skriver kort $X \in P(\lambda)$.



Vi antyder parametrenes indflydelse på fordelingen ved at skitsere nogle grafer af frekvensfunktioner for forskellige værdier af λ .



Det ses, at skævheden aftager meget med større værdier af λ . Samtidig spredtes sandsynlighedsmassen mere jævnt ud over de naturlige tal.

Vi samler nogle resultater om Poisson fordelingen i

SÆTNING 1.11. Lad $X \in P(\lambda)$. Da er den karakteristiske funktion lig

$$\phi(t) = \exp(\lambda(e^{it} - 1)),$$

og middelværdi og varians er

$$E(X) = \lambda$$

$$V(X) = \lambda$$

Bevis. Vi benytter sætning 0.10. Den karakteristiske funktion findes direkte ud fra

definitionen:

$$\begin{aligned}
 \phi(t) &= E(e^{iXt}) \\
 &= \sum_{x=0}^{\infty} e^{itx} \frac{\lambda^x}{x!} e^{-\lambda} \\
 &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{1}{x!} (\lambda e^{it})^x \\
 &= e^{-\lambda} \exp(\lambda e^{it})
 \end{aligned}$$

Vi finder middelværdien ved at differentiere ϕ m.h.t. t . Vi har

$$\begin{aligned}
 \phi'(t) &= e^{-\lambda} \exp(\lambda e^{it}) \cdot i \cdot \lambda e^{it} \\
 E(X) &= -i\phi'(0) = \lambda.
 \end{aligned}$$

Analogt fås resultatet for variansen v.h.a. $\phi''(t)$. ■

Vi bemærker, at middelværdi og varians er ens.

SÆTNING 1.12 (REPRODUKTIVITETSSÆTNING). Lad X og Y være uafhængige stokastiske variable. Da gælder

$$X \in P(\lambda_1) \wedge Y \in P(\lambda_2) \Rightarrow X + Y \in P(\lambda_1 + \lambda_2).$$

Bevis. Den karakteristiske funktion for $X + Y$ er, jfr. sætning 0.11

$$\begin{aligned}
 \psi(t) &= \exp(\lambda_1(e^{it} - 1)) \exp(\lambda_2(e^{it} - 1)) \\
 &= \exp((\lambda_1 + \lambda_2)(e^{it} - 1)),
 \end{aligned}$$

hvilket netop er den karakteristiske funktion for en $P(\lambda_1 + \lambda_2)$ -fordeling. ■

Vi giver nu Poisson's sætning, der viser, at Poisson fordelingen kan fremkomme ved en passende grænseovergang i binomialfordelingen.

SÆTNING 1.13 (POISSON). Lad $X_n \in B(n, p_n)$. Hvis

$$n \cdot p_n \rightarrow \lambda, \quad n \rightarrow \infty,$$

da vil for $x = 0, 1, 2, \dots$

$$P\{X_n = x\} \rightarrow \frac{\lambda^x}{x!} e^{-\lambda}, \quad n \rightarrow \infty.$$

Bevis. Se [21][p. 140]. ■

Denne sætning forklarer, hvorfor man i praksis ofte møder Poisson fordelingen, når man har et meget stort antal forsøg og en meget lille sandsynlighed for, at en bestemt hændelse indtræffer i det enkelte forsøg (jfr. [10][p. 55]). Et eksempel, hvor man præciserer disse verbale formuleringer, har vi i

SÆTNING 1.14. Der er givet en proces, der tilfældigt genererer punkter i et område. Processen tilfredsstiller:

1. Sandsynligheden for, at der er et punkt i et givet delområde, afhænger kun af områdets størrelse (volumen, areal etc.) og **ikke** af områdets form.
2. Sandsynligheden for, at der er et punkt i et område af størrelsen v , er

$$\lambda v + o(v),$$

og sandsynligheden for, at der er mere end ét punkt, er $o(v)$ ¹.

3. Antallene af punkter i ikke-overlappende områder er stokastisk uafhængige.

Da er sandsynligheden for, at et område af størrelsen v netop indeholder k punkter lig

$$\frac{1}{k!} (\lambda v)^k e^{-\lambda v}, \quad k = 0, 1, \dots,$$

d.v.s. antallet af punkter i et område af størrelsen v er en stokastisk variabel $\in P(\lambda v)$.

Bevis. Vi inddeler området A i n lige store dele. Sandsynligheden for at få mindst ét udfald i en givet del er ifølge 2. (idet $o(v) + o(v) = o(v)$)

$$p_n = \lambda \frac{v}{n} + o\left(\frac{v}{n}\right).$$

Ifølge 3. har vi nu givet n Bernoulli forsøg med sandsynligheden p_n for succes (hvor succes altså betyder mindst ét punkt i et delområde. Antallet af succeser er derfor $\in$

¹En funktion $f \in o(v)$, $v \rightarrow 0$, hvis $\frac{f(v)}{v} \rightarrow 0$, $v \rightarrow 0$ d.v.s. f går hurtigere mod 0 end v .

$B(n, p_n)$. Da sandsynlighed for at få 2 eller flere udfald i et delområde er forsvindende (for store n), har vi, at

$$P\{k \text{ punkter i } A\} \approx \binom{n}{k} p_n^k (1 - p_n)^{n-k}.$$

Da

$$np_n = \lambda v + n o\left(\frac{v}{n}\right) = \lambda v + \frac{o\left(\frac{v}{n}\right)}{\frac{v}{n}} v \rightarrow \lambda v$$

for $n \rightarrow \infty$, kan vi anvende Poisson's sætning, og resultatet følger. ■

EKSEMPEL 1.10. I forbindelse med partikeltælling i mikroskop møder man ofte Poisson fordelingen. Deler man synsfeltet i mikroskopet op i n lige store dele, og tæller man derefter antallet af partikler i hvert delfelt, vil man, hvis fordelingen er "tilfældig" (d.v.s. tilfredsstillende 1)-3) i sætningen), have n realisationer af Poisson fordelte stokastiske variable. Dette er f.eks. ofte tilfældet, når man tæller bakteriekolonier eller blodplader etc. Eksempler på rumlige fordelinger af tilfældigt placerede punkter er f.eks. fordelingen af sten af en vis størrelse i beton eller fordelingen af "fejl" i faste stoffer (e.g. små brud o.lign. i metalstykker etc.).

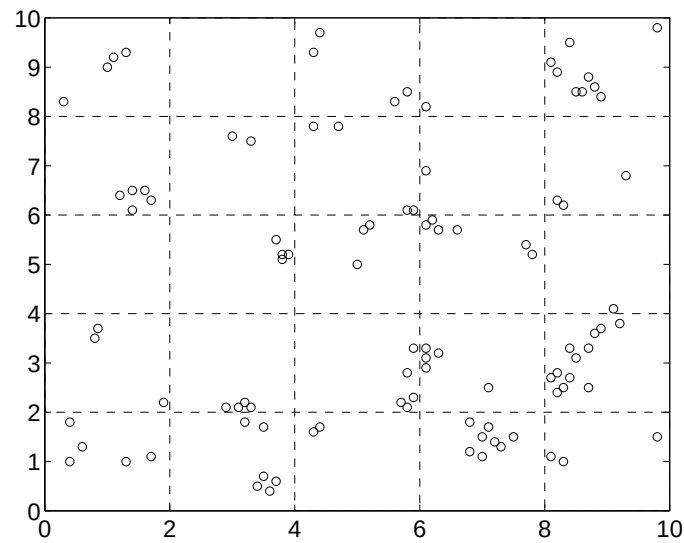
Den af forudsætningerne, som i praksis er mest kritisk, er nok 3), uafhængighedsforudsætningen. Således vil partikler opløst i væsker ofte enten tiltrække eller frastøde hinanden, således at 3) ikke længere kan antages opfyldt. Hvis partikelkoncentrationen er stor, vil de øvrige betingelser ofte ikke være til stede, således at man må forvente, at Poisson fordelingen vil være en dårlig approksimation til fordelingen. I sådanne tilfælde vil man hyppigt kunne rede situationen ved at fortynde opløsningen. ♦

Vi giver dernæst et eksempel, der kan illustrere ovenstående problemstilling.

EKSEMPEL 1.11. I nedenstående figur har vi anført 100 punkters placering i et kvadrat. Vi vil søge at afgøre, om placeringen kan antages at være foregået efter en proces, der tilfredsstillende i sætning 1.20 anførte betingelser, eller - mere enkelt - om punkterne synes tilfældigt fordelt i planen efter en lovmæssighed, som angivet ved Poisson fordelingen.

Vi har inddelt kvadratet i 25 mindre delkvadrater og talt antallet af punkter i hvert kvadrat. Herved er følgende tal fremkommet.

4,	0,	4,	1,	9,	5,	2,	4,	1,	3,	0,	4, 3,
6,	1,	3,	4,	5,	5,	12,	5,	6,	2,	8,	3,

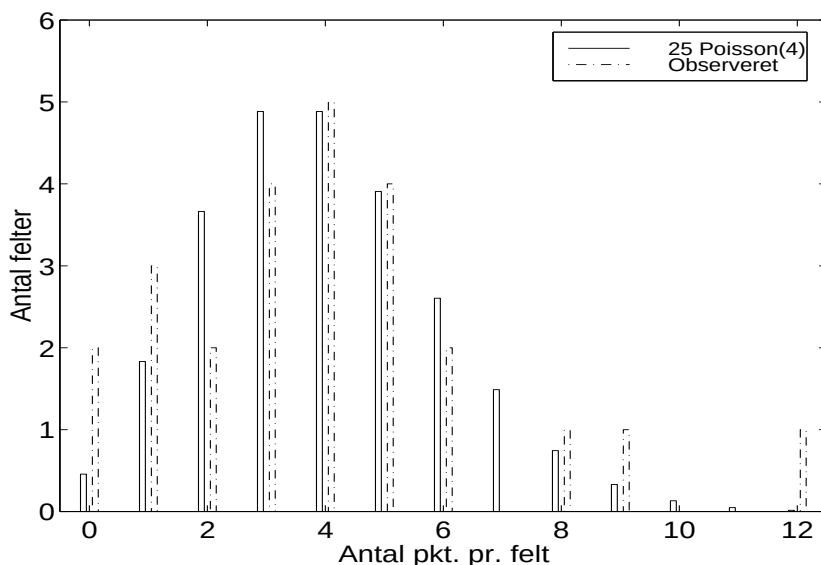


Hvis den punktgenererede proces virkelig spredte punkterne tilfældigt udover kvadratet, måtte disse tal kunne opfattes som realiserede udfald $x_1, \dots, x_{25}$ af indbyrdes uafhængige identisk $P(\lambda)$ -fordelte stokastiske variable. Vi vil derfor sammenligne den empiriske fordeling svarende til ovenstående tal med en $P(\lambda)$ -fordeling. Det vises i kapitel 2, at den (i en vis forstand) bedst approksimerende Poisson fordeling er en $P(4)$ -fordeling (idet der gennemsnitligt er 4 punkter/kvadrat).

Vi finder først den empiriske fordeling.

Antal pkt.f. kvadr.	0	1	2	3	4	5	6	7	8	9	10	11	12
Antal kvadr.	2	3	2	4	5	4	2	0	1	1	0	0	1
Rela- tive antal kvadr.	0.08	0.12	0.08	0.16	0.20	0.16	0.08	0.00	0.04	0.04	0.00	0.00	0.04

Denne empiriske fordeling er på nedenstående figur sammenlignet med en $P(4)$ -fordeling.



Det ses, at den empiriske fordeling har tykkere haler end Poisson fordelingen. Dette svarer til, at der er større sandsynlighed for at finde helt tomme kvadrater og for at finde kvadrater med mange observationer i, end der ville være, hvis punkterne var fordelt tilfældigt ud over planen. Der synes med andre ord at være en tendens til "klumpning".



De samme forudsætninger som 1)-3) i sætning 1.20 møder vi i postulaterne for **Poisson processen**. Der er givet en proces, der til tilfældige tidspunkter genererer en hændelse. Antallet af hændelser i tidsintervallet $[0, t)$ kaldes $X(t)$. Da er $X(t)$ en stokastisk proces, d.v.s. at vi til ethvert tidspunkt $t \in [0, \infty)$ har en stokastisk variabel $X(t)$. Vi anfører nu postulaterne i

SÆTNING 1.15. Hvis processen tilfredsstiller:

- a. sandsynligheden for at få mindst én hændelse i et interval af længden h er

$$\lambda h + o(h),$$

- b. sandsynligheden for at få 2 eller flere hændelser i intervallet er $o(h)$,

- c. antallene af hændelser i ikke-overlappende intervaller er stokastisk uafhængige,

da er $X(t) \in P(\lambda t)$.

Bevis. Beviset for dette resultat kan naturligvis føres ganske som beviset for sætning 1.20. Imidlertid kan man ved at foretage nogle grænseovergange komme frem til et system af differentiaalligninger, som ved løsning umiddelbart giver $X(t) \in P(\lambda t)$. Disse overvejelser er, hvad angår den matematiske stringens, mere tilfredsstillende. For nærmere detaljer henvises til [32][p. 65], [19], [21]. ■

DEFINITION 1.10. Det i sætningen forekommende λ kaldes **intensiteten**. Den angiver middeltallet af hændelser i et tidsinterval af længden 1 (én). ▲

EKSEMPEL 1.12. Et af de kendteste tilfælde på tilfældige hændelser, der fører til Poisson processer, er udsendelsen af α -partikler fra et radioaktivt stof med en ikke for kort halveringstid ($\sim$ konstant intensitet). ◆

EKSEMPEL 1.13. Betragter man telefontrafikken i en central, vil man meget ofte have, at antallet af opkald til centralen følger en Poisson proces, i det mindste i den såkaldte travle time (den time i døgnet, hvor trafikken er størst). Hvis man betragter et længere tidsrum, vil intensiteten λ ikke længere være konstant. Da centraler oftest planlægges og udbygges efter trafikken i den travle time, betyder dette ikke nogen indskrænkning i anvendeligheden af Poisson fordelingen som beskrivelse af forholdene. I [27] findes en mængde eksempler på anvendelser af Poisson processen i teletrafikken. ◆

EKSEMPEL 1.14. I tabellen side 117 er anført kundeankomsttiderne for en kasse i et supermarked (i min.) og afstandene mellem disse tider (i $\frac{1}{100}$ min.).

For at vurdere, om ankomstprocessen tilfredsstiller Poisson postulaterne, kan vi betragte antallet af kundeankomster i en fast tidsenhed, f.eks. et minut. Dette giver værdierne

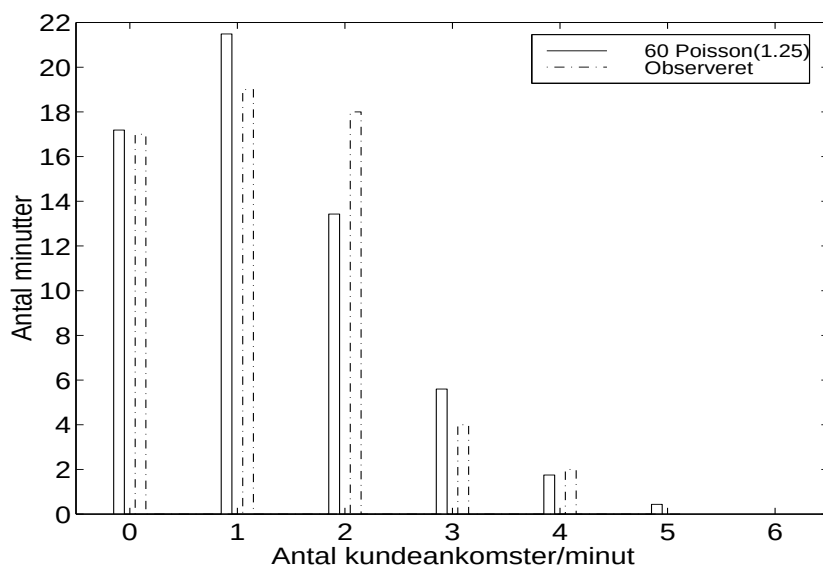
1,	3,	1,	1,	0,	1,	2,	1,	1,	2,	2,	1,
1,	3,	0,	1,	2,	2,	3,	1,	4,	0,	1,	0,
0,	1,	2,	1,	2,	1,	2,	0,	0,	0,	1,	2,
3,	0,	1,	2,	0,	0,	1,	2,	2,	2,	4,	2,
0,	1,	1,	0,	2,	0,	0,	2,	2,	2,	0,	0.

Disse skal kunne opfattes om realiserede udfald af indbyrdes uafhængige $P(\lambda)$ -fordelte variable, hvis antagelserne om ankomstprocessen holder. Vi finder derfor de relative hyppigheder

Antal kunde- ank./min.	Antal minutter	Rel. antal minutter
0	17	0.28
1	19	0.32
2	18	0.30
3	4	0.07
4	2	0.03

I grafen side 116 er de relative hyppigheder sammenlignet med frekvensfunktionen for en $P(1.25)$ -fordeling.

Der ses at være en pæn overensstemmelse mellem de to frekvensfunktioner, og Poisson modellen synes derfor velegnet til at beskrive de betragtede ankomsttider. ♦



EKSEMPEL 1.15. Ved et busstoppested ankommer busser efter en Poisson proces med en intensitet på 0.4 busser pr. minut. Hvad er da fordelingen af ventetiden mellem 2 busser? Ventetiden er større end t , netop hvis der ingen busser ankommer i intervallet $(0, t)$. Dette sker med sandsynligheden $e^{-0.4t}$. Ventetiden T har altså fordelingen

$$P\{T \leq t\} = 1 - P\{T > t\} = 1 - e^{-0.4t}, \quad t > 0$$

for t målt i minutter. ♦

Ank.	Diff	Ank.	Diff	Ank.	Diff	Ank.	Diff
0.93	11	15.60	106	29.10	92	46.61	0
1.04	4	16.66	27	30.02	52	46.61	2
1.08	52	16.93	52	30.54	350	46.63	61
1.60	43	17.45	38	34.04	121	47.24	35
2.03	186	17.83	43	35.25	55	47.59	178
3.89	128	18.26	68	35.80	71	49.37	121
5.17	127	18.94	1	36.51	26	50.58	163
6.44	46	18.95	64	36.77	10	52.21	53
6.90	41	19.59	53	36.87	197	52.74	269
7.31	157	20.22	22	38.84	35	55.43	13
8.88	36	20.44	0	39.19	71	55.56	69
9.24	71	20.44	13	39.90	271	56.25	33
9.25	29	20.57	147	42.61	64	56.58	84
10.24	62	22.04	342	43.25	61	57.42	51
10.86	58	25.46	86	43.86	17	57.93	227
11.44	127	26.32	26	44.03	42		
12.71	45	26.58	65	44.45	124		
13.16	16	27.23	140	45.69	2		
13.32	17	28.63	12	45.71	29		
13.49	211	28.75	35	46.00	61		

Tabel 1.1: Kundeankomsttider ved kasse i et supermarked og differenser mellem disse (i min. og $\frac{1}{100}$ min.).

Med dette eksempel er vi ovre i næste afsnit, hvor vi skal diskutere ventetidsfordelinger i Poisson processen.

1.4.2 Erlang- og Γ -fordelingen

Vi generaliserer ovenstående eksempel til problemet med ventetiden indtil den k 'te hændelse i en Poisson proces.

SÆTNING 1.16 (ERLANG-FORDELING). Lad der være givet en Poisson proces med intensiteten λ . Da er ventetiden mellem k hændelser en stokastisk variabel T med tæthed

$$f(t) = \frac{1}{\Gamma(k)} \lambda^k t^{k-1} e^{-\lambda t} \quad t \geq 0 \quad (1.11)$$

og 0 ellers.

Bevis. Beviset følger ganske analogt til bestemmelse af ventetiden i eksempel 1.15. Vi har

$$\begin{aligned} F(t) &= P\{T \leq t\} \\ &= P\{\text{mindst } k \text{ hændelser inden tiden } t\} \\ &= \sum_{i=k}^{\infty} P\{i \text{ hændelser inden tiden } t\} \\ &= \sum_{i=k}^{\infty} e^{-\lambda t} \frac{(\lambda t)^i}{i!} \end{aligned} \quad (1.12)$$

Tætheden fremkommer nu ved differentiation af (1.12), d.v.s.

$$\begin{aligned} f(t) &= -\lambda e^{-\lambda t} \sum_{i=k}^{\infty} \frac{(\lambda t)^i}{i!} + e^{-\lambda t} \sum_{i=k}^{\infty} \lambda \frac{(\lambda t)^{i-1}}{(i-1)!} \\ &= \lambda e^{-\lambda t} (\lambda t)^{k-1} \frac{1}{(k-1)!} \\ &= \frac{1}{\Gamma(k)} \lambda^k t^{k-1} e^{-\lambda t}, \end{aligned}$$

hvor vi har anvendt resultatet $\Gamma(k) = (k-1)!$ ((0.66) p. 75). ■

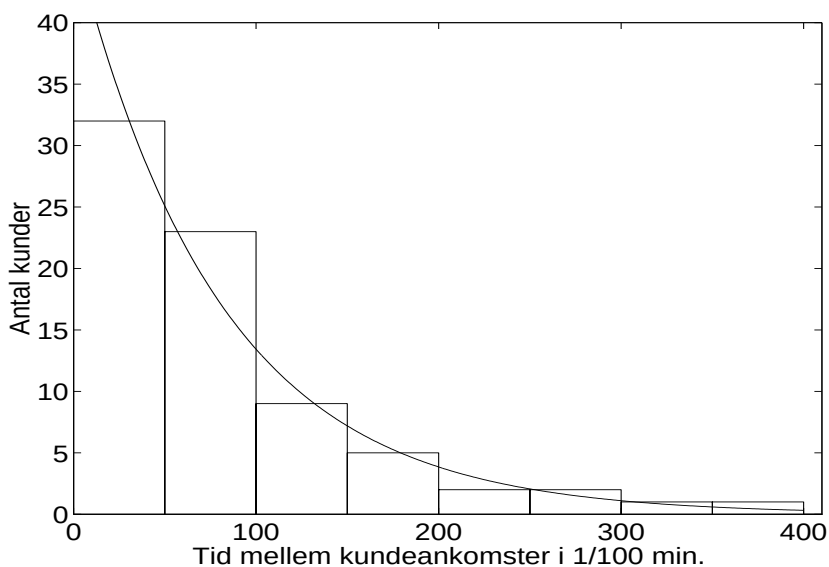
Ventetidsfordelingen (1.11) i Poisson processen kaldes **Erlang** fordelingen. Den er en analog til den negative binomialfordeling i en række af Bernoulli forsøg, hvor man finder ventetiden udtrykt i antal enkeltforsøg til den k 'te succes (se p. 99).

EKSEMPEL 1.16. Vi betragter de i eksempel 1.14 anførte differenser mellem kundeankomster. Man ville - med resultatet i eksemplet in mente - forvente, at disse fulgte en Erlang fordeling med parameter 1.25 min^{-1} .

Vi har foretaget en gruppering af materialet, og nedenfor er anført de relative hyppigheder af de forskellige tidsafstande

Klasse	Antal	Rel. antal
[0, 50[	32	0.43
[50,100[	23	0.31
[100,150[	9	0.12
[150,200[	5	0.07
[200,250[	2	0.03
[250,300[	2	0.03
[300,350[	1	0.01
[350,400[	1	0.01

Histogrammet svarende til disse målinger er sammen med 50×75 tætheden for en Erlang fordeling med parametre $(1, 1.25/100)$ indtegnet i nedenstående figur.



Der ses at være en pæn overensstemmelse, og vi har igen fået en indikation af, at Poisson modellen er velegnet til at beskrive fænomener som det anførte. ♦

Parameteren k er heltallig ifølge sin oprindelse, men man ser, at udtrykket (1.11) også er defineret for andre værdier af k . En sådan generaliseret Erlang fordeling kaldes en Γ -fordeling. Det er dog almindeligt her at anvende $\lambda^{-1} = \beta$ som parameter i stedet for λ . Vi får da

DEFINITION 1.11 (GAMMA-FORDELING). En stokastisk variabel X er Γ -fordelt med parametre (k, β) , $k > 0$, $\beta > 0$, hvis tætheden for X er

$$f(x) = \frac{1}{\Gamma(k)\beta^k} x^{k-1} e^{-x/\beta}, \quad x > 0 \quad (1.13)$$

og 0 ellers. Vi skriver kort $X \in G(k, \beta)$. ▲

Tilfældet $k = 1$ er særlig vigtigt:

DEFINITION 1.12 (EKSPONENTIALFORDELING). En $G(1, \beta)$ -fordelt stokastisk variabel X siges også at være **eksponentialt fordelt** med parameter β . Vi skriver kort $X \in \text{Ex}(\beta)$ ($\Leftrightarrow X \in G(1, \beta)$). Af (1.13) fås, at tætheden for en $\text{Ex}(\beta)$ -fordelt variabel er

$$f(x) = \frac{1}{\beta} e^{-x/\beta}, \quad x > 0$$

og 0 ellers. ▲

Vi anfører graferne for tætheder af nogle Γ -fordelinger

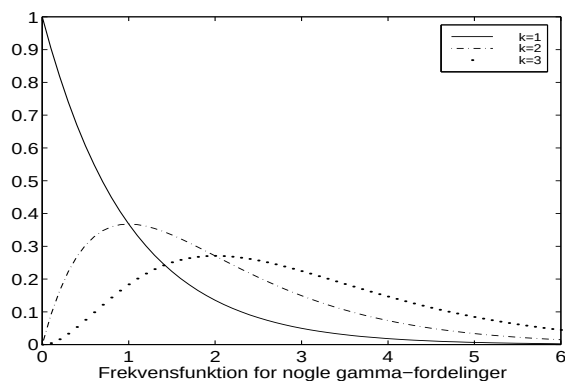
For $k < 1$ har tæthederne både x-aksen og y-aksen som asymptote.

Især i ældre fremstillinger benævnes Γ -fordelingen ofte som Pearson Type-III fordelingen.

Vi ser af graferne, at parameteren k er af stor betydning for tæthedens udseende. Vi kalder k for **formparameteren** i fordelingen. Om parameteren β har vi

SÆTNING 1.17. Parameteren β er en skalaparameter, d.v.s. vi har

$$X \in G(k, \alpha) \Rightarrow \beta X \in G(k, \alpha\beta).$$



Bevis. Af (0.30) p. 43 fås, at tætheden for βX er

$$\begin{aligned} f_{\beta X}(x) &= \frac{1}{\beta} f_X\left(\frac{x}{\beta}\right) \\ &= \frac{1}{\beta} \frac{1}{\Gamma(k)} \frac{1}{\alpha^k} \left(\frac{x}{\beta}\right)^{k-1} e^{-\frac{x}{\alpha\beta}} \\ &= \frac{1}{\Gamma(k)} \frac{1}{(\alpha\beta)^k} x^{k-1} e^{-\frac{x}{\alpha\beta}}, \quad x > 0, \end{aligned}$$

hvilket netop er tætheden for en $G(k, \beta)$ -fordeling. ■

SÆTNING 1.18. Lad $X \in G(k, \beta)$. Da er den karakteristiske funktion

$$\phi(t) = \frac{1}{(1 - i\beta t)^k},$$

og middelværdien og variansen er

$$\begin{aligned} E(X) &= k\beta \\ V(X) &= k\beta^2 \end{aligned}$$

Bevis. Sætningen vises ved at finde $\phi(t)$, som er givet i definition 0.9, side 66, og benytte sætning 0.10. ■

Da Erlang fordelingen er en ventetidsfordeling, må man forvente, at der eksisterer en additionssætning. Det gør der, og det endda også for Γ -fordelingen. Den er som følger

SÆTNING 1.19 (REPRODUKTIVITETSSÆTNING). Lad X og Y være uafhængige stokastiske variable. Da gælder

$$X \in G(k_1, \beta) \wedge Y \in G(k_2, \beta) \Rightarrow X + Y \in G(k_1 + k_2, \beta).$$

Bevis. For Erlang fordelinger er resultatet trivielt, idet $X + Y$ kan tolkes som ventetiden, indtil vi har observeret $k_1 + k_2$ hændelser i en Poisson proces med intensitet $\lambda = \frac{1}{\beta}$. I det generelle tilfælde fås resultatet umiddelbart ved at danne den karakteristiske funktion for $X + Y$. ■

SÆTNING 1.20. Lad et system have levetiden T , og lad

$$P\{T > t + s | T > t\} = P\{T > s\}, \quad (1.14)$$

d.v.s. lad sandsynligheden for, at systemet endnu lever mindst tiden s , givet, det har nået alderen t , være den samme som sandsynligheden for, at systemet i alt lever tiden s . Da er $T \in \text{Ex}(\beta)$ for et eller andet β .

Bevis. Se [32][p. 30] ■

På grund af sætningen siger vi, at eksponentialfordelingen ingen hukommelse har (**lack of memory**).

EKSEMPEL 1.17. Elektroniske komponenters levetid tilfredsstiller ofte betingelsen i sætning 1.20, hvilket forklarer, hvorfor deres levetidsfordeling hyppigt er eksponentiel. Den approksimation holder dog hyppigt kun i den midterste (og længste) del af en komponents levetid. I begyndelsen vil en hel del svigte på grund af, at fabrikations- og materialefejl viser sig ret hurtigt, og i den sidste del stiger "dødeligheden" igen på grund af almindelig "udslidthed". ♦

1.5 Den normale fordeling

Den vigtigste af de fordelinger, vi møder i statistikken, er den normale fordeling eller Gauss fordelingen. Før vi begrundet dette, anfører vi

1.5.1 Analytiske egenskaber

Vi starter med

DEFINITION 1.13. Vi siger, at X er **standardiseret (eller normeret) normalt fordelt**, hvis X har tætheden

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

og dermed fordelingsfunktionen

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt$$

Betegnelserne φ og Φ benyttes ofte som standardbetegnelser for den standardiserede normalfordeling. ▲

Hvis X følger fordelingen med positionsparameter μ og skalaparameter σ , siger vi, at X er normalt fordelt med parametre μ og σ^2 (NB! σ^2) og skriver $X \in N(\mu, \sigma^2)$.

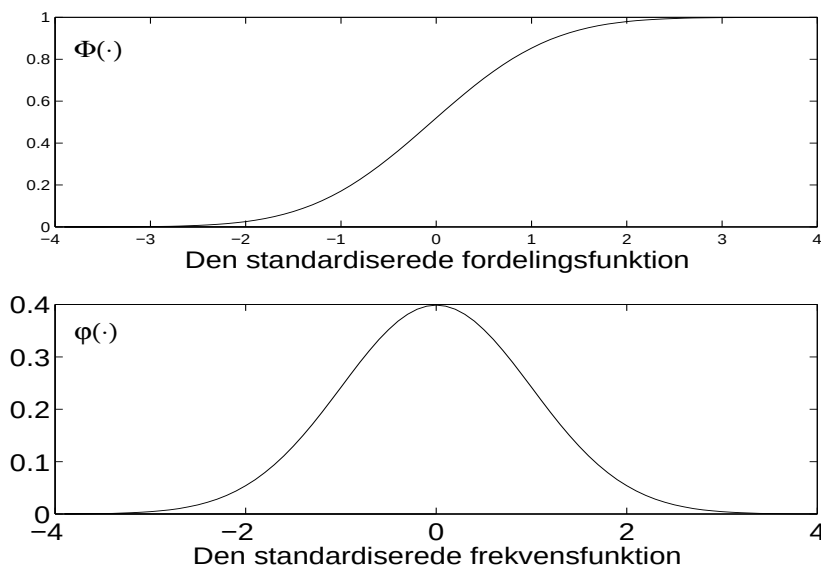
Tætheden for $X \in N(\mu, \sigma^2)$ er ifølge (0.30) p. 43

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

DEFINITION 1.14. Fraktillerne i den standardiserede normalfordeling benævnes **u-fraktiler**. p -fraktilen skrives u_p . ▲

Vi anfører grafen for fordelingsfunktion og tæthed for et $X \in N(0, 1)$

Forklaringen på, hvorfor vi vælger at parametrisere med σ^2 i stedet for σ , får vi af (sidste halvdel af)



SÆTNING 1.21. Lad $X \in N(\mu, \sigma^2)$. Da er den karakteristiske funktion

$$\psi(t) = \exp(i\mu t - \frac{1}{2}\sigma^2 t^2).$$

Middelværdi og varians er

$$\begin{aligned} E(X) &= \mu \\ V(X) &= \sigma^2 \end{aligned}$$

Bevis. Sætningen vises ved at benytte definition 0.9, side 66, og sætning 0.10, side 67. ■

Vi parametriserer som regel normalfordelingen med middelværdien μ og variansen σ^2 .

Normalfordelingen er reproduktiv i begge parametre, d.v.s. vi har

SÆTNING 1.22. Lad X og Y være stokastisk **uafhængige**. Da gælder

$$X \in N(\mu_1, \sigma_1^2) \wedge Y \in N(\mu_2, \sigma_2^2) \Rightarrow X+Y \in N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2).$$

Bevis. Følger umiddelbart af sætning 0.10, side 67. ■

BEMÆRKNING 1.4. Sætningen udvides umiddelbart til at omfatte en vilkårlig linearkombination af normalt fordelte variable. Vi har altså, at hvis $X_1, \dots, X_n$ er stokastisk uafhængige, da vil

$$X_i \in N(\mu_i, \sigma_i^2), \quad i = 1, \dots, n \Rightarrow a_0 + a_1 X_1 + \dots + a_n X_n \in N\left(a_0 + \sum a_i \mu_i, \sum a_i^2 \sigma_i^2\right),$$

hvor $a_0, a_1, \dots, a_n$ er givne konstanter. ▼

BEMÆRKNING 1.5. Der findes en slags omvendt sætning, idet det kan vises, at hvis X og Y er stokastisk uafhængige, og hvis $X + Y$ er normalt fordelt, da er såvel X som Y normalfordelte. ▼

Vi vil nu nævne nogle af de egenskaber ved den normale fordeling, der har givet fordelingen dens fremtrædende rolle i anvendt sandsynlighedsregning.

1.5.2 Den centrale grænseværdisætning

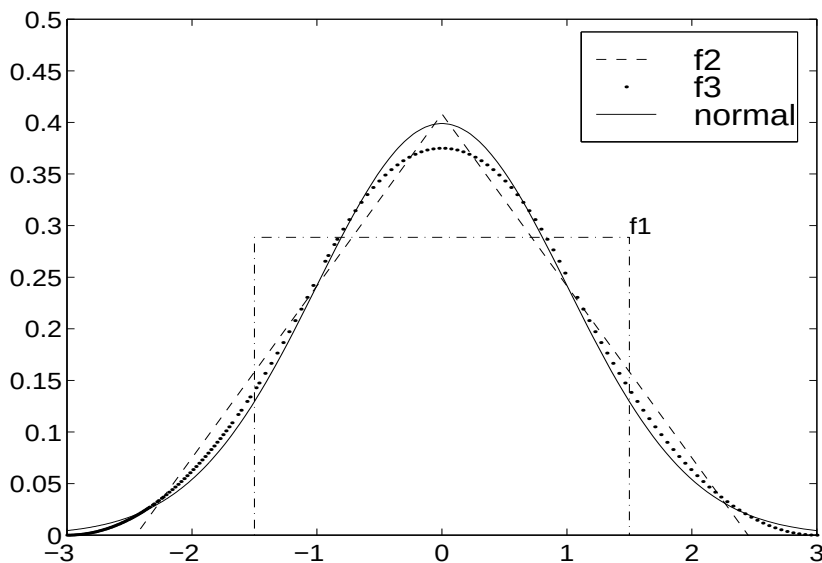
Vi indleder dette afsnit med

EKSEMPEL 1.18. Lad der være givet et emne, hvis vægt vi bestemmer på en nøjagtig skålvægt. Vi ønsker at angive resultatet i hele gram og runder derfor af til nærmeste hele antal gram. Derved introduceres en fejl X , og det vil være rimeligt at antage, at enhver værdi af X mellem $-\frac{1}{2}$ og $\frac{1}{2}$ vil være lige sandsynlig, d.v.s. at X har tætheden

$$f(x) = 1 \quad \text{for } -\frac{1}{2} < x < \frac{1}{2}.$$

Denne tæthed kaldes den rektangulære tæthedsfordeling, se også afsnit 1.8.1. side 180.

Er der nu givet n emner, og bestemmer vi den totale vægt af disse som summen af enkeltvægtene, har vi en fejl $X_1 + \dots + X_n$, hvor X_i er fejlen introduceret ved afrundingen af det i 'te emne. Vi kan nu succesivt bestemme tæthederne for $Y_1 = X_1$, $Y_2 = X_1 + X_2$, $Y_3 = X_1 + X_2 + X_3$ etc. (jfr. [32][p. 10]). Det vises let, at $E(Y_i) = 0$ og $V(Y_i) = i \cdot V(X_1) = \frac{i}{12}$. I nedenstående figur er anført tæthederne f_1, f_2, f_3 for $\sqrt{12}Y_1$, $\sqrt{\frac{12}{2}}Y_2$, $\sqrt{\frac{12}{3}}Y_3$. Konstanterne bevirker, at de 3 variable alle har samme middelværdi 0 og varians 1, således at vi lettere kan sammenligne tæthederne



Det ses, at tætheden f_3 er smukt klokkeformet. Endvidere fremgår det, at allerede tætheden f_3 ligger meget tæt ved $N(0, 1)$ -fordelingen. Tætheden for summen $(X_1 + \dots + X_n) \sqrt{\frac{12}{n}}$ synes altså at nærme sig tætheden for $N(0, 1)$ -fordelingen. ♦

SÆTNING 1.23 (CENTRAL GRÆNSEVÆRDISÆTNING (LINDBERG-LÉVY)). Hvis $X_1, X_2, \dots$ er en følge af uafhængige, identisk fordelte stokastiske variable med middelværdi μ og varians σ^2 , er $X_1 + \dots + X_n, n = 1, 2, \dots$, **asymptotisk** $\in N(n\mu, n\sigma^2)$ for $n \rightarrow \infty$, d.v.s.

$$P \left\{ \frac{\sum X_i - n\mu}{\sqrt{n\sigma^2}} \leq x \right\} \rightarrow \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt \text{ for } n \rightarrow \infty.$$

Bevis. Se [21][p. 197]. ■

Kravet om identisk fordeling er ikke nødvendig for at sikre, at summen er asymptotisk normalt fordelt for $n \rightarrow \infty$, idet vi har følgende mere generelle sætning:

SÆTNING 1.24 (CENTRAL GRÆNSEVÆRDISÆTNING I LJAPOUNOV'S VERSION). Lad $X_1, X_2, \dots$ være en følge af indbyrdes uafhængige, stokastiske variable, og lad

$$\begin{aligned} E(X_i) &= \mu_i \\ V(X_i) &= \sigma_i^2 \\ E(|X_i - \mu_i|^3) &= \gamma_i^3 \end{aligned}$$

være endelige for alle $i = 1, 2, \dots$. Såfremt

$$\frac{\sum_{i=1}^n \gamma_i^3}{\left(\sum_{i=1}^n \sigma_i^2\right)^{3/2}} \rightarrow 0 \text{ for } n \rightarrow \infty,$$

da vil $X_1 + \dots + X_n$ asymptotisk være $\in N(\sum \mu_i, \sum \sigma_i^2)$.

Bevis. Se [21][p. 203.] ■

Den centrale grænseværdisætning findes i endnu mere generelle versioner, der ikke involverer eksistensen af tredje momenter. Dette skal vi ikke komme ind på her, men blot henvise til [21][p. 206].

BEMÆRKNING 1.6. Normalfordelingen er gerne tabelleret for $\mu = 0$ og $\sigma^2 = 1$. For $X \in N(\mu, \sigma^2)$ gælder imidlertid

$$\frac{X - \mu}{\sigma} \in N(0, 1)$$

således at

$$P\{N(\mu, \sigma^2) \leq x\} = P\{N(0, 1) \leq \frac{x - \mu}{\sigma}\} = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

hvor den standardiserede normale fordelingsfunktion $\Phi(\cdot)$ findes ved tabelopslag.

Heraf følger at tabellen for en $N(0,1)$ -fordeling kan bruges for enhver $N(\mu, \sigma^2)$ -fordeling. ▼

Fremgangsmåden illustreres i

EKSEMPEL 1.19. Ved dimensioneringen af gulvet i en virksomhed regner man med, at vægten, der hviler på gulvet, er summen af vægten X af maskiner og inventar plus vægten af Y af varer, personale og kunder etc. Da hver af disse er summen af mange relativt små vægte, antages det, at de hver er normalt fordelte. Da der ikke kan antages at være nogen væsentlig korrelation mellem dem, antages det, at de er uafhængige.

Erfaringer fra tidligere konstruktioner af samme art foreslår, at man anvender - målt i 100 kg resp. $(100 \text{ kg})^2$ -

$$\begin{array}{ll} E(X) &= 100 & V(X) &= 50 \\ E(Y) &= 25 & V(Y) &= 18 \end{array}$$

Som design vægt vil man nu forvente en vægt, der under ovenstående forudsætninger kun overstiges i 5% af tilfældene. Vi har, at $X + Y \in N(100 + 25, 50 + 18) = N(125, 68) = N(125, 8.25^2)$. Man skal nu bestemme x , så

$$P\{N(125, 8.25^2) > x\} = 5\%$$

eller ensbetydende hermed

$$P\left\{N(0, 1) \leq \frac{x - 125}{8.25}\right\} = 95\%$$

Af tabel fås

$$\frac{x - 125}{8.25} = 1.645$$

i.e.:

$$x = 138.57$$

Som designvægt anvendes altså 138.57 (100 kg) = 13857 kg. ♦

Den centrale grænseværdisætning forklarer, hvorfor mange typer fejl (eller afvigelser) følger en normal fordeling, idet man ofte kan antage, at en fejl opstår som summen af mange små bidrag. Disse overvejelser er især gyldige, hvor det drejer sig om resultatet af **fysiske målinger**, hvor fejlen meget hyppigt ses at følge en normal fordeling. Der findes selvsagt mange fremstillinger af, hvilke fejlstrukturer, der fører til den normale fordeling. Vi skal indskrænke os til at anføre en enkelt, nemlig

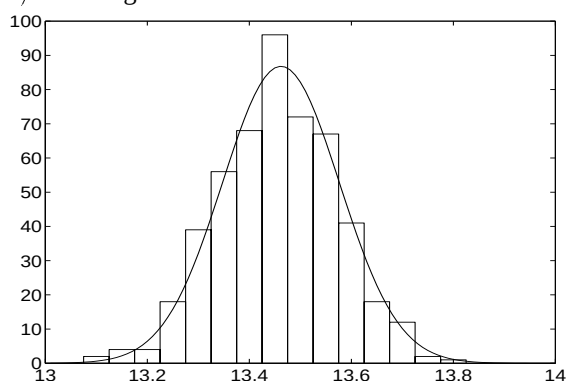
Hagens fejhhypoteser

- a. En fejl er **summen** af et stort antal infinitesimale fejl, alle af samme størrelse, og de er alle en følge af forskellige årsager.

