TEKST NR 167a 1988

BASISSTATISTIK
1. Diskrete modeller

Jørgen Larsen

IMFUFA
Roskilde Universitetscenter

September 1988

TEKSTER fra

IMFUFA ROSKILDE UNIVERSITETSCENTER
INSTITUT FOR STUDIET AF MATEMATIK OG FYSIK SAMT DERES FUNKTIONER I UNDERVISNING, FORSKNING OG ANVENDELSER

IMFUFA, Roskilde Universitetscenter, Postboks 260, DK-4000 Roskilde.

Jørgen Larsen: **Basisstatistik, 1 + 2**

IMFUFA-tekst 167 A&B/1988 (ii+156) + 165 sider ISSN 0106-6242

Basisstatistik er en lærebog der giver en introduktion til et grundlag for matematisk-statistisk tankegang og til de traditionelle simple statistiske modeller. Overalt er det likelihoodmetoden der er den paradigmatiske metode for analyse af statistiske modeller.

De enkelte kapitler i fremstillingen er i nogen grad uafhængige af hinanden, men det anbefales afgjort at man læser kapitlerne nogenlunde i rækkefølge. Kapitlerne er forsynede med resumeer, der i kort form giver nogle konkrete hovedpointer. De forskellige metoder illustreres omhyggeligt med gennemregnede eksempler.

Der forudsættes et vist beskedent kendskab til grundlæggende begreber og terminologi fra sandsynlighedsregningen. Der henvises i denne forbindelse til *Grundbegreber i Sandsynlighedsregningen*, IMFUFA-tekst 166/1988.

Indhold

Indledning: om statistik

Faget statistik beskæftiger sig med analyse og præsentation af talmaterialer. Engang drejede det sig, hvad man endnu kan se af navnet, især om tal vedrørende statens tilstand og økonomiske og militære styrke (og om de fjendtlige staters mangel på samme). I vore dage er der stadig dele af statistikfaget der har med (indsamling,) præsentation og formidling af numeriske oplysninger at gøre, men faget består nu fortrinsvis af aktiviteter der har til formål i talmaterialer at skille det væsentlige fra det uvæsentlige. Som det slagordsmæssigt blev formuleret (i 1922) af R. A. Fisher (1890-1962), en af den klassiske statistiks grundlæggere, *målet med statistiske metoder er datareduktion*: man søger at reducere talmaterialet til nogle få størrelser der udtrykker det væsentlige, og resten, det uvæsentlige, anses så for at være tilfældigt. Det er i den forbindelse en vigtig pointe, at det uvæsentlige ikke anses for at være blot uspecificeret tilfældigt, men for at være tilfældigt i henhold til en nærmere angivet sandsynlighedsfordeling. Ved således også at påtage sig at beskrive den såkaldte tilfældige variation bliver statistikeren nemlig i stand til at vurdere, om tilsyneladende forskelle, f.eks. mellem virkningerne af visse behandlinger, er signifikante, dvs. om de er så store at de ikke med rimelighed kan tænkes at være fremkommet ved tilfældigheder.

En statistisk model for et talmateriale er kort fortalt en beskrivelse af, hvilke træk ved talmaterialet man p.t. anser for at være væsentlige, systematiske, og hvilke man p.t. anser for at være tilfældige. (Den statistiske model foregiver altså *ikke* at være en model af den del af virkeligheden som talmaterialet vedrører!) Ordentlige statistiske modeller bliver sædvanligvis kun til gennem et (ofte langvarigt) samarbejde mellem statistikeren og den fagmand der har "produceret" talmaterialet. Statistiske modeller formuleres i matematiksprog, hvor beskrivelsen af det tilfældige sker ved hjælp af sandsynlighedsregningen. Takket være

den matematiske formulering er man i stand til at deducere egenskaber ved de statistiske modeller og de måder de analyseres på. Ved brug af forskellige statistiske principper kan man fra en statistisk model udlede den statistiske metode der bør benyttes i et konkret tilfælde; metodens praktiske udførelse kræver undertiden en del regnearbejde, hvilket dog takket være de moderne regnetekniske hjælpemidler ikke er noget problem (i modsætning til tidligere, hvor man i høj grad søgte efter regnemæssige tilnærmelser til komplicerede udtryk og efter statistiske modeller der godt nok ikke var helt rigtige, men til gengæld gav lette beregninger).

Er statistik et håndværk eller en videnskab? Mange introducerende statistikbøger præsenterer faget som et ugennemskueligt sammensurium af metoder og tommelfingerregler med en besynderlig vekslen mellem på den ene side tilsyneladende løse og ubegrundede antagelser, og på den anden side stringent matematik og numerik; det kan næsten synes som om læseren skal i lære hos håndværksmesteren, den praktiserende statistiker, for at indføres til nogle af dette esoteriske fags hemmeligheder! — Faget statistik som vi kender det i dag har næppe hundrede år på bagen, men takket være den gren af faget som går under betegnelsen matematisk statistik eller teoretisk statistik er der alligevel sket en betydelig begrebsafklaring og -præcisering og en deraf muliggjort systematisk behandling af centrale problemer på måder som gør i hvert fald dele af statistikken fortjent til at benævnes en videnskab.

Lad os udskille tre typer hovedingredienser i et statistisk problem:

1. den aktuelle praktiske problemstilling,
2. den statistiske model,
3. den statistiske metode (dvs. den måde man analyserer tallene på).

Den statistiske videnskab har langt overvejende beskæftiget sig med den opgave der består i at udlede adækvate statistiske metoder ud fra en given statistisk model for det praktiske problem. På det seneste beskæftiger man sig også en del med den "modsatte" opgave: givet den statistiske metode man vil benytte i den foreliggende problemstilling, hvad er da den underliggende statistiske model? Derimod ligger det betydelig tungere med en systematisk behandling af de for alvor interessante opgaver for den statistiske videnskab: hvordan bærer man sig *egentlig* ad med at komme fra den praktiske problemstilling til den

"rigtige" statistiske model (og dermed en statistisk metode), eller alternativt, hvordan bærer man sig *egentlig* ad med at komme fra den praktiske problemstilling til den "rigtige" statistiske metode (og dermed statistiske model); og de tilsvarende normative opgaver: hvad vil det sige at en statistisk model/metode er "den rigtige"?

Hvad er det da man skal lære, når man lærer statistik? Man skal lære de grundlæggende principper (for så vidt de findes) og lære hvordan de anvendes. Man skal altså ikke lære seks forskellige statistiske metoder som anvendes i seks forskellige situationer — for hvad så når det er en helt syvende situation der foreligger! Den der kan og kender de grundlæggende principper vil i nogen grad også være den nye situations herre, mens den der kun kan de seks metoder tværtimod ender med at blive slave af denne sin kunnen.

Det man skal lære er ikke hvordan man laver regnestykkerne, ejheller hele det formelle matematiske apparatur der i større eller mindre omfang måtte bringes i anvendelse; skønt ofte besværlige og tidskrævende er der kun tale om højst nødvendige hjælpediscipliner, som bare er noget man kan! Det man egentlig skal lære er

- hvordan finder man frem til en statistisk model for en given problemstilling?
- hvordan bestemmer man de statistiske metoder der skal anvendes for at analysere data i forhold til den statistiske model?
- hvad er det for en slags svar man får i forhold til sin problemstilling?

Kapitel 1

Præsentation af et talmateriale

I dette kapitel præsenteres et større talmateriale, som vil blive benyttet til eksempler i nogle af de følgende kapitler. Desuden gøres nogle indledende tilløb til at vise, hvad en statistisk model for et talmateriale er og hvad den kan bruges til.

Talmaterialet

I forbindelse med en undersøgelse af studieforløb ved Roskilde Universitetscenter har man oprettet en database som omfatter alle studerende der har været indskrevet ved centeret et kortere eller længere tidsrum inden for perioden 1972-84. Oplysningerne i databasen vedrører personernes køn, alder, valg af basisuddannelse, valg af overbygningsudddannelse, startår, samlet studietid etc. Databasen blev oprettet i forbindelse med et overbygningsprojekt[1]. — I dette og i nogle af de følgende kapitler vil vi studere eksempler der beskæftiger sig med en del af dette talmateriale, nemlig den del der handler om basisuddannelser. Denne del af databasen vedrører de 5665 personer der har været indskrevet ved en basisuddannelse inden for perioden 1972-84, og for hver af disse personer har vi fire oplysninger, nemlig personens *køn*, *basis*, *startår* og *alder*; et eksempel er vist i Figur 1.1.

[1] M. W. Johansen, P. Kattler og T. Andreasen (1985): *COX I STUDIETIDEN - Cox's regressionsmodel anvendt på studenteroplysninger fra RUC.* IMFUFA-tekst 109/85.

Figur 1.1: En personoplysning fra databasen. Denne person er en pige der i en alder af 21 år begyndte på SAM-BAS i 1980.

KØN	K
BASIS	SAM
STARTÅR	80
ALDER	21

Man har almindeligvis ikke nogen særlig fornøjelse af en opremsning af KØN, BASIS, STARTÅR og ALDER for hver af de 5665 personer. Det er statistikkens og statistikerens opgave at formidle indholdet af databasen. I første omgang kan man foretage simple optællinger der præsenteres i tabeller og grafer.