- b. De forskellige fejlkomponenter er uafhængige.
- c. Hver fejlkomponent har lige stor sandsynlighed for at være positiv eller negativ.

Under disse forudsætninger vil fejlfordelingen være normal, som det let ses ved anvendelse af den centrale grænseværdisætning.

EKSEMPEL 1.20. I nedenstående figur er anført et histogram over hoveddiameteren for 500 nittehoveder. Data stammer fra [25]. Endvidere er indtegnet $500 \times$ klassebredden $\times$ frekvensfunktionen for en $N(13.426, 0.0132)$ -fordeling.

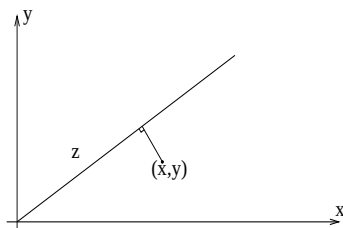


Det er vel ikke urimeligt at antage, at afvigelsen mellem nittehoveder skyldes en mængde små, additive bidrag, således at man skulle vente, at fordelingen af hovedets diameter kan beskrives ved en normalfordeling. Dette bekræftes af ovenstående figur. ♦

Vi vil senere se en mængde eksempler på praktiske situationer, hvor den normale fordeling optræder som modelfordeling for tilfældige afvigelser.

1.5.3 Andre hypoteser, der fører til den normale fordeling

Det er, som vi skal se, ikke kun den centrale grænseværdisætning, der begrunder, at man så hyppigt møder den normale fordeling. Lad os f.eks. betragte en fordeling af skud affyret mod et punktformigt mål. Vi lægger et ortogonalt koordinatsystem med begyndelsespunkt i målet. Endvidere indlægger vi en vilkårlig orienteret linie gennem origo.



Vi projicerer punktet (x, y) ind på linien, og projektionens afstand fra $(0, 0)$ benævnes z . Lad nu følgende være opfyldt.

- De marginale tætheder for fejlene X og Y er kontinuerte.
- Tætheden i punktet (x, y) afhænger kun af afstanden $r = (x^2 + y^2)^{\frac{1}{2}}$ fra centrum.
- Fejlene i x - og y -retning er uafhængige.

Da er afvigelsen $Z \in N(0, \sigma^2)$ for et eller andet σ^2 . Se også [46][p. 127].

Også i den statistiske mekanik finder man den normale fordeling stor anvendelse. Således kan nævnes, at Maxwell's hypoteser om molekylehastigheder fører til, at hastighederne er normalt fordelte.

Beviserne for denne type sætninger bygger ofte på, at man ud fra hypoteserne opstiller nogle funktionalligninger, som tæthederne må tilfredsstille, og det vises derefter, at de eneste løsninger er normale tætheder.

1.5.4 Den normale fordeling som tilnærmelse til andre fordelinger

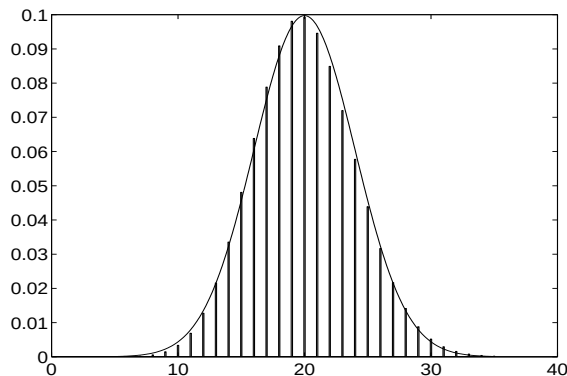
Ved hjælp af den centrale grænseværdisætning ses, at vi for fordelinger, der besidder reproduktivitetsegenskaber (d.v.s. der eksisterer additionssætninger), ofte kan anvende tilnærmelser med normalfordelingen. Vi anfører nogle eksempler.

EKSEMPEL 1.21 (APPROKSIMATIONER TIL BINOMIALFORDELINGEN). Lad $X \in B(n, p)$. Da kan X opfattes som en sum $X = X_1 + \dots + X_n$ af n uafhængige, identisk fordelte Bernoulli variable. Af Lindeberg-Lévy's sætning følger da, at X er asymptotisk $N(np, np(1-p))$ -fordelt, for $n \rightarrow \infty$. For store n kan vi derfor finde binomialsandsynligheder ved hjælp af tabeller over den normale fordeling. Lad f.eks. $X \in B(100, 0.20)$. Da er $np = 20$ og $np(1-p) = 16$. Anvendes normalapproximationen, fås e.g.

$$P\{X \leq 30\} = P\left\{\frac{X - np}{\sqrt{np(1-p)}} \leq 2.5\right\} \simeq 0.9938.$$

Den eksakte værdi er 0.99394, hvilket jo ligger meget tæt på det approksimativt beregnede.

På nedenstående figur er vist frekvensfunktionerne for en $B(100, 0.20)$ - og for en $N(20, 16)$ -fordeling. Man bemærker også her den gode overensstemmelse.



Nu antager en binomialt fordelt variabel jo kun heltallige værdier, så hvis vi vil anvende en normalfordelingstilnærmelse, kan vi ofte opnå forbedrede resultater ved at ændre grænserne for den binomialt fordelte variabel, idet ethvert x erstattes med intervallet $(x - \frac{1}{2}, x + \frac{1}{2})$ (den såkaldte **kontinuitetskorrektion**). Eksempelvis har vi (med samme X som før)

$$\begin{aligned} P\{10 \leq X \leq 25\} &= P\{9.5 \leq X \leq 25.5\} \\ &= P\left\{-\frac{10.5}{4} \leq \frac{X-np}{\sqrt{np(1-p)}} \leq \frac{5.5}{4}\right\} \\ &\simeq P\{-2.63 \leq N(0, 1) \leq 1.38\} \\ &= 0.9119. \end{aligned}$$

Ved hjælp af et program findes, at sandsynligheden faktisk er 0.9102, d.v.s. igen en fin overensstemmelse. For at approksimationen skal være anvendelig, skal n være stor og p ikke nær 0 eller 1. Som tommelfingerregel kan man anføre, at man kan anvende approksimationen, hvis $np > 5$ og $n(1-p) > 5$. Hvis p eller $1-p$ er nær ved 0, er binomialfordelingen ret skæv. Det er da bedre at anvende Poisson's sætning og approksimere med en Poisson fordeling. Vi har eksempelvis

$$P\{B(25, 0.01) \leq 1\} \simeq P\{P(0.25) \leq 1\} = 0.974.$$

Den eksakte værdi er 0.97424, hvilket viser en meget fin overensstemmelse. Hvis p er stor, går vi frem som følger

$$\begin{aligned} P\{B(10, 0.95) \leq 9\} &= P\{10 - B(10, 0.95) \geq 1\} \\ &= P\{B(10, 0.05) \geq 1\} \\ &\simeq P\{P(0.5) \geq 1\} \\ &= 0.393. \end{aligned}$$

Den eksakte sandsynlighed er 0.40126, en tålelig overensstemmelse. Som tommelfingerreglen anfører vi, at man især kan anvende Poisson tilnærmelse, hvis p eller $1 - p < 0.1$. ♦

EKSEMPEL 1.22 (APPROKSIMATIONER TIL POISSON FORDELINGEN). Lad

$X \in P(\lambda)$. Da kan X ifølge sætning 1.12 opfattes som en sum $X = X_1 + \cdots + X_n$ af uafhængige $P(\mu)$ -fordelte stokastiske variable, hvor μ er givet ved $n\mu = \lambda$. Derfor vil X være asymptotisk $N(\lambda, \lambda)$ -fordelt for $\lambda \rightarrow \infty$ ifølge Lindeberg-Lévy's sætning. Vi kan derfor anvende den normale fordeling som en tilnærmelse til Poisson fordelingen. Idet tilsvarende betragtninger vedrørende kontinuitetskorrektion gør sig gældende her som ved binomialfordelingen, har vi e.g. for $X \in P(25)$

$$\begin{aligned} P\{X \leq 30\} &= P\{X \leq 30.5\} \\ &= P\left\{\frac{x-\lambda}{\sqrt{\lambda}} \leq \frac{30.5-25}{5}\right\} \\ &\simeq P\{N(0, 1) \leq 1.1\} \\ &= 0.8643. \end{aligned}$$

Den eksakte værdi er 0.8633, d.v.s. en rimelig overensstemmelse. Approksimationer er selvsagt bedst anvendelige for store λ . En tommelfingerregel er, at $\lambda > 10$. ♦

EKSEMPEL 1.23 (APPROKSIMATIONER TIL Γ -FORDELINGEN). Som i de 2 foregående eksempler indses, at en stokastisk variabel $X \in G(k, \beta)$ er asymptotisk $N(k\beta, k\beta^2)$ -fordelt for $k \rightarrow \infty$. Dette kan som før anvendes til at approksimere fordelingen med en normal fordeling. Lad f.eks. $k = 25$ og $\beta = 1$. Da er

$$\begin{aligned} P\{X \leq 31.6\} &= P\left\{\frac{X - k\beta}{\sqrt{k\beta^2}} \leq \frac{31.6 - 25}{5}\right\} \\ &\simeq P\{N(0, 1) \leq 1.32\} \\ &= 0.9066. \end{aligned}$$

Den eksakte sandsynlighed er 0.9000, d.v.s. overensstemmelsen er tilfredsstillende. Vi skal senere se, hvorledes tabeller over den såkaldte χ^2 -fordeling kan anvendes til at finde sandsynligheder i Γ -fordelingen. ♦

1.6 Den logaritmiske normale fordeling

Den logaritmiske normale fordeling spiller samme rolle i situationer, hvor multiplikation er den naturlige komposition, som den normale fordeling gør der, hvor addition er naturlig. Vi indleder med

1.6.1 Analytiske egenskaber

Vi anvender som hidtil betegnelsen $\log_e$ for den naturlige logaritme.

DEFINITION 1.15. Vi siger, at en stokastisk variabel X er **logaritmisk normalt fordelt**, hvis $\log_e X$ er normalt fordelt. Mere præcist

$$X \in \text{LN}(\alpha, \beta^2) \Leftrightarrow \log_e X \in \text{N}(\alpha, \beta^2).$$

▲

SÆTNING 1.25. Lad $X \in \text{LN}(\alpha, \beta^2)$. Da har X tætheden

$$f(x) = \frac{1}{x\beta\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log_e x - \alpha}{\beta}\right)^2\right), \quad x > 0$$

og 0 ellers.

Bevis. Vi har, at

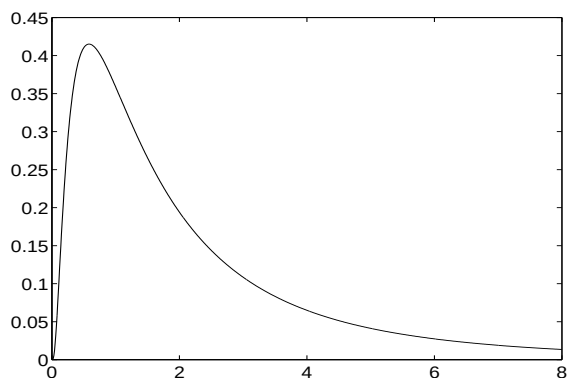
$$\begin{aligned} P\{X \leq x\} &= P\{\log_e X \leq \log_e x\} \\ &= \int_{-\infty}^{\log_e x} \frac{1}{\beta\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\alpha}{\beta}\right)^2} dt. \end{aligned}$$

Ved differentiation fås tætheden

$$f(x) = \frac{1}{\beta\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log_e x - \alpha}{\beta}\right)^2\right) \frac{1}{x}$$

■

Grafen for en logaritmisk normal tæthed er skæv. Den har et udseende som skitseret i nedenstående graf, hvor $\alpha = 0.46$ og $\beta = 1$.



Vi bemærker, at parametrene ikke er middelværdi og varians. Vi har

SÆTNING 1.26. Lad $X \in \text{LN}(\alpha, \beta^2)$. Da er

$$\begin{aligned} E(X) &= e^{\alpha + \frac{1}{2}\beta^2} \\ V(X) &= e^{2\alpha + \beta^2} (e^{\beta^2} - 1). \end{aligned}$$

Bevis. Forbigås. Fås ved direkte regning. ■

Sammenhængen mellem parametre α og β^2 og middelværdi og varians udtrykker vi i

SÆTNING 1.27. Lad $X \in \text{LN}(\alpha, \beta^2)$ og sæt $E(X) = \mu$ og $V(X) = \sigma^2$. Da er

$$\begin{aligned} \alpha &= \log_e \mu - \frac{1}{2} \log_e \left(\frac{\sigma^2}{\mu^2} + 1 \right) \\ \beta^2 &= \log_e \left(\frac{\sigma^2}{\mu^2} + 1 \right). \end{aligned}$$

Bevis. Fås ved løsning af ligningerne i sætning 1.26. ■

I den tekniske litteratur møder man ofte begreberne geometrisk middelværdi og geometrisk standardafvigelse. Disse er defineret i

DEFINITION 1.16. Den geometriske middelværdi μ_g og den geometriske standardafvigelse σ_g i en $\text{LN}(\alpha, \beta^2)$ -fordeling er givet ved

$$\begin{aligned}\mu_g &= e^\alpha \\ \sigma_g &= e^\beta\end{aligned}$$

▲

Vi kan nu uden videre overføre alle sætninger om addition af normalt fordelte stokastisk variable til sætninger om produkter af logaritmisk normalt fordelte variable. Vi har således

SÆTNING 1.28. Lad $X_1, \dots, X_n$ være uafhængige stokastiske variable, og lad $X_i \in \text{LN}(\alpha_i, \beta_i^2)$, $i = 1, \dots, n$. Da er for $b > 0$ og $a_1, \dots, a_n \in \mathbb{R}$

$$b \prod_{i=1}^n X_i^{a_i} \in \text{LN} \left(\log_e b + \sum a_i \alpha_i, \sum a_i^2 \beta_i^2 \right).$$

Bevis. Beviset følger af definitionen og af bemærkning 1 p. 125. ■

EKSEMPEL 1.24. Lad f.eks. $X \in \text{LN}(0.5, 1)$ og $Y \in \text{LN}(0.75, 1.25)$ være uafhængige stokastiske variable. Det følger af sætning 1.28 at

$$3X \cdot Y^2 \in \text{LN}(\log_e 3 + 0.5 + 2 \cdot 0.75, 1 + 2^2 \cdot 1.25) = \text{LN}(3.1, 6),$$

idet $b = 3$, $a_1 = 1$ og $a_2 = 2$. ◆

Vi får endvidere analoge generaliseringer til den centrale grænseværdisætning. Lindeberg-Lévy's sætning bliver til

SÆTNING 1.29. Lad $X_1, X_2, \dots$ være en følge af uafhængige, identisk fordelte positive stokastiske variable. Lad det endvidere være givet, at $\log_e X_i$ har endelig middelværdi μ og varians σ^2 . Da er $\prod X_i$ asymptotisk $\text{LN}(n\mu, n\sigma^2)$ -fordelt for $n \rightarrow \infty$.

Liapounov's sætning går over i

SÆTNING 1.30. Lad $X_1, X_2, \dots$ være en følge af uafhængige, positive stokastiske variable, således at

$$\begin{aligned} E(\log_e X_i) &= \mu_i \\ V(\log_e X_i) &= \sigma_i^2 \\ E(|\log_e X_i - \mu_i|^3) &= \gamma_i^3 \end{aligned}$$

alle eksisterer. Hvis

$$\frac{\sum_{i=1}^n \gamma_i^3}{\left(\sum_{i=1}^n \sigma_i^2\right)^{3/2}} \rightarrow 0 \text{ for } n \rightarrow \infty,$$

er $\prod X_i$ asymptotisk fordelt som $LN(\sum \mu_i, \sum \sigma_i^2)$.

Ved hjælp af disse sætninger er vi i den følgende paragraf i stand til at formulere modeller, der fører til logaritmisk normalt fordelte data.

1.6.2 Loven om proportional effekt

Vi betragter en genstand, hvis størrelse (f.eks. vægt, volumen etc.) underkastes forandringer i n skridt. Vi går ud fra, at størrelsen efter det r 'te skridt kan betragtes som en stokastisk variabel X_r . Med X_0 betegnes begyndelsesværdien. Lad os nu antage, at tilvæksten (der kan være en reduktion!) fra skridt $i - 1$ til skridt i er en stokastisk variabel af den aktuelle størrelse, d.v.s.

$$X_i - X_{i-1} = C_i X_{i-1} \quad (1.15)$$

eller

$$X_i = (1 + C_i) X_{i-1}. \quad (1.16)$$

Heraf fås

$$X_n = X_0(1 + C_1)(1 + C_2) \cdots (1 + C_n).$$

Bemærk, at C_i 'erne er stokastisk variable.

DEFINITION 1.17. Et system siges at følge **loven om proportional effekt**, hvis det tilfredsstiller (1.15). ▲

Hvis følgen af stokastiske variable $X_0, 1+C_1, \dots, 1+C_n, \dots$ tilfredsstiller betingelserne i sætning 1.29 eller sætning 1.30, da vil størrelsen af genstanden være asymptotisk LN-fordelt.

Hvis man derfor har en fysisk proces, som følger loven om proportional effekt, vil man derfor kunne beskrive udfaldet af processen ved en stokastisk variabel, der er LN-fordelt.

Som eksempel kan vi nævne **biologisk vækst**, idet tilvæksten her ofte vil være proportional med antallet af celler. LN-fordelte stokastiske variable anvendes derfor ofte til beskrivelse af vægt, højde, længde etc. af biologiske individer.

I kemien møder man ofte tilsvarende forhold f.eks. ved **krystallers vækst**. Eksempelvis anvendes LN-fordelingen ofte til beskrivelse af størrelsen af sølvpartikler i en fotografisk emulsion.

Endvidere kan LN-fordelingen ofte anvendes til beskrivelse af partikelstørrelser i kornede materialer, der er resultatet af gentagne **knusningsprocesser**. Som eksempler på sådanne materialer kan nævnes sand og grus ("naturlig" knusning) og malm ("kunstig" knusning). Begrundelsen er som før loven om proportional effekt, denne gang blot med negative tilvækster. Udsættes partiklen for et slag, vil den formindskes med en vis brøkdel. Udsættes den for nok et slag, vil den igen formindskes, denne gang formentlig en anden brøkdel etc., således at vi kan antage (1.15) opfyldt med negative C 'er.

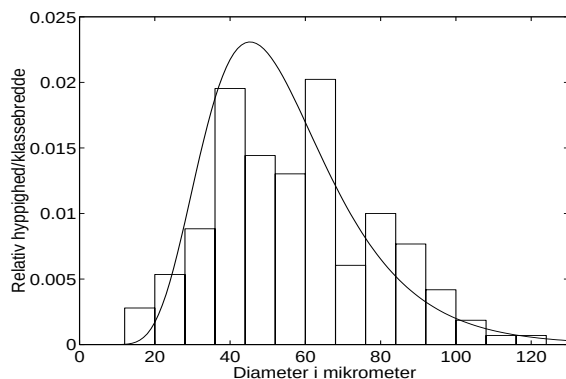
Det må også nævnes, at [11] har godtgjort, at en **hydrologisk** hændelse af en vis størrelse X kan opfattes om et produkt

$$X = X_1 \cdots X_r$$

af r uafhængige størrelser, der skyldes meteorologiske og geografiske faktorer. Da r er et stort tal, vil man igen forvente, at X er LN-fordelt.

EKSEMPEL 1.25. I nedenstående figur er vist et histogram for fordelingen af diametre af alit-krystaller brændt i 18 timer ved 1500° . Alit er et cementklinkemineral. Data stammer fra [49].

Efter loven om proportional effekt skulle man forvente en logaritmisk normal fordeling, og dette bekræftes af den indtegnede frekvensfunktion for en LN-fordeling. ♦



EKSEMPEL 1.26. Lad os antage, at diameteren af sandpartikler kan beskrives ved en stokastisk variabel $X \in \text{LN}(\alpha, \beta^2)$. Hvis massetætheden af partiklerne er ρ , vil vægten af en partikel være $Z = \gamma \cdot \rho \cdot X^3$, hvor γ er en konstant. Ifølge sætning 1.28 er $Z \in \text{LN}(\log_e(\gamma \cdot \rho) + 3\alpha, 9\beta^2)$, d.v.s. vægtfordelingen er igen logaritmisk normal. Dette resultat retfærdiggør den noget ubestemte vending "partikelstørrelse". ♦

I tilknytning til ovenstående indfører vi begrebet en momentfordeling i

DEFINITION 1.18 (MOMENTFORDELING). Lad den stokastiske variabel X have tætheden $f(x)$, som forudsættes at være 0 for $x < 0$. Ved den j 'te momentfordeling for X forstås da fordelingen med fordelingsfunktion

$$G_j(x) = \frac{1}{E(X^j)} \int_0^x t^j f(t) dt.$$

▲

Læg mærke til, at $G_j(x)$ er defineret for enhver tæthedsfunktion $f(x)$ og således, at den automatisk bliver en fordelingsfunktion med $G_j(\infty) = 1$.

Der gælder nu

SÆTNING 1.31. Lad $X \in \text{LN}(\alpha, \beta^2)$. Da er den j 'te momentfordeling for X en $\text{LN}(\alpha + j\beta^2, \beta^2)$ -fordeling.

Bevis. Beviset er ligefremt. Vi har

$$G_j(x) = c \int_0^x t^j \frac{1}{t} \exp\left(-\frac{1}{2\beta^2}[\log_e t - \alpha]^2\right) dt.$$

Da

$$\begin{aligned} & t^j \exp\left(-\frac{1}{2\beta^2}[\log_e t - \alpha]^2\right) \\ &= \exp\left(-\frac{1}{2\beta^2}(-2\beta^2 j \log_e t + [\log_e t - \alpha]^2)\right) \\ &= c_1 \exp\left(-\frac{1}{2\beta^2}[\log_e t - \alpha - j\beta^2]^2\right), \end{aligned}$$

hvor c_1 (og c) er konstanter uafhængige af t .

Vi ser nu, at integranden er proportional med frekvensfunktionen for en $\text{LN}(\alpha + j\beta^2, \beta^2)$ -fordeling. Heraf følger resultatet umiddelbart. Bemærk i øvrigt, at det i ovenstående ikke er nødvendigt at holde rede på konstanterne, da vi ved, at $G_j(\infty) = 1$. ■

EKSEMPEL 1.27. Ved analysen af kornede materialer som f.eks. sand anvender man ofte et sold med forskellige maskestørrelse. På denne måde kan man få separeret materialet i en række grupper svarende til forskellige diametre. Ved sådanne undersøgelser er man som regel interesseret i fordelingen af korndiameteren. Man kan derfor "blot" tælle antallet af partikler i hver klasse. På denne måde får man et histogram over kornfordelingen. Nu vil der jo ofte være et endog meget stort antal partikler i hver klasse, således at dette ikke er praktisk muligt.

I stedet kan man **veje** materialet i hver klasse og anvende vægte som udtryk for den relative hyppighed i de forskellige klasser. Lad der være n_i partikler i klassen med midtpunkt i x_i . Ved vejemetoden får vi "hyppigheden"

$$cx_i^3 n_i \simeq c_1 x_i^3 f(x_i),$$

hvor $f(x_i)$ er tætheden for fordelingen af korndiameteren. Vi ser nu, at vejemetoden i stedet for at give den søgte fordeling giver den tredie momentfordeling.

Hvis (tælle-)fordelingen af korndiametre er $\text{LN}(\alpha, \beta^2)$, bliver vægtfordelingen derfor en $\text{LN}(\alpha + 3\beta^2, \beta^2)$ -fordeling, og det fremgår, at det umiddelbart er muligt at komme fra karakteristika i den ene fordeling til karakteristika i den anden. ♦

Sluttelig må nævnes, at LN-fordelingen har været anvendt til beskrivelse af **elektricitetsforbruget** f.eks. i en by og til beskrivelse af **indkomstfordelinger**.

EKSEMPEL 1.28. Fra et elektricitetsværk oplyses, at middelforbruget i et døgn i en bydel er 5000 kWh, og standardafvigelsen er 2500 kWh. Man er interesseret i at erfare, hvad sandsynligheden er for, at forbruget i et enkelt døgn overstiger 10000 kWh. For at besvare dette spørgsmål må vi gøre en antagelse om **fordelingen** af forbruget. Vi **antager**, at forbruget kan beskrives ved en stokastisk variabel $X \in \text{LN}(\alpha, \beta^2)$, og vi søger $P\{X > 10000\}$. Vi bestemmer først α og β^2 . Sætning 1.27 giver umiddelbart

$$\begin{aligned}\alpha &= \log_e 5000 - \frac{1}{2} \log_e \left(\frac{625 \cdot 10^4}{25 \cdot 10^6} + 1 \right) = 8.40559 \\ \beta^2 &= \log_e \left(\frac{625 \cdot 10^4}{25 \cdot 10^6} + 1 \right) = 0.22314 = 0.472^2\end{aligned}$$

Vi har nu

$$\begin{aligned}P\{X > 10000\} &= P\{\log_e X > 4 \log_e 10\} \\ &= P\{N(8.40559, 0.22314) > 9.21034\} \\ &= P\{N(0, 1) > 1.70\} \\ &= 0.0446\end{aligned}$$

Under antagelsen om, at $X \in \text{LN}(\alpha, \beta^2)$, har vi altså fundet, at sandsynligheden for et forbrug på over 10000 kWh er 4.46%. $\blacklozenge$

1.7 Ekstremværdiproblemer

Statistiske problemer, som vedrører den mindste eller den største værdi blandt et antal stokastiske variable, benævnes ekstremværdi problemer. Sådanne problemer optræder hyppigt i praksis i forbindelse med bestemmelse af eksempelvis levetider, brudstyrker eller i forbindelse med vurdering af såkaldte 'outliers', dvs. observationer, som ligger betænkeligt langt fra de øvrige observationer i en stikprøve.

EKSEMPEL 1.29. Betragt en almindelig kæde, som består af n led. Kæden brydes, når det svageste led bryder. Det kan være rimeligt at antage, at brudstyrken for de enkelte led, X_i , $i = 1, 2, \dots, n$, er uafhængige stokastiske variable med fordelingsfunktion $F(x)$. Ordnes de n stokastiske variable (brudstyrken af kædens enkelte led) efter størrelse, benævnes den mindste $X_{(1)}$, den næstmindste $X_{(2)}$ osv, indtil den største, som benævnes $X_{(n)}$.

En kædes brudstyrke er således netop givet ved $X_{(1)}$. Hvis man ønsker at vurdere kædens brudstyrke, må man derfor tage i betragtning, at brudstyrken afhænger af, hvor mange led, der er i kæden, såvel som variationen mellem de enkelte led. Ekstremværdistatistikken beskæftiger sig hermed. ♦

1.7.1 Største og mindste observations fordeling

Lad nu generelt $X_1, \dots, X_n$ være uafhængige stokastiske variable med samme fordelingsfunktion F . Vi ordner de stokastiske variable efter stigende størrelse og får $X_{(1)}, \dots, X_{(n)}$. Her betegner $X_{(i)}$ altså den i 'te mindste af $X_1, \dots, X_n$.

DEFINITION 1.19. Ekstremværdierne i sættet $X_1, \dots, X_n$ er

$$\begin{aligned} X_{(1)} &= X_{n,(1)} = \min_{1 \leq i \leq n} X_i \\ X_{(n)} &= X_{n,(n)} = \max_{1 \leq i \leq n} X_i, \end{aligned}$$

hvor de dobbeltindicerede betegnelser anvendes, når der kan være tvivl om stikprøvestørrelsen n . ▲

SÆTNING 1.32. Fordelingsfunktionerne for $X_{(1)}$ og $X_{(n)}$ er givet, henholdsvis, ved

$$\begin{aligned} L_n(x) &= P\{X_{(1)} \leq x\} = 1 - (1 - F(x))^n \\ &\text{og} \\ H_n(x) &= P\{X_{(n)} \leq x\} = (F(x))^n. \end{aligned}$$

Bevis. Vi har

$$\begin{aligned}
 P\{X_{(1)} \leq x\} &= 1 - P\{X_{(1)} > x\} \\
 &= 1 - P\{X_1 > x, \dots, X_n > x\} \\
 &= 1 - (1 - F(x))^n, \\
 P\{X_{(n)} \leq x\} &= P\{X_1 \leq x, \dots, X_n \leq x\} = (F(x))^n.
 \end{aligned}$$

Beviset for $H_n(x)$ indses let ved at huske, at $F(x)$ angiver sandsynligheden for, at en enkelt værdi, X_i , er mindre end eller lig x . Sandsynligheden for, at alle værdierne i en stikprøve på n er mindre end eller lig x er da netop $(F(x))^n$. ■

På lignende måde kan man finde fordelingen af en vilkårlig af de ordnede observationer ved hjælp af binomialfordelingen.

Betragt en stikprøve $X_1, \dots, X_n$, og lad N_x betegne antallet af observationer, som er mindre end eller lig med x . Den stokastiske variabel N_x vil være binomialfordelt med parameteren $p = F(x)$. Som ovenfor betegner $X_{(i)}$ den i 'te mindste observation, og vi finder:

$$P\{X_{(i)} \leq x\} = P\{N_x \geq i\} = P\{B(n, F(x)) \geq i\}$$

På den måde kan den eksakte fordeling for en vilkårlig af de ordnede observationer – og herunder specielt $X_{(1)}$ og $X_{(n)}$ – beregnes ud fra den underliggende fordeling $F(x)$ ved hjælp af binomialfordelingen. I praksis får man dog ofte ret uhåndterlige udtryk.

EKSEMPEL 1.30. Man kan opskrive fordelingsfunktionen for $X_{(1)}$ og $X_{(n)}$ fra en $N(\mu, \sigma^2)$ -fordeling generelt:

$$L_n(x) = 1 - [1 - \Phi((x - \mu)/\sigma)]^n$$

$$H_n(x) = [\Phi((x - \mu)/\sigma)]^n$$

hvor $\Phi(\cdot)$ angiver $N(0, 1)$ fordelingsfunktionen:

$$\Phi(x) = \int_{-\infty}^x (\sqrt{2\pi})^{-1} \exp(-s^2/2) ds$$

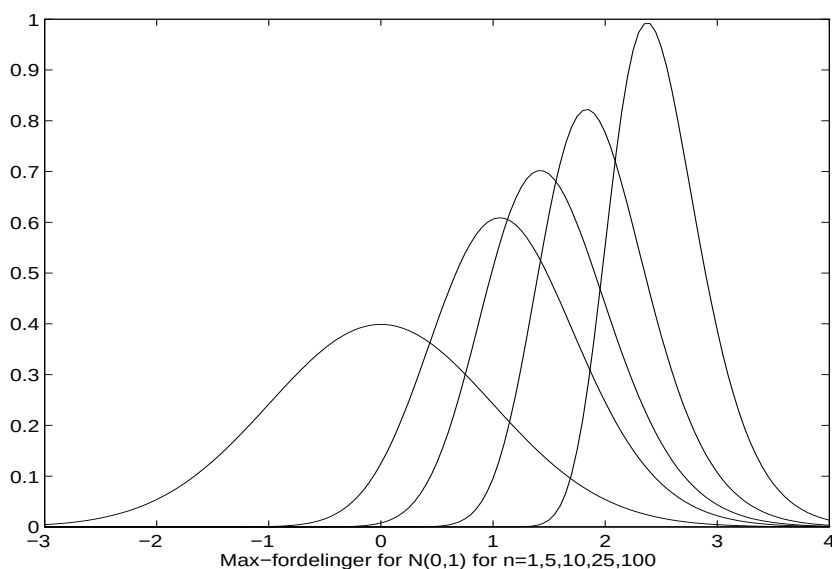
Ved differentiation mht x findes tæthedsfunktionen for $X_{(1)}$:

$$l_n(x) = n \cdot \varphi((x - \mu)/\sigma) \cdot [1 - \Phi((x - \mu)/\sigma)]^{n-1}$$

For størsteværdien $X_{(n)}$ fås tilsvarende

$$h_n(x) = n \cdot \varphi((x - \mu)/\sigma) \cdot [\Phi((x - \mu)/\sigma)]^{n-1}$$

hvor $\varphi(\cdot)$ angiver $N(0, 1)$ tætheden. ♦



EKSEMPEL 1.31. Vi fortsætter eksempel 1.29 og betragter en kæde med 10 led. De enkelte leds brudstyrker kaldes $X_i, i = 1, \dots, 10$. Antag videre, at de enkelte leds brudstyrker følger en normal fordeling med middelværdi $\mu = 110$ kp og varians $\sigma^2 = 5^2$ kp². Man kan nu spørge om sandsynligheden for, at en kæde med 10 led har en brudstyrke på mindst 100 kp.

Fordelingen af den mindste observation blandt n kan udtrykkes som angivet i sætning 1.32.

Benyttes nu $F(x) = N(110, 5^2)$, fås

$$P\{X_{(1)} \leq 100\} = 1 - (1 - F(100))^{10}$$

hvor

$$F(100) = \Phi\left(\frac{100 - \mu}{\sigma}\right) = \Phi\left(\frac{100 - 110}{5}\right) = 0.0228$$

idet værdien af $\Phi(\cdot)$ findes ved hjælp af en normalfordelingstabel.

$$P\{X_{(1)} \leq 100\} = 1 - (1 - 0.0228)^{10} = 0.2060$$

Dvs, at ca 20 % af alle kæder har brudstyrker under 100 kp.

Man kunne spørge om, hvor meget middelbrudstyrken af de enkelte led skal hæves for at sikre, at mindst 95% af alle kæder har en brudstyrke på mindst 100 kp. Kravet kan formuleres:

$$P\{X_{(1)} \leq 100\} = 1 - (1 - F(100))^{10} \leq 0.05 \Leftrightarrow F(100) \leq 0.0051$$

hvoraf fås

$$F(100) = \Phi\left(\frac{100 - \mu}{5}\right) \leq 0.0051$$

$$\Leftrightarrow \frac{100 - \mu}{5} \leq -2.57 \Leftrightarrow \mu \geq 112.85$$

De enkelte led skal altså have en brudstyrke på gennemsnitligt mindst ca 113 kp (under forudsætningen om $\sigma^2 = 5^2$). Hvis kæden er dobbelt så lang, dvs med $n = 20$ led, ville kravet til den gennemsnitlige brudstyrke for det enkelte led blive ca 114 kp. ♦

Vi giver endnu et konkret eksempel på anvendelse af en ekstremværdifordeling.

EKSEMPEL 1.32. Eksemplet vedrører levetiden af et system bestående af et antal ens komponenter, hvor levetiden for de enkelte komponenter antages at være eksponentialfordelt. Tænker man sig, at komponenterne er monteret parallelt således, at systemet virker så længe, blot én af komponenterne virker, vil systemets levetid være levetiden for den længstlevende komponent. Et sådant system kaldes et *parallel-redundant* system. Vi ønsker at finde fordeling for levetiden for den længstlevende komponent blandt de n komponenter.

Vi antager altså $X_i \in \text{Ex}(\beta)$. Da er

$$F(x) = 1 - e^{-x/\beta} \quad (1.17)$$

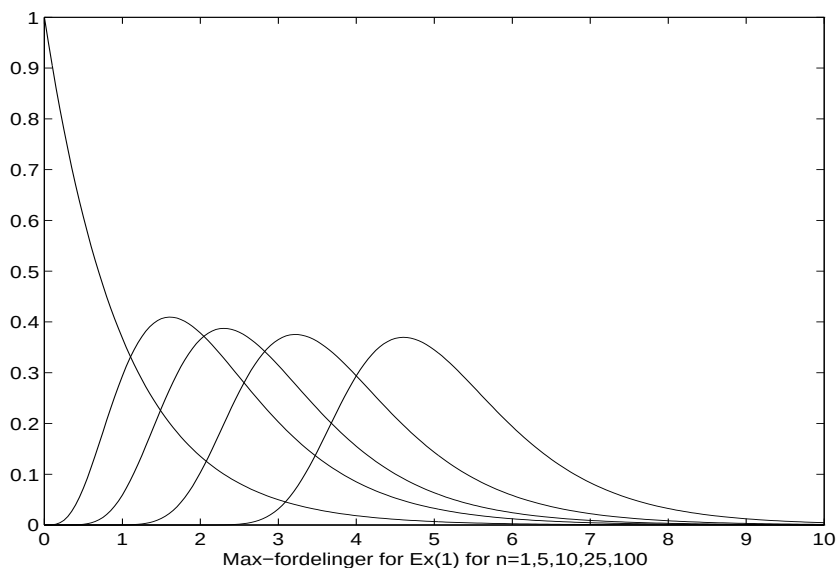
hvorfor

$$H_n(x) = \left(1 - e^{-x/\beta}\right)^n \quad (1.18)$$

Tætheden findes ved differentiation af $H_n(x)$ m.h.t. x :

$$h_n(x) = \frac{n \cdot e^{-x/\beta}}{\beta} \left(1 - e^{-x/\beta}\right)^{n-1} \quad (1.19)$$

◆



Vi har i eksemplerne 1.30 og 1.32 set, hvorledes fordelingen for den største observation i en stikprøve rykker mod højre, når stikprøven øges. Imidlertid vil det under ret generelle antagelser være sådan, at såvel L_n som H_n for $n \rightarrow \infty$ nærmer sig bestemte enkle fordelinger, der kaldes *asymptotiske fordelinger* for ekstremværdier. Disse generelle fordelinger skal vi beskæftige os med i det følgende.

1.7.2 Asymptotiske ekstremværdifordelinger

EKSEMPEL 1.33. Vi fortsætter eksempel 1.32. Det ses umiddelbart, at for fast x er $\lim H_n(x) = 0$, $n \rightarrow \infty$. Det viser, at fordelingen af $X_{(n)}$ flytter sig mod højre, når $n \rightarrow \infty$.

Ser vi imidlertid i stedet på størrelsen $Y_{(n)} = \frac{1}{\beta}(X_{(n)} - \beta \log_e n)$, fås

$$\begin{aligned} P\{Y_{(n)} \leq y\} &= P\left\{\frac{X_{(n)} - \beta \log_e n}{\beta} \leq y\right\} \\ &= P\{X_{(n)} \leq \beta y + \beta \log_e n\} \\ &= (1 - e^{-y - \log_e n})^n \\ &= \left(1 - \frac{1}{n}e^{-y}\right)^n \end{aligned}$$

Endelig kan man vise, at

$$\left(1 - \frac{1}{n}e^{-y}\right)^n \rightarrow \exp(-e^{-y}), \quad n \rightarrow \infty.$$

Vi har hermed fundet nogle konstanter, $\alpha_n = \beta \log_e n$ og $\beta_n = \beta$, således, at den transformerede størrelse $Y_{(n)} = (X_{(n)} - \alpha_n)/\beta_n$ har en grænsefordeling for $n \rightarrow \infty$. Det vil konkret sige, at

$$P\{X_{(n)} \leq x\} = H_n(x) \rightarrow \exp(-\exp(-(x - \alpha_n)/\beta_n))$$

når $n \rightarrow \infty$.

Man ser, at den asymptotiske fordeling af $X_{(n)}/\beta_n$ flytter sig "mod højre" i takt med $(\log_e n)/\beta_n$. ♦

I det følgende vil vi generalisere det i eksempel 1.33 opnåede resultat. Dvs at vi vil angive de forudsætninger, F skal tilfredsstille, for at der eksisterer en følge (α_n, β_n) af konstanter og en fordelingsfunktion G , så

$$\lim_{n \rightarrow \infty} P\left\{\frac{X_{(n)} - \alpha_n}{\beta_n} \leq x\right\} = G(x), \quad (1.20)$$

På tilsvarende måde kan der defineres en grænsefordeling for $X_{(1)}$, og det noteres, at de generelle resultater, som kan udledes for fordelingen af størsteværdier, umiddelbart kan overføres til mindsteværdier (se afsnit 1.7.4 side 156).

Vi anfører nu det væsentligste resultat mht bestemmelse af grænsefordelingen for størsteværdien $X_{(n)}$. Resultatet skyldes Fisher og Tippett samt Gnedenko [22].

SÆTNING 1.33. Lad $X_1, X_2, \dots$ være en følge af uafhængige stokastiske variable. Hvis der eksisterer en følge af konstanter, (α_n, β_n) og en asymptotisk fordelingsfunktion G , således, at (1.20) er opfyldt, da vil funktionen, G , være en af følgende tre typer

$$\begin{aligned} i) \quad \Lambda_1(x) &= \exp(-e^{-x}) \\ ii) \quad \Lambda_{2,k}(x) &= \begin{cases} \exp(-x^{-k}) & x > 0 \\ 0 & x < 0 \end{cases} \\ iii) \quad \Lambda_{3,k}(x) &= \begin{cases} \exp(-(-x)^k) & x < 0 \\ 1 & x > 0 \end{cases} \end{aligned}$$

idet k er en positiv konstant.

Bevis. Se [22]. ■

Fordelinger, som leder til en af de tre mulige asymptotiske ekstremværdifordelinger siges at være af 1): **Eksponentiel type**, 2): **Cauchy type** eller 3): **Tredie type**.

De tre mulige asymptotiske ekstremværdifordelinger, $\Lambda_1(x)$, $\Lambda_2(x)$ og $\Lambda_3(x)$, kaldes også **Max₁**, **Max₂** og **Max₃**, henholdsvis.

For en given fordelingsfunktion F skal vi nedenfor anføre tilstrækkelige betingelserne for, at G eksisterer, og at G er en af de tre anførte typer.

Vi bemærker, at der i sætningen **ikke** står, at fordelingen af maximumsværdien altid kan approksimeres ved en af de 3 typer. Der findes fordelinger, for hvilke (1.20) ikke er opfyldt for noget G , uanset valget af (α_n, β_n) .

1.7.3 Maximumsfordeling for eksponentiel type

Vi indleder med

DEFINITION 1.20. Lad der være givet en fordelingsfunktion F , der er < 1 for alle x , og lad F have en differentiabel tæthed f . Vi siger da, at F er af **eksponentiel type** med

hensyn til maximumsfordeling, hvis

$$\lim_{x \rightarrow \infty} \frac{d}{dx} \left(\frac{1 - F(x)}{f(x)} \right) = 0 \quad (1.21)$$

▲

Der gælder nu følgende hovedresultat:

SÆTNING 1.34. Lad $X_1, X_2, \dots$ være en følge af uafhængige, identisk fordelte stokastiske variable med fordelingsfunktion F . Hvis F er af eksponentiel type, da vil for $n \rightarrow \infty$

$$P \left\{ \frac{X_{(n)} - \alpha_n}{\beta_n} \leq x \right\} \rightarrow \Lambda_1(x) = \exp(-e^{-x}),$$

hvor α_n og β_n kan bestemmes af

$$\begin{aligned} F(\alpha_n) &= 1 - \frac{1}{n} \\ \beta_n &= \frac{1}{nF'(\alpha_n)}. \end{aligned}$$

Bevis. Se [13][p. 205].

■

BEMÆRKNING 1.7. α_n og β_n er ikke entydigt bestemte. Andre værdier end de i sætning 1.34 definerede kan komme på tale.

▼

Forklaringen på navnet "eksponentiel type" fås af

EKSEMPEL 1.34. Lad F være en $\text{Ex}(\beta)$ -fordelingsfunktion, dvs

$$F(x) = 1 - e^{-x/\beta}$$

Da er

$$\frac{1 - F(x)}{f(x)} = \frac{e^{-x/\beta}}{\frac{1}{\beta}e^{-x/\beta}} = \beta,$$

og trivielt

$$\frac{d}{dx} \frac{1 - F(x)}{f(x)} = 0,$$

således, at grænseovergangen i betingelsen (1.21) er opfyldt for eksponentialfordelingen. ♦

Normalfordelingen er endnu et eksempel på en fordeling af eksponentiel type. Her er grænseovergangen (1.21) dog mere kompliceret.

EKSEMPEL 1.35. Lad $X \in N(0, 1)$. Da er

$$\begin{aligned} \frac{1 - F(x)}{f(x)} &= \frac{\int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt}{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}}, \\ \text{d.v.s.} \\ \frac{d}{dx} \frac{1 - F(x)}{f(x)} &= e^{x^2} \left(-e^{-\frac{1}{2}x^2} e^{-\frac{1}{2}x^2} + x e^{-\frac{1}{2}x^2} \int_x^\infty e^{-\frac{1}{2}t^2} dt \right) \\ &= -1 + x e^{\frac{1}{2}x^2} \int_x^\infty e^{-\frac{1}{2}t^2} dt. \end{aligned}$$

Vi betragter udtrykket

$$x e^{\frac{1}{2}x^2} \int_x^\infty e^{-\frac{1}{2}t^2} dt = \frac{\int_x^\infty e^{-\frac{1}{2}t^2} dt}{\frac{1}{x} e^{-\frac{1}{2}x^2}} = \frac{f(x)}{h(x)}.$$

Det er klart, at såvel f og $h \rightarrow 0$ for $x \rightarrow \infty$. Vi anvender derfor l'Hôpital's regel og får

$$\frac{f'(x)}{h'(x)} = \frac{-e^{-\frac{1}{2}x^2}}{-\frac{1}{x^2} e^{-\frac{1}{2}x^2} - e^{-\frac{1}{2}x^2}} = \frac{1}{\frac{1}{x^2} + 1}.$$

Denne størrelse har klart grænseværdien 1 for $x \rightarrow \infty$, og ifølge l'Hôpital's regel har da også $f(x)/h(x)$ grænseværdien 1. Derfor vil

$$\frac{d}{dx} \frac{1 - F(x)}{f(x)} \rightarrow 0, \quad x \rightarrow \infty,$$

hvormed vi har vist, at en $N(0, 1)$ -fordeling er af eksponentiel type. $\blacklozenge$

Det vises let, at hvis $X \in F(x)$, som tilhører den eksponentielle type, da vil, eksempelvis, $(X - \mu)/\sigma \in F(\frac{x-\mu}{\sigma})$, og $F(\frac{x-\mu}{\sigma})$ vil også tilhøre typen. Hermed har vi vist, at en generel $N(\mu, \sigma^2)$ -fordeling er indeholdt i den eksponentielle type.

På samme måde kan man vise, at f.eks. log-normalfordelingen, $LN(\alpha, \beta^2)$, og gammafordelingen, $G(k, \beta)$, også er indeholdt i den eksponentielle type med hensyn til maximumsfordeling (men ikke med hensyn til minimumsfordeling).

DEFINITION 1.21. Vi siger, at en stokastisk variabel $X \in \text{Max}_1(0, 1)$, hvis X har fordelingsfunktionen $\Lambda_1(x)$. Fordelingen med positionsparameter α og skalaparameter β benævnes $\text{Max}_1(\alpha, \beta)$. Dvs, at

$$X \in \text{Max}_1(\alpha, \beta) \Leftrightarrow \frac{X - \alpha}{\beta} \in \text{Max}_1(0, 1)$$

$\blacktriangle$

Hvis man udtager et antal (n) stikprøver fra en Max_1 -fordeling og deriblandt igen udtager den største, vil denne igen - ikke uventet - følge en Max_1 -fordeling. Dette udtrykkes i

SÆTNING 1.35. Klassen Max_1 er afsluttet over for maximumsdannelse, idet der for uafhængige variable gælder

$$X_i \in \text{Max}_1(\alpha, \beta), \quad i = 1, \dots, n, \quad \Rightarrow X_{(n)} \in \text{Max}_1(\alpha + \beta \log_e n, \beta).$$

Bevis. Beviset er ligefremt. Vi finder

$$\begin{aligned} & P \left\{ \frac{X_{(n)} - \alpha - \beta \log_e n}{\beta} \leq x \right\} \\ &= P \{ X_{(n)} \leq \alpha + \beta(x + \log_e n) \} \\ &= \Lambda_1^n \left(\frac{\alpha + \beta(x + \log_e n) - \alpha}{\beta} \right) \\ &= \Lambda_1^n(x + \log_e n) = [\exp(-e^{-x - \log_e n})]^n \\ &= \exp(-e^{-x}) = \Lambda_1(x). \end{aligned}$$

$\blacksquare$

De analytiske egenskaber ved Max_1 -fordelinger fremgår af

SÆTNING 1.36. Lad $X \in \text{Max}_1(\alpha, \beta)$. Da har X tætheden

$$f(x) = \frac{1}{\beta} \lambda_1 \left(\frac{x - \alpha}{\beta} \right) = \frac{1}{\beta} \left(-\frac{x - \alpha}{\beta} - \exp \left(-\frac{x - \alpha}{\beta} \right) \right),$$

hvor $\lambda_1(x) = \Lambda_1'(x)$. Middelværdien er

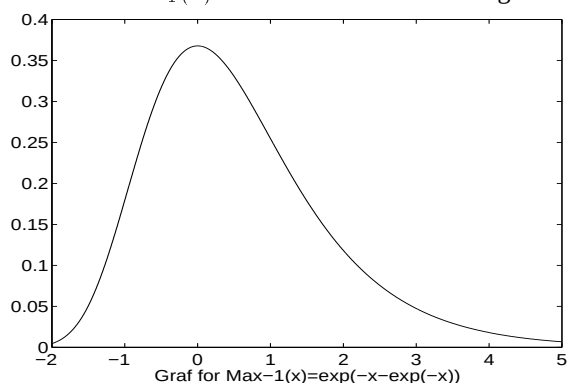
$$E(x) = \alpha + \beta\gamma, \quad \gamma = 0.57722,$$

hvor γ her angiver Eulers konstant. Variansen er

$$V(x) = \frac{\beta^2 \pi^2}{6}.$$

Bevis. Forbigås. ■

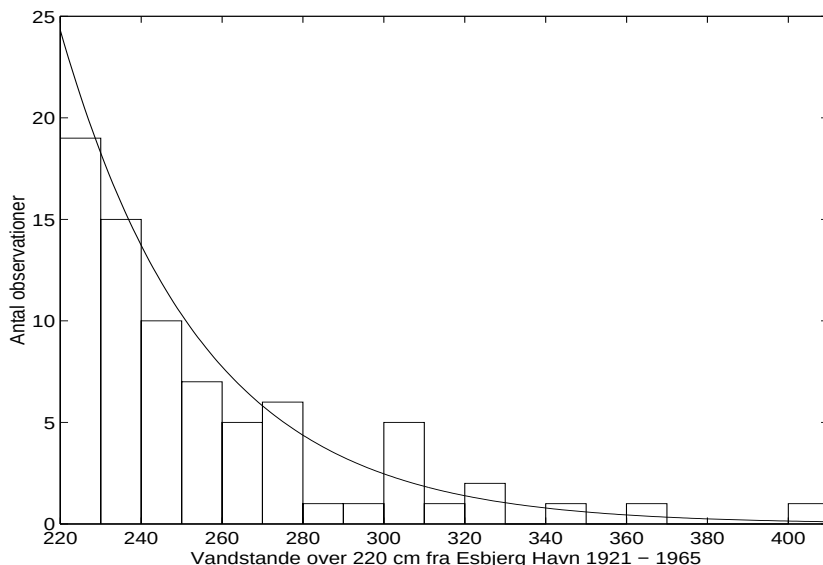
Grafen for tæthedsfunktionen $\lambda_1(x)$ er anført i nedenstående figur.



EKSEMPEL 1.36 (EKSTREME VANDSTANDE). I tabellen side 154 er anført målte maksimale vandstandshøjder fra perioder med særligt høje vandstande. Kun de perioder, hvor vandstanden nåede over 220 cm over det normale ved Esbjerg havn er medtaget. I nedenstående figur er tegnet et histogram svarende til disse målinger. Endvidere er anført en estimeret tæthedsfunktion ($\times$ antal obs. $\times$ klassebredde) givet ved

$$f(x) = \frac{1}{35} \exp \left(-\frac{x - 220}{35} \right), \quad x > 220, \quad (1.22)$$

og 0 ellers. Der ses at være en god overensstemmelse, og vi vil derfor regne med, at vandstandshøjder på mindst 220 cm over normalen kan beskrives ved en stokastisk variabel X med tætheden (1.22)



Vi ser endvidere, at sådanne ekstreme vandstande optræder med hyppighed på gennemsnitligt ca. 1.7 pr. år. Eksempelvis ved opførelsen af diger til afværgelse af oversvømmelser rejser der sig nu naturligt spørgsmålet: Hvad er sandsynligheden for, at vandstanden vil overstige en given højde (f.eks. 6 m) inden for en bestemt tidsperiode (f.eks. 50 år), og hvad er forventningsværdien af den højeste vandstand i en sådan periode?

Inden for en 50 års periode vil vi forvente $50 \cdot 1.7 = 85$ observationer af vandstandshøjder på mindst 220 cm over normalen. Dette tal vil vi for nemheds skyld regne for givet (selv om det naturligvis er behæftet med usikkerhed).

I den matematiske model har vi altså identisk fordelte stokastiske variable $X_1, \dots, X_n$ ($n = 85$), hvor X_i har tætheden (1.22). Vi forudsætter, at disse er stokastisk uafhængige - en forudsætning, der forekommer yderst rimelig. Vi er interesseret i fordelingen af $X_{(n)}$ for $n = 85$, og vi vil anvende sætning 1.34. Vi bestemmer først α_n og β_n :

$$F(\alpha_n) = 1 - \exp\left(-\frac{\alpha_n - 220}{35}\right) = 1 - \frac{1}{n}$$

Dato			Vand-stand			Dato			Vand-stand			Dato			Vand-stand		
18/12	1921	226	5/12	1936	226	13/11	1954	246									
21/12	21	261	30/1	38	262	5/12	54	232									
4/11	23	247	5/10	39	261	23/12	54	255									
16/11	23	265	24/11	38	360	13/1	55	258									
10/9	24	221	27/11	39	284	29/12	55	255									
3/1	25	274	3/12	39	224	21/1	56	319									
31/12	25	230	1/11	40	222	13/12	56	236									
9/10	26	309	13/11	40	230	19/10	57	265									
16/11	26	237	8/12	41	233	4/11	57	238									
29/10	27	232	16/10	42	258	8/12	57	224									
17/11	28	230	13/2	43	220	23/1	57	222									
24/11	28	400	16/2	43	241	19/10	59	244									
12/12	28	325	23/1	44	274	21/12	59	241									
12/12	29	244	5/11	44	229	5/12	60	275									
30/12	29	221	21/1	49	237	29/1	61	225									
13/1	30	232	10/2	49	276	6/12	61	292									
23/11	30	247	24/10	49	276	23/1	62	233									
24/1	31	227	26/10	49	307	12/2	62	248									
6/12	31	252	20/12	49	229	16/2	62	321									
7/1	32	224	9/12	51	232	27/9	63	224									
11/10	33	253	31/12	51	222	13/11	63	273									
23/9	34	255	16/1	52	232	17/11	64	221									
18/10	36	303	1/2	53	224	13/12	64	234									
27/10	36	343	16/1	54	306	14/2	65	240									
1/12	36	244	17/9	54	226	2/11	65	308									

Tabel 1.2: Vandstandshøjder større end 220 cm ved Esbjerg.

dvs.

$$\begin{aligned}\alpha_n &= 220 + 35 \log_e n \\ \beta_n &= \left(n \frac{1}{35} \exp\left(-\frac{220 + 35 \log_e n - 220}{35}\right) \right)^{-1} = 35.\end{aligned}$$

For $n = 85$ fås altså

$$\alpha_{85} = 220 + 35 \cdot 4.44265 = 375.49 \text{ og } \beta_{85} = 35.$$

Vi har derfor, idet $n = 85$:

$$\begin{aligned}P\{X_{(n)} \leq 600\} &= P\left\{\frac{X_{(n)} - \alpha_n}{\beta_n} \leq \frac{600 - 375.49}{35}\right\} \\ &= P\left\{\frac{X_{(n)} - \alpha_n}{\beta_n} \leq 6.41\right\} \\ &\simeq \exp(-\exp(-6.41)) \\ &= 0.9983,\end{aligned}$$

dvs. med $n = 85$ er

$$P\{X_{(85)} > 600\} \simeq 0.17\%.$$

Sandsynligheden for, at der i en 50-års periode optræder en vandstand, der er større end eller lig en eventuel påtænkt digehøjde på 6 meter, skønnes derfor til ca. 0.17%.

Forventningsværdien af $X_{(n)}$ for $n = 85$ er

$$\begin{aligned}E(X_{(n)}) &= \alpha_n + \beta_n \gamma \simeq 375.49 + 35 \cdot 0.57722 \\ &= 395.69,\end{aligned}$$

d.v.s. ca. 396 cm. Variansen er

$$V(X_{(n)}) = \frac{\beta_n^2 \pi^2}{6} \simeq 2015.04 = (44.9 \text{ cm})^2$$

Standardafvigelsen er altså ca. 45 cm.

Anvendeligheden af ovenstående resultater bygger på, at konvergens mod ekstremværdifordelingen Λ_1 sker ret hurtigt for eksponentialfordelingen. Til sammenligning kan anføres det eksakte resultat

$$E(X_{(n)}) = 220 + 35 \left(1 + \frac{1}{2} + \cdots + \frac{1}{n} \right),$$

hvilket for $n = 85$ er

$$E(X_{(85)}) = 220 + 35 \left(1 + \frac{1}{2} + \cdots + \frac{1}{85} \right) = 395.90 \quad ,$$

d.v.s. en afvigelse på kun 2.1 mm. ♦

1.7.4 Minimumsfordeling for eksponentiel type

Der gælder selvfølgelig ganske analoge resultater for minimumsværdier som for maksimumsværdier. Minimumsfordelingernes egenskaber kan findes ved at betragte maksimumsfordelingen for følgen $-X_1, -X_2, \dots$, idet vi jo har

$$\min\{X_1, \dots, X_n\} = -\max\{-X_1, -X_2, \dots, -X_n\}.$$

Uden nærmere kommentarer giver vi nu resultaterne svarende til den eksponentielle type.

SÆTNING 1.37. Lad $X_1, X_2, \dots$ være en uafhængige stokastiske variable med fordelingsfunktionen F og differentiabel tæthed f . Vi forudsætter, at $F(x) > 0$ for alle x . Hvis

$$\lim_{x \rightarrow -\infty} \left(\frac{d F(x)}{dx f(x)} \right) = 0,$$

da vil

$$P \left\{ \frac{X_{(1)} - \gamma_n}{\delta_n} \leq x \right\} \xrightarrow{n \rightarrow \infty} 1 - \Lambda_1(-x) = 1 - \exp(-e^{-x}),$$

hvor konstanterne γ_n og δ_n bestemmes af

$$\begin{aligned} F(\gamma_n) &= \frac{1}{n} \\ \delta_n &= \frac{1}{nf(\gamma_n)}. \end{aligned}$$

Bevis. Fremgår af sætning 1.34 ved at anvende den ovenfor beskrevne fremgangsmåde. ■

Analogt til definitionen af Max_1 -fordelingen for størsteværdien (se p. 151) definerer vi klassen Min_1 ved

DEFINITION 1.22. Vi siger, at

$$X \in \text{Min}_1(\alpha, \beta) \stackrel{d}{\Leftrightarrow} F(x) = 1 - \Lambda_1\left(-\frac{x - \alpha}{\beta}\right).$$

▲

Vi får da umiddelbart

SÆTNING 1.38. Klassen Min_1 er afsluttet over for minimumsdannelse. Der gælder nemlig

$$X_i \in \text{Min}_1(\alpha, \beta), \quad i = 1, \dots, n \Rightarrow X_{(1)} \in \text{Min}_1(\alpha - \beta \log_e n, \beta),$$

hvor det stadig er forudsat, at $X_1, \dots, X_n$ er stokastisk uafhængige.

Bevis. Ligeftremt. Forbigås. ■

Endvidere gælder

SÆTNING 1.39. I ovenstående parametrisering er α og β positions- og skalaparametre. Tætheden for $X \in \text{Min}_1(\alpha, \beta)$ er

$$f(x) = \frac{1}{\beta} \lambda_1\left(-\frac{x - \alpha}{\beta}\right) = \frac{1}{\beta} \exp\left(\frac{x - \alpha}{\beta} - \exp\left(\frac{x - \alpha}{\beta}\right)\right).$$

Middelværdi og varians er

$$\begin{aligned} E(x) &= \alpha - \beta\gamma \text{ hvor } \gamma = 0.57722 \\ V(x) &= \frac{\beta^2 \pi^2}{6}. \end{aligned}$$

Bevis. Umiddelbar følge af sætning 1.36. ■

EKSEMPEL 1.37 (BRUDSTYRKE). Vi antager, at grunden til, at der observeres forskelle mellem den beregnede styrke af et materiale og den observerede styrke, skyldes tilstedeværelsen af "fejl", som svækker det pågældende emne. Dette betyder, at der skal anvendes forskellige (i størrelse) kræfter til at bryde emnet forskellige steder. Vi antager nu, at "fejlene" er tilfældigt fordelt i emnet (jfr. p. 122). Det er da klart, at brudstyrken bestemmes af det svageste punkt i emnet, eller med andre ord, brudstyrken bestemmes som minimumsværdien af n stokastiske variable, hvor n altså er antallet af fejl i emnet. Hvis man kan antage, at fordelingen af styrken et bestemt sted i emnet er af den eksponentielle type, da vil styrken af hele emnet følge ekstremværdifordelingen $\text{Min}_1(\alpha, \beta)$. ♦

Det er klart, at ovenstående anvendelse af princippet om **det svageste led**, som a vi indledte dette afsnit med i det simple eksempel om styrken af en kæde med n led, finder anvendelse i mange andre sammenhænge. Man kan f.eks. nævne problemet om en metaltråds **elektriske ledningsevne** eller en plades evne til at **bremse (lys-)stråler** etc. Eller det kan være **levetiden** (eller **tiden mellem reparationer**) af et system, der svigter, hvis en enkelt af en række vitale dele bryder sammen.

EKSEMPEL 1.38 (GENNEMSLAGSSPÆNDING AF KONDENSATORER). Nedenstående data stammer fra målinger af den spænding, hvorved nogle papirkondensatorer bryder sammen (kortsluttes), se [17]. Elektroderne havde et areal på 337.5 sq.in., og papiret var olieimprægneret 0.4-mil papir.

Vi vil forsøge at formulere en model. Det er rimeligt at forsøge at anvende princippet om det svageste led. Fejlene viser sig ved, at der optræder ledende partikler. Hvis fordelingen af størrelsen af en sådan partikel er af eksponentiel type, da vil også fordelingen af den dielektriske "styrke" i omegnen af partiklen være af eksponentiel type. Dette forekommer rimeligt at antage, og det er eksperimentelt godtgjort. Man må derfor forvente, at den i 'te måling kan beskrives ved en stokastisk variabel $X_i \in \text{Min}_1(\alpha, \beta)$ for et eller andet valg af α og β .

Den enkleste metode til at bestemme α og β på er som i foregående eksempel at beregne stikprøvens gennemsnit og empiriske varians og sætte disse lig med den valgte

Sammenbrudsspænding (i Volt)									
1500	1620	1620	1740	1740	1680	1740	1620	1740	1440
1740	1620	1740	1620	1440	1620	1620	1800	1800	1740
1500	1620	1800	1740	1650	1620	1680	1680	1680	1680
1440	1380	1740	1740	1500	1680	1800	1440	1860	1620
1620	1740	1680	1740	1680	1440	1740	1800	1680	1500
1620	1500	1500	1500	1740	1560	1440	1680	1740	1800
1860	1740	1680	1740	1740	1620	1620	1740	1740	1560
1740	1560	1740	1740	1740	1620	1680	1740	1800	1740
1560	1620	1680	1800	1500	1620	1620	1800	1800	1440
1560	1440	1740	1740	1680	1740	1680	1680	1740	1560
1680	1680	1740	1620	1740	1740	1740	1740	1320	1740
1620	1680	1500	1740	1500	1740	1740	1740	1680	1800
1620	1740	1560	1740	1680	1740	1680	1620	1740	1440
1680	1680	1740	1800	1740	1080	1620	1740	1740	1380
1620	1620	1620	1740	1800	1680	1620	1740	1800	1500
1740	1620	1500	1620	1620	1800	1800	1740	1800	1620
1620	1740	1560	1680	1680	1620	1740	1680	1200	1440
1740	1560	1380	1560	1500	1740	1620	1800	1740	1740
1680	1680	1620	1800	1800	1620	1740	1740	1800	900
1620	1380	1800	1380	1560	1740	1740	1680	1500	1740
1680	1620	1500	1740	1800	1620	1500	1260	1560	1740
1800	1680	1740	1680	1800	1800	1800	1560	1680	1560
1560	1560	1680	1680	1680	1740	1740	1800	1620	1620
1860	1560	1740	1620	1740	1620	1620	1680	1620	1740
1800	1740	1680	1500	1800	1800	1680	1740	1440	1800

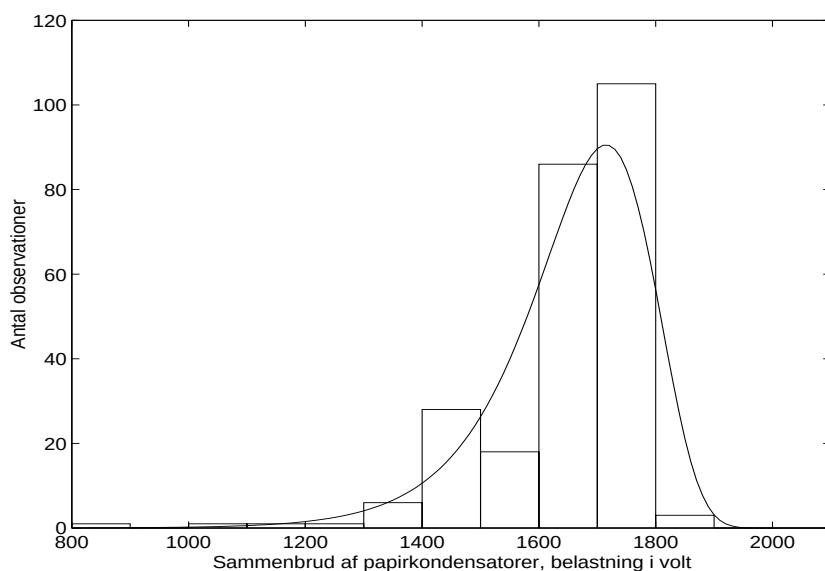
fordelings teoretiske middelværdi og varians. For de viste data kunne man altså sætte

$$E\{X\} = \alpha - \beta \cdot 0.57722 \simeq \bar{x} = \frac{1}{250} \sum_{i=1}^{250} x_i = 1656 ,$$

$$V\{X\} = \beta^2 \pi^2 / 6 \simeq s_*^2 = \frac{1}{250} \sum_{i=1}^{250} (x_i - \bar{x})^2 = 16960$$

Sammenhængen mellem parametrene α og β og middelværdi og varians for Min_1 -fordelingen er givet i sætning 1.39. Dermed findes de to skøn:

$$\alpha \simeq 1714.9 \text{ og } \beta \simeq 101.54$$



I ovenstående figur har vi anført et histogram over målingerne og indtegnet tætheden for en $\text{Min}_1(1714.9, 101.54)$ -fordeling $\times$ antal obs. $\times$ klassebredde. Overensstemmelse mellem de to funktioner er god, og den valgte model synes derfor velegnet til at beskrive det observerede fænomen. ♦

Det væsentlige, man skal hæfte sig ved med dette eksempel, er, at man som grundlæggende model for de observerede data antager, at der er tale om et minimumsværdiproblem. Derfor vælger man som modelfordeling en fordeling, der generelt er egnet til at

beskrive fordelingen af den mindste blandt mange, og det er en Min_1 -fordeling netop et eksempel på. Modsætningsvis egner en normalfordeling sig bestemt ikke som modelfordeling for ekstreme værdier. Den egner sig derimod til at beskrive fænomener, som er udtryk for en akkumulering af mange ikke stærkt forskellige bidrag (den centrale grænseværdisætning).

1.7.5 Fordelinger af Cauchy type

Ovenfor har vi beskæftiget os med de asymptotiske fordelinger af eksponentiel type. Som anført i sætning 1.33, kan der være tale om yderligere to typer. En nærmere undersøgelse af betingelsen (1.21) i definition 1.20 viser, at denne i det store og hele går ud på, at $1 - F(x)$ skal gå mod 0 lige så hurtigt som en eksponentialfunktion. I denne paragraf vil vi betragte fordelinger F , hvor $1 - F(x)$ går mod 0 omtrent som en potensfunktion. Vi specificerer dette i

DEFINITION 1.23. En fordelingsfunktion F siges at være af **Cauchy type**, hvis $F(x) < 1$ for alle x , og hvis der eksisterer et $k > 0$ således, at vi for ethvert $a > 0$ har

$$\frac{1 - F(x)}{1 - F(ax)} \rightarrow a^k, \quad x \rightarrow \infty.$$

▲

Vi viser nu, at en fordeling, der nærmer sig 1 som en potensfunktion, er af Cauchy type.

SÆTNING 1.40. Lad fordelingen F være givet ved

$$F(x) = 1 - cx^{-k}(1 + \varepsilon(x)),$$

hvor c og k positive konstanter, og $\varepsilon(x) \rightarrow 0$, $x \rightarrow \infty$. Da er F af Cauchy type.

Bevis. Beviset er en triviell følge af definitionen. Vi har for $a > 0$

$$\frac{1 - F(x)}{1 - F(ax)} = \frac{cx^{-k}(1 + \varepsilon(x))}{c(ax)^{-k}(1 + \varepsilon(ax))} = a^k \frac{1 + \varepsilon(x)}{1 + \varepsilon(ax)}.$$

Det ses umiddelbart, at denne størrelse har grænseværdien a^k for $x \rightarrow \infty$. ■

Ekstremværdifordelingen for denne type af variable fremgår af den næste sætning.

SÆTNING 1.41. Lad $X_1, X_2, \dots$ være en følge af uafhængige, identisk fordelte stokastiske variable, hvis fordelingsfunktion F er af Cauchy type. Da vil

$$P \left\{ \frac{X_{(n)}}{\beta_n} \leq x \right\} \rightarrow \Lambda_{2,k}(x) = \exp(-x^{-k}),$$

for $n \rightarrow \infty$. Her er k det samme k som det i definitionen anførte, og β_n er givet ved

$$F(\beta_n) = 1 - \frac{1}{n}.$$

Bevis. Se f.eks. [22]. ■

BEMÆRKNING 1.8. β_n er ikke entydigt bestemt, jfr. bemærkningen side 149. ▼

DEFINITION 1.24. Vi siger, at $X \in \text{Max}_2(k, 1)$, hvis X har fordelingsfunktionen $\Lambda_{2,k}(x)$. Fordelingen med skalaparameter β benævnes $\text{Max}_2(k, \beta)$. ▲

I fortsættelse af sætning 1.41 anfører vi

SÆTNING 1.42. Klassen Max_2 er afsluttet over for maximumsdannelse, idet der for uafhængige variable gælder

$$X_i \in \text{Max}_2(k, \beta), \quad i = 1, \dots, n \Rightarrow X_{(n)} \in \text{Max}_2\left(k, n^{1/k}\beta\right).$$

Bevis. Resultatet fremgår ved direkte regning. ■

De analytiske egenskaber ved Max_2 -fordelingen fremgår af

SÆTNING 1.43. Lad $X \in \text{Max}_2(k, \beta)$. Da har X tætheden

$$f(x) = \frac{1}{\beta} \lambda_{2,k}\left(\frac{x}{\beta}\right) = \frac{k}{\beta} \left(\frac{\beta}{x}\right)^{k+1} \exp\left(-\left(\frac{\beta}{x}\right)^k\right),$$

for $x > 0$ og 0 ellers. Her er $\lambda_{2,k}(x) = \Lambda'_{2,k}(x)$. Middelværdien i fordelingen er

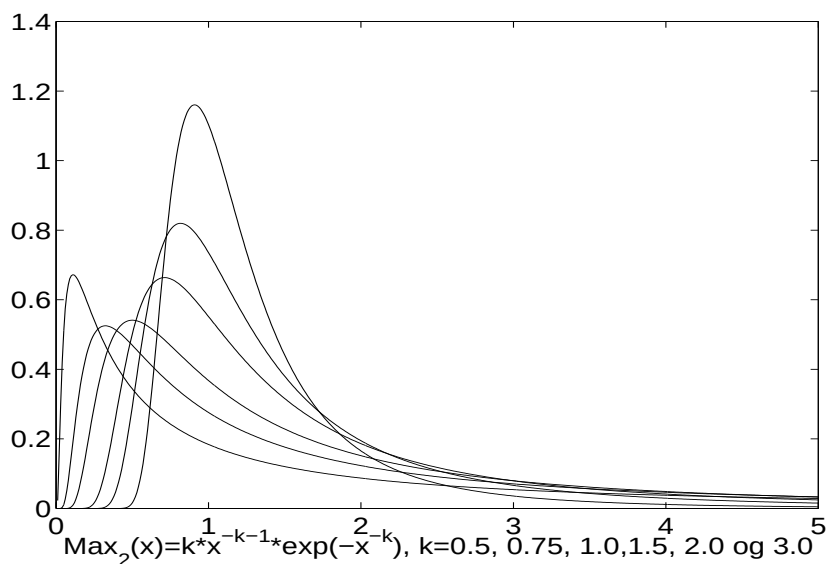
$$E(X) = \beta \Gamma\left(1 - \frac{1}{k}\right), \quad k > 1$$

og variansen

$$V(X) = \beta^2 \left(\Gamma\left(1 - \frac{2}{k}\right) - \Gamma^2\left(1 - \frac{1}{k}\right) \right), \quad k > 2.$$

Bevis. Forbigås. Følger ved direkte regning. ■

Parameteren k 's betydning for grafens udseende fremgår af nedenstående figur.



Vi viser nu, at der er en sammenhæng mellem ekstremværdifordelingerne Λ_1 og $\Lambda_{2,k}$. Vi har nemlig

SÆTNING 1.44. Lad $X \in \text{Max}_2(k, \beta)$. Da er $\log_e X \in \text{Max}_1(\log_e \beta, \frac{1}{k})$.

Bevis. Vi går direkte frem.

$$\begin{aligned}
 P\{\log_e X \leq x\} &= P\{X \leq e^x\} \\
 &= \Lambda_{2,k}\left(\frac{e^x}{\beta}\right) \\
 &= \exp(-\beta^k e^{-kx}) \\
 &= \exp(-e^{k \log_e \beta - kx}) \\
 &= \Lambda_1\left(\frac{x - \log_e \beta}{k^{-1}}\right)
 \end{aligned}$$

■

Vi giver nu de analoge resultater om den asymptotiske fordeling af minimumsværdien.

SÆTNING 1.45. Lad $X_1, X_2, \dots$ være en følge af uafhængige stokastiske variable med samme fordelingsfunktion F . Lad der eksistere et $k > 0$, således at vi for alle $a > 0$ har

$$\frac{F(x)}{F(ax)} \rightarrow a^k, \quad x \rightarrow -\infty. \quad (1.23)$$

Der gælder da, at

$$P\left\{\frac{X_{(1)}}{\delta_n} \leq x\right\} \rightarrow 1 - \Lambda_{2,k}(-x) = 1 - \exp(-(-x)^{-k}),$$

for $n \rightarrow \infty$. Her bestemmes δ_n af

$$F(-\delta_n) = \frac{1}{n}.$$

Bevis. Sætningen følger direkte af sætning 1.41 ved overgang til $-X_1, -X_2, \dots$ etc.

■

Vi giver grænsefordelingen et navn i

DEFINITION 1.25. Vi siger, at $X \in \text{Min}_2(k, 1)$, hvis X har fordelingsfunktionen $1 - \Lambda_{2,k}(-x)$. Fordelingen med skalaparameter β benævnes $\text{Min}_2(k, \beta)$. ▲

SÆTNING 1.46. Klassen Min_2 er afsluttet over for minimumsdannelse, idet der for uafhængige stokastiske variable gælder

$$X_i \in \text{Min}_2(k, \beta) \Rightarrow X_{(1)} \in \text{Min}_2\left(k, n^{1/k}\beta\right).$$

Bevis. Forbigås. ■

Vi giver de analytiske egenskaber ved Min_2 -fordelingen i

SÆTNING 1.47. Lad $X \in \text{Min}_2(k, \beta)$. Da er tætheden

$$f(x) = \frac{1}{\beta} \lambda_{2,k} \left(-\frac{x}{\beta} \right) = \frac{k}{\beta} \left(-\frac{\beta}{x} \right)^{k+1} \exp \left(- \left(-\frac{\beta}{x} \right)^k \right),$$

for $x < 0$ og 0 ellers. Middelværdi og varians er

$$\begin{aligned} E(X) &= -\beta \Gamma\left(1 - \frac{1}{k}\right) & k > 1 \\ V(X) &= \beta^2 \left(\Gamma\left(1 - \frac{2}{k}\right) - \Gamma^2\left(1 - \frac{1}{k}\right) \right) & k > 2. \end{aligned}$$

Bevis. Beviset er umiddelbart. ■

Der gælder selvsagt også en analog sætning til sætning 1.44, nemlig

SÆTNING 1.48. Lad $X \in \text{Min}_2(k, \beta)$. Da er

$$-\log_e(-X) \in \text{Min}_1\left(-\log_e \beta, \frac{1}{k}\right)$$

og

$$\log_e(-X) \in \text{Max}_1\left(\log_e \beta, \frac{1}{k}\right)$$

Bevis. Forbigås. ■

EKSEMPEL 1.39. På en given lokalitet kan det antages, at middelværdien og spredningen af den maksimale årlige vindhastighed er

$$\mu = 66 \text{ km/h og } \sigma = 15.4 \text{ km/h.}$$

Den maksimale vindhastighed inden for et år kan opfattes som maksimum af en mængde vindhastigheder. Antages det, at fordelingen af de enkelte hastigheder er af Cauchy type, vil det årlige maksimum (med tilnærmelse) følge af en $\text{Max}_2(k, \beta)$ -fordeling.

Det er almindeligt, at bygningskonstruktioner etc. designes med henblik på at kunne modstå en vindhastighed, der kun overstiges med en sandsynlighed på 0.02 inden for et år. Vi vil nu beregne den (mindste) vindhastighed, der tilfredsstiller dette krav.

Vi bestemmer først parametrene k og β . Vi får ligningerne

$$\begin{aligned} 66 &= \beta \Gamma \left(1 - \frac{1}{k} \right) \\ 15.4^2 &= \beta^2 \left(\Gamma \left(1 - \frac{2}{k} \right) - \Gamma^2 \right) 1 - \frac{1}{k} \Bigg) . \end{aligned}$$

Ved numerisk løsning af disse fås

$$k = 6.5, \quad \beta = 58.9.$$

Vi skal nu bestemme x , så

$$P\{X \leq x\} = \Lambda_{2,6.5} \left(\frac{x}{58.9} \right) = 0.98.$$

Indsættes i udtrykket for $\Lambda_{2,6.5}$ fås

$$\exp \left(-\frac{58.9^{6.5}}{x^{6.5}} \right) = 0.98.$$

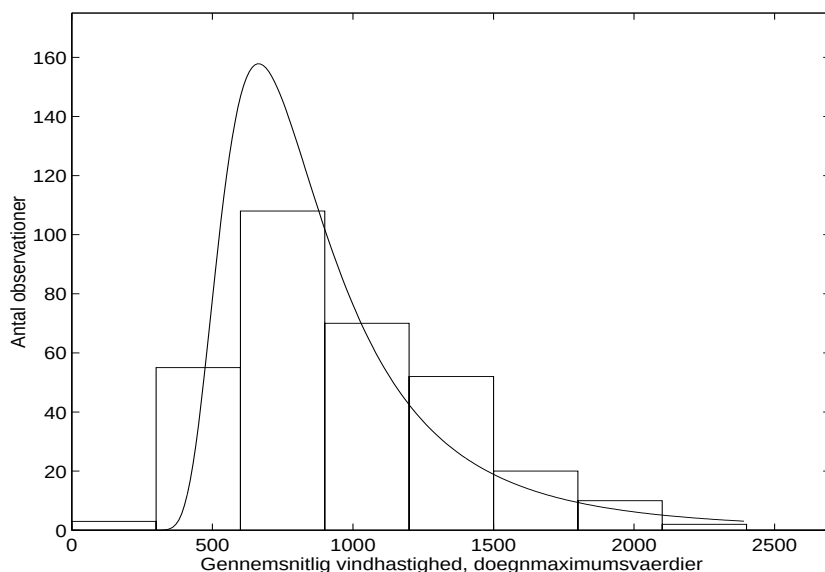
Denne ligning giver ved løsning $x = 107$, d.v.s. at konstruktionerne skal designes med henblik på at kunne modstå vindhastigheder på 107 km/h. ♦

EKSEMPEL 1.40. Ved Stignæs har man gennem hele 1973 målt den gennemsnitlige vindhastighed over hver halve time. I tabellen side 167 og 168 er anført de maksimale

Maksimale $\frac{1}{2}$ h vindhastigheder pr. døgn i 1973.						
Dato	Jan.	Feb.	Marts	April	Maj	Juni
1	605	406	1078	1417	1282	647
2	863	856	1085	1206	692	855
3	849	950	878	1304	771	1216
4	741	993	720	1505	879	505
5	641	(1306)	830	1500	1286	848
6	616	1709	991	1402	1256	(843)
7	669	(-1)	906	759	575	372
8	545	1100	764	713	1143	891
9	354	676	435	747	604	988
10	(472)	737	990	1291	1124	1444
11	(-1)	1340	768	(2134)	1096	1453
12	372	(1980)	708	904	976	1021
13	702	(1666)	483	587	1386	1284
14	819	(1256)	482	979	972	1489
15	(843)	558	409	692	1091	1347
16	(689)	868	1020	11335	887	514
17	334	496	1369	1563	593	525
18	343	333	1201	1183	527	623
19	366	677	805	571	738	890
20	755	(848)	1291	374	811	1035
21	874	1352	1104	774	431	753
22	670	1498	916	1100	478	(557)
23	395	1137	950	940	491	757
24	859	576	803	824	624	568
25	1075	305	625	1025	657	(421)
26	1062	849	551	1329	689	315
27	1097	841	692	807	342	884
28	704	984	862	547	338	(758)
29	1027		812	428	620	919
30	937		553	1036	658	882
31	651		942		661	

Maksimale $\frac{1}{2}$ h vindhastigheder pr. døgn i 1973.						
Dato	Juli	Aug.	Sept.	Okt.	Nov.	Dec.
1	414	651	1284	699	1198	(-1)
2	(408)	701	1211	1039	962	1313
3	663	714	(1017)	392	432	1521
4	597	1143	554	675	814	1801
5	634	1554	793	756	1662	1507
6	1024	1242	601	730	(1843)	(2028)
7	965	1809	1072	(515)	1605	1403
8	1245	(-1)	843	(-1)	1343	957
9	1178	(854)	1569	(-1)	1449	1218
10	(1107)	1044	1338	1270	759	1462
11	850	762	1172	871	1293	1638
12	701	806	1303	513	(1948)	1551
13	742	467	743	940	(2052)	1490
14	563	268	851	851	1544	1802
15	963	266	583	571	(1327)	805
16	1030	695	866	476	(-1)	(967)
17	836	806	1206	1209	(1698)	1304
18	1339	1041	1217	1533	(0)	1179
19	1129	(575)	1129	1346	(1821)	888
20	1308	(-1)	(1111)	660	(-1)	903
21	824	(1432)	1046	927	1065	1003
22	747	954	1395	(762)	1590	601
23	1011	855	1030	782	1747	694
24	742	875	1311	1070	(2015)	977
25	730	735	1216	1231	2239	549
26	1037	639	425	876	(-1)	929
27	987	780	1034	624	(-1)	1005
28	739	558	1730	872	(-1)	(846)
29	733	568	(1216)	1618	(-1)	883
30	1117	866	1122	1536	(-1)	901
31	1057	1132		928		1127

Tabel 1.3: Maksimale vindhastigheder ($\frac{1}{2}$ h-gennemsnit) i 1973 ved Stigsnæs. Et tal i parentes angiver, at der ikke foreligger målinger fra hele døgnnet. Et (-1) angiver, at der ikke foreligger målinger fra det pågældende døgn.



værdier for hvert døgn. I overensstemmelse med eksempel 1.39 vil vi sammenligne de relative hyppigheder for forskellige vindhastigheder med en Max_2 -fordeling.