Klassifikation efter ét kriterium

Variablen STARTÅR kan antage en af de 12 værdier 72, 73, ..., 83. Vi kan inddele de 5665 personer i grupper, *klasser*, efter hvornår de er startet. Man siger at man *klassificerer* personerne efter STARTÅR. Vi kan så tælle op hvor mange der starter i hvert af de 12 forskellige år, dvs. hvor mange der kommer i hver klasse. Resultatet ses i Tabel 1.1.

En tabel giver de nøjagtige tal, men det er ikke altid at det nøjagtige er så forenеligt med det overskuelige, så derfor er det undertiden en hjælp hvis man supplerer en tabel med en (overskuelig) grafisk fremstilling. Figur 1.2 er et forsøg på en grafisk fremstilling af indholdet i Tabel 1.1[2]. Vi kan naturligvis også klassificere efter f.eks. KØN (Tabel 1.2 og Figur 1.3) eller efter BASIS (Tabel 1.3 og Figur 1.4).

Klassifikation efter to kriterier

Det er velkendt, at der på NAT-BAS er forholdsvis færre piger end på de to andre basisuddannelser, men det er ikke noget man kan se af de

[2]En storartet introduktion til moderne principper for grafisk fremstilling af statistiske data er W. S. Cleveland: *The Elements of Graphing Data*. Monterey, California. 1985.

Tabel 1.1: 5665 studerende fordelt efter startår.

STARTÅR	antal
1972	755
73	688
74	413
75	397
76	354
77	226
78	433
79	387
80	495
81	537
82	537
83	443
i alt	5665

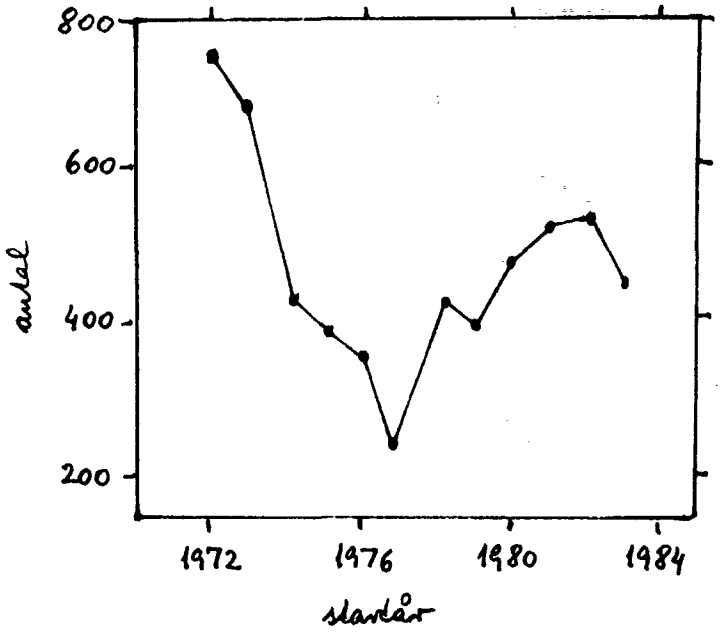
Tabel 1.2: 5665 studerende fordelt efter køn.

KØN	antal
M	3170
K	2495
i alt	5665

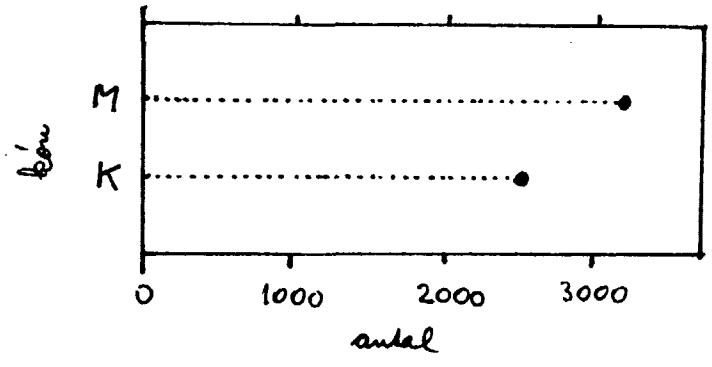
Tabel 1.3: 5665 studerende fordelt efter basisuddannelse.

BASIS	antal
HUM	1893
SAM	2643
NAT	1129
i alt	5665

Figur 1.2: Antal personer der starter på en basisuddannelse i hvert af årene 1972-83 jf. Tabel 1.1.