Dette er gjort i ovenstående figur, hvor der dels er anført et histogram over målingerne og dels frekvensfunktionen ($\times$ antal obs. $\times$ klassebredde) for en $\text{Max}_2(3.10, 727.4)$ -fordeling. Overensstemmelsen er rimeligt god, og det må anses for rimeligt at beskrive fordelingen af de maksimale vindhastigheder ved Max_2 -fordelinger. ♦

1.7.6 Fordelinger af tredje type

I denne paragraf vil vi beskæftige os med ekstremværdier fra fordelinger, der er begrænsede, enten til højre, hvis det er maximumsværdier, vi er interesserede i, eller til venstre, hvis det er minimumsværdier, vi undersøger.

DEFINITION 1.26. En fordelingsfunktion F siges at være af **tredie type**, såfremt der eksisterer et x_0 således, at

$$F(x_0) = 1 \quad \text{og} \quad F(x_0 - \varepsilon) < 1 \quad \forall \varepsilon > 0,$$

og der tillige eksisterer en positiv konstant k således, at der for ethvert $a > 0$ gælder

$$\frac{1 - F(ax + x_0)}{1 - F(x + x_0)} \rightarrow a^k \quad \text{for } x \rightarrow 0. \quad (1.24)$$



SÆTNING 1.49. Lad fordelingsfunktionen F være givet ved

$$F(x) = 1 - c(x_0 - x)^k (1 + \varepsilon(x - x_0)), \quad x \leq x_0,$$

hvor $\varepsilon \rightarrow 0$, $x \rightarrow 0$. Da er $F(x)$ af tredje type.

Bevis. Vi har

$$\begin{aligned} \frac{1 - F(ax + x_0)}{1 - F(x + x_0)} &= \frac{c(-ax)^k (1 + \varepsilon(ax))}{c(-x)^k (1 + \varepsilon(x))} \\ &= a^k \frac{1 + \varepsilon(ax)}{1 + \varepsilon(x)} \\ &\rightarrow a^k, \quad x \rightarrow 0. \end{aligned}$$



Grænsefordelingen for maksimumsværdien af **type 3** fordelte stokastiske variable fremgår af den næste sætning.

SÆTNING 1.50. Lad $X_1, X_2, \dots$ være en følge af indbyrdes uafhængige, identisk fordelte stokastiske variable med fordelingsfunktion F af type 3. Da vil

$$P \left\{ \frac{X_{(n)} - x_0}{\beta_n} \leq x \right\} \rightarrow \Lambda_{3,k}(x) = \begin{cases} \exp(-(-x)^k) & x < 0 \\ 1 & x > 0, \end{cases}$$

for $n \rightarrow \infty$. Her er k og x_0 de i definition 1.26 omtalte størrelser, og β_n defineres ved

$$F(x_0 - \beta_n) = 1 - \frac{1}{n}.$$

Bevis. Se [22].



BEMÆRKNING 1.9. β_n er ikke entydigt bestemt, jfr. bemærkningen side 162.



DEFINITION 1.27. Vi siger, at $X \in \text{Max}_3(k; 0, 1)$, hvis X har fordelingsfunktionen $\Lambda_{3,k}(x)$. Fordelingen med skalaparameter β og positionsparamter α benævnes $\text{Max}_3(k; \alpha, \beta)$.



I fortsættelse af sætning 1.50 har vi

SÆTNING 1.51. For indbyrdes uafhængige stokastiske variable $X_1, \dots, X_n$ gælder

$$X_i \in \text{Max}_3(k; \alpha, \beta) \Rightarrow X_{(n)} \in \text{Max}_3\left(k; \alpha, \beta n^{-1/k}\right),$$

d.v.s. klassen Max_3 er afsluttet over for maximumsdannelse.

Bevis. Forbigås. ■

De analytiske egenskaber ved Max_3 -fordelingen er samlet i

SÆTNING 1.52. Lad $X \in \text{Max}_3(k; \alpha, \beta)$. Da er tætheden for X

$$f(x) = \frac{1}{\beta} \lambda_{3,k} \left(\frac{x - \alpha}{\beta} \right) = \frac{k}{\beta} \left(-\frac{x - \alpha}{\beta} \right)^{k-1} \exp \left(\left(-\frac{x - \alpha}{\beta} \right)^k \right),$$

for $x < \alpha$ og 0 ellers. $\lambda_{3,k}$ er lig $\Lambda'_{3,k}$. Middelværdien og varians i fordelingen er

$$\begin{aligned} E(X) &= \alpha - \beta \Gamma \left(1 + \frac{1}{k} \right) \\ V(X) &= \beta^2 \left(\Gamma \left(1 + \frac{2}{k} \right) - \Gamma^2 \left(1 + \frac{1}{k} \right) \right) \end{aligned}$$

Bevis. Forbigås. Følger ved direkte regning. ■

Denne fordeling er ligeledes knyttet til Λ_1 , idet vi har

SÆTNING 1.53. Lad $X \in \text{Max}_3(k; \alpha, \beta)$. Da er

$$-\log_e(\alpha - X) \in \text{Max}_1 \left(\log_e \frac{1}{\beta}, \frac{1}{k} \right).$$

Bevis. Fuldstændig analogt til beviset for sætning 1.44. ■

De interessante ekstremværdifordelinger i forbindelse med type 3 fordelinger er oftest fordelingen af minimum, hvorfor vi ikke anfører nogen graf for Max_3 -fordelingen, men i stedet for graferne for nogle Min_3 -tætheder (side 174). Vi må dog først nævne

SÆTNING 1.54. Lad $X_1, X_2, \dots$ være en følge af uafhængige, identisk fordelte stokastiske variable med fordelingsfunktion F . Lad der eksistere et x_0 , så $F(x_0) = 0$ og $F(X_0 + \varepsilon) > 0$ for $\varepsilon > 0$. Hvis der eksisterer en positiv konstant k således, at der for alle $a > 0$, gælder

$$\frac{F(ax + x_0)}{F(x + x_0)} \rightarrow a^k, \quad x \rightarrow 0,$$

da vil

$$P\left\{\frac{X_{(1)} - x_0}{\delta_n} \leq x\right\} \rightarrow 1 - \Lambda_{3,k}(-x) = \exp(-x^k)$$

for $x > 0$ og 0 ellers. δ_n bestemmes af

$$F(x_0 + \delta_n) = \frac{1}{n}.$$

Bevis. Sætningen følger umiddelbart af sætning 1.50. ■

DEFINITION 1.28. Vi siger, at $X \in \text{Min}_3(k; 0, 1)$, hvis X har fordelingsfunktionen $1 - \Lambda_{3,k}(-x)$. Fordelingen med positionsparameter α og skalaparameter β benævnes $\text{Min}_3(k; \alpha, \beta)$. Særlig vigtig er den klasse af Min_3 -fordelinger, der har positionsparameteren $\alpha = 0$. Disse kaldes Weibull-fordelinger efter den person, der først anvendte dem. ▲

DEFINITION 1.29. Den stokastiske variabel X siges at være **Weibull-fordelt** med parametre (k, β) , hvis $X \in \text{Min}_3(k; 0, \beta)$. Vi skriver kort $X \in \text{We}(k, \beta)$. ▲

Såvel Min_3 -fordelinger som Weibull-fordelinger er afsluttet over for minimumsdannelse, idet der gælder følgende

SÆTNING 1.55. Lad $X_1, \dots, X_n$ være stokastisk uafhængige. Da gælder

$$\begin{aligned} X_i &\in \text{Min}_3(k; \alpha, \beta) &\Rightarrow X_{(1)} &\in \text{Min}_3(k; \alpha, \beta n^{-1/k}) \\ X_i &\in \text{We}(k, \beta) &\Rightarrow X_{(1)} &\in \text{We}(k, \beta n^{-1/k}). \end{aligned}$$

Bevis. Forbigås. ■

De analytiske egenskaber ved Min_3 - og We-fordelingen fremgår af følgende to sætninger

SÆTNING 1.56. Lad $X \in \text{Min}_3(k; \alpha, \beta)$. Da er tætheden for X

$$f(x) = \frac{k}{\beta} \left(\frac{x - \alpha}{\beta} \right)^{k-1} \exp \left(- \left(\frac{x - \alpha}{\beta} \right)^k \right), \quad x > \alpha$$

og 0 ellers. Middelværdi og varians er

$$\begin{aligned} E(X) &= \alpha + \beta \Gamma \left(1 + \frac{1}{k} \right) \\ V(X) &= \beta^2 \left(\Gamma \left(1 + \frac{2}{k} \right) - \Gamma^2 \left(1 + \frac{1}{k} \right) \right). \end{aligned}$$

SÆTNING 1.57. Lad $X \in \text{We}(k, \beta)$. Da er tætheden for X

$$f(x) = \frac{k}{\beta} \left(\frac{x}{\beta} \right)^{k-1} \exp \left(- \left(\frac{x}{\beta} \right)^k \right), \quad x > 0$$

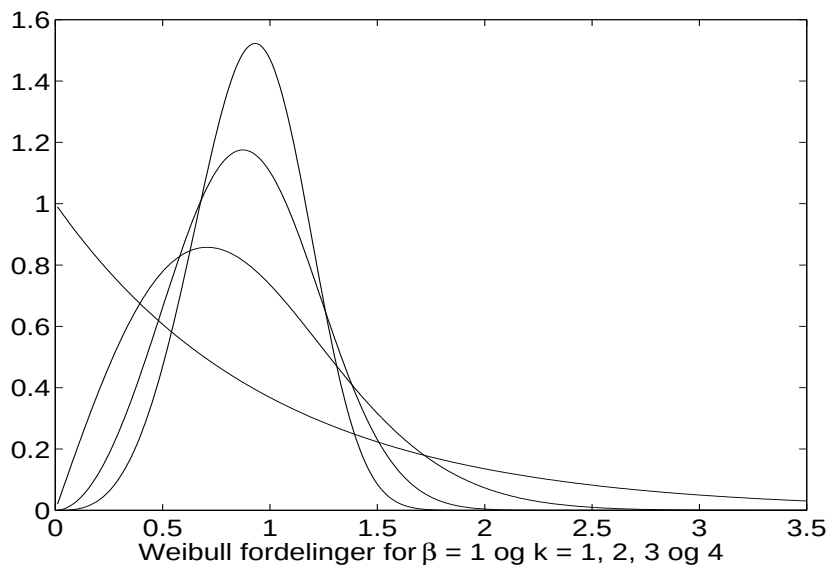
og 0 ellers. Middelværdi og varians er

$$\begin{aligned} E(X) &= \beta \Gamma \left(1 + \frac{1}{k} \right) \\ V(X) &= \beta^2 \left(\Gamma \left(1 + \frac{2}{k} \right) - \Gamma^2 \left(1 + \frac{1}{k} \right) \right). \end{aligned}$$

Bevis. Beviset er umiddelbart. ■

I nedenstående figur har vi anført nogle tætheder for Weibullfordelinger. Det ses, at formparameteren k er af væsentlig betydning for grafens udseende.

Sammenhængen mellem Min_3 - og Weibull-fordelingerne og Max_1 -fordelingen fremgår af



SÆTNING 1.58. Lad $X \in \text{Min}_3(k; \alpha, \beta)$. Da er

$$\begin{aligned} \log_e(X - \alpha) &\in \text{Min}_1\left(\log_e \beta, \frac{1}{k}\right) \\ -\log_e(X - \alpha) &\in \text{Max}_1\left(-\log_e \beta, \frac{1}{k}\right). \end{aligned}$$

For Weibull-fordelingen gælder med $Y \in \text{We}(k, \beta)$

$$\begin{aligned} \log_e Y &\in \text{Min}_1\left(\log_e \beta, \frac{1}{k}\right) \\ -\log_e Y &\in \text{Max}_1\left(-\log_e \beta, \frac{1}{k}\right). \end{aligned}$$

Bevis. Umiddelbart. ■

EKSEMPEL 1.41 (METALTRÆTHED). Vi udsætter et metallisk emne for en bestemt type belastninger. Dette gentages, indtil emnet brister. Da vil antallet af gentagelser, der kan foretages, inden det brister, være en stokastisk variabel. Denne vil kun antage positive værdier, eventuelt endog kun værdier over en vis størrelse. Der kan nu argumenteres for, at metalstykket brister på det svageste sted, og at antallet af gentagelser derfor er en minimumsværdi, nemlig minimum af de antal forsøg (e.g. bøjninger), der skal foretages de forskellige steder, før emnet brister. Det kunne derfor være rimeligt

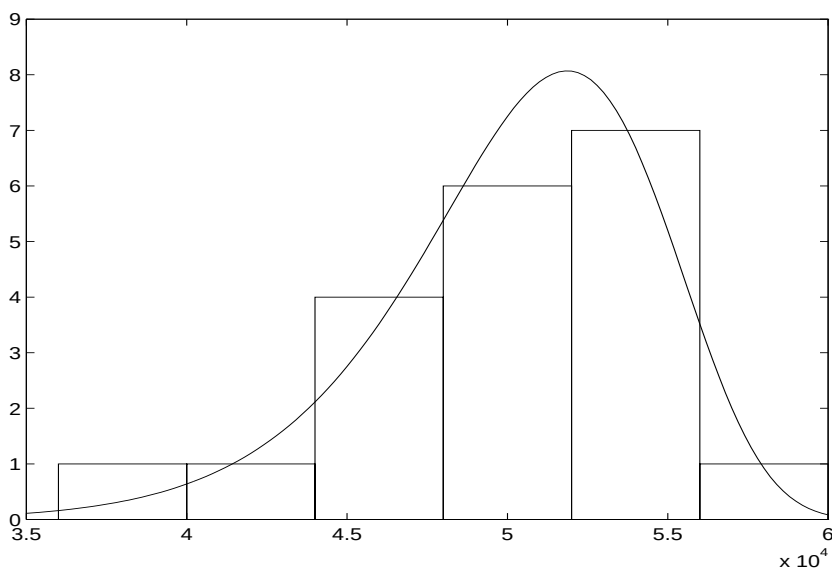
at antage, at antallet af påvirkninger, som emnet kan tåle før et træthedsbrud, er en Weibull-fordelt stokastisk variabel. ♦

Weibull anvendte også fordelingen til at beskrive brudstyrker af metaller m.v.

EKSEMPEL 1.42. I nedenstående tabel er anført antallet af torsioner, en række nikkel-tråde har været udsat for, før de bristede (se [23]). Trådene var udsat for en belastning på $\pm 37.5 \text{ kgmm}^{-2}$.

Emne nr.	Antal torsioner	Emne nr.	Antal torsioner
1	40000	11	51200
2	43900	12	51500
3	45400	13	52300
4	46500	14	52500
5	47100	15	53700
6	48000	16	54500
7	48200	17	54800
8	48500	18	55000
9	49300	19	55700
10	49900	20	57000

I den følgende figur er angivet et histogram for disse målinger og der er anført grafen for frekvensfunktionen for en $We(14.26, 52129)$ -fordeling. ($\times \text{antal obs} \times \text{klassebredde}$).



Overensstemmelsen ser udmærket ud, og Weibull-fordelingen må derfor siges at være egnet til at beskrive fænomener som ovenstående.

I en del situationer vil en Min_3 -fordeling dog komme på tale, idet der ofte vil være en vis nedre grænse (større end 0) for det antal på virkninger, et emne udsættes for, før der optræder brud. ♦

EKSEMPEL 1.43 (LEVETIDER). Vi vil nu undersøge minimumsfordelingen for Γ -fordelingen, der ofte selv benyttes som modelfordeling for levetider (ventetiden mellem k hændelser, der hver indtræffer med $Ex(\beta)$ -fordelte tidsafstande vil være $G(k, \beta)$ -fordelt). Lad $X \in G(k, \beta)$, d.v.s.:

$$f(x) = \frac{1}{\Gamma(k)\beta^k} x^{k-1} e^{-x/\beta}.$$

Heraf fås

$$\frac{F(ax)}{F(x)} = \frac{\int_0^{ax} t^{k-1} e^{-t/\beta} dt}{\int_0^x t^{k-1} e^{-t/\beta} dt} = \frac{g(x)}{h(x)}.$$

Da såvel tæller som nævner går mod 0 for $x \rightarrow 0$, kan vi anvende l'Hôpital's regel og får

$$\begin{aligned} \frac{g'(x)}{h'(x)} &= \frac{a(ax)^{k-1} e^{-\frac{ax}{\beta}}}{x^{k-1} e^{-\frac{x}{\beta}}} \\ &= a^k e^{-\frac{x}{\beta}(a-1)} \rightarrow a^k, \quad x \rightarrow 0. \end{aligned}$$

Heraf ses, at en $G(k, \beta)$ -fordeling tilfredsstiller kravene i sætning 1.54, hvorfor minimumsværdien af mange $G(k, \beta)$ fordelte variable med tilnærmelse vil være $\in We(k, \beta')$. (NB! Samme k). ♦

Dette anskueliggør, hvorfor man i **pålidelighedsteorien** så hyppigt beskriver levetider ved Weibull fordelinger. Vi betragter nemlig ofte apparater, som er sammensat af et stort antal komponenter, og som svigter, når blot én komponent går ud. Da levetiden for en enkelt komponent ofte kan antages at følge en Γ -fordeling, vil den totale levetid være minimumsværdien af de Γ -fordelte levetider, d.v.s. approksimativt Weibull fordelt.

1.7.7 Oversigter over asymptotiske ekstremværdifordelinger

Nedenstående gives et antal oversigter, som viser hovedtrækkene af de beskrevne ekstremværdifordelingers definitioner og oprindelse samt matematiske udtryk for fordelingsfunktioner og parametre.

Definition af Max_j	Middelværdi	Varsians
$X \in \text{Max}_1(\alpha, \beta)$ $\Leftrightarrow F(x) = \Lambda_1(\frac{x-\alpha}{\beta})$	$\alpha + \beta\gamma$	$\frac{\beta^2 \pi^2}{6}$
$X \in \text{Max}_2(k, \beta)$ $\Leftrightarrow F(x) = \Lambda_{2,k}(\frac{x}{\beta})$	$\beta\Gamma(1 - \frac{1}{k})$	$\beta^2\Gamma(1 - \frac{2}{k})$ $-\beta^2\Gamma^2(1 - \frac{1}{k})$
$X \in \text{Max}_3(k, \alpha, \beta)$ $\Leftrightarrow F(x) = \Lambda_{3,k}(\frac{x-\alpha}{\beta})$	$\alpha - \beta\Gamma(1 - \frac{1}{k})$	$\beta^2\Gamma(1 + \frac{2}{k})$ $-\beta^2\Gamma^2(1 + \frac{1}{k})$
Definition af Min_j	Middelværdi	Varsians
$X \in \text{Min}_1(\alpha, \beta)$ $\Leftrightarrow F(x) = 1 - \Lambda_1(-\frac{x-\alpha}{\beta})$	$\alpha - \beta\gamma$	$\frac{\beta^2 \pi^2}{6}$
$X \in \text{Min}_2(k, \beta)$ $\Leftrightarrow F(x) = 1 - \Lambda_{2,k}(-\frac{x}{\beta})$	$-\beta\Gamma(1 - \frac{1}{k})$	$\beta^2\Gamma(1 - \frac{2}{k})$ $-\beta^2\Gamma^2(1 - \frac{1}{k})$
$X \in \text{Min}_3(k, \alpha, \beta)$ $\Leftrightarrow F(x) = 1 - \Lambda_{3,k}(-\frac{x-\alpha}{\beta})$	$\alpha + \beta\Gamma(1 + \frac{1}{k})$	$\beta^2\Gamma(1 + \frac{2}{k})$ $-\beta^2\Gamma^2(1 + \frac{1}{k})$
$X \in \text{We}(k, \beta)$ $\Leftrightarrow F(x) = \Lambda_{3,k}(-\frac{x}{\beta})$	$\beta\Gamma(1 + \frac{1}{k})$	$\beta^2\Gamma(1 + \frac{2}{k})$ $-\beta^2\Gamma^2(1 + \frac{1}{k})$

Definition hvis fordeling for største observation søges	Definition hvis fordeling for mindste observation søges	Heuristisk formulering af definition
Eksponentiel type		
$\frac{d}{dx} \frac{1-F(x)}{f(x)} \rightarrow 0$ for $x \rightarrow \infty$	$\frac{d}{dx} \frac{F(x)}{f(x)} \rightarrow 0$ for $x \rightarrow -\infty$	F har ubegrænset støtte opadtil resp. nedadtil, $F \rightarrow 1$ resp. 0 mindst så hurtigt som en eksponentialfkt.
Cauchy type		
$\exists k \left(\frac{d}{dx} \frac{1-F(x)}{1-F(ax)} \rightarrow a^k \right)$ for $x \rightarrow \infty$ og $\forall a$	$\exists k \left(\frac{d}{dx} \frac{F(x)}{F(ax)} \rightarrow a^k \right)$ for $x \rightarrow -\infty$ og $\forall a$	F har ubegrænset støtte opadtil resp. nedadtil, $F \rightarrow 1$ resp. 0 mindst så hurtigt som en potensfunktion.
Tredie type		
$F(x_0) = 1 \wedge \exists k$ $\left(\frac{d}{dx} \frac{1-F(ax+x_0)}{1-F(x+x_0)} \rightarrow a^k \right)$ for $x \rightarrow \infty$ og $\forall a$	$F(x_0) = 0 \wedge \exists k$ $\left(\frac{d}{dx} \frac{1-F(ax+x_0)}{F(x+x_0)} \rightarrow a^k \right)$ for $x \rightarrow -\infty$ og $\forall a$	F har <u>begrænset</u> støtte opadtil resp. nedadtil, $F \rightarrow 1$ resp. 0 mindst så hurtigt som en potensfunktion.

Asymptotisk fordeling af største observation	Def. af parametre
Eksponentiel type	
$P \left\{ \frac{x_{(n)} - \alpha_n}{\beta_n} \leq x \right\} \rightarrow \exp(-e^{-x})$ for $n \rightarrow \infty$ $= \Lambda_1(x)$	$F(\alpha_n) = 1 - \frac{1}{n}$ $\beta_n = \frac{1}{n \cdot f(\alpha_n)}$
Cauchy type	
$P \left\{ \frac{x_{(n)} - \alpha_n}{\beta_n} \leq x \right\} \rightarrow \exp(-x^{-k})$ for $n \rightarrow \infty$ $= \Lambda_{2,k}(x)$	$F(\beta_n) = 1 - \frac{1}{n}$
Tredie type	
$P \left\{ \frac{x_{(n)} - x_0}{\beta_n} \leq x \right\} \rightarrow \begin{cases} \exp(-(-x)^k), & x < 0 \\ 1 & x > 0 \end{cases}$ for $n \rightarrow \infty$ $= \Lambda_{3,k}(x)$	$F(x_0 - \beta_n)$ $= 1 - \frac{1}{n}$

Asymptotisk fordeling af mindste observation	Def. af parametre
Eksponentiel type	
$P \left\{ \frac{x_{(1)} - \gamma_n}{\delta_n} \leq x \right\} \rightarrow 1 - \exp(-e^x)$ for $n \rightarrow \infty$ $= 1 - \Lambda_1(-x)$	$F(\gamma_n) = \frac{1}{n}$ $\delta_n = \frac{1}{n \cdot f(\gamma_n)}$
Cauchy type	
$P \left\{ \frac{x_{(1)} - \gamma_n}{\delta_n} \leq x \right\} \rightarrow 1 - \exp(-(-x)^k)$ for $n \rightarrow \infty$ $= 1 - \Lambda_{2,k}(-x)$	$F(-\delta_n) = \frac{1}{n}$
Tredie type	
$P \left\{ \frac{x_{(1)} - x_0}{\delta_n} \leq x \right\} \rightarrow \begin{cases} 1 - \exp(-x^k), & x > 0 \\ 0 & x < 0 \end{cases}$ for $n \rightarrow \infty$ $= 1 - \Lambda_{3,k}(-x)$	$F(x_0 + \delta_n)$ $= \frac{1}{n}$

1.8 Andre sandsynlighedsteoretiske modeller

I dette afsnit vil vi omtale nogle fordelinger, som ikke alle har en naturlig fysisk fortolkning, men som alligevel ofte omtales (og anvendes) i statistikken. Der er derfor få interessante praktiske eksempler at anføre, og beviserne vil derfor også i nogen grad være uinteressante. Disse forbigås derfor i endnu højere grad end tidligere, og der bliver tale om en noget summarisk opremsning.

1.8.1 Den rektangulære fordeling

er især vigtig til beskrivelse af f.eks. afrundingsfejl.

DEFINITION 1.30. Vi siger, at den stokastiske variable X er **rektangulært** eller **uniformt fordelt over intervallet** (a, b) , hvis X har tætheden

$$f(x) = \frac{1}{b-a}, \quad a < x < b$$

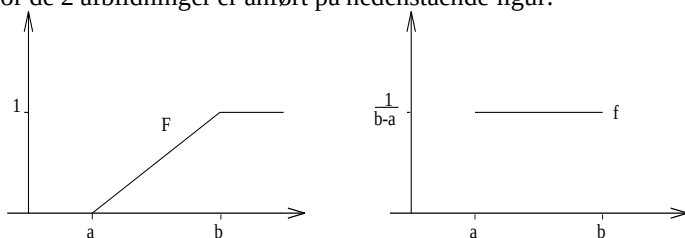
og 0 ellers. Vi skriver $X \in U(a, b)$. ▲

BEMÆRKNING 1.10. Der findes en diskret analog til den rektangulære fordeling, nemlig **ligefordelingen**, se evt. side 187. ▼

Det ses umiddelbart, at fordelingsfunktionen er

$$F(x) = \begin{cases} 0 & x \leq a \\ \frac{1}{b-a}(x-a) & a < x < b \\ 1 & b \leq x \end{cases}$$

Graferne for de 2 afbildninger er anført på nedenstående figur:



Middelværdi og varians er anført i

SÆTNING 1.59. Lad $X \in U(a, b)$. Da er den karakteristiske funktion

$$\phi(t) = \frac{1}{b-a} \frac{1}{it} (e^{itb} - e^{ita}).$$

Middelværdien og variansen er

$$E(X) = \frac{a+b}{2}$$

og

$$V(X) = \frac{1}{12}(b-a)^2$$

Bevis. Forbigås. ■

Vi anfører endvidere, hvorledes forskellige rektangulære fordelinger kan fremgå af hinanden.

SÆTNING 1.60. Lad $X \in U(0, 1)$. Da vil

$$a + (b-a)X \in U(a, b).$$

Bevis. Beviset er umiddelbart. ■

Den rektangulære fordeling anvendes altså, når man har en situation, hvor udfaldet af den datagenererende proces har den samme sandsynlighed for at falde i et delinterval af (a, b) af en vilkårlig, fast mængde. Sandsynligheden bliver da nødvendigvis proportional med delintervallets længde h . Når man taler om tilfældigt at vælge et punkt i et endeligt interval, mener man, at punktets placering skal følge en rektangulær fordeling over intervallet.

1.8.2 Beta-fordelingen

Vi minder om definitionen af Beta-integralet. For positive α og β har vi ifølge p. ??

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1} (1-t)^{\beta-1} dt.$$

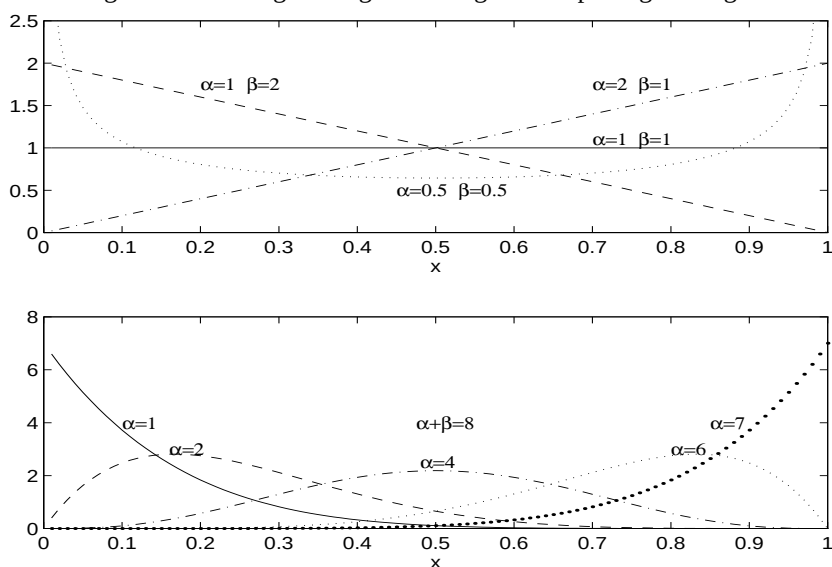
Dette udtryk giver anledning til

DEFINITION 1.31. Vi siger, at den stokastiske variable X er **Beta-fordelt** med parametre (α, β) , hvis X har tætheden

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 < x < 1,$$

og 0 ellers. Vi skriver kort $X \in \text{Be}(\alpha, \beta)$. ▲

Beta-fordelingen er en fordeling koncentreret på intervallet $(0, 1)$, og den er beskrevet ved 2 parametre. Dette bevirker, at tæthederne for forskellige parameterværdier kan se vidt forskellige ud. Vi har angivet nogle af mulighederne på følgende figur.



SÆTNING 1.61. Lad $X \in \text{Be}(\alpha, \beta)$. Da er

$$E(X) = \frac{\alpha}{\alpha + \beta}$$

og

$$V(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Bevis. Umiddelbart. ■

Som det synes at fremgå af graferne, er en $\text{Be}(1, 1)$ -fordeling en uniform fordeling over $(0, 1)$, d.v.s. en $\text{U}(0, 1)$ -fordeling.

I testteorien vil vi endvidere få brug for

SÆTNING 1.62. Lad X og Y være uafhængige stokastiske variable. Da gælder

$$X \in G(\alpha_1, \beta) \wedge Y \in G(\alpha_2, \beta) \Rightarrow \frac{X}{X+Y} \in \text{Be}(\alpha_1, \alpha_2).$$

Beta-fordelingen kan, som det fremgår af graferne over tæthederne, anvendes som en approksimation til de fleste kontinuerte fordelinger på $(0, 1)$. Den anvendes endvidere i testteorien og i beslutningsteorien.

1.8.3 Cauchy fordelingen

DEFINITION 1.32. En stokastisk variabel X siges at være **Cauchy fordelt** med parametre $(0, 1)$, hvis tætheden er

$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}.$$

Vi skriver $X \in \text{Ca}(0, 1)$. Fordelingen med positionsparameter α og skalaparameter β benævnes $\text{Ca}(\alpha, \beta)$. ▲

Man ser let, at middelværdien **ikke** eksisterer og derfor naturligvis heller ikke variansen. $\text{Ca}(\alpha, \beta)$ -fordelingen er symmetrisk om α , og vi har

SÆTNING 1.63. Lad $X \in \text{Ca}(\alpha, \beta)$. Da har X tætheden

$$f(x) = \frac{\beta}{\pi} \frac{1}{\beta^2 + (x - \alpha)^2}.$$

Den karakteristiske funktion er

$$\phi(t) = \exp(it\alpha - \beta|t|).$$

Bevis. Forbigås. Fås ved direkte regning. ■

Cauchy fordelingen besidder en reproduktivitetssætning.

SÆTNING 1.64. (Reproduktivitetssætning) Lad X og Y være stokastiske uafhængige. Da gælder

$$X \in \text{Ca}(\alpha_1, \beta_1) \wedge Y \in \text{Ca}(\alpha_2, \beta_2) \Rightarrow X+Y \in \text{Ca}(\alpha_1+\alpha_2, \beta_1+\beta_2).$$

Bevis. Fås umiddelbart ved at danne den karakteristiske funktion for $X + Y$. ■

Vi har et interessant corollar til sætningen, nemlig

KOROLLAR 1.1. Lad $X_1, \dots, X_n$ være indbyrdes uafhængige, $\text{Ca}(\alpha, \beta)$ -fordelte stokastiske variable. Da er $\bar{X} = \frac{1}{n} \sum X_i$ igen $\text{Ca}(\alpha, \beta)$ -fordelt.

Bevis. Corollaret er en direkte følge af sætningen, når man erindrer, at α og β er positions- og skalaparametre. ■

1.8.4 LaPlace fordelingen

DEFINITION 1.33. Den stokastiske variable X siges at være LaPlace fordelt med parametre $(0, 1)$, hvis tætheden for X er

$$f(x) = \frac{1}{2} e^{-|x|}, \quad -\infty < x < \infty.$$

Vi skriver $X \in \text{La}(0, 1)$. Fordelingen med positionsparameter α og skalaparameter β benævnes $\text{La}(\alpha, \beta)$. ▲

Ved direkte regning fås

SÆTNING 1.65. Lad $X \in \text{La}(\alpha, \beta)$. Da er tætheden for X lig

$$f(x) = \frac{1}{2\beta} e^{-\frac{1}{\beta}|x-\alpha|}.$$

Den karakteristiske funktion er

$$\phi(t) = e^{i\alpha t} (1 + \beta^2 t^2)^{-1},$$

og middelværdi og varians

$$\begin{aligned} E(X) &= \alpha \\ V(X) &= 2\beta^2 \end{aligned}$$

Bevis. Forbigås. ■

Det ses, at fordelingen er symmetrisk om α . Tætheden er ikke differentiabel i α .

1.8.5 Den logistiske fordeling

I bio-assay problemer anvender man hyppigt den logistiske fordeling. Den defineres i følgende

DEFINITION 1.34. En stokastisk variabel X siges at være **logistisk fordelt** med parametre $(0, 1)$, hvis tætheden for X er

$$f(x) = \frac{\exp(-x)}{(1 + \exp(-x))^2}, \quad -\infty < x < \infty.$$

Vi skriver kort $X \in L(0, 1)$. Fordelingen med positionsparameter α og skalaparameter β benævnes $L(\alpha, \beta)$. ▲

Man finder, at fordelingsfunktionen for en $L(\alpha, \beta)$ -fordelt variabel er

$$F(x) = \frac{1}{1 + \exp\left(-\frac{x-\alpha}{\beta}\right)}.$$

Vi nævner

SÆTNING 1.66. Lad $X \in L(\alpha, \beta)$. Da er

$$\begin{aligned} E(X) &= \alpha \\ V(X) &= \frac{\beta^2 \pi^2}{3}. \end{aligned}$$

Bevis. Forbigås. ■

1.8.6 Pareto fordelingen

Pareto fordelingen anvendes ofte ved beskrivelse af indkomstfordelinger (især fordelinger af indkomster over et vist niveau) og lignende. Den defineres i følgende

DEFINITION 1.35. En stokastisk variabel X siges at være Pareto fordelt med parametre $(k, 1)$, hvis tætheden er

$$f(x) = \frac{k}{x^{k+1}}, \quad x \geq 1$$

og 0 ellers. Vi skriver kort $X \in \text{Par}(k, 1)$. Fordelingen med skalaparameter β , benævnes $\text{Par}(k, \beta)$. ▲

Vi giver de analytiske egenskaber ved $\text{Par}(k, \beta)$ -fordelingen i

SÆTNING 1.67. Lad $X \in \text{Par}(k, \beta)$. Da er frekvensfunktionen

$$f(x) = \frac{k\beta^k}{x^{k+1}}, \quad x \geq \beta,$$

og 0 ellers. Fordelingsfunktionen er

$$F(x) = 1 - \left(\frac{\beta}{x}\right)^k, \quad x \geq \beta,$$

og 0 ellers. Middelværdi og varians er

$$\begin{aligned} E(X) &= \beta \frac{k}{k-1} & (k > 1) \\ V(X) &= \beta^2 \frac{k}{(k-1)^2(k-2)} & (k > 2). \end{aligned}$$

Bevis. Forbigås. ■

1.8.7 Ligefordelingen på $\{0, 1, \dots, n\}$

Den diskrete analog til den rektangulære fordeling over et interval $[0, \theta]$ er ligefordelingen på $\{0, 1, \dots, n\}$. Definitionen er umiddelbar.

DEFINITION 1.36. En stokastisk variabel X siges at være **ligefordelt** over $\{0, 1, \dots, n\}$, hvis frekvensfunktionen for X er

$$f(x) = \begin{cases} \frac{1}{n+1} & x = 0, 1, \dots, n \\ 0 & \text{ellers.} \end{cases}$$

Vi skriver kort $X \in \text{UD}(n)$. ▲

SÆTNING 1.68. Middelværdi og varians for en $\text{UD}(n)$ -fordelt stokastisk variabel X er

$$\begin{aligned} E(X) &= \frac{1}{2}n \\ V(X) &= \frac{n(n+2)}{12}. \end{aligned}$$

Bevis. Trivielt. ■

1.8.8 Den logaritmiske fordeling

Den logaritmiske fordeling er en diskret fordeling og må ikke forveksles med den logaritmiske normalfordeling.

DEFINITION 1.37. En stokastisk variabel X siges at være logaritmisk fordelt med parameter $\theta \in]0, 1[$, hvis frekvensfunktionen for X er

$$f(x) = -\frac{1}{\log_e(1-\theta)} \frac{\theta^x}{x}, \quad x = 1, 2, \dots$$

og 0 ellers. Vi skriver $X \in \text{LoD}(\theta)$. ▲

SÆTNING 1.69. Middelværdi og varians for et $X \in \text{LoD}(\theta)$ er

$$\begin{aligned} E(X) &= -\frac{\theta}{(1-\theta) \log_e(1-\theta)} \\ V(X) &= -\frac{\theta^2 + \theta \log_e(1-\theta)}{(1-\theta)^2 \log_e^2(1-\theta)}. \end{aligned}$$

Bevis. Trivielt. ■

1.9 Compound fordelinger

Ved undersøgelser af data, som man mener genereres af mekanismer, der f.eks. fører til Poisson fordelte stokastiske variable, opdager man somme tider en tilsyneladende uforklarlig afvigelse fra det forventede, som består i, at der observeres relativt mange ret store og mange ret små værdier på bekostning af observationer i 'midten' af fordelingen. Blandt mange andre ting kan dette skyldes, at intensiteten ikke er konstant, men varierende. Den resulterende fordeling vil derved få en tilsyneladende for stor varians i forhold til middelværdien.

For at løse det derved fremkomne problem indfører vi de såkaldte **compound** fordelinger (compound=sammensat).

Vi giver først et generelt resultat.

SÆTNING 1.70. Lad Λ være en stokastisk variabel med tæthedsfunktion (eller frekvensfunktion) $g(\lambda)$. Lad endvidere X være en stokastisk variabel, der for givet $\Lambda = \lambda$ har frekvensfunktionen $f(x|\lambda)$. Da har den ubetingede fordeling af X frekvensfunktionen

$$h(x) = \begin{cases} \int_{-\infty}^{\infty} f(x|\lambda)g(\lambda) d\lambda & \text{hvis } \Lambda \text{ er kontinuert fordelt} \\ \sum_{\lambda} f(x|\lambda)g(\lambda) & \text{hvis } \Lambda \text{ er diskret fordelt} \end{cases} \quad (1.25)$$

DEFINITION 1.38. h kaldes en **compound f fordeling**. ▲

Bevis. for sætningen. Vi indskrænker os til det tilfælde, hvor Λ kun antager endeligt mange værdier $\lambda_1, \dots, \lambda_n$, og hvor X er en diskret stokastisk variabel. Vi har da

$$\begin{aligned} P\{X = x\} &= \sum_{i=1}^n P(\{X = x\} \cap \{\Lambda = \lambda_i\}) \\ &= \sum_{i=1}^n P(\{X = x\} | \{\Lambda = \lambda_i\}) P\{\Lambda = \lambda_i\} \\ &= \sum_{i=1}^n f(x|\lambda_i)g(\lambda_i). \end{aligned}$$

Beviset forløber ganske analogt i de øvrige tilfælde. ■

Vi går nu over til specialtilfældet, hvor Λ er Γ -fordelt ($\Lambda \in G(\nu, \beta)$), og X for givet

$\Lambda = \lambda$ er Poissonfordelt ($\in P(\lambda)$). Vi har da, at Λ har tætheden

$$g(\lambda) = \frac{1}{\Gamma(\nu)} \frac{1}{\beta^\nu} \lambda^{\nu-1} e^{-\lambda/\beta}, \quad \lambda > 0 \quad (1.26)$$

X har frekvensfunktionen (for givet Λ)

$$f(x|\lambda) = \frac{1}{x!} \lambda^x e^{-\lambda} \quad x = 0, 1, 2, \dots \quad (1.27)$$

Vi indsætter nu (1.26) og (1.27) i (1.25) og får

$$\begin{aligned} h(x) &= \int_0^\infty \frac{1}{x!} \lambda^x e^{-\lambda} \frac{1}{\Gamma(\nu)} \frac{1}{\beta^\nu} \lambda^{\nu-1} e^{-\lambda/\beta} d\lambda \\ &= \frac{1}{x! \Gamma(\nu) \beta^\nu} \int_0^\infty \lambda^{x+\nu-1} e^{-\lambda(1+1/\beta)} d\lambda. \end{aligned}$$

Vi anvender substitutionen $u = \lambda(1 + \frac{1}{\beta})$ og får

$$h(x) = \frac{1}{x! \Gamma(\nu) \beta^\nu} \frac{1}{(1 + \frac{1}{\beta})^{x+\nu}} \int_0^\infty u^{x+\nu-1} e^{-u} du$$

Integralet er lig $\Gamma(x + \nu)$, hvorfor vi har

$$\begin{aligned} h(x) &= \frac{\Gamma(x+\nu)}{x! \Gamma(\nu)} \left(\frac{1}{1+\beta}\right)^\nu \left(\frac{\beta}{\beta+1}\right)^x \\ &= \binom{x+\nu-1}{x} \left(\frac{1}{1+\beta}\right)^\nu \left(\frac{\beta}{\beta+1}\right)^x \quad x = 0, 1, 2, \dots, \end{aligned}$$

d.v.s. $X \in \text{NB}(\nu, p)$, $p = \frac{1}{1+\beta}$. En negativ binomialfordeling opstået på denne måde kaldes altså en **compound Poisson fordeling**.

Vi vil klargøre disse begreber med nogle eksempler.

EKSEMPEL 1.44. Ved udvidelsen af en telefoncentral bliver man nødt til at lukke den i et vist tidsrum. Dette er ikke en kendt tid, men erfaringer fra tilsvarende arbejder viser, at tiden målt i minutter kan antages at være en stokastisk variabel $T \in \text{Ex}(80)$. Antal opkald til centralen i en periode af længden t minutter følger en $P(2.5t)$ -fordeling. Ifølge sætning 1.17 er $2.5T \in \text{Ex}(2.5 \cdot 80) = \text{Ex}(200) = G(1, 200)$. Derfor er det totale antal opkald

$$X \in \text{NB}\left(1, \frac{1}{1+200}\right) = \text{NB}\left(1, \frac{1}{201}\right).$$

Ifølge sætning 1.5 er

$$E(X) = 1 \cdot \frac{200}{201} \cdot \frac{201}{1} = 200,$$

et resultat, der forekommer intuitivt rigtigt. Vi har altså fundet, at det forventede antal abonnenter, der "afvises", medens centralen er lukket for udvidelse er 200. ♦

EKSEMPEL 1.45. Ved undersøgelse af ulykkesstatistikker for fabriksarbejdere viser det sig ofte, at antallet af ulykker, der overgår en arbejder i et vist tidsinterval, **ikke** følger en Poisson fordeling. Hvis alle arbejdere har samme sandsynlighed for at komme ud for en ulykke, og hvis ulykkerne optræder uafhængigt, ville man på forhånd vente, at sådanne observationer ville følge en Poisson fordeling.

Afvigelserne fra Poisson fordelingen kan dels skyldes, at arbejderne ikke har ens ulykkesintensitet, og dels, at intensiteten for den enkelte arbejder ændres med tiden. Antager vi, at intensiteten er en stokastisk variabel, der følger en Γ -fordeling, vil ulykkestallet følge en negativ binomialfordeling. Vi vil senere vende tilbage til denne problematik. ♦

Vi slutter dette afsnit med at nævne 3 andre compound fordelinger.

SÆTNING 1.71. Lad N være en Poisson fordelt stokastisk variabel ($\in P(\lambda)$), og lad for givet $N = n$ X være binomialt fordelt med parametre (n, p) . Da vil den ubetingede fordeling af X være en $P(\lambda p)$ -fordeling.

Bevis. Beviset er ligefremt. Vi har ifølge (1.25)

$$\begin{aligned} f(x) &= \sum_{n=x}^{\infty} \binom{n}{x} p^x (1-p)^{n-x} \frac{\lambda^n}{n!} e^{-\lambda} \\ &= \frac{(p\lambda)^x}{x!} e^{-\lambda} \sum_{n=x}^{\infty} \frac{1}{(n-x)!} (\lambda(1-p))^{n-x} \\ &= \frac{(p\lambda)^x}{x!} e^{-\lambda} \sum_{n=0}^{\infty} \frac{1}{n!} (\lambda(1-p))^n \\ &= \frac{(p\lambda)^x}{x!} e^{-\lambda} e^{\lambda(1-p)} \\ &= \frac{(p\lambda)^x}{x!} e^{-\lambda p} \end{aligned}$$

i.e.: $X \in P(\lambda p)$. ■

Den derved fremkomne Poisson fordeling er altså en **compound binomial fordeling**. Fordelingen beskriver altså den situation, at man udtager (eller observerer) N hændelser, $N \in P(\lambda)$, og blandt disse viser gennemsnitligt andelen p f.eks. succes og resten fiasko. Det endelige antal succeser der observeres bliver ikke uventet $P(p\lambda)$.

Den næste sætning har en vigtig forbindelse til stikprøveudtagning. Vi formulerer og viser først sætningen og tolker den derefter.

SÆTNING 1.72. Lad M og X være stokastiske variable. Hvis $M \in B(N, p)$, og hvis X for givet $M = m$ er $\in H(n, m, N)$, da er $X \in B(n, p)$.

Bevis. Beviset følger igen ved direkte anvendelse af (1.25). Den hypergeometriske sandsynlighed er

$$\frac{\binom{m}{x} \binom{N-m}{n-x}}{\binom{N}{n}}$$

For fast x er denne sandsynlighed $\neq 0$ for $x \leq m$ og $n-x \leq N-m$, d.v.s. for $x \leq m \leq N-n+x$. Vi får derfor

$$\begin{aligned} f(x) &= \sum_{m=x}^{N-n+x} \frac{\binom{m}{x} \binom{N-m}{n-x}}{\binom{N}{n}} \binom{N}{m} p^m (1-p)^{N-m} \\ &= \binom{n}{x} p^x (1-p)^{n-x} \sum_{m=x}^{N-n+x} \binom{N-n}{m-x} p^{m-x} (1-p)^{N-m-n+x} \\ &= \binom{n}{x} p^x (1-p)^{n-x} \sum_{m=0}^{N-n} \binom{N-n}{m} p^m (1-p)^{N-n-m} \\ &= \binom{n}{x} p^x (1-p)^{n-x} \end{aligned}$$

d.v.s. $X \in B(n, p)$. ■

Vi illustrerer sætningens indhold i

EKSEMPEL 1.46. Vi betragter en fabrikationsproces, hvor der er en sandsynlighed p for, at et emne er defekt. Vi forudsætter endvidere, at emner er stokastisk uafhængige.

Vi har derfor givet en række Bernoulli forsøg med "succes"-sandsynligheden p . De fabrikerede dele forsendes i pakninger med N emner. Lad der nu i en pakning være m defekte. Vi udtager nu en stikprøve på n **uden** tilbagelægning. Da vil antallet af defekte i stikprøven fra **denne specielle** pakning følge en hypergeometrisk fordeling ifølge Afsnit 1.3.1. Hvis man imidlertid løbende aftager varer fra den pågældende fabrik, er det ikke så meget antallet af fejlemner i en specifik pakning, som er interessant, men mere fordelingen af antal fejl i en stikprøve fra en **vilkårlig** pakning, idet det er processens fejlsandsynlighed p , man da har brug for. Antallet af fejlemner i en stikprøve fra en vilkårlig pakning er så ifølge sætningen binomialt fordelt $B(n, p)$ på trods af, at der er tale om stikprøveudtagning **uden** tilbagelægning. Det forhold, at der er opdelt i pakninger à N emner inden udtagningen af stikprøven på n emner influerer (naturligvis) ikke på fordelingen af X . ♦

Endelig nævner vi

SÆTNING 1.73. Lad P være rektangulært over intervallet $[0, 1]$ og lad X for givet $P = p$ være binomialt fordelt med parametre (n, p) . Da er den ubetingede fordeling af X ligefordelingen over $(0, 1, \dots, n)$. Udtrykt i formler

$$P \in U(0, 1) \wedge (X|P = p) \in B(n, p) \Rightarrow X \in UD(n).$$

Bevis. Direkte anvendelse af sætning 1.70. ■

1.10 Fordelinger afledt af den normale fordeling

I forbindelse med statistisk analyse af normalt fordelte observationer optræder der en række nye afledte fordelinger. Vi indleder med

1.10.1 χ^2 -fordelingen

Vi introducerer χ^2 -fordelingen via

EKSEMPEL 1.47. Lad $X_1, \dots, X_n$ være stokastisk uafhængige $N(0, 1)$ fordelte variable. Vi vil bestemme fordelingen af $Z = X_1^2 + \dots + X_n^2$. Vi starter med at finde fordelingen for $Y = X_1^2$. Vi har for $y > 0$

$$\begin{aligned} P\{Y \leq y\} &= P\{X_1^2 \leq y\} \\ &= P\{-\sqrt{y} \leq X_1 \leq \sqrt{y}\} \\ &= 2P\{0 \leq X_1 \leq \sqrt{y}\} \\ &= 2 \frac{1}{\sqrt{2\pi}} \int_0^{\sqrt{y}} e^{-\frac{1}{2}t^2} dt \end{aligned}$$

Tætheden for Y fås ved differentiation af dette udtryk, d.v.s.

$$\begin{aligned} g(y) &= \frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}y} \frac{1}{2\sqrt{y}} \\ &= \frac{1}{\Gamma(\frac{1}{2}) 2^{\frac{1}{2}}} y^{-\frac{1}{2}} e^{-y/2}, \end{aligned}$$

d.v.s.: $Y \in G(\frac{1}{2}, 2)$. Ifølge sætning 1.19 er da $Z \in G(\frac{n}{2}, 2)$. ◆

En Γ -fordeling fremkommet på denne måde kaldes en χ^2 -fordeling. Vi præciserer dette i

DEFINITION 1.39. En stokastisk variabel $X \in G(\frac{n}{2}, 2)$ siges at være χ^2 -fordelt med n **frihedsgrader**. Vi skriver kort $X \in \chi^2(n)$. ▲

Det mærkelige navn frihedsgrader vil vi komme tilbage til i et senere kapitel. Med ovenstående definition kan vi reformulere resultatet i eksempel 1.47 til

SÆTNING 1.74. Lad $X_1, \dots, X_n$ være uafhængige $N(0, 1)$ -fordelte stokastiske variable. Da er

$$\sum_{i=1}^n X_i^2 \in \chi^2(n).$$

Vi kan uden besvær overføre resultaterne fra afsnittet om Γ -fordelingen. Vi nævner

SÆTNING 1.75 (REPRODUKTIVITETSSÆTNING). Lad X og Y være stokastisk uafhængige. Da gælder

$$X \in \chi^2(n_1) \wedge Y \in \chi^2(n_2) \Rightarrow X + Y \in \chi^2(n_1 + n_2).$$

Bevis. Sætningen følger direkte af sætning 1.19 eller af sætning 1.74. ■

Af sætning 1.18 fås umiddelbart

SÆTNING 1.76. Lad $X \in \chi^2(n)$. Da er middelværdi og varians lig

$$\begin{aligned} E(X) &= n \\ V(X) &= 2n. \end{aligned}$$

Vi anfører graferne for nogle χ^2 -tætheder

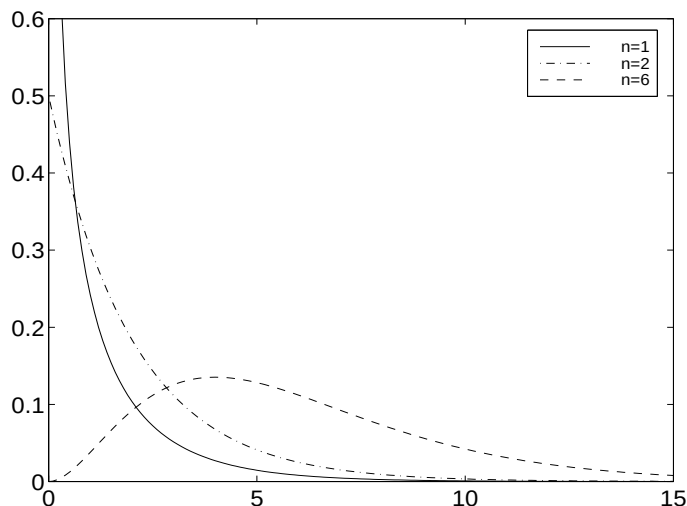
I de senere kapitler vil vi gøre udpræget brug af

SÆTNING 1.77 (SPALTNINGSSÆTNINGEN (COCHRANS SÆTNING)). Lad $Z_i, i = 1, \dots, n$, være uafhængige $N(0, 1)$ -fordelte stokastiske variable og

$$\sum_{i=1}^n Z_i^2 = Q_1 + Q_2 + \dots + Q_s \tag{1.28}$$

hvor $s \leq n$, og Q_i har n_i frihedsgrader. Da er $Q_1, Q_2, \dots, Q_s$ uafhængige χ^2 -fordelte størrelser med henholdsvis $n_1, n_2, \dots, n_s$ frihedsgrader, hvis og kun hvis

$$n = n_1 + n_2 + \dots + n_s \tag{1.29}$$



Bevis. Overspringes. ■

BEMÆRKNING 1.11. Da enhver stokastisk variabel $X_i \in N(\mu, \sigma^2)$ kan normeres

$$Z_i = \frac{X_i - \mu}{\sigma} \in N(0, 1)$$

jvnf. Bemærkningen side 127 ses det, at spaltningssætningen muliggør en opspaltning af **kvadratafvigelsessummen**

$$\sigma^2 \sum_{i=1}^n Z_i^2 = \sum_{i=1}^n (X_i - \mu)^2,$$

såfremt frihedsgraderne af de enkelte led opfylder (1.29). ▼

Vi har indført χ^2 -fordelingen som fordeling for en kvadratsum af $N(0, 1)$ -fordelte variable. Hvis middelværdierne i normalfordelingerne ikke længere er 0, får vi en anden fordeling, nemlig den såkaldte ikke-centrale χ^2 -fordeling. Vi definerer denne ved

DEFINITION 1.40. Vi siger, at den stokastiske variabel X er ikke-central χ^2 -fordelt med n frihedsgrader og skævhedsparameter γ^2 , hvis ($\mathcal{L}$ = law = fordelingen) for

$$\mathcal{L}(X) = \mathcal{L}(X_1^2 + \cdots + X_n^2),$$

hvor $X_1, \dots, X_n$ er stokastisk uafhængige, og $X_i \in N(\mu_i, 1)$, $i = 1, \dots, n$, hvor parametrene $\mu_1, \dots, \mu_n$ tilfredsstiller $\sum_{i=1}^n \mu_i^2 = \gamma^2$. Vi skriver kort $X \in \chi^2(n, \gamma^2)$. ▲

BEMÆRKNING 1.12. Som mnemoteknisk regel kan vi anvende skrivemåden

$$\chi^2 \left(n, \sum \mu_i^2 \right) = \sum_{i=1}^n N(\mu_i, 1)^2. \quad (1.30)$$
▼

Det må tilføjes, at definitionen først er tilladelig, når man har godtgjort, at fordelingen $X_1^2 + \dots + X_n^2$ alene afhænger af $\mu_1, \dots, \mu_n$ og ikke af størrelsen af de enkelte μ 'er. Dette forbigås dog.

En triviell følge af definitionen og sætning 1.74 er

SÆTNING 1.78. Der gælder, at $\chi^2(n, 0) = \chi^2(n)$ ($= G((\frac{n}{2}, 2))$).

BEMÆRKNING 1.13. Der findes mange tabeller over fraktiler i χ^2 -fordelinger. Disse kan også bruges til at bestemme sandsynligheder i Γ -fordelinger med halvtallig parameter, idet vi har

$$\begin{aligned} P \left\{ G \left(\frac{n}{2}, \beta \right) \leq x \right\} &= P \left\{ \frac{2}{\beta} G \left(\frac{n}{2}, \beta \right) \leq \frac{2}{\beta} x \right\} \\ &= P \left\{ G \left(\frac{n}{2}, 2 \right) \leq \frac{2}{\beta} x \right\} \\ &= P \left\{ \chi^2(n) \leq \frac{2}{\beta} x \right\}. \end{aligned}$$
▼

Til sidst nævner vi en nyttig sætning om sammenhængen mellem Poisson- og χ^2 -fordelingen, nemlig

SÆTNING 1.79. Lad $X \in P(\lambda)$. Da er

$$P\{X \leq x\} = 1 - P\left\{\chi^2(2(x+1)) \leq 2\lambda\right\}.$$

Bevis. Sætningen følger af sætning 1.16 om Erlang fordelingen som ventetidsfordeling i en Poisson proces. En Erlang fordeling med parametre (k, λ) er jo en $G(k, \frac{1}{\lambda})$ -fordeling $= \frac{1}{2\lambda} \cdot \chi^2(2k)$ -fordeling, og heraf følger resultatet. Detaljerne forbigås. ■

1.10.2 Rayleigh fordelingen

I dette afsnit skal vi betragte en fordeling, der optræder ret hyppigt i teoretisk fysik. Vi indleder med et

EKSEMPEL 1.48. Vi betragter en partikel med koordinater (X, Y) , hvor X og Y er indbyrdes uafhængige $N(0, \sigma^2)$ -fordelte stokastiske variable. Vi søger nu fordelingen af partiklens afstand Z til origo.

Vi finder

$$\begin{aligned} P\{Z \leq z\} &= P\left\{\sqrt{X^2 + Y^2} \leq z\right\} \\ &= P\{X^2 + Y^2 \leq z^2\} \\ &= P\{\sigma^2 \chi^2(2) \leq z^2\} \\ &= G\left(\frac{z^2}{\sigma^2}\right), \end{aligned}$$

hvor G er fordelingsfunktionen for en $\chi^2(2)$ -fordeling. Sætter vi $G' = g$ finder vi tætheden

$$\begin{aligned} f(z) &= \frac{2z}{\sigma^2} g\left(\frac{z^2}{\sigma^2}\right) \\ &= \frac{1}{\sigma^2} z e^{-\frac{1}{2\sigma^2} z^2}, \quad z > 0. \end{aligned} \tag{1.31}$$



DEFINITION 1.41. En stokastisk variabel Z med tætheden (1.31) siges at være Rayleigh fordelt med parameter σ^2 , og vi skriver $Z \in \text{Ray}(\sigma^2)$. ▲

SÆTNING 1.80. Middelværdi og varians i en $\text{Ray}(\sigma^2)$ -fordeling er

$$\begin{aligned} E(Z) &= \sigma \frac{\sqrt{2}}{2} \sqrt{\pi} \\ V(Z) &= \sigma^2 \left(2 - \frac{1}{2}\pi\right). \end{aligned}$$

Bevis. Forbigås. ■

BEMÆRKNING 1.14. Vi ser, at når $Z \in \text{Ray}(\sigma^2)$, vil $Z^2 \in \sigma^2 \chi^2(2)$. Man anvender derfor også ofte betegnelsen χ -fordelingen i stedet for Rayleigh fordelingen. ▼

1.10.3 Student's t-fordeling

Vi indfører t-fordelingen ved

DEFINITION 1.42. Vi siger, at den stokastiske variable Z er **ikke centralt Student fordelt** eller **t-fordelt** med n **frihedsgrader** og **skævhedsparameter** μ , hvis ($\mathbb{E}$ =law=fordelingen for)

$$\mathbb{E}(Z) = \mathbb{E} \left(\frac{X}{\sqrt{Y/n}} \right),$$

hvor X og Y er stokastisk uafhængige, og $X \in N(\mu, 1) \wedge Y \in \chi^2(n) = G(\frac{n}{2}, 2)$. Vi skriver kort $Z \in t(n, \mu)$. ▲

BEMÆRKNING 1.15. Som mnemoteknisk skrivemåde kan vi anvende

$$t(n, \mu) = \frac{N(\mu, 1)}{\sqrt{\chi^2(n)/n}} \quad (1.32)$$

▼

Det meget vigtige specialtilfælde $\mu = 0$ omtales specielt.

DEFINITION 1.43. En stokastisk variabel $Z \in t(n, 0)$ siges at være **Student-** eller **t-fordelt** med n **frihedsgrader**. Vi skriver $Z \in t(n)$. ▲

BEMÆRKNING 1.16. Mnemoteknisk regel.

$$t(n) = \frac{N(0, 1)}{\sqrt{\chi^2(n)/n}} \quad (1.33)$$

▼

Tætheden for en ikke-central t-fordeling er meget kompliceret at skrive op. Det er noget lettere for den centrale fordeling, og vi har

SÆTNING 1.81. Lad $X \in t(n)$. Da er tætheden

$$f(x) = \frac{\Gamma(\frac{1}{2}(n+1))}{\sqrt{n\pi} \Gamma(\frac{n}{2})} \left(\frac{x^2}{n} + 1 \right)^{-\frac{n+1}{2}}.$$

Middelværdien eksisterer for $n > 1$ og er

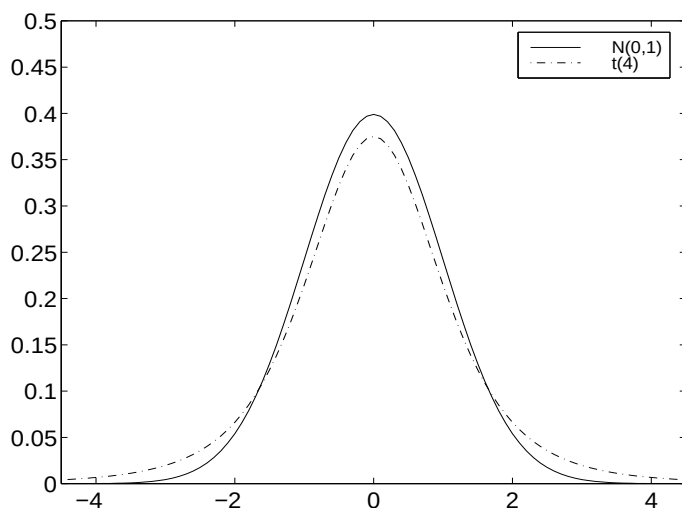
$$E(X) = 0 \quad (n > 1).$$

Variansen eksisterer for $n > 2$; den er

$$V(X) = \frac{n}{n-2} \quad (n > 2).$$

Bevis. Forbigås. ■

t-fordelingen er symmetrisk om 0 og minder om normalfordelingen. I nedenstående figur har vi anført grafen for tætheden for en t-fordeling og for en normalfordeling.



Af figuren fremgår, at $t(4)$ -fordelingen er "fladere" end normalfordelingen. Dette er selvfølgelig blot et udtryk for, at $V(t(4)) = 2 > V(N(0,1)) = 1$.

Forskellen mellem t-fordelingen og normalfordelingen mindskes med voksende antal frihedsgrader, idet vi har

SÆTNING 1.82. Tætheden for en $t(n)$ -fordeling konvergerer for $n \rightarrow \infty$ mod tætheden for en $N(0, 1)$ -fordeling.

Bevis. Forbigås. Fås direkte ved anvendelse af Stirling's formel. ■

SÆTNING 1.83. En Student fordeling med 1 frihedsgrad er en Cauchy fordeling med parametre $(0, 1)$.

Bevis. Trivielt. ■

1.10.4 F-fordelingen

Vi indleder med

DEFINITION 1.44. Vi siger, at en stokastisk variabel Z er **ikke centralt F-fordelt** med (n, m) **frihedsgrader** og **skævhedsparemeter** γ^2 , hvis

$$\mathbb{E}(Z) = \mathbb{E}\left(\frac{X/n}{Y/m}\right),$$

hvor X og Y er stokastisk uafhængige, og $X \in \chi^2(n, \gamma^2) \wedge Y \in \chi^2(m)$. ▲

Vi skriver kort $Z \in F(n, m; \gamma^2)$.

BEMÆRKNING 1.17. Mnemoteknisk skrives definitionen

$$F(n, m; \gamma^2) = \frac{\chi^2(n, \gamma^2)/n}{\chi^2(m)/m}. \quad (1.34)$$

▼

Hvis $\gamma^2 = 0$ har vi

DEFINITION 1.45. Hvis $Z \in F(n, m; 0)$, siges Z at være **F-fordelt** med (n, m) **frihedsgrader**, og vi skriver $Z \in F(n, m)$. ▲

BEMÆRKNING 1.18. Ovenstående mnemotekniske regel giver da

$$F(n, m) = \frac{\chi^2(n)/n}{\chi^2(m)/m}. \quad (1.35)$$

▼

Det må her indskræpes, at når vi i (1.30)-(1.35) har givet nogle hukommelsesvenlige skriveformer, så er disse ufuldstændige. Det er nemlig ganske fundamentalt, at **alle** de involverede stokastiske variable er **stokastisk uafhængige**. Dette fremgår ikke af huskereglene.

Vi anfører tætheden for den centrale F-fordeling i

SÆTNING 1.84. Lad $X \in F(n, m)$. Da har X tætheden

$$f(x) = \frac{1}{B\left(\frac{1}{2}n, \frac{1}{2}m\right)} \left(\frac{n}{m}\right)^{\frac{1}{2}n} \frac{x^{\frac{1}{2}n-1}}{\left(1 + \frac{n}{m}x\right)^{\frac{1}{2}(n+m)}}, \quad x > 0$$

og 0 ellers. Middelværdien eksisterer for $m > 2$ og er da

$$E(X) = \frac{m}{m-2} \quad (m > 2).$$

Variansen eksisterer for $m > 4$ og er da

$$V(X) = \frac{m^2(2n+2m-4)}{(m-2)^2(m-4)n} \quad (m > 4).$$

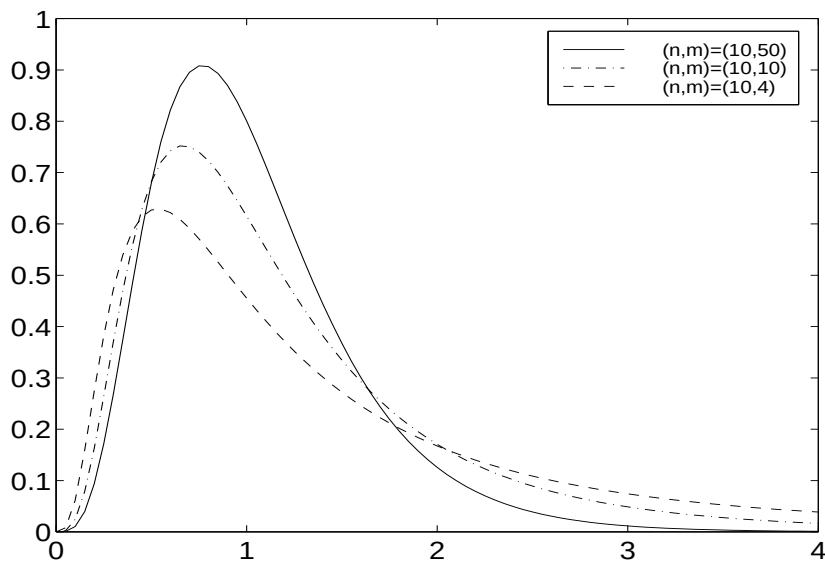
Bevis. Forbigås. ■

F-fordelingen indeholder 2 parametre, og disses betydning for tæthedens udseende fremgår af grafen.

BEMÆRKNING 1.19. Ved udregninger kan følgende relation ofte være nyttig:

$$F(n, m)_\alpha = \frac{1}{F(m, n)_{1-\alpha}},$$

idet F-fordelingen typisk kun er tabelleret for $\alpha \geq 50\%$. ▼



Der findes en sammenhæng mellem t-fordelingen og F-fordelingen, idet vi har

SÆTNING 1.85. Lad T være t-fordelt med n frihedsgrader. Da er $T^2 \sim F(1, n)$ -fordelt, d.v.s.

$$T \in t(n) \Rightarrow T^2 \in F(1, n).$$

Bevis. Lad X og Y være stokastisk uafhængige. Hvis $X \in N(0, 1)$ og $Y \in \chi^2(n)$ er

$$T = \frac{X}{\sqrt{Y/n}} \in t(n).$$

Endvidere er

$$T^2 = \frac{X^2}{Y/n} = \frac{X^2/1}{Y/n} \in F(1, n)$$

ifølge sætning 1.74 og definition 1.45. ■

Der er også en tæt sammenhæng imellem B-fordelingen og F-fordelingen. Vi udtrykker den i

SÆTNING 1.86. Lad X og Y være stokastiske variable. Da gælder

$$\begin{aligned} X \in \text{Be}(n, m) &\Rightarrow \frac{m}{n} \frac{X}{1-X} \in F(n, m) \\ Y \in F(n, m) &\Rightarrow \frac{nY}{nY+m} \in \text{Be}(n, m). \end{aligned}$$

Bevis. Forbigås. ■

Vi nævner til slut en nyttig sætning om sammenhængen mellem tætheden for en F-fordeling og en binomialfordeling.

SÆTNING 1.87. Lad $X \in B(n, p)$. Da er

$$P\{X \leq x\} = 1 - P\left\{F(2(x+1), 2(n-x)) \leq \frac{n-x}{x+1} \frac{p}{1-p}\right\}.$$

Bevis. Forbigås. ■

Navn Symbol	Støtte Parameter- område	Frekvensfunktion	Middelværdi Varsians	Side
Bernoulli $B(1, p)$	$\{0, 1\}$ $[0, 1]$	$p^x(1-p)^{1-x}$	p $p(1-p)$	91
Beta $Be(\alpha, \beta)$	$]0, 1[$ $]0, \infty[\times]0, \infty[$	$\frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$ $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	181
Binomial $B(n, p)$	$\{0, 1, \dots, n\}$ $N \times [0, 1]$	$\binom{n}{x} p^x (1-p)^{n-x}$	np $np(1-p)$	92
Cauchy $C(\alpha, \beta)$	R $R \times]0, \infty[$	$\frac{\beta}{\pi} \frac{1}{\beta^2 + (x-\alpha)^2}$	- -	183
Chi i anden $\chi^2(n), G(\frac{n}{2}, 2)$	$]0, \infty[$ N	$\frac{1}{\Gamma(\frac{n}{2}) 2^{\frac{n}{2}}} x^{\frac{n}{2}-1} e^{-x/2}$	n $2n$	194
Chi i anden (ikke central) $\chi^2(n, \gamma^2)$	$]0, \infty[$ $N \times]0, \infty[$			194
Erlang $G(k, \frac{1}{\lambda})$	$]0, \infty[$ $N \times]0, \infty[$	$\frac{1}{\Gamma(k)} \lambda^k x^{k-1} e^{-\lambda x}$	$\frac{k}{\lambda}$ $\frac{k}{\lambda^2}$	118
Exponential $Ex(\beta), G(1, \beta)$	$]0, \infty[$ $]0, \infty[$	$\frac{1}{\beta} e^{-\frac{x}{\beta}}$	β β^2	120
F-fordeling $F(n, m)$	$]0, \infty[$ $N \times N$	$\frac{(n/m)^{\frac{1}{2}n}}{B(\frac{1}{2}n, \frac{1}{2}m)} \frac{x^{\frac{1}{2}n-1}}{(1+\frac{n}{m}x)^{\frac{1}{2}(n+m)}}$	$\frac{m}{m-2}, (m > 2)$ $\frac{m^2(2n+2m-4)}{(m-2)^2(m-4)n}, (m > 4)$	201
F-fordeling (ikke central) $F(n, m; \gamma^2)$	$]0, \infty[$ $N \times N \times]0, \infty[$			201

Navn Symbol	Støtte Parameter- område	Frekvensfunktion	Middelværdi Varians	Side
Gamma $G(k, \beta)$	$]0, \infty[$ $]0, \infty[\times]0, \infty[$	$\frac{1}{\Gamma(k)\beta^k} x^{k-1} e^{-\frac{x}{\beta}}$	$k\beta$ $k\beta^2$	120
Geometrisk $NB(1, p)$	$\{0, 1, 2, \dots\}$ $]0, 1[$	$p(1-p)^x$	$\frac{1-p}{p}$ $\frac{1-p}{p^2}$	99
Hypergeometrisk $H(n, M, N)$	$\{\max(0, n+M-N), \dots, \min(n, M)\}$ $(n, M, N) \in \mathbb{N} \times \mathbb{N} \times \mathbb{N}, M \leq N, n \leq N$	$\frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}$	$n \cdot \frac{M}{N}$ $n \cdot \frac{M}{N} \left(1 - \frac{M}{N}\right) \cdot \frac{N-n}{N-1}$	102
LaPlace $La(\alpha, \beta)$	$\mathbb{R}$ $\mathbb{R} \times]0, \infty[$	$\frac{1}{2\beta} e^{-\frac{1}{\beta} x-\alpha }$	α $2\beta^2$	184
Ligefordeling på $\{0, 1, \dots, n\}$ $UD(n)$	$\{0, 1, \dots, n\}$ $\mathbb{N}$	$\frac{1}{n+1}$	$\frac{1}{2}n$ $\frac{n(n+2)}{12}$	187
Logaritmisk $LoD(\theta)$	$\mathbb{N}$ $]0, 1[$	$-\frac{1}{\log_e(1-\theta)} \frac{\theta^x}{x}$	$\frac{-\theta}{(1-\theta)\log_e(1-\theta)}$ $-\frac{\theta^2 + \theta \log_e(1-\theta)}{(1-\theta)^2 \log_e^2(1-\theta)}$	187
Logaritmisk normal $LN(\alpha, \beta^2)$	$]0, \infty[$ $\mathbb{R} \times]0, \infty[$	$\frac{\exp\left(-\frac{1}{2}\left(\frac{\log_e x - \alpha}{\beta}\right)^2\right)}{x\beta\sqrt{2\pi}}$	$e^{\alpha + \frac{1}{2}\beta^2}$ $e^{2\alpha + \beta^2} \left(e^{\beta^2} - 1\right)$	134
Logistisk $L(\alpha, \beta)$	$\mathbb{R}$ $\mathbb{R} \times]0, \infty[$	$\frac{\exp\left(-\frac{x-\alpha}{\beta}\right)}{\left(1 + \exp\left(-\frac{x-\alpha}{\beta}\right)\right)^2}$	α $\frac{\beta^2 \pi^2}{3}$	185

Navn Symbol	Støtte Parameter- område	Frekvensfunktion	Middelværdi Varsians	Side
Max-1 (dobbel exponential) $\text{Max}_1(\alpha, \beta)$	$\mathbb{R}$ $\mathbb{R} \times]0, \infty[$	$\frac{1}{\beta} \exp\left(-\frac{x-\alpha}{\beta} - \exp\left(-\frac{x-\alpha}{\beta}\right)\right)$	$\alpha + \beta\gamma \simeq \alpha + 0.57722\beta$ $\frac{\beta^2 \pi^2}{6}$	151
Max-2 $\text{Max}_2(k, \beta)$	$]0, \infty[$ $]0, \infty[\times]0, \infty[$	$\frac{k}{\beta} \left(\frac{\beta}{x}\right)^{k+1} \exp\left(-\left(\frac{\beta}{x}\right)^k\right)$	$\beta\Gamma\left(1 - \frac{1}{k}\right) \quad (k > 1)$ $\beta^2 \left(\Gamma\left(1 - \frac{2}{k}\right) - \Gamma^2\left(1 - \frac{1}{k}\right)\right) \quad (k > 2)$	162
Max-3 $\text{Max}_3(k; \alpha, \beta)$	$] - \infty, \alpha[$ $]0, \infty[\times \mathbb{R} \times]0, \infty[$	$\frac{k}{\beta} \left(-\frac{x-\alpha}{\beta}\right)^{k-1} \exp\left(-\left(-\frac{x-\alpha}{\beta}\right)^k\right)$	$\alpha - \beta\Gamma\left(1 + \frac{1}{k}\right)$ $\beta^2 \left(\Gamma\left(1 + \frac{2}{k}\right) - \Gamma^2\left(1 + \frac{1}{k}\right)\right)$	170
Min-1 (dobbel exponential) $\text{Min}_1(\alpha, \beta)$	$\mathbb{R}$ $\mathbb{R} \times]0, \infty[$	$\frac{1}{\beta} \exp\left(\frac{x-\alpha}{\beta} - \exp\left(\frac{x-\alpha}{\beta}\right)\right)$	$\alpha - \beta\gamma \simeq \alpha - 0.57722\beta$ $\frac{\beta^2 \pi^2}{6}$	157
Min-2 $\text{Min}_2(k, \beta)$	$] - \infty, 0[$ $]0, \infty[\times]0, \infty[$	$\frac{k}{\beta} \left(-\frac{\beta}{x}\right)^{k+1} \exp\left(-\left(-\frac{\beta}{x}\right)^k\right)$	$-\beta\Gamma\left(1 - \frac{1}{k}\right) \quad (k > 1)$ $\beta^2 \left(\Gamma\left(1 - \frac{2}{k}\right) - \Gamma^2\left(1 - \frac{1}{k}\right)\right) \quad (k > 2)$	164
Min-3 $\text{Min}_3(k; \alpha, \beta)$	$]\alpha, \infty[$ $]0, \infty[\times \mathbb{R} \times]0, \infty[$	$\frac{k}{\beta} \left(\frac{x-\alpha}{\beta}\right)^{k-1} \exp\left(-\left(\frac{x-\alpha}{\beta}\right)^k\right)$	$\alpha + \beta\Gamma\left(1 + \frac{1}{k}\right)$ $\beta^2 \left(\Gamma\left(1 + \frac{2}{k}\right) - \Gamma^2\left(1 + \frac{1}{k}\right)\right)$	172
Negativ binomial (Pascal) $\text{NB}(k, p)$	$\{0, 1, 2, \dots\}$ $\mathbb{N} \times [0, 1]$	$\binom{k+x-1}{x} p^k (1-p)^x$	$\frac{k(1-p)}{p}$ $\frac{k(1-p)}{p^2}$	99
Normal (Gauss) $\text{N}(\mu, \sigma^2)$	$\mathbb{R}$ $\mathbb{R} \times]0, \infty[$	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$	μ σ^2	104

Navn Symbol	Støtte Parameter- område	Frekvensfunktion	Middelværdi Varsians	Side
Pareto $\text{Par}(k, \beta)$	$] \beta, \infty[$ $] 0, \infty[\times] 0, \infty[$	$\frac{k\beta^k}{x^{k+1}}$	$\frac{\beta k}{(k-1)}, k > 1$ $\frac{\beta^2 k}{(k-1)^2(k-2)}, (k > 2)$	186
Poisson $P(\lambda)$	$\{0, 1, 2, \dots\}$ $] 0, \infty[$	$\frac{\lambda^x}{x!} e^{-\lambda}$	λ λ	108
Polynomial $\text{Pol}(n, p_1, \dots, p_k)$	$\{0, 1, \dots, n\}^k, \sum_{i=1}^k x_i = n$ $N \times [0, 1]^k, \sum_{i=1}^k p_i = 1$	$\binom{n}{x_1 \dots x_k} p_1^{x_1} \dots p_k^{x_k}$		104
Rayleigh $\text{Ray}(\sigma^2)$	$] 0, \infty[$ $] 0, \infty[$	$\frac{1}{\sigma^2} x e^{-\frac{1}{2\sigma^2} x^2}$	$\sigma \frac{\sqrt{2}}{2} \sqrt{\pi}$ $\sigma^2 (2 - \frac{1}{2}\pi)$	198
Rektangulær (Uniform) $U(a, b)$	$] a, b[$ $R^2, a < b$	$\frac{1}{b-a}$	$\frac{a+b}{2}$ $\frac{1}{12}(b-a)^2$	180
t-fordeling (Student's t) $t(n)$	R N	$\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi} \Gamma(\frac{n}{2})} \left(\frac{x^2}{n} + 1 \right)^{-\frac{n+1}{2}}$	$0 \quad (n > 1)$ $\frac{n}{n-2} \quad (n > 2)$	199
t-fordeling (ikke central) $t(n, \mu)$	R			199
Uniform, se Rektangulær	$N \times R$			180
Weibull $\text{We}(k, \beta)$ ($\text{Min}_3(k; 0, \beta)$)	$] 0, \infty[$ $] 0, \infty[\times] 0, \infty[$	$\frac{k}{\beta} \left(\frac{x}{\beta} \right)^{k-1} \exp \left(- \left(\frac{x}{\beta} \right)^k \right)$	$\beta \Gamma \left(1 + \frac{1}{k} \right)$ $\beta^2 \left(\Gamma \left(1 + \frac{2}{k} \right) - \Gamma^2 \left(1 + \frac{1}{k} \right) \right)$	172

Kapitel 2

Estimationsteori

I det foregående kapitel har vi primært beskæftiget os med sandsynlighedsteoretiske modeller. Det er vigtigt, at man gør sig klart, at disse modeller er **abstraktioner**, sålænge de ikke er knyttet til **målinger** af fysiske fænomener, d.v.s. sålænge de ikke på en eller anden måde baseres på observerede data. I dette kapitel vil vi se, hvorledes vi på basis af realiserede udfald af en stokastisk variabel kan danne et skøn over nogle af de parametre, som indgår i fordelingen af den stokastiske variable.

2.1 Generelt om estimationsteori

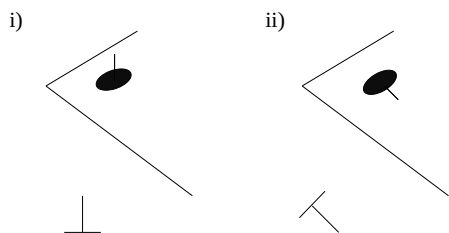
Vi indleder med et afsnit om inferens.

2.1.1 Statistisk inferens

Vi skal ikke her prøve på at give en udtømmende beskrivelse af begrebet statistisk inferens, men blot prøve at give en fornemmelse af det gennem nogle meget simple eksempler.

Lad os betragte et meget enkelt fysisk eksperiment, nemlig det at kaste en tegnestift op i luften og observere, hvorledes den lander. Der er kun to mulige stillinger for tegnestiften.

Vi kalder den første hændelse for succes (1) og den anden for fiasko (0). Forudsætter vi, hvad der må forekomme rimeligt, at udfaldet af forskellige kast er uafhængige, og



at sandsynligheden for succes er konstant fra kast til kast, har vi en følge af Bernoulli forsøg. Hvis vi nu kender sandsynligheden p for succes, kan vi, som det fremgår af afsnit 1.2, beregne sandsynligheden for at observere hændelser såsom det at vente 5 gange, før man får en succes, eller finde sandsynligheden for at få 4 succeser i 10 kast etc.

Sådanne betragtninger er ren sandsynlighedsregning. Systemets sandsynlighedsteoretiske adfærd er givet fra starten. Vore undersøgelser er ren **deduktion**, d.v.s. i det omfang forudsætningerne er rigtige, er konklusionerne det også.

Lad os nu antage, at vi ikke kender p , men er interesserede i denne størrelse. Hvis vi **observerer** en sekvens

0 1 0 1 1 0 ,

ville de fleste sikkert sige, at p nok er lig $\frac{1}{2}$, fordi den relative hyppighed af succeser er lig $\frac{1}{2}$. Havde vi i stedet fået

0 0 0 0 0 1 ,

ville man vel gætte på $p = \frac{1}{6}$, idet den relative hyppighed nu er lig $\frac{1}{6}$.

Ovenstående udsagn er eksempler på **statistisk inferens** eller **induktive slutninger**, i.e. slutninger fra det specielle (vore observationer) til det generelle (den fordelingslov, der ligger bag disse og bag fremtidige observationer). I og med at vi har sagt, at vi tror, at $p = \frac{1}{2}$, har vi også sagt, at vi tror, at sandsynligheden for at få succes i **fremtidige** kast, er $\frac{1}{2}$.

De inferensproblemer, vi vil beskæftige os med i dette kapitel, svarer til problemet med at gætte (danne et skøn) på en værdi af p . I det næste afsnit vil vi formalisere denne problemstilling.

2.1.2 Estimationsproblematikken

Lad der være givet n stokastiske variable $X_1, \dots, X_n$, hvis simultane fordeling er specificeret fuldstændigt på nær en ukendt **parameter** θ . Vi kalder den simultane frekvensfunktion $f(x_1, \dots, x_n; \theta)$.