Figur 1.3: 5665 studerende fordelt efter køn, jf. Tabel 1.2.



Tabel 1.4: 5665 studerende fordelt efter køn og basisuddannelse.

		BASIS:			
		HUM	SAM	NAT	sum
KØN:	M	918	1482	770	3170
	K	975	1161	359	2495
	sum	1893	2643	1129	5665

Tabel 1.5: 5665 studerende fordelt efter basisuddannelse og køn.

		KØN		
		M	K	sum
BASIS:	HUM	918	975	1893
	SAM	1482	1161	2643
	NAT	770	359	1129
	sum	3170	2495	5665

tabeller og grafer vi indtil nu har fremstillet. Det kan man derimod se ved at foretage en klassifikation efter de to variable KØN og BASIS på én gang, se Tabel 1.4.

Tabel 1.4 er en såkaldt KØN×BASIS-*tabel*, dvs. en tabel hvor de forskellige *rækker* svarer til de forskellige værdier af KØN og de forskellige *søjler* svarer til de forskellige værdier af BASIS. Tabeller af denne slags hedder *kontingenstabeller*. Vi kunne også stille tallene op i en BASIS×KØN-*tabel*, Tabel 1.5. Man kalder ofte en tabel som Tabel 1.4 for en 2×3 -tabel og en tabel som Tabel 1.5 for en 3×2 -tabel. At Tabel 1.5 er en 3×2 -tabel betyder, at tabellens første inddelingskriterium (BASIS) har tre niveauer og dens andet inddelingskriterium (KØN) har to niveauer. Vi kan se at 3×2 -tabellen har tre rækker (plus rækken med de to søjlesummer og totalsummen) og to søjler (plus søjlen med de tre rækkesummer og totalsummen).

Tabeller som Tabel 1.4 og Tabel 1.5 er *tosidede* eller *todimensionale* kontingenstabeller, fordi der er to inddelingskriterier. I en todimensional tabel som Tabel 1.5 kalder man søjlen bestående af rækkesummerne for tabellens *rækkemarginaler* og rækken bestående af søjlesummerne for *søjlemarginalerne*. I Tabel 1.5 udgør rækkemarginalerne netop den

Tabel 1.6: 5665 studerende fordelt efter basisuddannelse, startår og køn.

	HUM			SAM			NAT		
år	M	K	sum	M	K	sum	M	K	sum
72	197	132	329	190	97	287	92	47	139
73	137	102	239	223	125	348	84	17	101
74	72	58	130	121	97	218	48	17	65
75	49	88	137	117	86	203	43	14	57
76	49	90	139	93	74	167	36	12	48
77	72	66	138	12	13	25	44	19	63
78	65	65	130	117	121	238	46	19	65
79	43	50	93	101	105	206	53	35	88
80	61	83	144	102	130	232	76	43	119
81	84	85	169	123	116	239	87	42	129
82	53	105	158	148	99	247	83	49	132
83	36	51	87	135	98	233	78	45	123

endimensionale tabel Tabel 1.3 og søjlemarginalerne netop den endimensionale tabel Tabel 1.2. Tabellens rækkemarginaler bestemmes ved, som man siger, at *summere over søjlerne*, og søjlemarginalerne bestemmes ved at *summere over rækkerne*.

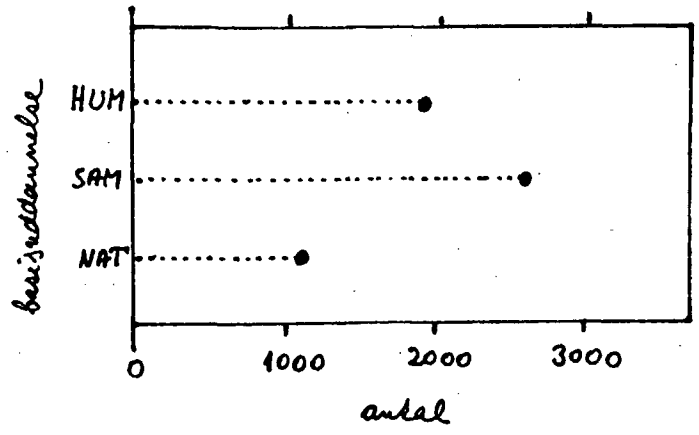
Figur 1.5 og 1.6 er to grafiske fremstillinger af vores todimensionale tabel.