Hvis vi på basis af et realiseret udfald af $(X_1, \dots, X_n)$ ønsker at udtale os om θ 's værdi, siger vi, at vi ønsker at **estimere** θ .

Med henblik på en mere stringent formulering af estimationsproblemet giver vi

DEFINITION 2.1. En funktion t af $X_1, \dots, X_n$ kaldes en **stikprøvefunktion** eller et **observat** (engelsk: **statistic**). Vi bemærker, at stikprøvefunktionen

$$T = t(X_1, \dots, X_n)$$

er en stokastisk variabel. ▲

Som eksempler på stikprøvefunktioner kan vi nævne størsteværdien:

$$T_1 = X_{(n)}$$

eller gennemsnittet:

$$T_2 = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

DEFINITION 2.2. Lad $(x_1, \dots, x_n)$ være et realiseret udfald af $(X_1, \dots, X_n)$. Hvis vi anvender $t(x_1, \dots, x_n)$ som et "gæt" på værdien af θ , kaldes den konstaterede værdi

$$t(x_1, \dots, x_n) \quad \text{et \textbf{estimat} af } \theta$$

og den tilsvarende stokastiske variabel

$$T = t(X_1, \dots, X_n) \quad \text{en \textbf{estimator} for } \theta.$$

▲

Vi har altså, at estimatet er et realiseret udfald af estimatoren, som er en stokastisk variabel. Med ovenstående er der næsten ikke lagt nogle bånd på, hvad vi skal forstå ved en estimator. F.eks. vil den konstante afbildning

$$t(X_1, \dots, X_n) = 7$$

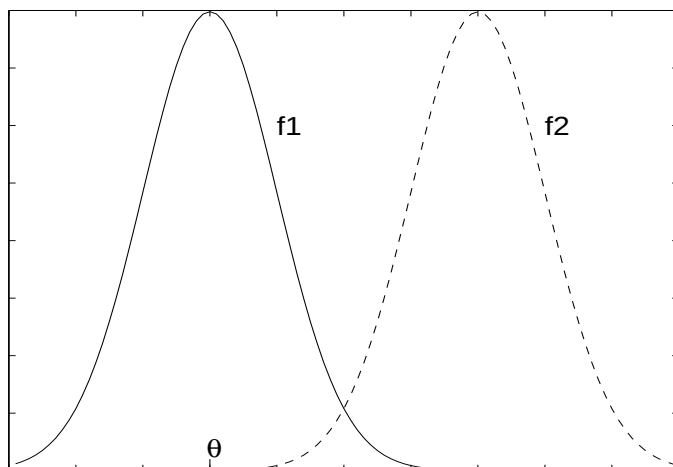
være en estimator for θ , men den er selvfølgelig ikke særlig "god", hvis θ ligger langt fra 7. Problemet i estimationsteorien er derfor dels at definere, hvad vi skal forstå ved en "god" estimator, dels at bestemme stikprøvefunktioner, som tilfredsstiller disse kriterier.

2.2 Estimatorers egenskaber

Det første krav, man naturligt formulerer, er ønsket om, at estimatet ikke må afvige systematisk fra parameteren. Vi indleder derfor med

2.2.1 Centrale estimatorer

Lad os betragte 2 estimatorer T_1 og T_2 . De har som stokastiske variable frekvensfunktioner f_1 og f_2 . Lad disse være som skitseret herunder.



Hvis vi på basis af mange uafhængige gentagelser af vort eksperiment beregner T_1 og T_2 , vil vi forvente, at værdierne af T_1 grupperer sig omkring θ , hvorimod værdierne af T_2 vil gruppere sig omkring et tal til højre for θ . Man vil derfor normalt foretrække T_1 , og vi siger, at T_1 er central. Den præcise definition er

DEFINITION 2.3. Lad der være givet stokastiske variable $X_1, \dots, X_n$ med simultan frekvensfunktion $f(x_1, \dots, x_n; \theta)$. En estimator

$T = t(X_1, \dots, X_n)$ siges at være en **central estimator** (engelsk: unbiased estimator) for θ , hvis middelværdien af T er θ , når den sande parameterværdi er θ , i.e. såfremt $E(T) = \theta$ d.v.s.:

$$\int \cdots \int t(x_1, \dots, x_n) f(x_1, \dots, x_n; \theta) dx_1 \cdots dx_n = \theta$$

respektivt

$$\sum \cdots \sum t(x_1, \dots, x_n) f(x_1, \dots, x_n; \theta) = \theta,$$

hvis X er af kontinuert, respektivt af diskret type. ▲

EKSEMPEL 2.1. Lad $X_1, \dots, X_n$ være stokastiske variable med samme middelværdi μ . Det er da nærliggende at anvende gennemsnittet $\bar{X}$ som estimator for μ . Lad os undersøge, om gennemsnittet er en central estimator for μ . Vi har

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

d.v.s. $\bar{X}$ er en central estimator for μ . ♦

EKSEMPEL 2.2. Lad $X_1, \dots, X_n$ være indbyrdes uafhængige stokastiske variable med samme middelværdi μ og varians σ^2 . Man kunne foreslå at anvende

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

som estimator for σ^2 . Lad os finde den matematiske forventning af S_*^2 . Vi betragter først summen. Vi har

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X})^2 &= \sum_{i=1}^n (X_i^2 + \bar{X}^2 - 2X_i\bar{X}) \\ &= \sum_{i=1}^n X_i^2 + n\bar{X}^2 - 2\bar{X} \sum_{i=1}^n X_i \\ &= \sum_{i=1}^n X_i^2 + n\bar{X}^2 - 2\bar{X}n\bar{X} \end{aligned}$$

d.v.s.

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

Dette mellemresultat er en vigtig formel, som vil blive benyttet meget i det følgende.

Vi bemærker nu, at vi for en stokastisk variabel Z har

$$V(Z) = E(Z^2) - (E(Z))^2$$

i.e.:

$$E(Z^2) = V(Z) + (E(Z))^2.$$

Nu er $V(\bar{X}) = \frac{1}{n}\sigma^2$, hvorfor vi får

$$\begin{aligned} E\left(\sum_{i=1}^n (X_i - \bar{X})^2\right) &= \sum_{i=1}^n E(X_i^2) - n E(\bar{X}^2) \\ &= \sum_{i=1}^n (\sigma^2 + \mu^2) - n \left(\frac{1}{n}\sigma^2 + \mu^2\right) \\ &= (n-1)\sigma^2. \end{aligned}$$

Derfor er

$$E(S_*^2) = \frac{n-1}{n}\sigma^2,$$

d.v.s. S_*^2 er **ikke** en central estimator for σ^2 . Da afvigelsen kun er en proportionalitetsfaktor (uafhængig af σ^2) ses, at vi let kan danne en central estimator på basis af S_*^2 . Sætter vi nemlig

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

fås

$$E(S^2) = \frac{n}{n-1}E(S_*^2) = \frac{n}{n-1} \frac{n-1}{n}\sigma^2 = \sigma^2,$$

d.v.s. S^2 er en central estimator for variansen. I praksis anvendes S^2 hyppigst fremfor S_*^2 . ♦

Man kunne her måske fristes til at tro, at kravet om centralitet entydigt fastlægger en estimator. Dette er dog langt fra tilfældet, som det fremgår af

EKSEMPEL 2.3. Vi betragter nu igen situationen i eksempel 2.1. Vi sætter

$$\tilde{X} = \frac{1X_1 + 2X_2 + \cdots + nX_n}{1 + 2 + \cdots + n} = \frac{1}{\sum i} \sum iX_i.$$

Vi får umiddelbart

$$E(\tilde{X}) = \frac{1}{\sum i} \sum i E(X_i) = \frac{1}{\sum i} \mu \sum i = \mu,$$

d.v.s. at $\tilde{X}$ også er en central estimator for middelværdien μ . ♦

Kravet om centralitet er altså ikke tilstrækkeligt til at fastlægge en estimator, således at vi må søge andre kriterier.

2.2.2 Konsistente estimatorer

Et minimumskrav at stille til en estimator T for en parameter θ må være, at den giver et mere nøjagtigt skøn, når den beregnes på basis af mange observationer i forhold til få. Vi giver derfor følgende

DEFINITION 2.4. En følge af estimators $T_n = t_n(X_1, \dots, X_n)$ siges at være **konsistente** skøn over θ , såfremt

$$P\{|T_n - \theta| > \epsilon\} \rightarrow 0, \quad n \rightarrow \infty$$

for ethvert $\epsilon > 0$. ▲

BEMÆRKNING 2.1. I sandsynlighedsregningen benævner man den form for konvergens, der er anvendt i definitionen, for **konvergens i sandsynlighed**. Udtrykt verbalt står der, at sandsynligheden for at observere afvigelser fra θ , der er større end et vilkårligt positivt tal ϵ , konvergerer mod 0 for $n \rightarrow \infty$. ▼

Vi illustrerer begrebet i

EKSEMPEL 2.4. Vi vil ved hjælp af Čebyčev's ulighed godtgøre, at X/n er et konsistent skøn over p , hvis $X \in B(n, p)$.

Da $E(X) = np$, og $V(X) = np(1-p)$, har vi ifølge Čebyčev's ulighed, at

$$P \left\{ |X - np| > k \sqrt{np(1-p)} \right\} < \frac{1}{k^2},$$

eller

$$P \left\{ \left| \frac{X}{n} - p \right| > k \sqrt{\frac{p(1-p)}{n}} \right\} < \frac{1}{k^2}.$$

Heraf fås for et vilkårligt $\epsilon > 0$, at

$$P \left\{ \left| \frac{X}{n} - p \right| > \epsilon \right\} < \frac{p(1-p)}{n\epsilon^2},$$

d.v.s.

$$P \left\{ \left| \frac{X}{n} - p \right| > \epsilon \right\} \rightarrow 0, \quad n \rightarrow \infty.$$

Altså er X/n et konsistent skøn over p . ♦

Vi har følgende sætning om bestemmelse af konsistente estimators.

SÆTNING 2.1. Lad $X_1, \dots, X_n$ være uafhængige, identisk fordelte stokastiske variable, hvis fordeling har forventningsværdien μ og den entydigt bestemte p -fraktil θ . Der gælder da, at

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

er et konsistent skøn over μ , og

$$X_{([np]+1)}$$

er et konsistent skøn over θ .

Her betegner $X_{(i)}$, $i = 1, \dots, n$, som sædvanlig de ordnede observationer, og $[]$ er heltalsfunktionen (f.eks. $[\pi] = 3$).

Bevis. Forbigås. Fås af den såkaldte "de store tals lov", se e.g. [21][p. 219 og p. 378], cf. [32][p. 27]. ■

En anden sætning, der kan være nyttig i forbindelse med bestemmelse af konsistente estimatorer, er

SÆTNING 2.2. Lad T_n være et konsistent skøn over θ og φ en kontinuert afbildning $R \rightarrow R$. Da er $\varphi(T_n)$ et konsistent skøn over $\varphi(\theta)$.

Bevis. Forbigås. Fås efter nogen regning direkte af definitionen ved at anvende φ 's kontinuitet. ■

Efter omtalen af disse "symmetri-" og "nærhedsegenskaber" for estimatorer vender vi os i næste afsnit mod noget andet, nemlig spørgsmålet om, hvorvidt man ved valget af en estimator har bevaret (eller udnyttet fuldt ud) den tilgængelige information i form af data, man har om parameteren.

2.2.3 Sufficiens

En af de vigtigste opgaver for en statistiker er at skaffe sig et overblik over det måske meget omfattende observationsmateriale. Dette sker ofte ved datareduktion, f.eks. ved beregning af nogle stikprøvefunktioner, der kan siges at karakterisere materialet. Der opstår nu spørgsmålet, om der findes stikprøvefunktioner, som ikke giver et informationstab? De såkaldte sufficente stikprøvefunktioner har denne egenskab.

DEFINITION 2.5. Lad $X_1, \dots, X_n$ være stokastiske variable med simultan frekvensfunktion $f(x_1, \dots, x_n; \theta)$. Stikprøvefunktionen

$T = t(X_1, \dots, X_n)$ siges at være **sufficient** for θ , hvis den betingede fordeling af $X = (X_1, \dots, X_n)$ givet T er uafhængig af θ . ▲

BEMÆRKNING 2.2. En sufficient stikprøvefunktion kaldes også ofte **udtømmende**, idet den indeholder al relevant information om parameteren. Vi præciserer endvidere, at den i definitionen forekommende parameter θ kan være såvel en skalar θ som en vektor $\underline{\theta} = (\theta_1, \dots, \theta_n)$. ▼

Definitionen kan nok forekomme en anelse kryptisk. Før vi diskuterer den nærmere, giver vi derfor et konkret eksempel til illustration af begrebet.

EKSEMPEL 2.5. Lad der være givet tre uafhængige Bernoulli fordelte variable, X_1 , X_2 og X_3 , alle med sandsynlighed p for succes, dvs.

$$f(x; p) = P\{X_i = x\} = p^x(1-p)^{1-x}, \quad x = 0, 1.$$

X_1, X_2 og X_3 har simultan fordeling

$$\begin{aligned} f(x_1, x_2, x_3; p) &= P\{X_1 = x_1 \wedge X_2 = x_2 \wedge X_3 = x_3\} \\ &= p^{x_1}(1-p)^{1-x_1} p^{x_2}(1-p)^{1-x_2} p^{x_3}(1-p)^{1-x_3} \\ &= p^{x_1+x_2+x_3} (1-p)^{3-x_1-x_2-x_3} \end{aligned}$$

Lad nu $S = X_1 + X_2 + X_3$. S er definitionsmæssigt binomialfordelt $B(3, p)$, dvs. med tætheden

$$g(s; p) = P\{S = s\} = \binom{3}{s} p^s (1-p)^{3-s}$$

Den betingede fordeling af (X_1, X_2, X_3) for givet $S = s$ findes ved at udregne

$$h(x_1, x_2, x_3 | S = s) = \frac{f(x_1, x_2, x_3; p)}{g(s; p)} = \binom{3}{s}^{-1}, \quad s = 0, 1, 2, 3$$

idet $s = x_1 + x_2 + x_3$ benyttes.

Denne fordeling udtaler sig altså om de mulige værdier, som X_1 , X_2 og X_3 kan antage, hvis man kender deres sum:

Givet s	$P\{S = s\}$	Mulige (X_1, X_2, X_3)	Sandsynlighed
0	$(1-p)^3$	(0, 0, 0)	1
1	$3p(1-p)^2$	(1, 0, 0), (0, 1, 0), (0, 0, 1)	$\frac{1}{3}$ for alle 3
2	$3p^2(1-p)$	(1, 1, 0), (0, 1, 1), (1, 0, 1)	$\frac{1}{3}$ for alle 3
3	p^3	(1, 1, 1)	1

Alle andre (X_1, X_2, X_3) kombinationer har sandsynligheden 0, når s har de angivne mulige værdier.

Vi ser, at alle kombinationer af (X_1, X_2, X_3) , som leder til **samme sum** (f.eks. $s = 1$), er **lige sandsynlige** (uanset p), hvorimod sandsynlighederne for forskellige s -værdier afhænger af p .

Hvis man derfor ønsker at udtale sig om p på basis af X_1 , X_2 og X_3 , kan man åbenbart nøjes med at se på s , medens det ikke giver mere information om p at vide, hvilke værdier de enkelte X 'er i summen har. Derfor kaldes S sufficient for p .

Eksemplet udvides let til et vilkårligt antal ensfordelte Bernoulli-, binomial- eller Poisson-fordelte variable, eksempelvis, hvor summerne alle er sufficente stikprøvefunktioner. $\blacklozenge$

Det er ofte byrdefuldt at bestemme betingede fordelinger. Vi giver derfor følgende meget vigtige

SÆTNING 2.3 (NEYMAN'S KRITERIUM). Lad $X_1, \dots, X_n$ være stokastiske variable med simultan frekvensfunktion $f(x_1, \dots, x_n; \theta)$. Stikprøvefunktionen $T = t(X_1, \dots, X_n)$ er sufficient for θ , hvis og kun hvis vi kan finde 2 ikke-negative funktioner k og h , så

$$f(x_1, \dots, x_n; \theta) = k(t(x_1, \dots, x_n); \theta)h(x_1, \dots, x_n). \quad (2.1)$$

Bevis. Forbigås. $\blacksquare$

I eksempel 2.5 betragtedes 3 observationer fra en Bernoulli fordeling med parameter p . Benyttes i samme eksempel generelt n observationer, findes

$$f(x_1, \dots, x_n; p) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^s (1-p)^{n-s}.$$

Denne funktion afhænger alene af $s = (x_1 + x_2 + \dots + x_n)$ og parameteren p . Som afbildning $h(x)$ i (2.1) kan vi derfor vælge et 1-tal, og deraf følger, at S er sufficient.

Hvis man f.eks. ønsker at betragte en funktion af en sufficient stikprøvefunktion, $T = t(X_1, \dots, X_n)$, f.eks. $\log_e(t)$ eller $\sqrt{t}$, gælder følgende

SÆTNING 2.4. Lad t være en sufficient stikprøvefunktion og $u(\cdot)$ en injektiv funktion defineret på t 's værdimængde. Da er $u(t)$ også sufficient.

Bevis. Forbigås. $\blacksquare$

Det erindres, at en injektiv funktion, $u(\cdot)$, er en funktion, som har en omvendt funktion, $u^{-1}(\cdot)$, så f.eks. $u^{-1}(u(t)) = t$. Vi nævner nu et meget vigtigt

EKSEMPEL 2.6. Lad $X_1, \dots, X_n$ være indbyrdes uafhængige og $N(\mu, \sigma^2)$ -fordelte. Da er den simultane frekvensfunktion for $X_1, \dots, X_n$ lig

$$\prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{1}{2} \left(\frac{x_i - \mu}{\sigma} \right)^2} = \left(\frac{1}{\sqrt{2\pi}} \right)^n \frac{1}{\sigma^n} \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right)$$

Vi ser på leddet

$$\begin{aligned} \sum (x_i - \mu)^2 &= \sum (x_i - \bar{x} + \bar{x} - \mu)^2 \\ &= \sum (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \\ &\quad + 2(\bar{x} - \mu) \sum (x_i - \bar{x}) \\ &= \sum (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2. \end{aligned}$$

Vi har altså, at

$$\begin{aligned} f(x_1, \dots, x_n; \mu, \sigma^2) &= \left(\frac{1}{\sqrt{2\pi}} \right)^n \frac{1}{\sigma^n} \exp \left(-\frac{1}{2\sigma^2} \sum (x_i - \bar{x})^2 \right) \cdot \\ &\quad \exp \left(-\frac{n}{2\sigma^2} (\bar{x} - \mu)^2 \right). \end{aligned}$$

Heraf aflæses, at $\bar{X} = \frac{1}{n} \sum X_i$ er sufficient for μ , idet kun den sidste faktor indeholder μ og $\bar{x}$, og den første **ikke** indeholder μ . Resultatet følger da af Neyman's kriterium.

Af sætning 2.4 følger endvidere, at $n \cdot \bar{X} = \sum_{i=1}^n X_i$ er sufficient for parameteren μ . Vi bemærker endvidere, at den todimensionale stokastiske variabel

$$\left(\bar{X}, \sum_{i=1}^n (X_i - \bar{X})^2 \right)$$

er sufficient for (μ, σ^2) . Dette er igen en følge af Neyman's kriterium. Derimod er $\sum (X_i - \bar{X})^2$ **ikke** sufficient for σ^2 . Det er den todimensionale stokastiske variabel, der er sufficient for den todimensionale parameter (μ, σ^2) . Da

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

ses, at også parret

$$\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right)$$

er sufficient for (μ, σ^2) .

Dette resultat muliggør selvsagt meget store datareduktioner. Hvis man i en statistisk analyse beskæftiger sig med uafhængige $N(\mu, \sigma^2)$ -fordelte stokastiske variable, hvor parametrene μ og σ^2 er ukendte, da man kan nøjes med at "opbevare" sum og kvadratsum af ens observationer, og man vil alligevel besidde al relevant information om de ukendte parametre. ♦

At ovenstående resultat ikke er specielt, men har generel gyldighed for sufficente stikprøvefunktioner, følger af

SÆTNING 2.5. Lad de stokastiske variable $X_1, \dots, X_n$ have simultan frekvensfunktion $f(x_1, \dots, x_n; \theta)$ og de stokastiske variable $Y_1, \dots, Y_m$ have simultan frekvensfunktion $g(y_1, \dots, y_m; \theta)$. Lad endvidere $(X_1, \dots, X_n)$ og $(Y_1, \dots, Y_m)$ være stokastisk uafhængige. Hvis $T_1 = t_1(X_1, \dots, X_n)$ er sufficient for θ i stikprøven $X_1, \dots, X_n$ og hvis $T_2 = t_2(Y_1, \dots, Y_m)$ er sufficient for θ i stikprøven $Y_1, \dots, Y_m$, da er (T_1, T_2) sufficient for θ i den totale stikprøve $X_1, \dots, X_n, Y_1, \dots, Y_m$.

Bevis. På trods af sætningens voldsomme udseende er beviset næsten trivielt. Ifølge Neyman's kriterium er nemlig

$$\begin{aligned} f(x_1, \dots, x_n; \theta) \cdot g(y_1, \dots, y_m; \theta) = \\ k_1(t_1(x_1, \dots, x_n); \theta) h_1(x_1, \dots, x_n) \\ k_2(t_2(y_1, \dots, y_m); \theta) h_2(y_1, \dots, y_m). \end{aligned}$$

Hvis vi derfor sætter $k = k_1 \cdot k_2$ og $h = h_1 \cdot h_2$, følger sætningen ved fornyet anvendelse af Neyman's kriterium. ■

EKSEMPEL 2.7. I nedenstående tabel er der angivet sufficente stikprøvefunktioner for uafhængige stokastiske variable $X_1, \dots, X_n$ fra forskellige fordelinger. Endvidere er der i visse tilfælde anført fordelingen for stikprøvefunktionen. ♦

Fordeling af X_i	Ukendt parameter	Sufficient stikprøvefunktion T	Fordeling af T
$B(m_i, p)$	p	$\sum_{i=1}^n X_i$	$B\left(\sum_{i=1}^n m_i, p\right)$
$NB(k_i, p)$	p	$\sum_{i=1}^n X_i$	$NB\left(\sum_{i=1}^n k_i, p\right)$
$\text{Pol}(m_i, p_1, \dots, p_k)$	$(p_1, \dots, p_k)$	$\left(\sum_{i=1}^n X_{1i}, \dots, \sum_{i=1}^n X_{ki}\right)$	$\text{Pol}\left(\sum_{i=1}^n m_i, p_1, \dots, p_n\right)$
$P(\lambda)$	λ	$\sum_{i=1}^n X_i$	$P(n\lambda)$
$\text{Ex}(\beta)$	β	$\sum_{i=1}^n X_i$	$G(n, \beta)$
$G(k, \beta)$	(k, β)	$\left(\prod_{i=1}^n X_i, \sum_{i=1}^n X_i\right)$	
$G(k, \beta)$	$\beta(k \text{ kendt})$	$\sum_{i=1}^n X_i$	$G(nk, \beta)$
$G(k, \beta)$	$k(\beta \text{ kendt})$	$\prod_{i=1}^n X_i$	

Fordeling af X_i	Ukendt parameter	Sufficient stikprøvefunktion T	Fordeling af T
$N(\mu, \sigma^2)$	(μ, σ^2)	$\left(\sum_{i=1}^n X_i, \sum_{i=1}^n (X_i - \bar{X})^2 \right)$	sætning 2.6
$N(\mu, \sigma^2)$	μ	$\sum_{i=1}^n X_i$	$N(n\mu, n\sigma^2)$
$N(\mu, \sigma^2)$	σ^2 (μ kendt)	$\sum_{i=1}^n (X_i - \mu)^2$	$\sigma^2 \chi^2(n)$
$LN(\alpha, \beta^2)$	(α, β^2)	$\left(\sum_{i=1}^n \log_e X_i, \sum_{i=1}^n \log_e^2 X_i \right)$	
$U(a, b)$	(a, b)	$(X_{(1)}, X_{(n)})$	
$Be(\alpha, \beta)$	(α, β)	$\left(\prod_{i=1}^n X_i, \prod_{i=1}^n (1 - X_i) \right)$	

Sufficiensen følger i hvert enkelt tilfælde direkte af Neyman's kriterium. Fordelingen af T følger i de angivne tilfælde af reproduktivitetssætningen i kapitel 1.

Vi bemærker, at vi ikke har anført fordelingen for $\sum (X_i - \bar{X})^2$ i det normale tilfælde, hvilket skyldes, at den ikke kan udledes ved elementære metoder. Da der er tale om en overordentlig vigtig stikprøvefunktion, giver vi fordelingen af den i

SÆTNING 2.6. Lad $X_1, \dots, X_n$ være uafhængige stokastiske variable med $X_i \in N(\mu, \sigma^2)$, $i = 1, \dots, n$. Da gælder, at

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

og

$$\sum_{i=1}^n (X_i - \bar{X})^2 \in G\left(\frac{n-1}{2}, 2\sigma^2\right) = \sigma^2 \chi^2(n-1)$$

d.v.s.

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

Endvidere er $\bar{X}$ og $\sum (X_i - \bar{X})$ stokastisk uafhængige.

Bevis. Forbigås. Se f.eks. [21][p. 347]. ■

BEMÆRKNING 2.3. Resultatet om $\bar{X}$'s fordeling er klart ifølge reproduktivitetssætningen for den normale fordeling (sætning 1.22 p. 124). Da

$$\frac{1}{\sigma} (X_i - \mu) \in N(0, 1)$$

er

$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \in \chi^2(n)$$

Ifølge sætning 1.74, p 195. Hvis vi i stedet for middelværdien μ indsætter $\bar{X}$, fås

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

Ifølge sætningen er denne størrelse stadig χ^2 -fordelt, men med **1 frihedsgrad mindre**. Det kan vises, at dette beror på, at de stokastiske variable $X_1 - \bar{X}, \dots, X_n - \bar{X}$ ikke længere kan variere frit, men må tilfredsstille relationen

$$(X_1 - \bar{X}) + \dots + (X_n - \bar{X}) = 0 \quad (2.2)$$

Vi har altså, at hvor vektoren $(X_1 - \mu, \dots, X_n - \mu)$ frit kan variere i $\mathbb{R}^n$, er vektoren $(X_1 - \bar{X}, \dots, X_n - \bar{X})$ bundet til det $n-1$ dimensionale underrum bestemt af (2.2). Vi taler om et **lineært bånd** mellem $X_1 - \bar{X}, \dots, X_n - \bar{X}$. Med disse betragtninger er benævnelsen **frihedsgrad** for parameteren i χ^2 -fordelingen også blevet mere naturlig. ▼

Med dette resultat forlader vi behandlingen af de sufficente stikprøvefunktioner og deraf afledte estimatorer, og går over til den sidste af de egenskaber, vi her opstiller som ønskelige for en estimator.

2.2.4 Efficiens

Vi betragter indledningsvis et

EKSEMPEL 2.8. Lad der være givet en række stokastiske variable $X_1, \dots, X_n$, der er uafhængige og har samme middelværdi μ og varians σ^2 . I afsnit 2.2.1 blev der foreslået følgende 2 estimatorer for μ

$$\begin{aligned}\bar{X} &= \frac{1}{n} \sum_{i=1}^n X_i \\ \tilde{X} &= \frac{X_1 + 2X_2 + \dots + nX_n}{1 + 2 + \dots + n} = \frac{1}{\sum i} \sum_{i=1}^n iX_i,\end{aligned}$$

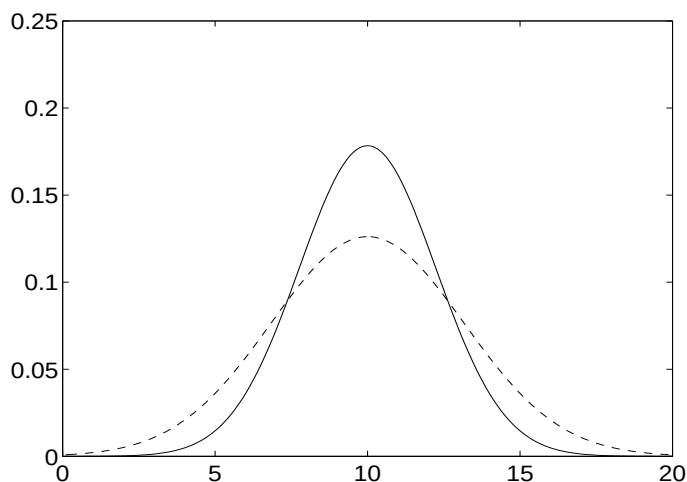
og det blev godtgjort, at de begge var centrale. Vi vil nu bestemme varianserne for de 2 estimatorer. Vi har

$$V(\bar{X}) = V\left(\frac{1}{n} \sum X_i\right) = \frac{1}{n^2} \sum \sigma^2 = \frac{\sigma^2}{n}$$

og

$$\begin{aligned}V(\tilde{X}) &= \frac{1}{(\sum i)^2} \sum i^2 \sigma^2 \\ &= \frac{\sum i^2}{(\sum i)^2} \sigma^2 \\ &= \frac{2}{3} \frac{2n+1}{n+1} \frac{1}{n} \sigma^2.\end{aligned}$$

Da faktoren $\frac{2}{3} \frac{2n+1}{n+1} > 1$ for $n > 1$, ses, at $\tilde{X}$ har en større varians end $\bar{X}$. Graferne for deres frekvensfunktioner kan derfor være som antydnet på nedenstående figur, hvor den fuldt optrukne tæthed svarer til $\bar{X}$ og den stiplede til $\tilde{X}$.



I normale situationer vil de fleste nok foretrække $\bar{X}$ fremfor $\tilde{X}$ som estimator for middelværdien μ , idet vi må forvente, at de realiserede udfald af $\bar{X}$ vil fordele sig snævrere om den **sande værdi** μ end udfaldene af $\tilde{X}$ vil. På figuren er $\mu = 10$. ♦

Ovenstående eksempel giver anledning til følgende

DEFINITION 2.6. Lad $T_1 = t_1(X_1, \dots, X_n)$ og $T_2 = t_2(X_1, \dots, X_n)$ være centrale estimatorer for parameteren θ . Vi siger da, at T_1 er mere **efficient** end T_2 , hvis $V(T_1) \leq V(T_2)$ for alle værdier af den ukendte parameter θ . ▲

Det er nu oplagt, at man vil forsøge at finde den mest efficiente estimator for den ukendte parameter θ . Det er her meget interessant, at der findes en nedre grænse for, hvor lille en varians en sådan kan have. Før vi formulerer dette resultat, må vi nævne

DEFINITION 2.7. Vi siger, at en fordeling med frekvensfunktion $f(x; \theta)$ er **regulær** med hensyn til θ , hvis

1. mængden $\{x | f(x; \theta) > 0\}$ er uafhængig af θ
2. $E\left(\frac{\partial}{\partial \theta} \log_e f(X; \theta)\right)$ og $E\left(\frac{\partial^2}{\partial \theta^2} \log_e f(X; \theta)\right)$ eksisterer for alle θ .

Denne definition overføres uændret til en flerdimensional stokastisk variabel $\underline{X} = (X_1, \dots, X_n)$. ▲

Hvis man f.eks. betragter en stokastisk variabel $X \in U(0, \theta)$ (rektangulært fordelt med ukendt højre afgrænsning), er $U(0, \theta)$ **ikke** regulær mht. θ . En log-normal fordeling $LN(\alpha, \beta^2)$ er regulær mht. såvel α som β^2 , idet $f(x; \alpha, \beta^2) > 0$ for alle $x > 0$ uanset α og β^2 .

LEMMA 2.1. Lad $\underline{X} = (X_1, \dots, X_n)$ have den regulære frekvensfunktion $f(x_1, \dots, x_n; \theta)$. Sætter vi

$$i_n(\theta) = E \left[\left\{ \frac{\partial}{\partial \theta} \log_e f(X_1, \dots, X_n; \theta) \right\}^2 \right]$$

gælder følgende identitet

$$i_n(\theta) = -E \left[\frac{\partial^2}{\partial \theta^2} \log_e f(X_1, \dots, X_n; \theta) \right]$$

Bevis. Forbigås. Se f.eks. [12][p. 108.]. ■

Er specielt $X_1, \dots, X_n$ stokastisk uafhængige med samme, regulære frekvensfunktion $f(x; \theta)$ fås

$$i_n(\theta) = n E \left[\left\{ \frac{\partial}{\partial \theta} \log_e f(X; \theta) \right\}^2 \right] = -n E \left[\frac{\partial^2}{\partial \theta^2} \log_e f(X; \theta) \right],$$

hvor X generelt angiver en stokastisk variabel med frekvensfunktion $f(x; \theta)$. Bevises ved at benytte, at uafhængigheden medfører, at

$$\begin{aligned} \log_e(f(X_1, \dots, X_n; \theta)) &= \log_e(f(X_1; \theta) \cdots f(X_n; \theta)) \\ &= \sum_{i=1}^n \log_e(f(X_i; \theta)) \end{aligned}$$

Ved herefter at differentiere og tage forventningsværdien for hvert led for sig i denne sum, følger resultatet umiddelbart.

Vi er nu i stand til at formulere Cramér-Rao's ulighed:

SÆTNING 2.7 (CRAMÉR-RAO'S ULIGHED). Lad $X_1, \dots, X_n$ have den regulære frekvensfunktion $f(x_1, \dots, x_n; \theta)$ og lad $T = t(X_1, \dots, X_n)$ være estimator for θ med

$$E(T) = E_\theta(T) = \theta + b(\theta)$$

idet $E_\theta(T)$ betegner forventningsværdien som funktion af θ .

Der gælder da, idet $V_\theta(T)$ betegner T 's varians som funktion af θ , at

$$V(T) = V_\theta(T) \geq \frac{[1 + b'(\theta)]^2}{i_n(\theta)}$$

hvor $i_n(\theta)$ er som i ovenstående lemma.

Bevis. Forbigås. Se f.eks. [12][p. 254]. ■

Er specielt T central, d.v.s. $b(\theta) = 0$, fås $b'(\theta) = 0$ og

$$V(T) = E_\theta((T - \theta)^2) \geq \frac{1}{i_n(\theta)} \quad (2.3)$$

Sætningen giver mulighed for at definere et "absolut mål" for, hvor "god" en estimator er, idet vi kan anføre

DEFINITION 2.8. Forholdet mellem variansen af en central estimator T og højresiden i (2.3) kaldes **efficiensen** af estimatoren. ▲

Vi skal ikke fordybe os i disse begreber; men vi anfører et eksempel til illustration.

EKSEMPEL 2.9. Lad $X_1, \dots, X_n$ være uafhængige $P(\lambda)$ -fordelte stokastiske variable. Vi ved da, at $\bar{X}$ er en central estimator for middelværdien λ (eksempel 2.1). Vi vil da bestemme efficiensen af $\bar{X}$. Vi har

$$\frac{\partial}{\partial \lambda} \log f(x; \lambda) = \frac{\partial}{\partial \lambda} \log \frac{\lambda^x}{x!} e^{-\lambda} = \frac{\partial}{\partial \lambda} (x \log \lambda - \log x! - \lambda) = \frac{x}{\lambda} - 1 \quad .$$

Følgelig er

$$\begin{aligned} E\left(\left\{\frac{\partial}{\partial \lambda} \log f(X; \lambda)\right\}^2\right) &= E\left(\left(\frac{X}{\lambda} - 1\right)^2\right) \\ &= \frac{1}{\lambda^2} E((X - \lambda)^2) \\ &= \frac{1}{\lambda^2} V(X) = \frac{1}{\lambda}, \end{aligned}$$

Den nedre grænse i Cramér-Rao's ulighed er derfor

$$\frac{1}{n \frac{1}{\lambda}} = \frac{\lambda}{n}.$$

Nu er

$$V(\overline{X}) = V\left(\frac{1}{n} \sum X_i\right) = \frac{\lambda}{n},$$

d.v.s. $V(\overline{X})$ er identisk med nedre grænse i uligheden, hvorfor $\overline{X}$ har efficiensen 1, og vi siger, at $\overline{X}$ er en **efficient estimator** for λ . ♦

BEMÆRKNING 2.4. Vi bemærker, at i ovenstående eksempel er den efficiente estimator $\overline{X}$ sufficient for λ . Dette er ikke noget specielt, idet det generelt kan vises, at hvis en central estimator T antager nedre grænse i Cramér-Rao's ulighed, da er T sufficient. Vi skal ikke komme nærmere ind på disse ting, men blot henvise til f.eks. [34]. ▼

2.3 Estimationsmetoder

Vi går nu over til at beskrive nogle generelle metoder for konstruktion af estimator. Vi indleder med den såkaldte maximum likelihood metode. Denne metode bygger på, at man tilpasser en **kendt** tæthedsfunktion til data ved at vælge dens parametre udfra et optimalitetsprincip:

2.3.1 Maximum likelihood metoden

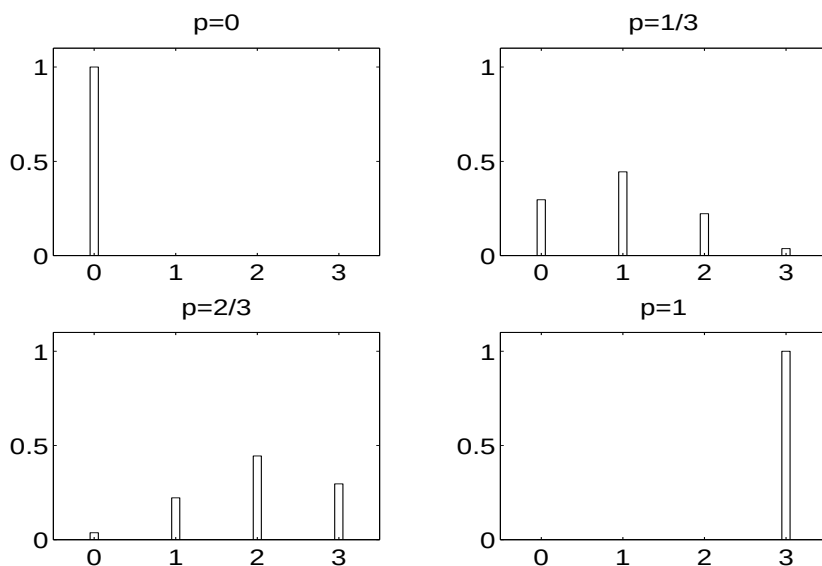
Vi giver først et konstrueret eksempel, der til gengæld er så simpelt, at man skulle forstå den generelle idé.

EKSEMPEL 2.10. Vi betragter en urne, der indeholder 3 kugler, der enten er røde eller hvide. Det vides ikke, hvor mange der er af hver slags. Endvidere foreligger der, hvor mange røde kugler man har fået ved at udtrække 3 gange med tilbagelægning. Kalder vi antallet af røde kugler X , vil $X \in B(3, p)$, hvor p er brøkdelen af røde kugler i urnen, d.v.s. p antager en af værdierne $0, \frac{1}{3}, \frac{2}{3}, 1$. Dette giver følgende tabel over de mulige fordelinger af X :

$f(x; p)$	$x :$	0	1	2	3
$p :$	0	$\frac{1}{27}$	$\frac{4}{27}$	$\frac{6}{27}$	$\frac{1}{27}$
	$\frac{1}{3}$	$\frac{8}{27}$	$\frac{12}{27}$	$\frac{6}{27}$	$\frac{1}{27}$
	$\frac{2}{3}$	$\frac{1}{27}$	$\frac{12}{27}$	$\frac{12}{27}$	$\frac{1}{27}$
	1	0	0	0	1

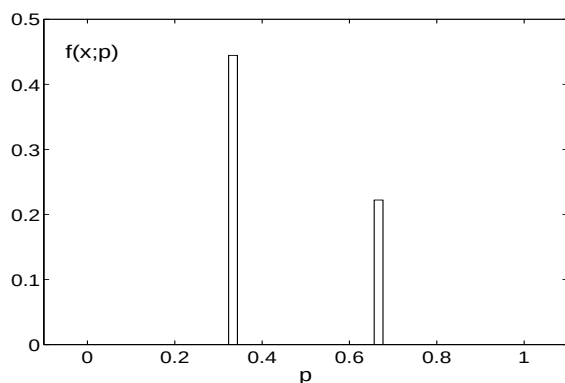
Af tabellen ses f.eks., at $P\{X = 1|p = \frac{1}{3}\} = \frac{4}{9}$, medens $P\{X = 1|p = \frac{2}{3}\} = \frac{2}{9}$.

Vi tegner graferne for de 4 frekvensfunktioner i figuren side 234.



Hvis vi nu har observeret $X = 1$, ser vi altså, at denne hændelse har den største sandsynlighed for $p = \frac{1}{3}$. Da hændelsen $\{X = 1\}$ jo er indtruffet, vil vi regne med, at den sandsynlighedsfordeling, der har genereret den, tilordner den en stor sandsynlighed. Det vil derfor være rimeligt at estimere den ukendte parameter ved det p , der maximaliserer denne sandsynlighed, i.e. maximaliserer sandsynligheden for den observerede hændelse. Denne måde at estimere en parameter på kaldes **maximum likelihood metoden**.

For lettere at kunne danne sig et overblik over de mulige værdier af p , kan man afbilde $f(x; p)$ som funktion af p . Med $x = 1$ fås i det aktuelle tilfælde grafen herunder.



Maximumlikelihood estimatoren for p fås nu blot ved at bestemme maximum for en funktion af 1 variabel. ♦

Ovenstående eksempel vil danne grundlaget for de mere præcise definitioner, vi vil give i det følgende. For at lette fremstillingen indfører vi dog først et nyt begreb i

DEFINITION 2.9. Lad $X_1, \dots, X_n$ være stokastiske variable med simultan frekvensfunktion $f(x_1, \dots, x_n; \theta)$. Ved **likelihood funktionen** L forstås afbildningen givet ved

$$L(\theta) = f(x_1, \dots, x_n; \theta). \quad (2.4)$$

eller udførligt

$$L(\theta; x_1, \dots, x_n) = f(x_1, \dots, x_n; \theta).$$



BEMÆRKNING 2.5. Vi opfatter altså blot f som funktion af θ for fastholdt $x_1, \dots, x_n$. Hvis $X_1, \dots, X_n$ er stokastisk **uafhængige** med marginale frekvensfunktioner $f_i(x_i; \theta)$, fås umiddelbart, at

$$L(\theta) = \prod_{i=1}^n f_i(x_i; \theta). \quad (2.5)$$



Vi er nu i stand til at give den nøjagtige definition af en maximum likelihood estimator.

DEFINITION 2.10. Lad likelihood-funktionen for de stokastiske variable $X_1, \dots, X_n$ være $L(\theta)$. Hvis der findes en afbildning $t(x_1, \dots, x_n)$ med den egenskab, at der for alle θ og alle $(x_1, \dots, x_n)$ gælder

$$L(t(x_1, \dots, x_n); x_1, \dots, x_n) \geq L(\theta; x_1, \dots, x_n),$$

da kaldes stikprøvefunktionen $\hat{\theta} = t(X_1, \dots, X_n)$ en **maximum likelihood estimator** for θ . ▲

BEMÆRKNING 2.6. Vi har mere overskueligt, at $\hat{\theta}$ er bestemt ved

$$L(\hat{\theta}) = \sup_{\theta} L(\theta). \quad (2.6)$$

Hvis vi er i det regulære tilfælde, d.v.s. hvis $\{x | f(x; \theta) > 0\}$ ikke afhænger af θ , bestemmes $\hat{\theta}$ lettest ved at løse den såkaldte **likelihood-ligning**

$$\frac{d}{d\theta} \log_e L(\theta) = 0. \quad (2.7)$$

Hvis θ er en vektor $\theta = (\theta_1, \dots, \theta_k)$, antager (2.7) formen

$$\frac{\partial}{\partial \theta_1} \log_e L(\theta) = 0, \dots, \frac{\partial}{\partial \theta_k} \log_e L(\theta) = 0. \quad (2.8)$$

At $L(\theta)$ og $\log_e L(\theta)$ har maximum for de samme værdier af θ følger trivielt af, at logaritmen er en voksende funktion. ▼

Inden vi går videre med behandlingen af maximum likelihood estimatorer, giver vi et

EKSEMPEL 2.11. Lad $X_1, \dots, X_n$ være stokastiske uafhængige og $P(\lambda)$ -fordelte. Vi vil bestemme maximum likelihood estimatoren $\hat{\lambda}$. Vi har, at

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^n \frac{1}{x_i!} \lambda^{x_i} e^{-\lambda} \\ &= \frac{1}{\prod x_i!} \lambda^{\sum x_i} e^{-n\lambda}, \end{aligned}$$

d.v.s.

$$\log_e L(\lambda) = k + \left(\sum x_i \right) \log_e \lambda - n\lambda,$$

hvor $k = -\log_e (\prod x_i!)$ er en konstant, der er uafhængig af λ , og som derfor ikke har betydning for bestemmelsen af λ . Følgelig er

$$\frac{d}{d\lambda} \log_e L(\lambda) = \frac{1}{\lambda} \sum x_i - n,$$

og det ses, at denne størrelse netop er 0 for

$$\lambda = \frac{1}{n} \sum x_i.$$

Da

$$\frac{d^2}{d\lambda^2} \log_e L(\lambda) = -\frac{1}{\lambda^2} \sum x_i < 0,$$

er ovenstående lokale extremumværdi et globalt maximum for likelihood-funktionen, hvorfor maximum likelihood estimatoren for λ er

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}.$$



Vi bemærker, at maximum likelihood estimatoren i ovenstående eksempel er konsistent og sufficient, og dens varians antager nedre grænse i Cramér-Rao's ulighed. Disse egenskaber gælder ikke i alle tilfælde, men vi har dog

SÆTNING 2.8. Lad $X_1, \dots, X_n$ være stokastisk uafhængige og identisk fordelte med frekvensfunktion $f(x; \theta)$. Hvis der eksisterer en sufficient stikprøvefunktion $T = t(X_1, \dots, X_n)$ for θ , og hvis der eksisterer en maximum likelihood estimator $\hat{\theta}$ for θ , da er $\hat{\theta}$ en funktion af T .

Bevis. Forbigås. Følger direkte ved en anvendelse af Neyman's kriterium. ■

Denne vigtige sætning siger med andre ord, at man roligt kan foretage en sufficient reduktion af sine data, hvis man er interesseret i senere at beregne maximum likelihood skøn over de ukendte parametre i de fordelinger, man betragter. Vi nævnte ovenfor, at maximum likelihood estimatoren ikke i alle tilfælde var efficiente. Vi har dog følgende

SÆTNING 2.9. Lad $X_1, \dots, X_n$ være uafhængige og identisk fordelte stokastiske variable med frekvensfunktion $f(x; \theta)$, hvor θ er en skalar, og f er regulær. Da gælder, at en maximum likelihood estimator $\hat{\theta}$ for θ er asymptotisk fordelt med parametre

$$\begin{aligned} \mu_n &= \theta \\ \sigma_n^2 &= \frac{1}{i_n(\theta)}, \end{aligned}$$

hvor $i_n(\theta)$ er defineret som i lemma p. 230. Af resultatet fås specielt, at maximum likelihood estimatoren er konsistent.

Bevis. Forbigås. Se f.eks. [35][p. 42.]. ■

BEMÆRKNING 2.7. Den asymptotiske varians i $\hat{\theta}$'s fordeling er lig den nedre grænse i Cramér-Rao's ulighed, d.v.s. at sætningen også udtrykker, at en maximum likelihood estimator er asymptotisk efficient i en meget stor klasse af fordelinger. Det skal endvidere nævnes, at sætning 2.9 har en umiddelbar generalisering til det tilfælde, hvor θ er en vektor. ▼

BEMÆRKNING 2.8. Sætningen er også gyldig i tilfældet med uafhængige, men ikke identisk fordelte stokastiske variable. I dette tilfælde må dog en række yderligere regularitetsbetingelser kræves opfyldt, se f.eks. [7]. I [3] vises nogle udvidelser til tilfældet med afhængige observationer. ▼

Vi nævner endelig en sidste enkel, men vigtig sætning om maximum likelihood estimators.

SÆTNING 2.10. Lad T være en maximum likelihood estimator for θ , og lad φ være en afbildning defineret på parameterrummet. Da er estimatoren $\varphi(T)$ er en maximum likelihood estimator for parameteren $\varphi(\theta)$.

Bevis. Forbigås. Fås direkte af definitionen. ■

Vi giver et eksempel på bestemmelse af maximum likelihood estimators.

EKSEMPEL 2.12. I nedenstående tabel er anført et uddrag af målingerne i eksempel 1.14 vedrørende tidsintervaller mellem succesive kundeankomster ved kasserne i et supermarked. Tallene er angivet i $\frac{1}{100}$ min.

Tidsafstand i $\frac{1}{100}$ min.			
11	128	36	127
4	127	71	45
52	46	29	16
43	41	62	17
186	157	58	211

Man antager som anført, at kunder ankommer ved kasseapparaterne efter en Poisson proces. Dette giver ifølge sætning 1.16, p. 118, at ovenstående observationer er realiserede udfald af uafhængige eksponentialt fordelte stokastiske variable. Vi skal senere se, hvorledes en sådan antagelse kan undersøges.

Den teoretiske model, vi vil arbejde med, er altså: Der er givet realiserede udfald af

uafhængige, identisk fordelte stokastiske variable $X_1, \dots, X_n$, $n = 20$, med tæthedsfunktionen

$$f(x; \beta) = \frac{1}{\beta} e^{-x/\beta}, \quad x > 0.$$

Her er $\beta > 0$ den ukendte parameter. Vi vil bestemme et maximum likelihood skøn over β .

Likelihood funktionen er

$$L(\beta) = \prod_{i=1}^n \frac{1}{\beta} e^{-x_i/\beta} = \beta^{-n} \exp\left(-\frac{1}{\beta} \sum x_i\right),$$

d.v.s.

$$\log_e L(\beta) = -n \log_e \beta - \frac{1}{\beta} \sum x_i.$$

Dette giver

$$\frac{d}{d\beta} \log_e L(\beta) = -\frac{n}{\beta} + \frac{1}{\beta^2} \sum x_i.$$

Denne størrelse er 0 netop, hvis

$$\beta = \frac{1}{n} \sum x_i = \bar{x}.$$

Da endvidere

$$\frac{d^2}{d\beta^2} \log_e L(\beta) = +\frac{n}{\beta^2} - \frac{2}{\beta^3} \sum x_i,$$

som i punktet $\beta = \bar{x}$ antager værdien

$$\frac{n}{\bar{x}^2} - \frac{2n\bar{x}}{\bar{x}^3} = -\frac{n}{\bar{x}^2} < 0,$$

hvorfor $\beta = \bar{x}$ er et globalt maximum for $L(\beta)$. Derfor er maximum likelihood estimatoren for β lig

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}.$$

Summen af disse i tabellen anførte data er

$$11 + 4 + \cdots + 17 + 211 = 1467$$

og gennemsnittet af dem

$$\frac{1467}{20} = 73.35 \quad .$$

Et maximum likelihood skøn over parameteren β er derfor

$$\hat{\beta} = 0.7335 \text{ min.}$$

Da middelværdien i fordelingen er β , er ovenstående også et maximum likelihood skøn over middelværdien. Variansen i fordelingen er β^2 ifølge sætning 1.18, p. 121. Ifølge ovenstående sætning 2.10 er imidlertid

$$(\widehat{\beta^2}) = (\hat{\beta})^2,$$

hvorfor maximum likelihood skønnet over variansen bliver

$$0.7335^2 \text{ min}^2 = 0.538022 \text{ min}^2.$$

Idet vi nu vedtager, at vi vil betragte ovenstående værdi af β som givet, kan vi løse problemer som at finde sandsynligheden for, at der går mere end 1 min. mellem 2 på hinanden følgende kundeankomster, eller mere korrekt - vi kan estimere denne. Den er nemlig

$$\begin{aligned} P\{\text{Ex}(\beta) > 1\} &= 1 - F(1; \beta) \\ &= \exp\left(-\frac{1}{\beta}\right) \\ &\simeq \exp\left(-\frac{1}{0.7335}\right) = 0.257 \end{aligned}$$

Ifølge sætning 2.10 er dette et ML-skøn over den ukendte sandsynlighed. ♦

EKSEMPEL 2.13. Vi betragter nu de i eksempel 1.1 anførte 36 målinger af diameteren på nittehoveder udtaget fra en løbende produktion.

Man er interesseret i at kunne komme med udsagn om, hvad sandsynligheden er for, at et tilfældigt udvalgt nittehoved holder visse mål.

For at løse dette problem må vi postulere en fordelingslov. Det forekommer rimeligt at antage, at variationerne i størrelsen af de enkelte nittehoveder skyldes mange små, additive "fejl", således at vi i overensstemmelse med overvejelserne i afsnit 1.5.2 formulerer følgende matematiske model. Der er givet realiserede udfald af uafhængige, identisk $N(\mu, \sigma^2)$ -fordelte stokastiske variable $X_1, \dots, X_n$, $n = 36$. Vi vil på basis af ovenstående data bestemme maximum likelihood skøn over (μ, σ^2) .

Ifølge sætning 2.8 kan vi starte med at foretage en sufficient reduktion af vore data. Ifølge eksempel 2.6 er summen og kvadratsummen af observationerne sufficente. Vi har

$$\begin{aligned}\sum x_i &= 13.62 + \dots + 13.33 = 481.54 \\ \sum x_i^2 &= 13.62^2 + \dots + 13.33^2 = 6441.6938\end{aligned}$$

Disse to tal indeholder altså al relevant information om de ukendte parametre μ og σ^2 , og vi ved, at maximum likelihood skønnene kan beregnes på basis af disse. Vi går nu i gang med at bestemme estimatorene.

Vi har likelihood-funktionen

$$\begin{aligned}L(\mu, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\ &= \left(\sqrt{2\pi\sigma^2}\right)^{-n} \exp\left(-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2\right).\end{aligned}$$

Vi har derfor

$$\log_e L(\mu, \sigma^2) = k - \frac{n}{2} \log_e \sigma^2 - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2,$$

hvor k er uafhængig af μ og σ^2 . En nødvendig betingelse for et ekstremum i det indre af parameterområdet er, at de partielle afledede med hensyn til μ og σ^2 er 0. Vi udregner

derfor

$$\begin{aligned}\frac{\partial}{\partial \mu} \log_e L(\mu, \sigma^2) &= -\frac{1}{2\sigma^2} (-2) \sum (x_i - \mu) \\ &= -\frac{1}{\sigma^2} \left(n\mu - \sum x_i \right),\end{aligned}$$

og den ses at være 0, netop hvis $\mu = \bar{x}$. Tilsvarende

$$\frac{\partial}{\partial \sigma^2} \log_e L(\mu, \sigma^2) = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum (x_i - \mu)^2.$$

Hvis vi indsætter $\mu = \bar{x}$, ses, at ovenstående størrelse er 0, netop hvis $\sigma^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$. En nærmere undersøgelse af funktionen $\log_e L(\mu, \sigma^2)$ viser, at dette er det globale maximum, således at vi har, at

$$(\widehat{\mu}, \widehat{\sigma^2}) = (\hat{\mu}, \hat{\sigma^2}) = \left(\bar{X}, \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right) = (\bar{X}, S_*^2)$$

er maximum likelihood estimatoren for (μ, σ^2) .

Ifølge eksempel 2.2 er

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 = \sum x_i^2 - \frac{1}{n} \left(\sum x_i \right)^2,$$

således får vi

$$\begin{aligned}\bar{x} &= \frac{1}{36} 481.54 = 13.376 \\ s_*^2 &= \frac{1}{36} \left(6441.6938 - \frac{1}{36} 481.54^2 \right) = \frac{1}{36} 0.5618 \\ &= 0.0156 = 0.125^2\end{aligned}$$

eller

$$(\bar{x}, s_*^2) = (13.376, 0.125^2).$$

Vi kan nu f.eks "bestemme" sandsynligheden for, at et tilfældigt valgt nittehoved overholder kravet om, at diameteren skal ligge i intervallet $(13.43 - 0.2, 13.43 + 0.2)$. (NB!

Under forudsætning af de ovenfor estimerede parametre). Den er

$$\begin{aligned}
 & P\{13.43 - 0.2 < N(13.376, 0.125^2) < 13.43 + 0.2\} \\
 &= P\left\{\frac{13.23 - 13.376}{0.125} < N(0, 1) < \frac{13.63 - 13.376}{0.125}\right\} \\
 &= P\{-1.17 < N(0, 1) < 2.03\} \\
 &= 0.9788 - 0.1210 \\
 &= 0.8578.
 \end{aligned}$$

Eftersom $(\bar{x}, s_*^2)$ er maximum likelihood estimator, er den fundne sandsynlighed faktisk også et maximum likelihood estimat. ♦

Vi giver nu endelig et eksempel, hvor vi ikke har identisk fordelte observationer.

EKSEMPEL 2.14. 6 ens stålplader udsættes for vejrligets påvirkninger. Hver dag udtages der herefter tilfældigt 1 af pladerne, og man måler vægttabet, som skyldes korrosion. Resultaterne af eksperimentet fremgår af følgende skema (vægttabene er målt i gram):

Dag	1	2	3	4	5	6
Vægttab	0.24	0.22	0.95	0.81	1.44	1.56

Man mener nu, at vægttabet vokser lineært med tiden inden for et tilpas kort tidsrum, og man ønsker at estimere hældningskoefficienten for en sådan linie.

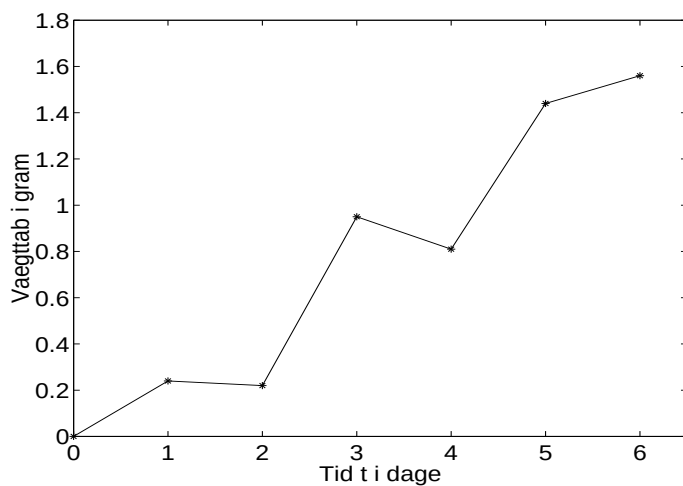
Før vi går i gang med selve modelformuleringen, afbilder vi ovenstående data i et retvinklet koordinatsystem.

Vi går ud fra, at vægttabet til tid t kan beskrives ved en stokastisk variabel $X_t \in N(\mu_t, \sigma^2)$. Grunden til, at vi vælger normalfordelingen, skal igen søges i afsnit 1.5.2 om den centrale grænseværdisætning. At variansen antages konstant, må begrundes med kendskab til tidligere forsøg af samme art. Vi kan nu formulere vendingen "vægttabet vokser lineært med tiden" mere stringent, nemlig som

$$\mu_t = \alpha \cdot t,$$

hvor α er en ukendt parameter. Vi vil bestemme et maximum likelihood skøn over α .

Vi har altså modellen, at de stokastiske variable $X_1, \dots, X_6$ er uafhængige, og $X_t \in$



$N(\alpha t, \sigma^2)$, $t = 1, \dots, 6$. Vi har derfor

$$L(\alpha, \sigma^2) = \prod_{t=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (x_t - \alpha t)^2\right).$$

Heraf fås

$$\log_e L(\alpha, \sigma^2) = k - \frac{n}{2} \log_e \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^n (x_t - \alpha t)^2.$$

Vi ser, at

$$\begin{aligned} \frac{\partial}{\partial \alpha} \log_e L(\alpha, \sigma^2) &= \frac{1}{\sigma^2} \sum_{t=1}^n (x_t - \alpha t) t \\ &= \frac{1}{\sigma^2} \left[\sum_{t=1}^n t x_t - \alpha \sum_{t=1}^n t^2 \right]. \end{aligned} \quad (2.9)$$

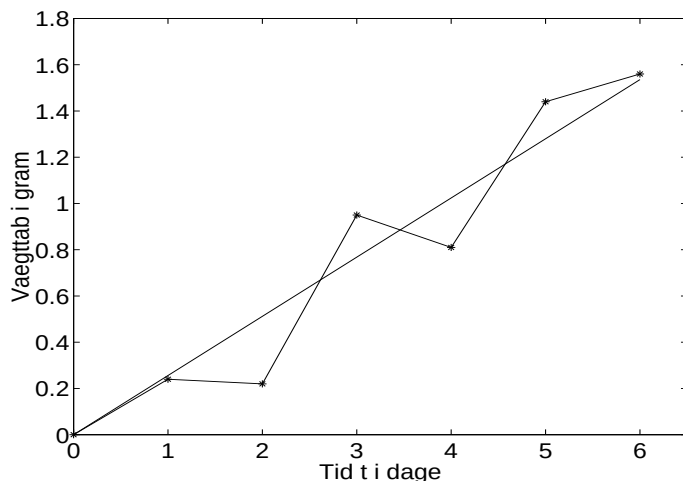
Det ses heraf, at maximum likelihood estimatoren for α er

$$\hat{\alpha} = \frac{\sum_{t=1}^n t X_t}{\sum_{t=1}^n t^2}.$$

Den observerede værdi af $\hat{\alpha}$ er

$$\frac{1 \cdot 0.24 + 2 \cdot 0.22 + \dots + 6 \cdot 1.56}{1 + 4 + \dots + 36} = \frac{23.33}{91} = 0.256 \quad ,$$

hvilket altså er maximum likelihood estimatet af hældningen. Vi indtegner linien med denne hældning i foranstående figur og får



Det ses, at linien tilnærmer de målte data "pænt". Der er ingen systematisk afvigelse fra linien (såsom at alle punkter ligger på den ene side af linien etc.).

Da vi har introduceret en sammenhæng mellem middelværdierne af de stokastiske variable X_t , vil det også være muligt at finde variansen på trods af, at vi kun har 1 observation for hver værdi af t . Hvis vi nemlig differentierer likelihood-funktionen $L(\alpha, \sigma^2)$ med hensyn til σ^2 , får vi

$$\begin{aligned} \frac{\partial}{\partial \sigma^2} \log_e L(\alpha, \sigma^2) &= -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2} \frac{1}{\sigma^4} \sum_{t=1}^n (x_t - \alpha t)^2 \\ &= -\frac{n}{2} \frac{1}{\sigma^4} \left(\sigma^2 - \frac{1}{n} \sum_{t=1}^n (x_t - \alpha t)^2 \right). \end{aligned}$$

Denne størrelse er 0, netop hvis

$$\sigma^2 = \frac{1}{n} \sum_{t=1}^n (x_t - \alpha t)^2.$$

Da $\frac{\partial}{\partial \alpha} \log_e L$ kun er 0, hvis $\alpha = \hat{\alpha}$ som i (2.9), vises det på samme måde som i eksempel 2.12, at

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{t=1}^n (X_t - \hat{\alpha}t)^2$$

er maximum likelihood estimator for σ^2 . Da

$$\begin{aligned} \sum_{t=1}^n (x_t - \hat{\alpha}t)^2 &= \sum_t x_t^2 + \hat{\alpha}^2 \sum_t t^2 - 2\hat{\alpha} \sum_t tx_t \\ &= \sum_t x_t^2 - \hat{\alpha}^2 \sum_t t^2 \end{aligned}$$

ifølge (2.9) (jfr. den tilsvarende opspaltning i eksempel 2.2), og da

$$\begin{aligned} \sum_t x_t^2 &= 6.1718 \\ \sum_t t^2 &= 91, \end{aligned}$$

er den observerede værdi af $\hat{\sigma}^2$ lig

$$\frac{1}{6}(6.1718 - 5.9812) = 0.03177 \simeq 0.18^2 \quad .$$

Vi resumerer: Under de p. 243 gjorde forudsætninger - bl.a. om fordelingen af X_t - har vi, at korrosionen til tid t beskrives ved en stokastisk variabel X_t med middelværdi $0.256 \cdot t$ gram/dag og varians 0.18^2 gram², i det mindste så længe t holder sig inden for det område, hvor vi har målt, d.v.s. $t \leq 6$ dage. Der er dog intet, der tyder på, at moderate ekstrapolationer ikke skulle være forsvarlige. ♦

Det ovenstående eksempel er et simpelt tilfælde af en generel type problemer (de såkaldte regressionsanalyser), vi senere skal beskæftige os nøjere med.

I det næste eksempel viser vi, hvorledes man kan gå frem, hvis fordelingen ikke er regulær, f.eks. hvis $\{x | f(x; \theta) > 0\}$ afhænger af θ .

EKSEMPEL 2.15. Vi betragter uafhængige stokastiske variable $X_1, \dots, X_n$, der alle

følger en $U(0, \theta)$ -fordeling, d.v.s.

$$f(x_i; \theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x_i \leq \theta \\ 0 & \text{ellers.} \end{cases}$$

Vi vil bestemme en maximum likelihood estimator over θ . For lettere at kunne behandle problemet minder vi om definitionen af **indikatorfunktionen** I_A ,

$$I_A(x) = \begin{cases} 1 & \text{hvis } x \in A \\ 0 & \text{hvis } x \notin A \end{cases}.$$

Vi har så

$$f(x; \theta) = \frac{1}{\theta} I_{[0, \theta]}(x).$$

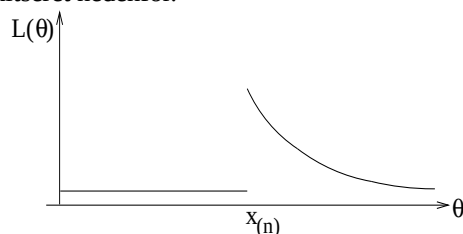
Likelihood-funktionen er nu

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta) = \theta^{-n} \prod_{i=1}^n I_{[0, \theta]}(x_i).$$

Vi ser, at $L(\theta)$ er forskellig fra 0 (og da lig θ^{-n}), netop hvis alle x_i 'er ligger i intervallet $[0, \theta]$. Da x_i 'erne er positive, er dette tilfældet, hvis og kun hvis det største x_i er indeholdt i $[0, \theta]$, d.v.s. hvis og kun hvis $x_{(n)} \in [0, \theta]$. Foreløbig er likelihood-funktionen

$$L(\theta) = \theta^{-n} I_{[0, \theta]}(x_{(n)}) = \theta^{-n} I_{[x_{(n)}, \infty)}(\theta),$$

hvor det sidste lighedstegn blot er udtryk for, at $0 \leq x_{(n)} \leq \theta$, hvis $\theta \in [x_{(n)}, \infty)$. Grafen for $L(\theta)$ er skitseret nedenfor.



Det ses umiddelbart, at $L(\theta)$ har maximum i $\theta = x_{(n)}$, hvorfor maximum likelihood estimatoren for θ er

$$\hat{\theta} = X_{(n)}.$$



BEMÆRKNING 2.9. Hvis vi i stedet havde anvendt tætheden

$$f(x; \theta) = \frac{1}{\theta} I_{(0, \theta)}(x),$$

ville likelihood-funktionen være blevet

$$L(\theta) = \theta^{-n} I_{(x_{(n)}, \infty)}(\theta).$$

(NB! $(0, \theta)$ betegner det åbne interval fra 0 til θ og $[0, \theta]$ det lukkede interval). ▼

Da $L(x_{(n)}) = 0$, har vi, at supremumsværdien af likelihood-funktionen ikke antages i noget punkt, hvorfor maximum likelihood estimatoren ikke er defineret. Ved at vælge værdier tilstrækkelig nær ved $x_{(n)}$ kan vi dog komme vilkårlig tæt til supremumsværdien $x_{(n)}^-$. Det vil derfor være naturligt også i dette tilfælde at definere $\hat{\theta} = X_{(n)}$. Disse vanskeligheder med definitionen kan omgås ved at indføje en række målteoretiske overvejelser; men da man i praksis aldrig er i tvivl om, hvad man skal forstå ved maximum likelihood estimatoren, vil vi ikke besværliggøre tilegnelsen af begrebet ved at give disse overvejelser.

EKSEMPEL 2.16. I dette eksempel betragter vi en række uafhængige stokastiske variable $X_1, \dots, X_n$, hvis fordelinger indeholder én eller flere ukendte parametre.

I nedenstående tabel giver vi en oversigt over maximum likelihood estimatorer af disse parametre. Estimatorerne er dannet på basis af de eksempel 2.7 anførte stikprøvefunktioner. I de fleste tilfælde kan fordelingen af estimatoren også udledes ved hjælp af det i eksempel 2.7 anførte. Hvis dette ikke er muligt, kan man i de fleste tilfælde bestemme den asymptotiske fordeling ved hjælp af sætning 2.9.

I tabellen har vi sat

$$\begin{aligned} k &= \sum_{i=1}^n k_i & m &= \sum_{i=1}^n m_i \\ Y_i &= \log_e X_i & Y &= \frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{n} \sum_{i=1}^n \log_e X_i \\ \bar{X} &= \frac{1}{n} \sum_{i=1}^n X_i \end{aligned}$$

Fordeling af X_i	Ukendt parameter	Maximum likelihood estimator
$B(m_i, p)$	p	$\frac{1}{m} \sum_{i=1}^n X_i$
$NB(k_i, p)$	p	$\frac{k}{k + \sum_{i=1}^n X_i}$
$\text{Pol}(m_i, p_1, \dots, p_k)$	$(p_1, \dots, p_k)$	$\frac{1}{m} \left(\sum_{i=1}^n X_{1i}, \dots, \sum_{i=1}^n X_{ki} \right)$
$P(\lambda)$	λ	$\frac{1}{n} \sum_{i=1}^n X_i$
$\text{Ex}(\beta)$	β	$\frac{1}{n} \sum_{i=1}^n X_i$
$G(k, \beta)$	β (k kendt)	$\frac{1}{kn} \sum_{i=1}^n X_i$
$N(\mu, \sigma^2)$	(μ, σ^2)	$\left(\frac{1}{n} \sum_{i=1}^n X_i, \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right)$
$N(\mu, \sigma^2)$	μ	$\frac{1}{n} \sum_{i=1}^n X_i$
$N(\mu, \sigma^2)$	σ^2 (μ kendt)	$\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$
$\text{LN}(\alpha, \beta^2)$	(α, β^2)	$\left(\frac{1}{n} \sum_{i=1}^n Y_i, \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2 \right)$
$U(a, b)$	(a, b)	$(X_{(1)}, X_{(n)})$



2.3.2 Mindste kvadraters metode

Lad der være givet n stokastiske variable $X_1, \dots, X_n$, som har middelværdier

$$E(X_i) = \mu_i(\theta_1, \dots, \theta_k).$$

Vi definerer da

DEFINITION 2.11. Ved **mindste kvadraters metode** for $(\theta_1, \dots, \theta_k)$ forstår vi den værdi $(\tilde{\theta}_1, \dots, \tilde{\theta}_k)$ som minimaliserer kvadratsummen

$$\sum_{i=1}^n \left(X_i - \mu_i(\theta_1, \dots, \theta_k) \right)^2. \quad (2.10)$$

▲

Metoden er meget anvendt, blandt andet fordi den ikke bygger på nogen antagelse om fordelingen af de stokastiske variable. For en vis type problemer (de såkaldte lineære modeller, i.e. modeller, hvor μ_i er en lineær funktion), kan det vises at mindste kvadraters estimatore er centrale; men ellers besidder metoden ingen generelle optimalitetssegenskaber.

Vi giver nogle eksempler.

EKSEMPEL 2.17. Lad $X_1, \dots, X_n$ have samme middelværdi μ . Vi vil bestemme mindste kvadraters estimatoren $\tilde{\mu}$. Vi betragter

$$f(\mu) = \sum_{i=1}^n (x_i - \mu)^2$$

og har

$$f'(\mu) = -2 \sum (x_i - \mu) = +2 \left(n\mu - \sum x_i \right)$$

og

$$f''(\mu) = 2n > 0,$$

således at f har det globale minimum $\mu = \frac{1}{n} \sum x_i$. Mindste kvadraters estimatoren er altså

$$\tilde{\mu} = \frac{1}{n} \sum_{i=1}^n X_i = \overline{X}.$$

◆

Hvis de foreliggende observationer ikke har samme varians, vil det ofte være hensigtsmæssigt at "normere" dem, så de bliver varianshomogene (hvis dette er muligt), inden man beregner et mindste kvadraters skøn. Dette giver hyppigt mere effektive estimatorer. Vi skal ikke komme nærmere ind på spørgsmålet, men blot illustrere det med et eksempel.

EKSEMPEL 2.18. Vi betragter en Poisson proces med intensitet λ . Vi måler antallet af hændelser i nogle (ikke overlappende) intervaller af længden $t_1, \dots, t_n$. Dette giver uafhængige stokastiske variable $X_1, \dots, X_n$, hvor $X_i \in P(\lambda t_i)$. Hvis vi nu vil bestemme en mindste kvadraters estimator for λ , kan vi gå frem som følger. Vi sætter - idet $E(X_i) = \lambda t_i$

$$g(\lambda) = \sum_{i=1}^n (x_i - \lambda t_i)^2$$

og har

$$\begin{aligned} g'(\lambda) &= -2 \sum (x_i - \lambda t_i) t_i \\ &= -2 \left(\sum x_i t_i - \lambda \sum t_i^2 \right). \end{aligned}$$

Det ses umiddelbart, at g har globalt minimum i

$$\lambda = \frac{\sum x_i t_i}{\sum t_i^2},$$

hvilket giver anledning til estimatoren

$$T_1 = \frac{\sum t_i X_i}{\sum t_i^2}.$$

Vi kunne alternativt definere nye variable $X_i/\sqrt{t_i}$. Vi ser, at $V(X_i/\sqrt{t_i}) = V(X_i)/t_i = \lambda$, d.v.s. de nye variable har samme varians. Da $E(X_i/\sqrt{t_i}) = \sqrt{t_i} \lambda$, skal vi altså nu minimalisere

$$\begin{aligned} f(\lambda) &= \sum_{i=1}^n \left(\frac{x_i}{\sqrt{t_i}} - \sqrt{t_i} \lambda \right)^2 \\ &= \sum_{i=1}^n \frac{1}{t_i} (x_i - t_i \lambda)^2, \end{aligned}$$

Vi får

$$\begin{aligned} f'(\lambda) &= -2 \sum (x_i - t_i \lambda) \\ &= -2 \left(\sum x_i - \lambda \sum t_i \right). \end{aligned}$$

Det ses, at dette giver estimatoren

$$T_2 = \frac{\sum X_i}{\sum t_i}.$$

For at afgøre, hvilken af de 2 alternative estimators T_1 og T_2 vi skal foretrække, bestemmer vi deres middelværdier og varianser. Vi har

$$E(T_1) = \frac{\sum t_i \lambda t_i}{\sum t_i^2} = \lambda$$

og

$$E(T_2) = \frac{\sum \lambda t_i}{\sum t_i} = \lambda,$$

d.v.s. at de begge er centrale. Ydermere er

$$\begin{aligned} V(T_1) &= \frac{1}{(\sum t_i^2)^2} \sum t_i^2 \lambda t_i = \lambda \cdot \sum t_i^3 / (\sum t_i^2)^2 \\ V(T_2) &= \frac{1}{(\sum t_i)^2} \sum \lambda t_i = \lambda / \sum t_i. \end{aligned}$$

Hvis e.g. $t_i = i$, får man efter nogen regning, at

$$\begin{aligned} V(T_1) &= \frac{9}{(2n+1)^2} \lambda \\ V(T_2) &= \frac{2}{n(n+1)} \lambda, \end{aligned}$$

og det ses, at $V(T_1) > V(T_2)$, hvorfor vi i dette tilfælde må foretrække T_2 , som er den mest effiente af de to estimators. ♦

BEMÆRKNING 2.10. Ovenstående må kun opfattes som et eksempel. Hvis man som her kender fordelingstypen, bør man bruge maximum likelihood metoden. Det kan i øvrigt nævnes, at maximum likelihood estimatoren netop er T_2 . ▼

Vi afslutter dette afsnit med at vise, at der er en sammenhæng mellem mindste kvadraters metode og maximum likelihood metoden i tilfældet uafhængige normalt fordelte observationer.

SÆTNING 2.11. Lad $X_1, \dots, X_n$ være uafhængige normalt fordelte stokastiske variable med fælles varians σ^2 og middelværdier $E(X_i) = \mu_i(\theta_1, \dots, \theta_k)$. Da er mindste kvadraters estimatorerne for $\theta_1, \dots, \theta_k$ også maximum likelihood estimatorer.

Bevis. Mindste kvadraters estimatorer bestemmes ved minimalisering af

$$f(\theta) = \sum_{i=1}^n \left(x_i - \mu_i(\theta_1, \dots, \theta_k) \right)^2. \quad (2.11)$$

Likelihood-funktionen er

$$L(\theta_1, \dots, \theta_k) \quad (2.12)$$

$$\begin{aligned} &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{1}{2\sigma^2} \left(x_i - \mu_i(\theta_1, \dots, \theta_k) \right)^2 \right) \\ &= k_1 \exp \left(-\frac{1}{2\sigma^2} \sum \left(x_i - \mu_i(\theta_1, \dots, \theta_k) \right)^2 \right) \\ &= k_1 \exp \left(-\frac{1}{2\sigma^2} f(\theta) \right). \end{aligned} \quad (2.13)$$

Her betegner k_1 en af $(\theta_1, \dots, \theta_k)$ uafhængig størrelse. Det ses da umiddelbart, at minimalisering af (2.11) er ensbetydende med maksimalisering af (2.13). ■

Det ses, at det er lettere at arbejde med udtrykket (2.11) end med (2.13). Derfor vil vi i de senere kapitler ofte beregne maximum likelihood estimator i den normale fordeling ved hjælp af mindste kvadraters metode.

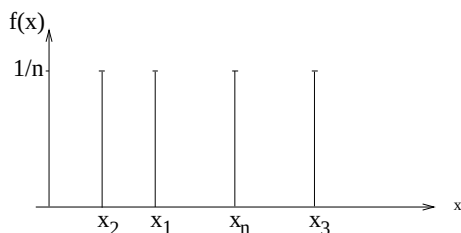
2.3.3 Momentmetoden

Det kan være vanskeligt at løse de ekstremumsbestemmelser, der kræves, for at finde enten maximum likelihood estimator eller mindste kvadraters estimator. Vi vil derfor omtale en mere ad hoc præget metode, som kan være ganske udmærket i mange situationer, nemlig den såkaldte momentmetode.

Vi betragter indledningsvis **ligefordelingen** over tallene $x_1, \dots, x_n$ (vilkaarlige, reelle tal). Den har frekvensfunktion

$$f(x) = \begin{cases} \frac{1}{n} & \text{hvis } \exists i \ (x = x_i) \\ 0 & \text{ellers} \end{cases}$$

d.v.s.



Vi vil bestemme middelværdi μ og varians σ^2 i denne fordeling. Af definitionen fås direkte

$$\mu = \sum_{i=1}^n x_i f(x_i) = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

og

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

ifølge (0.43), p. 55, har vi i øvrigt

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

Vi bemærker, at vi hermed har givet et nyt bevis for spaltningen i eksempel 2.2.

Det k 'te centrale moment i fordelingen er

$$\mu_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k.$$

Disse betragtninger over ligefordelingen over tallene $x_1, \dots, x_n$ giver anledning til følgende

DEFINITION 2.12. Lad $X_1, \dots, X_n$ være uafhængige stokastiske variable. Ved **den empiriske middelværdi** og **den empiriske varians** forstår vi størrelserne

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

henholdsvis

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

Det k 'te **empiriske, centrale moment** er

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

▲

BEMÆRKNING 2.11. Det indskræpes, at de empiriske momenter med ovenstående definition er stokastiske variable. Har vi imidlertid et realiseret udfald $x_1, \dots, x_n$ af $X_1, \dots, X_n$, giver dette selvfølgelig et realiseret udfald af f.eks. den empiriske varians, og denne størrelse er da netop variansen i ligefordelingen på tallene $x_1, \dots, x_n$. ▼

Hvis den simultane frekvensfunktion for $X_1, \dots, X_n$ er $f(x_1, \dots, x_n; \theta)$, hvor $\theta = (\theta_1, \dots, \theta_k)$ er ukendte parametre, vil vi ofte kunne udtrykke θ 'erne ved fordelingsmomenter. Eksempelvis er parametrene (μ, σ^2) lig (middelværdi, varians) i den normale fordeling.

BEMÆRKNING 2.12. Momentmetoden går da ud på at estimere parametrene $(\theta_1, \dots, \theta_k)$ ved at indsætte de tilsvarende empiriske momenter i en sådan relation. ▼

BEMÆRKNING 2.13. Momentmetoden er subjektiv i den forstand, at der som regel findes flere relationer mellem fordelings momenter og parametrene $\theta_1, \dots, \theta_k$, som f.eks. i nedenstående ▼

EKSEMPEL 2.19. Lad $X_1, \dots, X_n$ være stokastisk uafhængige $P(\lambda)$ -fordelte stokastiske variable. Da er

$$E(X_i) = \lambda$$

og

$$V(X_i) = \lambda,$$

d.v.s. såvel middelværdi som varians er lig den ukendte parameter λ . Følgelig vil såvel

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

som

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

være estimatorer for λ baseret på momentmetoden. I en sådan situation bør man **anvende den estimator, der er baseret på momenter af den laveste orden.** ♦

Momentmetoden anvendes meget ofte i tilfælde, hvor likelihood-ligningerne er vanskelige at løse. Et sådant tilfælde har vi i

EKSEMPEL 2.20. Nedenstående data stammer fra målinger over den spænding, hvorved nogle papirkondensatorer bryder sammen. Elektroderne havde et areal på 337.5 sq.in., og papiret var olieimprægneret 0.4 -mil papir.

Sammenbrudsspænding				
1440	1380	1500	1500	1440
1440	1440	1440	1680	1440
1620	1380	1380	1380	1500
1080	1620	1620	1200	900
1560	1500	1200	1440	1560

Vi må formulere en model. Det er åbenbart, at vi kan anvende princippet om det svageste led, som beskrevet i eksempel 1.37, p. 158. Fejlene viser sig ved, at der optræder ledende partikler. Hvis fordelingen af størrelsen af en sådan partikel er af eksponentiel type, da vil også fordelingen af den dielektriske "styrke" i omegnen af partiklen være af eksponentiel type. Det forekommer rimeligt at antage, og det er eksperimentelt godtgjort, at størrelsen af sådanne partikler er af eksponentiel type, således at vi har, at $X_i \in \text{Min}_1(\alpha, \beta)$. Vi antager ydermere, at X_i 'erne er stokastisk uafhængige.

Vi vil nu estimere (α, β) ved hjælp af momentmetoden. Sætter vi $\mu = E(X_i)$ og $\sigma^2 = V(X_i)$, har vi ved hjælp af sætning 1.39, p. 157

$$\begin{aligned}\mu &= \alpha - \beta\gamma \\ \sigma^2 &= \frac{\pi^2}{6} \cdot \beta^2\end{aligned}$$

eller

$$\begin{aligned}\alpha &= \mu + \frac{\sqrt{6}}{\pi} \sigma \cdot 0.57722 \\ \beta &= \frac{\sqrt{6}}{\pi} \sigma.\end{aligned}\tag{2.14}$$

Sum og kvadratsum af ovenstående observationer er

$$\begin{aligned}1440 &+ 1440 + \dots + 1560 = 35640 \\ 1440^2 &+ 1440^2 + \dots + 1560^2 = 51552000\end{aligned}$$

Heraf fås, at de observerede værdier af den empiriske middelværdi og den empiriske varians er:

$$\begin{aligned}\bar{x} &= \frac{1}{25} 35640 = 1425.6 \\ s_*^2 &= \frac{1}{25} (51552000 - 25 \cdot 1425.6^2) = 29744.7 = 172.5^2.\end{aligned}$$

Indsættes disse værdier for μ og σ^2 i (2.14), får vi følgende estimater dannet på basis

af momentmetoden

$$\alpha_M = 1503.2$$

$$\beta_M = 134.5.$$

Med bestemmelse af disse parametre er de sandsynlighedsteoretiske aspekter (principielt) fuldstændigt klarlagt. ♦

2.3.4 Intervalestimation (konfidensintervaller)

Som vi har set adskillige gange, er en estimator ikke entydigt bestemt ved f.eks. et krav om centralitet. Når man opgiver et estimat, vil det derfor være ønskeligt at få oplyst, hvad fordelingen af estimatoren er, eller i det mindste få oplyst, hvor stor en varians denne fordeling har. Det er da yderst almindeligt at inkorporere usikkerheden ved at anvende skrivemåder som, at en eller anden størrelse f.eks. er lig 0.27 ± 0.15 . Nu fremgår det jo desværre ikke helt klart, hvad meningen med denne skrivemåde er. Disse vanskeligheder kan omgås ved at indføre de såkaldte konfidensintervaller, i.e. intervaller, der med en vis sandsynlighed indeholder den ukendte parameter værdi.