Flerdimensionale tabeller

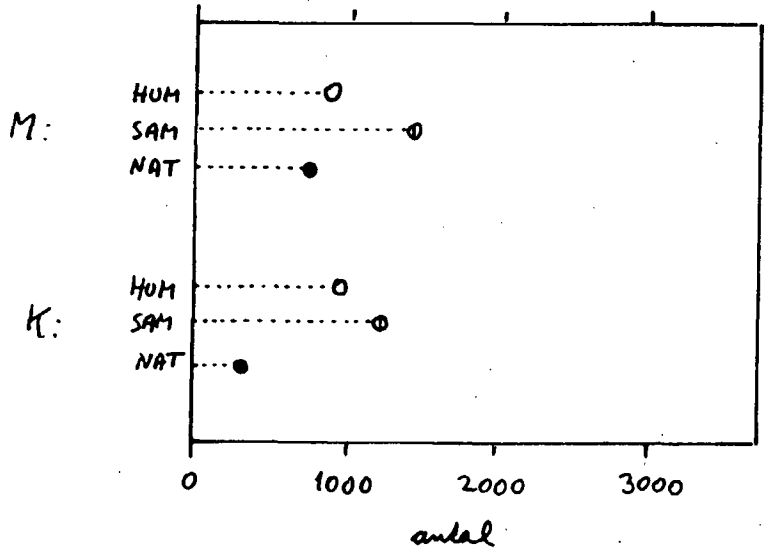
En *tredimensional* tabel er resultatet af at klassificere efter *tre* kriterier. Hvis vi klassificerer de 5665 personer efter BASIS, STARTÅR og KØN får vi den tredimensionale BASIS×STARTÅR×KØN-tabel, som bliver en $3 \times 12 \times 2$ -tabel og som består af tre *lag*, 12 *rækker* og to *søjler* som antydet i Figur 1.7. Man kan skrive en tredimensional tabel ved at skrive hvert lag for sig; vores $3 \times 12 \times 2$ -tabel kan således skrives som tre 12×2 -tabeller, hvilket vises i Tabel 1.6.

En tredimensional tabel har mange marginaler. I Tabel 1.6 er medtaget (lag,række)-marginalerne, som fremkommer ved summation over søjler (KØN). Derimod er (lag,søjle)-marginalerne, som fremkommer

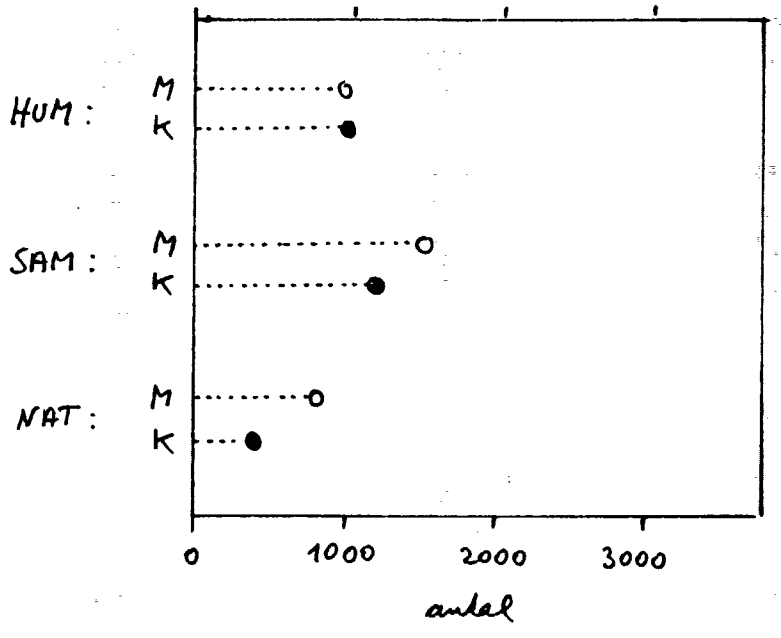
Figur 1.4: 5665 studerende fordelt efter basisuddannelse, jf. Tabel 1.3.



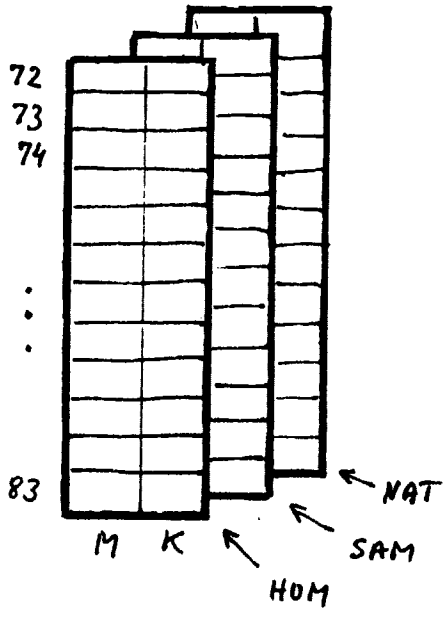
Figur 1.5: 5665 studerende fordelt efter køn og basis, jf. Tabel 1.4.