Først imidlertid nogle definitioner.

DEFINITION 2.13. Et **stokastisk interval** I er et interval, hvor endepunkterne er stokastiske variable, d.v.s. $I = [X, Y]$ (fx.), hvor X og Y er stokastiske variable. ▲

Lad der nu være givet stokastiske variable $X_1, \dots, X_n$, hvis simultane frekvensfunktion $f(x_1, \dots, x_n; \lambda)$ er bestemt på nær den ukendte parameter λ .

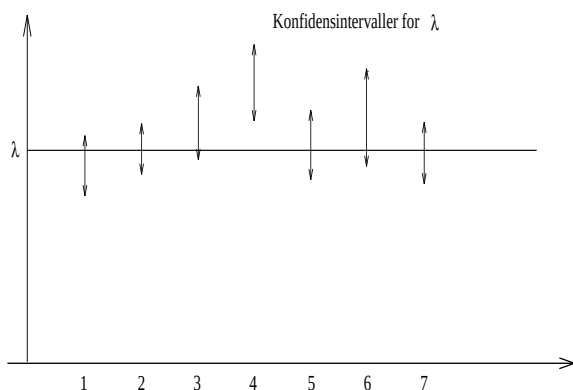
DEFINITION 2.14. Lad der være givet et stokastisk interval $I = [\underline{t}(X_1, \dots, X_n), \bar{t}(X_1, \dots, X_n)]$, hvorom der gælder

$$P\{\lambda \in I\} = 1 - \alpha$$

d.v.s.

$$P\{\underline{t}(X_1, \dots, X_n) \leq \lambda \leq \bar{t}(X_1, \dots, X_n)\} = 1 - \alpha.$$

Da kaldes $I = [\underline{t}(X_1, \dots, X_n), \bar{t}(X_1, \dots, X_n)]$ et **konfidensinterval** for λ med **konfidensgraden** $1 - \alpha$ eller blot et $(1 - \alpha)$ -**konfidensinterval** for λ . ▲



På ovenstående figur er vist 7 konfidensintervaller for λ beregnet på basis af 7 forskellige stikprøver. Vi ser, at et af konfidensintervallerne ikke indeholder λ . Hvis konfidensgraden er $1 - \alpha$, vil dette - i det lange løb - kun indtræffe i $100\alpha\%$ af tilfældene.

BEMÆRKNING 2.14. Sandsynlighedsudsagnet gælder sandsynligheden for, at I omslutter λ , og **ikke** sandsynligheden for, at λ ligger i I . Det er I , der varierer stokastisk, medens λ er et (fast) tal. ▼

BEMÆRKNING 2.15. En betegnelse som f.eks.

$$P\{\underline{l}(x_1, \dots, x_n) \leq \lambda \leq \bar{l}(x_1, \dots, x_n)\} = 95\%$$

har ingen mening. Enten ligger λ i det pågældende interval, eller også gør den ikke. I de respektive tilfælde er sandsynlighederne 1 og 0. ▼

NOTE 1. Hovedprincip når man søger konfidensinterval: Vi skal finde en funktion $h(X_1, \dots, X_n; \lambda)$ af $X_1, \dots, X_n$, der afhænger af λ , men hvis fordeling er uafhængig af λ . Derefter søges ulighederne

$$a \leq h(X_1, \dots, X_n; \lambda) \leq b$$

løst med hensyn til λ .

Princippet forekommer måske lidt dunkelt; men nogle eksempler vil forhåbentlig klare begreberne.

EKSEMPEL 2.21. Lad $X_1, \dots, X_n$ være indbyrdes uafhængige $N(\mu, \sigma^2)$ -fordelte stokastiske variable, hvor σ^2 antages at være kendt. Vi søger et $1 - \alpha$ konfidensinterval for μ . Som

$h(X_1, \dots, X_n; \mu)$ anvendes

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

(Den ses at tilfredsstille kravene). Hvis u_α betegner α -fraktilen i en $N(0, 1)$ -fordeling, har vi

$$P\left\{u_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < u_{1-\alpha/2}\right\} = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha.$$

Nu er

$$u_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < u_{1-\alpha/2} \Leftrightarrow \bar{X} - \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2} < \mu < \bar{X} - \frac{\sigma}{\sqrt{n}}u_{\alpha/2},$$

hvorfor intervallet

$$\left[\bar{x} - \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}, \bar{x} - \frac{\sigma}{\sqrt{n}}u_{\alpha/2}\right]$$

er et $(1 - \alpha)$ -konfidensinterval for μ . ♦

Vi betragter nu et eksempel med 2 normalt fordelte stikprøver med samme ukendte varians.

EKSEMPEL 2.22. Lad $X_1, \dots, X_n$ og $Y_1, \dots, Y_m$ være stokastisk uafhængige med

$$X_i \in N(\mu_1, \sigma^2) \quad i = 1, \dots, n$$

og

$$Y_i \in N(\mu_2, \sigma^2) \quad i = 1, \dots, m.$$

Vi søger et konfidensinterval for $\mu = \mu_1 - \mu_2$. Vi har, at

$$\bar{X} - \bar{Y} \in N\left(\mu_1 - \mu_2, \sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)\right),$$

d.v.s.

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} \in N(0, 1).$$

Nu er σ^2 ukendt, hvorfor den må estimeres. Vi vil her bruge et centralt skøn. Ifølge sætning 2.6 er

$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 + \frac{1}{\sigma^2} \sum_{i=1}^m (Y_i - \bar{Y})^2$$

fordelt som summen af 2 uafhængige $\chi^2(n-1)$ - og $\chi^2(m-1)$ -fordelte stokastiske variable. Af reproduktivitetssætningen for χ^2 -fordelingen fås

$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 + \frac{1}{\sigma^2} \sum_{i=1}^m (Y_i - \bar{Y})^2 \in \chi^2(n+m-2).$$

Da middelværdien af $\chi^2(\nu)$ -fordelt variabel er ν , er

$$\begin{aligned} S^2 &= \frac{1}{m+n-2} \left(\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^m (Y_i - \bar{Y})^2 \right) \\ &= \frac{1}{m+n-2} ((n-1)S_1^2 + (m-1)S_2^2) \end{aligned}$$

en central estimator for σ^2 . Her betegner S_1^2 og S_2^2 de centrale skøn for σ^2 dannet på basis af $X_1, \dots, X_n$ henholdsvis $Y_1, \dots, Y_m$. Vi ser i øvrigt, at **den nye varians-estimator er et vejte gennemsnit af de oprindelige med antallet af frihedsgrader som vægte**.

Indsætter vi nu S i stedet for σ , fås størrelsen

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S \sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} \cdot \frac{1}{S/\sigma}$$

der ses at være fordelt som

$$\frac{N(0, 1)}{\sqrt{\chi^2(n+m-2)/(n+m-2)}}.$$

Da tæller og nævner er stokastisk uafhængige, ifølge sætning 2.4, får vi af definition 1.43, p. 199, at $T \in t(n + m - 2)$. Vi har derfor, at T tilfredsstiller de krav, vi stillede til h : den afhænger af observationerne og $\mu_1 - \mu_2$, men dens fordeling er uafhængig af μ_1 og μ_2 . Benævnes α -fraktilen i en $t(n + m - 2)$ -fordeling t_α , haves

$$\begin{aligned} 1 - \alpha &= P\{t_{\alpha/2} \leq T \leq t_{1-\alpha/2}\} \\ &= P\left\{t_{\alpha/2} \leq \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} \leq t_{1-\alpha/2}\right\} \\ &= P\left\{\bar{X} - \bar{Y} - S\sqrt{\frac{1}{n} + \frac{1}{m}} t_{1-\alpha/2} \leq \right. \\ &\quad \left. \mu_1 - \mu_2 \leq \bar{X} - \bar{Y} - S\sqrt{\frac{1}{n} + \frac{1}{m}} t_{\alpha/2}\right\}, \end{aligned}$$

d.v.s.

$$\left[\bar{x} - \bar{y} - s\sqrt{\frac{1}{n} + \frac{1}{m}} t_{1-\alpha/2}, \bar{x} - \bar{y} - s\sqrt{\frac{1}{n} + \frac{1}{m}} t_{\alpha/2} \right]$$

er et $(1 - \alpha)$ -konfidensinterval for $\mu_1 - \mu_2$. ♦

Vi skal nu give et eksempel på anvendelse af et konfidensinterval.

EKSEMPEL 2.23. I nedenstående tabel er der givet resultatet af 10 målinger af klorindholdet i 2 batches af en polymer. Analyserne blev foretaget af den samme laborant, der anvendte samme metode ved de 2 undersøgelser.

Klorindhold i %	
Batch I	Batch II
58.59	55.71
59.64	56.72
58.45	56.65
58.64	57.56
58.00	58.27
57.03	56.58
57.33	57.08
57.80	57.13
58.04	57.92
58.41	56.21

Man er nu interesseret i at undersøge, om de 2 batches kan antages at være ens.

For at løse dette problem må vi først formulere en matematisk model. Da variationerne i de to stikprøver skyldes målefejl og "tilfældige" inhomogeniteter, **antager** vi, at observationerne er realiserede udfald af uafhængige, normalt fordelte stokastiske variable. Da analysemetoderne var ens, og da de undersøgte polymere stoffer a priori antages at være "næsten ens" (i.e. at klorinprocenterne er af samme størrelsesorden), vil vi endvidere antage, at varianserne er ens. Vi har altså de uafhængige stokastiske variable

$$X_i \in N(\mu_1, \sigma^2) \quad i = 1, \dots, 10$$

og

$$Y_i \in N(\mu_2, \sigma^2) \quad i = 1, \dots, 10.$$

En eventuel forskel i de 2 batches afspejles i størrelsen $\mu_1 - \mu_2$, hvorfor vi vil finde en intervalestimator for denne størrelse. Vi vil anvende konfidensintervallet fundet i eksempel 2.22. Vi har

$$\sum x_i = 581.9 \wedge \sum x_i^2 = 33869.1293.$$

Heraf fås

$$\bar{x} = 58.193$$

$$\sum (x_i - \bar{x})^2 = 33869.1293 - \frac{1}{10} 581.93^2 = 4.8768.$$

Analogt findes

$$\bar{y} = 56.983$$

$$\sum (y_i - \bar{y})^2 = 5.4588 \quad .$$

Følgelig har vi som skøn over σ^2

$$s^2 = \frac{4.8768 + 5.4588}{10 + 10 - 2} = 0.5742 = 0.758^2 \quad .$$

Hvis vi vælger $\alpha = 5\%$ får vi at $t(18)_{1-\alpha/2} = t(18)_{0.975} = -t(18)_{0.025} = 2.10$, d.v.s. at 95%-konfidensintervallet for $\mu_1 - \mu_2$ er

$$\begin{aligned} & \left[58.193 - 56.983 - 0.758 \sqrt{\frac{1}{10} + \frac{1}{10}} 2.10, \right. \\ & \quad \left. 58.193 - 56.983 + 0.758 \sqrt{\frac{1}{10} + \frac{1}{10}} 2.10 \right] \\ &= [1.21 - 0.71, 1.21 + 0.71] \\ &= [0.50, 1.92]. \end{aligned}$$

Hvordan skal dette resultat nu tolkes? Vi ved, at hvis vi på basis af nye observationer beregner et konfidensinterval på ovenstående måde, vil et sådant interval i det lange løb indeholde den sande værdi af $\mu_1 - \mu_2$ i 95% af tilfældene. Vi vil nu sige, at da vi **har** observeret intervallet $[0.50, 1.92]$, vil vi gå ud fra, at dette er en hændelse med en "stor sandsynlighed", og derfor slutte, at der er tale om et interval, der indeholder den sande værdi af $\mu_1 - \mu_2$. Da $0 \notin [0.50, 1.92]$, vil vi altså omvendt konkludere, at den sande værdi ikke er 0, d.v.s. at $\mu_1 - \mu_2 \neq 0$ eller $\mu_1 \neq \mu_2$. Vi konkluderer altså, at da 95%-konfidensintervallet ikke indeholder 0, må vi antage, at de to batches er forskellige.

Bemærk, at ovenstående konklusion bygger på det arbitrært valgte $\alpha = 0.05$. ♦

EKSEMPEL 2.24. I omstående 2 tabeller er anført konfidensintervaller for ukendte parametre i forskellige fordelinger baseret dels på en stikprøve $X_1, \dots, X_n$ af uafhængige variable, dels på 2 stikprøver $X_1, \dots, X_n$ og $Y_1, \dots, Y_m$ af uafhængige variable.

I tilfældet med binomial og Poisson fordelingen bemærker vi, at der er tale om konfidensintervaller med en konfidensgrad på **mindst** $1 - \alpha$. På grund af, at fordelingerne er diskrete, kan man ikke opnå eksakte intervaller.

Vi bemærker endvidere, at det ikke er nødvendigt at anvende $\alpha/2$ - og $(1 - \alpha/2)$ -fraktilerne i fordelingerne. Man kan uden videre erstatte disse med β og $(1 - (\alpha - \beta))$ -fraktilerne, hvor $0 \leq \beta \leq \alpha$. Dette giver stadig $(1 - \alpha)$ -konfidensintervaller, men muligvis med en længere middellængde.

I omstående skema p. 266 er anvendt betegnelserne

$$\begin{aligned} m &= \sum_{i=1}^n m_i \quad (\text{binomialfordeling}) \\ x &= \sum_{i=1}^n x_i \quad (\text{binomial- og Poisson fordeling}) \\ s'^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \\ s^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned}$$

og i skemaet p. 266 er anvendt betegnelserne

$$\begin{aligned} s'^2_1 &= \frac{1}{n} \sum_{i=1}^n (x_i - \mu_1)^2; & s'^2_2 &= \frac{1}{m} \sum_{i=1}^m (y_i - \mu_2)^2 \\ s^2_1 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2; & s^2_2 &= \frac{1}{m-1} \sum_{i=1}^m (y_i - \bar{y})^2 \\ s^2 &= \frac{(n-1)s^2_1 + (m-1)s^2_2}{n+m-2} \\ &= \frac{1}{n+m-2} \left(\sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^m (y_i - \bar{y})^2 \right), \end{aligned}$$

hvor naturligvis

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad ; \quad \bar{y} = \frac{1}{m} \sum_{i=1}^m y_i.$$



Forde- ling af X_i	Ukendt parameter	$(1 - \alpha)$ -konfidensinterval
$B(m_i, p)$	p	$\left[\frac{x}{x + (m - x + 1) F(2m - 2x + 2, 2x)_{1-\alpha/2}}, \frac{(x + 1) F(2x + 2, 2m - 2x)_{1-\alpha/2}}{m - x + (x + 1) F(2x + 2, 2m - 2x)_{1-\alpha/2}} \right]$
$P(\lambda)$	λ	$\left[\frac{1}{2n} \chi^2(2x)_{\alpha/2}, \frac{1}{2n} \chi^2(2x + 2)_{1-\alpha/2} \right]$
$G(k, \beta)$	β (k kendt)	$\left[\frac{\sum_{i=1}^n x_i}{G(nk, 1)_{1-\alpha/2}}, \frac{\sum_{i=1}^n x_i}{G(nk, 1)_{\alpha/2}} \right]$ eller, hvis $2nk$ hel $\left[\frac{2 \sum_{i=1}^n x_i}{\chi^2(2nk)_{1-\alpha/2}}, \frac{2 \sum_{i=1}^n x_i}{\chi^2(2nk)_{\alpha/2}} \right]$
$N(\mu, \sigma^2)$	μ (σ^2 kendt)	$\left[\bar{x} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, \bar{x} - \frac{\sigma}{\sqrt{n}} u_{\alpha/2} \right]$
$N(\mu, \sigma^2)$	μ (σ^2 ukendt)	$\left[\bar{x} - \frac{s}{\sqrt{n}} t(n-1)_{1-\alpha/2}, \bar{x} - \frac{s}{\sqrt{n}} t(n-1)_{\alpha/2} \right]$
$N(\mu, \sigma^2)$	σ^2 (μ kendt)	$\left[s'^2 \frac{n}{\chi^2(n)_{1-\alpha/2}}, s'^2 \frac{n}{\chi^2(n)_{\alpha/2}} \right]$
$N(\mu, \sigma^2)$	σ^2 (μ ukendt)	$\left[s^2 \frac{n-1}{\chi^2(n-1)_{1-\alpha/2}}, s^2 \frac{n-1}{\chi^2(n-1)_{\alpha/2}} \right]$

Angående den i skemaet anvendte notation: se p. 265

$X_i \in G(k_1, \beta_1) \quad i = 1, \dots, n \quad ; \quad Y_i \in G(k_2, \beta_2) \quad i = 1, \dots, m$	
Ukendt parameter	$(1 - \alpha)$ -konfidensinterval
$\frac{\beta_1}{\beta_2}$ (k_1 og k_2 kendte, $2nk_1$ og $2mk_2$ hele)	$\left[\frac{\sum x_i / nk_1}{\sum y_i / mk_2} F(2mk_2, 2nk_1)_{\alpha/2}, \frac{\sum x_i / nk_1}{\sum y_i / mk_2} F(2mk_2, 2nk_1)_{1-\alpha/2} \right]$
$X_i \in N(\mu_1, \sigma_1^2) \quad i = 1, \dots, n \quad ; \quad Y_i \in N(\mu_2, \sigma_2^2) \quad i = 1, \dots, m$	
$\mu_1 - \mu_2$ (σ_1^2 og σ_2^2 kendte)	$\left[\bar{x} - \bar{y} - \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} u_{1-\alpha/2}, \bar{x} - \bar{y} - \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} u_{\alpha/2} \right]$
$\mu_1 - \mu_2$ ($\sigma_1^2 = \sigma_2^2 = \sigma^2$ ukendt)	$\left[\bar{x} - \bar{y} - s \sqrt{\frac{1}{n} + \frac{1}{m}} t(n+m-2)_{1-\alpha/2}, \bar{x} - \bar{y} - s \sqrt{\frac{1}{n} + \frac{1}{m}} t(n+m-2)_{\alpha/2} \right]$
$\frac{\sigma_1^2}{\sigma_2^2}$ (μ_1 og μ_2 kendte)	$\left[\frac{s_1'^2}{s_2'^2} F(m, n)_{\alpha/2}, \frac{s_1'^2}{s_2'^2} F(m, n)_{1-\alpha/2} \right]$
$\frac{\sigma_1^2}{\sigma_2^2}$ (μ_1 ukendt, μ_2 kendt)	$\left[\frac{s_1^2}{s_2'^2} F(m, n-1)_{\alpha/2}, \frac{s_1^2}{s_2'^2} F(m, n-1)_{1-\alpha/2} \right]$
$\frac{\sigma_1^2}{\sigma_2^2}$ (μ_1 og μ_2 ukendte)	$\left[\frac{s_1^2}{s_2^2} F(m-1, n-1)_{\alpha/2}, \frac{s_1^2}{s_2^2} F(m-1, n-1)_{1-\alpha/2} \right]$

Angående den i skemaet anvendte notation: se p. 265

Bemærk, at vi i konfidensintervallet for $\mu_1 - \mu_2$ i tilfældet ukendt varians har forudsat, at varianserne er ens.

Bibliografi

- [1] Artin, E.: *Einführung in die Theorie der Gammafunktion*. Leipzig 1931.
- [2] Barnett, Vic: *Comparative Statistical Inference*. John Wiley & Sons, New York & London 1973.
- [3] Bar-Shalom, Yaakov: *On the Asymptotic Properties of the Maximum-Likelihood Estimate Obtained from dependent Observations*. J.R. Statist. Soc., B, vol. 33 1971, p. 72.
- [4] Bennet, C. A., & Franklin, N. L.: *Statistical Analysis in Chemistry and the Chemical Industri*. John Wiley & Sons, New York 1954.
- [5] Bourbaki, N.: *Element de Mathematique. Fonctions d'une variable réelle*. Hermann, Paris 1961.
- [6] Bradley, J. V.: *Distribution-free Statistical Tests*. Prentice-Hall, Englewood Cliffs 1968.
- [7] Bradley, Ralph A. and Cart, John H.: *The asymptotic properties of ML estimators when sampling from associated populations*. Biometrika, vol. 49 1962, p. 205.
- [8] Brøns, H. K.: *Forelæsninger i matematisk statistik og sandsynlighedsregning 9–15, 2. udgave*. Institut for Matematisk Statistik, Københavns Universitet 1968.
- [9] Brøns, H. et al.: *Statistik 1*. Institut for matematisk statistik. Københavns Universitet, 1974.
- [10] Buch, K.R. m.fl.: *Sandsynlighedsregning i grundrids*. Gyldendal, København 1967.
- [11] Chow, V.T.: *The Log-Probability Law and Its Engineering Applications*. Proc. ASCE, vol. 80, No. 536, 1954.
- [12] Cox, D.R. and Hinkley, D.V.: *Theoretical Statistics*. Chapman and Hall, London 1974.

- [13] David, H.A.: *Order Statistics*, John Wiley & Sons, New York 1970.
- [14] Davies, O. L. (Ed.): *Statistical Methods in Research and Production*. Oliver and Boyd, London 1961.
- [15] Davies, O. L. (Ed.): *Design and Analysis of Industrial Experiments*. Oliver and Boyd, London 1967.
- [16] Dorfman, R.: *The detection of defective members of large populations*. Annals of Math. Stat. 14 (1943), pp. 436-440.
- [17] Epstein, B.: *Statistical aspects of fracture problems*. Journal of Applied Physics, vol. 19, (1948), pp. 140-147.
- [18] Epstein, B. & Brooks, H.: *The Theory of Extreme values and Its Implications in the Study of the Dielectric Strength of Paper Capacitors*. Journal of Applied Physics 19, (1948), pp. 544-550.
- [19] Feller, W.: *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons, New York 1965.
- [20] Ferguson, T. S.: *Mathematical Statistics. A Decision Theoretic Approach*. Academic Press, New York 1967.
- [21] Fisz, M.: *Probability Theory and Mathematical Statistics*. John Wiley & Sons, New York 1963.
- [22] Gnedenko, B.: *Sur la distribution limits du terme maxime d'une série aléatoire*. Ann. of Math. 44 (1943), pp. 423-453.
- [23] Gumbel, E.J.: *Statistics of Extremes*. Columbia University Press, New York 1958.
- [24] Hájek, S. & Šidák, Z.: *Theory of Rank Tests*. Academia, Prague 1967.
- [25] Hald, A.: *Statistical Theory with engineering Applications*. John Wiley & Sons, New York & London 1952.
- [26] Hansen, N.H.: *Stokastiske modeller*. IMSOR, Lyngby 1975
- [27] Jensen, Arne: 1948.
- [28] Jensen, Arne: *A distribution model*. Munksgaard, København 1954.
- [29] Jensen, Arne: *An Elucidation of Erlangs Statistical Works through the Theory of Stochastic Processes*. Acta Polytechnica Scandinavica, Danish Contribution No. 12, København 1960.
- [30] Jensen, Arne: *Application of the Decision Theory to Law and Administration*. Bull. Inst. Int. de Stat., Paris 1961.
- [31] Jensen, Arne: *2. rapport til Thyborønudvalget af 1957 vedr. stormflod og digebygning*, 1963.

- [32] Jørsboe, Ole Groth: *Noter til sandsynlighedsregning*. DTH, 1972.
- [33] Kendall, M. G.: *Rank Correlation Methods (Third Ed.)*. Charles Griffin & Co., Ltd., London 1962.
- [34] Kendall, M. G., & Stuart, A.: *The advanced Theory of Statistics*, vol. 1, 2, 3, Charles Griffin & Co., London 1963, 1967, 1966.
- [35] Kendall, M.G. and Stuart, A.: *The Advanced Theory of Statistics*, vol. 2. Charles Griffin and Co., London 1967.
- [36] Khichin, A. I.: *Mathematical Foundations of Information Theory*. Dover, New York 1957.
- [37] Lehmann, E. L.: *Testing Statistical Hypotheses*. John Wiley & Sons, New York 1959.
- [38] Lehmann, E. L.: *Nonparametrics: Statistical Methods Based on ranks*. Holden Day, Inc., San Francisco 1975.
- [39] Luce, R. D., & Raiffa, H.: *Games and Decisions*. John Wiley & Sons, New York 1957.
- [40] Noether, G. E.: *Elements of Nonparametric Statistics*. John Wiley & Sons, New York 1967.
- [41] Owen, D. B.: *Handbook of Statistical Tables*. Addison-Wesley, Reading Mass. 1962.
- [42] Patil, G.P.: *On the evolution of the negative binomial distribution with examples*. Technometrics, vol. 2, No. 4 (1960), pp. 501-505.
- [43] Plackett, R. L.: *Principles of Regression Analysis*. Clarendon Press, Oxford 1960.
- [44] Puri, M. L. & Sen, P. K.: *Nonparametric Methods in Multivariate Analysis*. John Wiley & Sons, Inc., New York 1971.
- [45] Raiffa, H., & Schlaifer, R.: *Applied Statistical Decision Theory*. Graduate School of Business Administration, Harvard 1961.
- [46] Rao, C. Radhakrishna: *Linear Statistical Inference and Its Applications*. John Wiley & Sons, New York 1965.
- [47] Renyi, A.: *Wahrscheinlichkeitsrechnung*. VEB Deutscher Verlag der Wissenschaften, Berlin 1966.
- [48] Scheffé, H.: *The Analysis of variance*. John Wiley & Sons, New York 1959.
- [49] Simonsen, K.Aa.: *Om Portland-Cementens Konstitution*. Proc. 11 fra Nordiska Kemistmötet, 20.- 25. august 1962 i Åbo, pp. 193-214.

-
- [50] Sobel, M. & Groll, P.A.: *Group testing to eliminate efficiently all defectives in a binomial sample*. Bell Syst. Tech. Journal 38 (1959), pp. 1179-1252.
- [51] Sobel, M.: *Group testing to classify all all defectives in a binomial sample*. A Contribution in Information and Decision Processes. (Ed. R.E. Machol). McGraw Hill, New York 1960, pp. 127-161.
- [52] Spearman, C.: *The Proof and Measurement of Association Between Two Things*. Am j. Psych., vol. 15, 1904.
- [53] Walsh, J. E.: *Handbook of Nonparametric Statistics*, Vol. i-iii. D. van Nostrand Company, Inc., Princeton 1962, 1965, 1968.
- [54] Weibull, W.: *A Statistical distribution function of wide applicability*. J. Appl. Mech. 18 (1952), p. 293.
- [55] Wilks, S.S.: *Mathematical Statistics*. John Wiley & Sons, New York 1962.

Stikordsregister

A

A posteriori fordeling 519
A priori fordeling 519
Acceptområde 302, 310
Additionssætning 37
Additivitet 424, 424 f
Aktion 516
Approximationer til fordelinger 130 f

B

$B(n, p)$ 92
B-fordeling se betafordeling
B-funktion se betafunktion
Bartlett's test se test
Bayes estimator se estimator
Bayes regel 38, 519
Bayes strategi 521
Bayes test se test
Bayes-risiko 521
 $Be(\alpha, \beta)$ 182
Bedste kritisk område se kritisk område
Beregningsenhed 410, 439
Beregningsformler se endv. variationsanalyse, 410 f
Beregningsnulpunkt 410, 439
Bernoulli eksperiment 91
Bernoulli fordeling 91
Bernoulli forsøg 91
Bernoulli variabel 91
Beslutningsfunktion 516
Beslutningsteori 515
Betafordeling 181, 203, 226, 520

Betinget fordeling 45
Betinget middelværdi 59
Betinget sandsynlighed 36
Betinget varians 59
Bimodal 66
Binomialfordeling . 92, 103, 110, 130, 191, 193, 204 f, 225, 264, 266, 326 f, 333, 520
Binomialkoefficient 26
Bio-assay 185
Biologisk vækst 138
Brudstyrke 142, 158

C

$Ca(\alpha, \beta)$ 183
Cauchy fordeling 183 f, 201
Cauchy type 161 f
Čebyčev's ulighed 218
Central estimator se estimator
Central grænseværdisætning 125 f
Centralt moment 55
Chebychevs ulighed 70
 $\chi^2(n)$ 194
 $\chi^2(n, \gamma^2)$ 197
 χ^2 -fordeling .. se chi i anden fordeling
Chi i anden fordeling 194 f, 226
Cochran's test se test
Cochrans sætning 195
Compound fordelinger 189
Compound-fordeling 80
Covariansanalyse 469
Cramér-Rao's ulighed 230

D

Decisionsteori 515
 Deduktion 212
 Dispersionsmatricen 58

E

Efficiens se estimator
 Pitman- 372
 Eksponentialfordeling . 120, 156, 225,
 238, 257, 354, 399
 Eksponentiel type 148 f
 Ekstreme vandstande 152
 Ekstremværdier 142 f
 Ekstremværdifordelinger .. 142, 148 f,
 403
 Elektricitetsforbrug 141
 Elektrisk ledningsevne 158
 Empirisk fordelingsfunktion ... 9, 90 f
 Empirisk frekvensfunktion 9
 Empiriske momenter 255
 beregning af 410
 Ensided variansanalyse se
 variansanalyse
 Enstikprøveproblem 373
 Entropi 488
 Erlang fordeling 119
 Estimat 213
 Estimationsmetoder 233 f
 Estimationsteori 213 f
 Estimator 213
 Bayes 531
 central 215
 efficient 229 f, 238
 konsistent 218
 lineær 529
 minimax 528
 sufficient ... se stikprøvefunktion
 $Ex(\beta)$ 120

F

$F(n, m)$ 201
 $F(n, m; \gamma)$ 201
 F-fordeling 201 f
 Faktorforsøg 445
 Fakultetsfunktion 26
 Fejl af type I & II 302 f

Fejlophobningsloven 69
 Fermat 29
 Fiasko 91
 Flerdimensional stokastisk variabel 43
 foldning 50
 Fordelingsfrie tests se test
 Fordelingsfunktion 40
 Fordelingstype, test for 404
 Formparameter 120
 Forsøgsplaner . se variansanalyser, 413
 Forventning 53
 Fraktil 42, 74
 Fraktildiagram 397
 Frekvens 9
 Frekvensfunktion 41
 regulær 229
 Friedman's test 496, 507
 Frieman's test se test
 Frihedsgrader ... 194, 199, 201, 227 f,
 322, 363, 405, 406, 428 f, 443
 Fuldstændig homogenitet 415
 Fysiske målinger (fejlløve) 128

G

$G(k, \beta)$ 120
 Γ -funktionen 74
 Γ -fordeling se gammafordeling
 Γ -funktion se gammafunktion
 Gammafordeling 120 f, 132, 176, 183,
 194, 225, 249, 266, 354 f, 489,
 520
 Gammafunktion 74
 Gauss fordeling 123, se
 (normalfordeling)
 Gennemsnit af kvadrerede successive
 differenser 390
 Geometrisk fordeling 99
 Geometrisk middelværdi 56
 Geometrisk middelværdi ... 135, 489
 Geometrisk Standardafvigelse 56
 Geometrisk standardafvigelse 135, 489
 Grafiske metoder 394
 Group testing 96
 Gruppering 88, 401

H

$H(n, M, N)$ 102
 Hagens hypoteser 128
 Histogram 14, 88 f, 401 f
 Homogenitet af fordelinger
 test for 367
 Homogenitet inden for grupper .. 414,
 433
 Hydrologisk hændelser 138
 Hypergeometrisk fordeling .. 102, 192,
 193, 329
 Hypotese 310
 sammensat 310
 simpel 310
 Hypoteseprøvning 301 f
 Hyppighed 9
 Hændelse 31

I

Indikatorfunktion 86
 Indkomstfordeling 141, 186
 Inferens 211 f, 301 f
 Injektiv funktion 222
 Intensitet 115
 Intervalestimation 258
 intervalskala 489
 Invarians 490
 Invers normalvægtstest 379

J

Jacobi-determinant 46

K

Karakteristiske funktion 66
 Kendalls, τ 500 ff, 507
 Klassifikationsskala 487
 Knusningsprocesser 138
 Kolmogorov-Smirnov test 407
 Kombination 26
 Konfidensinterval 258 f, 312
 Konfidensintervaller 258
 Konfidensområde 409
 Konfidenssgrad 258
 Konjugerede fordelinger 521
 Konsistens se estimator
 Kontingenstabeller 363

Kontinuitetskorrektion 131, 506
 Konvergens 70
 Konvergens i fordeling 71
 Konvergens i sandsynlighed 72
 Korrelationskoefficient 57
 kovarians 57
 Kovariansanalyse 469
 Kovariansmatricen 58
 Kritisk område 311, 312
 bedste 311
 uniformt bedste 312
 Kritiske værdi 315
 Kruskal-Wallis' test se test
 Kumuleret 9
 Kurtosis 66
 Kvalitetskontrol 100
 Kvartil 16
 Kvotienttest 317 f
 Kvotientteststørrelse 322

L

$L(\alpha, \beta)$ 185
 $\mathcal{L}$ 74
 Lack of memory 122
 Λ_1 148
 $\Lambda_{2,k}$ 148
 $\Lambda_{3,k}$ 148
 LaPlace fordeling 184 f
 Latin square se romersk kvadrat
 Ledningsevne, elektrisk 158
 Lehmann's test se test
 Levetid 158
 Levetidsfordeling 122, 176
 Ligefordeling 254
 Ligefordeling på $\{0, 1, \dots, n\}$.. 187 f
 Likelihood ligning 236
 Likelihood-funktion 235
 Lindeberg-Lévy's sætning 130
 Lineær estimator se estimator
 Lineært bånd 227
 Ljapounov's sætning 126
 $LN(\alpha, \beta^2)$ 134
 $LoD(\theta)$ 187
 Logaritmisk fordeling 187 f
 Logaritmisk normalfordeling ... 134 f,
 226, 249, 489

Logistisk fordeling 185 f
 Loven om proportional effekt 137

M

Marginal fordeling 44
 Matematisk forventning 53
 $\text{Max}_1(\alpha, \beta)$ 151
 $\text{Max}_1(k, \beta)$ 162
 $\text{Max}_3(k; \alpha, \beta)$ 170
 Maximum likelihood metode . . . 233 f,
 253
 Maxwell's hypoteser 130
 Median 16, 42
 Metaltræthed 174
 Middelværdi 53
 empirisk 255
 Middelværdiestimator . 216, 218, 223,
 249, 250, 255
 Midtrange 377, 502, 506
 $\text{Min}_1(\alpha, \beta)$ 157, 257, 408
 $\text{Min}_2(k, \beta)$ 164
 $\text{Min}_3(k; \alpha, \beta)$ 172
 Mindste kvadraters metode 249 f
 mindste værdi 142
 Minimax estimator se estimator
 Minimax strategi 518
 Minimax test se test
 Modus 66
 Moment 53, 55
 Momentfordeling 139
 Momentmetode 254
 Multiplikationssætning 38
 Måleskalaer 487

N

$N(\mu, \sigma^2)$ 123
 $\text{NB}(k, p)$ 99
 Negativ binomialfordeling . . . 99, 190,
 225, 249, 324
 Neyman's kriterium 222
 Neyman-Pearson's fundamentale lemma
 314
 Niveau se signifikansniveau
 Nomialskala 487
 Normalfordeling 123 f, 194 f, 226, 227,
 243, 249, 253, 260, 318, 335,

398, 402, 403, 405

Normaligninger 475
 Normeret normalfordeling 123

O

Observat se stikprøvefunktion
 Opportunity loss 523
 Ordnet udvalg 26

P

p-fraktil 42
 $\text{Par}(k, \beta)$ 186
 Parameter 516
 Pareto fordelingen 186
 Parret t-test 350
 Partikelstørrelse 139
 Partikelstælling 112
 Pascal 29
 Pascal's fordeling se negativ binomial-
 fordeling, se negativ binomi-
 alfordeling
 Permutation 26
 Pitman-efficiens 372
 Poisson fordeling 108 f, 132, 190, 191,
 197, 225, 231, 238, 251, 264,
 266, 314, 332 f, 394, 404, 520
 Poisson's proces se Poisson fordeling,
 se Poisson fordeling
 Poisson's sætning 110
 $\text{Pol}(n, p_1, \dots, p_k)$ 104
 Polynomialfordelingen . 104, 225, 249,
 360 f, 404
 Polynomialkoefficient 26
 Polynomialkoefficienter 105
 Pooled variansskøn 352
 Positionsparameter 42
 Princippet om det svageste led . . . 158
 Proportional effekt, loven om 137
 Præ posteriori analyse 527
 Præcision 521

R

Rangkorrelationskoefficienter . . 500 ff
 Rangtests 488
 Ratioskala 489
 $\text{Ray}(\sigma^2)$ 198

Rayleigh fordelingen 198
 Regressionsanalyse 246, 451
 med 1 uafhængig variabel ... 451
 med 2 uafhængige variable ... 474
 sammenligning af 2 466
 Regressionslinie 454
 Regret 523
 Regulær frekvensfunktion se
 frekvensfunktion
 Rektangulære fordeling 180 f, 226, 246
 Residualvariation 442
 Risiko, Bayes- 521
 Risikofunktion 517
 Romersk kvadrat 442
 Run 386
 Run test 386 f

S

Sammenbrudsspænding 257
 Sammensat hypotese se hypotese
 Sandsynlighed 31
 Sandsynlighedsfelt 31
 Sandsynlighedsfelt 31
 Siegel-Tukey test se test
 σ -algebra 31
 Signifikansniveau 311
 Simpel hypotese se hypotese
 Simultan fordeling 43
 Skalaparameter 42
 Skævhedsparameter ... 196, 199, 201
 Spaltning af total variation se variation
 Spaltningssætningen 195
 Spearman's, r 504, 507
 Spredning 56
 største værdi 142
 Standardafvigelse 56
 Standardiseret normalfordeling ... 123
 Statistic se stikprøvefunktion
 Statistisk inferens se inferens
 Stikprøvefunktion 213
 sufficient 220 f, 237
 Stikprøveplan 100
 Stikprøvestørrelse 309
 Stikprøveudtagning 102 f
 med tilbagelægning 94
 uden tilbagelægning ... 103, 193

Stirling's formel 76
 Stokastisk interval 258
 Stokastisk variabel 32
 Store tals lov 73
 Store tals lov, De 220
 Strategi 516
 Student's t-fordeling se t-fordeling
 Styrke 306, 311
 Styrkefunktion 311
 Stærkeste test se test
 Størrelse se kritisk område
 Succes 91
 Sufficiens se stikprøvefunktion
 Sumpolygon 14, 88 f
 Svageste led, Princippet om det ... 158

T

$t(n)$ 199
 $t(n, \mu)$ 199
 t-fordeling 199 f, 203
 Tabsfunktion 516
 Test 310
 Bartlett's 482
 Bayes' 538
 Cochran's 486
 Fordelingsfrit 370, 487
 Fortegns- 370, 371
 Invers normalvægts-(van der Waerden) 379, 491
 Kruskal-Wallis' test .. 491 ff, 507
 Lehmann's 485
 Minimax 536
 Siegel-Tukey 381
 stærkeste 312
 uniformt stærkeste 312
 Varianshomogenitet, t. for ... 482
 Wilcoxon's ... 370, 371, 374, 491
 Test i ... se de respektive fordelinger
 Testprincipper 314 f
 Testteori se hypoteseprøvning
 Tilbagelægning 26
 Tilfældighed, test for 385 f
 Tilstand 516
 Tosidet variansanalyse se
 variansanalyse

Tostikpøveproblem 373
 Tredie type 169 f
 Type 43
 Tæthed 41

U

$U(a, b)$ 180
 u-fraktiler 123
 Uafhængige variable (i regr.-analyse) 474
 Uafhængighed 45
 $UD(n)$ 187
 Udfaldsrum 31
 Udtømmende 220
 Uniform fordeling 180 f, se
 rektangulær fordeling
 Uniformt bedste kritiske område . . . se
 kritisk område, 316
 Uniformt stærkeste test se test
 Unimodal 66
 Univers 31
 Uparret t-test 351

V

van der Waerden test se test
 Vandstande, ekstreme 152
 Varians 55
 Variansanalyse 413 f, 415
 beregningsformler . . . 418 f, 431,
 437 f
 ensidet 413 f
 tosidet 423 f
 Variansanalyseeskema . . 419, 430, 436,
 443, 447, 456, 475
 Variansestimator . . 216, 223, 227, 246,
 249, 254, 261, 335, 417, 419,
 431
 Varianshomogenitet, test for . . . se test
 Variation
 i faktorforsøg 446
 i regressionsanalyse 482
 inden for grupper . . 414, 433, 435
 mellem grupper 416
 mellem prøver 442
 mellem rækker 430, 435, 436, 443
 mellem søjler . 430, 435, 436, 443
 residual- 442, 443, 447

spaltning af totale . 417, 428, 435,
 442, 446, 454, 475
 vekselsvirknings- 443
 vekselvirknings- . . 428, 430, 435,
 436, 442, 447

Variationskoefficienten 56
 Vekselvirkning se additivitet
 Ventetidsfordeling 99, 118
 Vindhastigheder 166
 Vækst 138

W

Weibull fordeling 172
 Wilcoxon-test se test

Det græske alfabet

	<i>Bogstavnavn</i>	<i>Udtale</i>	<i>Gengivelse</i>
A α	alfa	[a][a:]	a
B β	bēta	[b]	b
Γ γ	gamma	[g]	g
Δ δ	delta	[d]	d
E ε	epsilon	[e]	e
Z ζ	zēta	[ts,s]	z
H η	ēta	[æ:]	ē
Θ θ θ	thēta	[θ,th,t]	th,t
I ι	iōta	[i][i:]	i
K κ	kappa	[k]	k
Λ λ	lambda	[l]	l
M μ	my	[m]	m
N ν	ny	[n]	n
Ξ ξ	ksi	[ks]	ks(x)
O o	omikron	[o]	o
Π π	pi	[p]	p
P ρ	ro	[r]	r
Σ σ ς	sigma	[s]	s
T τ	tau	[t]	t
Υ υ	ypsilon	[y][y:]	y
Φ φ	fi	[f]	f(ph)
X χ	khi	[x,ç,kh,k]	ch(kh)
Ψ ψ	psi	[ps]	ps
Ω ω	ōmega	[ā:]	ō