Figur 1.6: 5665 studerende fordelt efter basis og køn, jf. Tabel 1.5.



Figur 1.7: En $3 \times 12 \times 2$ -tabel.



ved summation over rækker (STARTÅR) ikke medtaget, men disse marginaler udgør præcis BASIS×KØN-tabellen Tabel 1.5. Endvidere er f.eks. rækkemarginalerne for Tabel 1.6 simpelt hen Tabel 1.1.

Figur 1.8 er en grafisk fremstilling af Tabel 1.6. Figuren (og tabellen) kan nu kommenteres på mange måder af uddannelsesforskere og sociologer, og læseren opfordres til at gøre sig bekendt med RUC's historie for eventuelt at kunne fornemme hvad der kan ligge bag kurvernes udseende.

Spørgsmål til talmaterialet

Af Figur 1.8 fremgår det ret tydeligt, at der på NAT-BAS altid er flere drenge end piger, hvorimod det måske er mere hip som hap på de to andre basisuddannelser. For nærmere at belyse dette forhold kan man f.eks. udregne brøkdelen af drenge på hver basisuddannelse og derved blive i stand til at tegne Figur 1.9, hvor de faktiske brøkdele sammenholdes med tallet 51.2% som er brøkdelen af mænd i aldersgruppen 20-29 år i hele landet. Figur 1.9 viser at der er en vis fluktuation[3] i forholdet mellem kønnene. Det kunne godt se ud som om M-procenten på SAM-BAS varierer omkring den "rigtige" værdi, mens M-procenten på HUM-BAS synes at være aftagende og at have temmelig store udsving, og endelig ligger M-procenten på NAT-BAS afgjort *over* den "rigtige" værdi 51.2%, dog synes afvigelsen at være aftagende. Man kan nu stille sig spørgsmål af typen

- Afviger M-procenten på SAM-BAS reelt ikke mere fra 51.2% end hvad der kan forklares som værende tilfældige afvigelser?

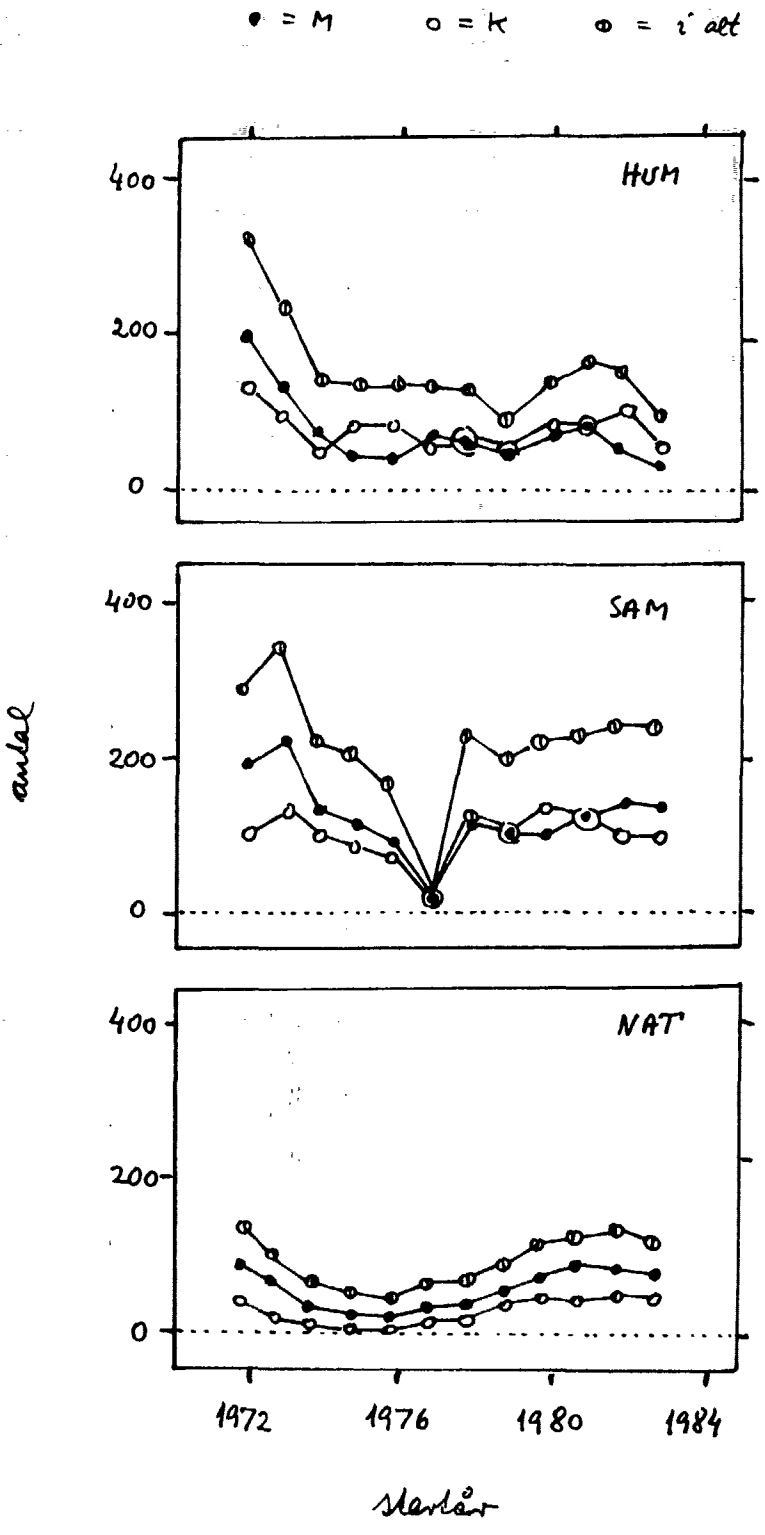
  Spørgsmålet kan (f.eks.) konkretiseres til: I årgang '83 er der 58% drenge på SAM-BAS, hvor man ellers kunne formode at der var 51.2%. Kan man forklare det som værende en tilfældighed at SAM-BAS '83 har så mange drenge?

- Kan de observerede forskelle mellem de tre basisuddannelser forklares som værende tilfældigheder?

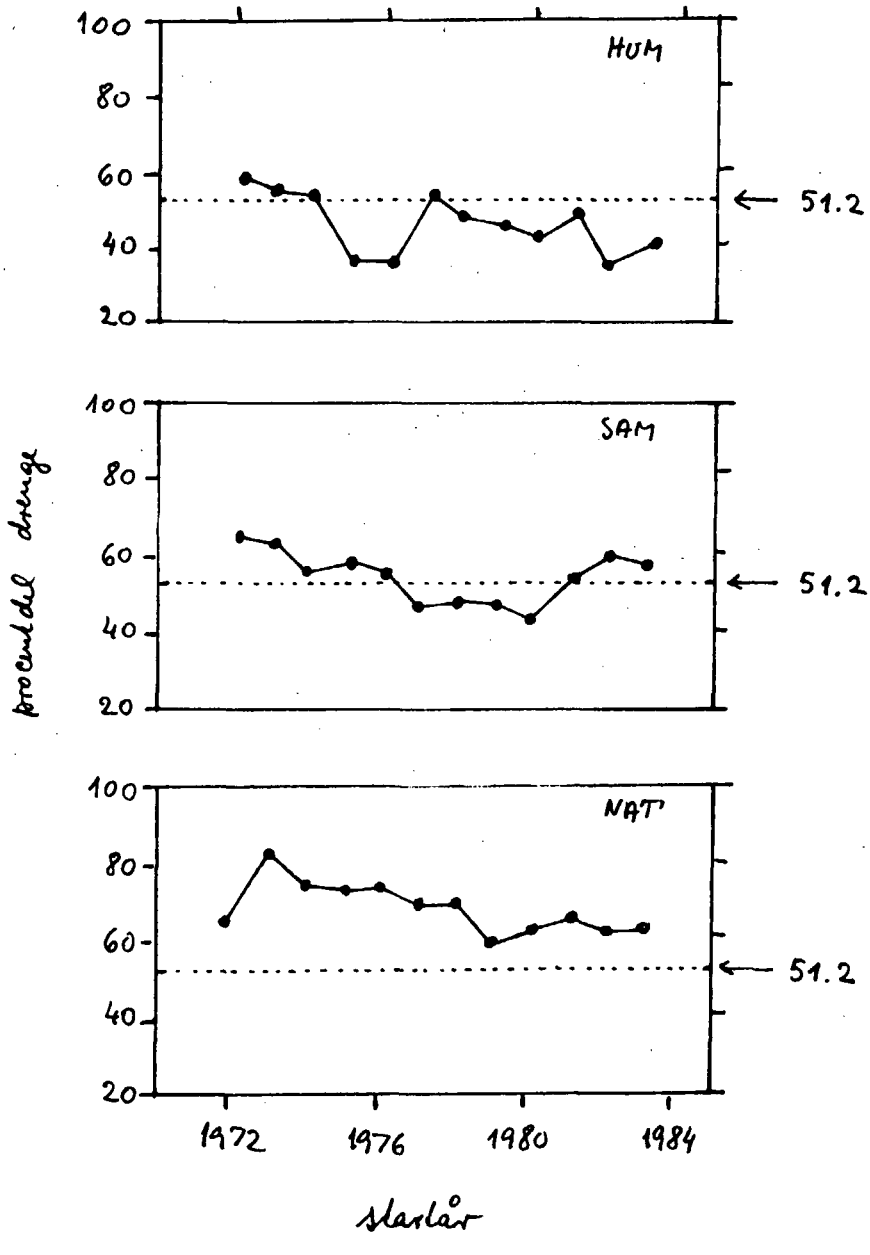
  Et konkret eksempel: I årgang '83 er der hhv. 41%, 58% og 63% drenge på de tre basisuddannelser. Kan man tillade sig at sige,

[3] dvs. (tilfældig) stigen og falden.

Figur 1.8: Det årlige optag på hver basisuddannelse, fordelt på køn.



Figur 1.9: Den årlige andel af drenge på hver basisuddannelse.



Tabel 1.7: Den procentvise fordeling på basisuddannelser for hvert år.

	procent			
år	HUM	SAM	NAT	sum
1972	44	38	18	100
1973	35	51	15	101
1974	31	53	16	100
1975	35	51	14	100
1976	39	47	14	100
1977	61	11	28	100
1978	30	55	15	100
1979	24	53	23	100
1980	29	47	24	100
1981	31	45	24	100
1982	29	46	25	100
1983	20	53	28	101

at der egentlig er samme andel drenge de tre steder og at de observerede forskelle blot skyldes tilfældigheder?

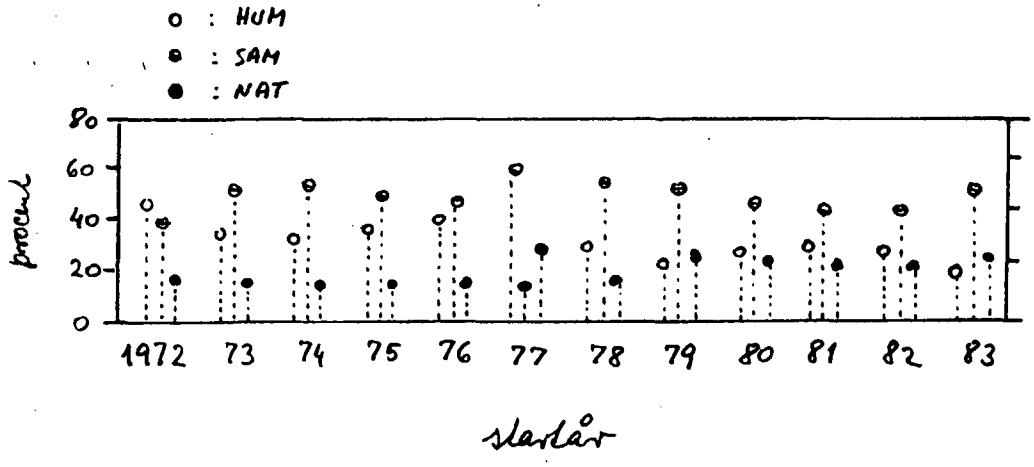
Det er den slags spørgsmål som statistikeren kan hjælpe med at besvare.

Man kunne også interessere sig for hvordan styrkeforholdet er mellem de tre basisuddannelser. Til det brug kunne man for hvert af de 12 år udregne den relative fordeling på hver af de tre basisuddannelser, se Tabel 1.7.

På grundlag af Tabel 1.7 kan man tegne figurerne 1.10 og 1.11. Figur 1.10 viser hvordan fordelingen på de tre basisuddannelser ændrer sig med tiden. Figur 1.11 viser hvordan hver enkelt basisuddannelses andel ændrer sig med tiden. Figurerne kunne antyde, at der måske er en tendens til at NAT-BAS vokse relativt, på bekostning af HUM-BAS. Man kan derfor stille sig spørgsmål af typen:

- Sker der en reel ændring i styrkeforholdet mellem de tre basisuddannelser? (I vurderingen heraf skal man naturligvis se bort fra året 1977 hvor der var lukket for tilgang til SAM-BAS.) Vi kan forsimple spørgsmålet til (f.eks.):

Figur 1.10: Den årlige fordeling på de tre basisuddannelser, jf. Tabel 1.7.

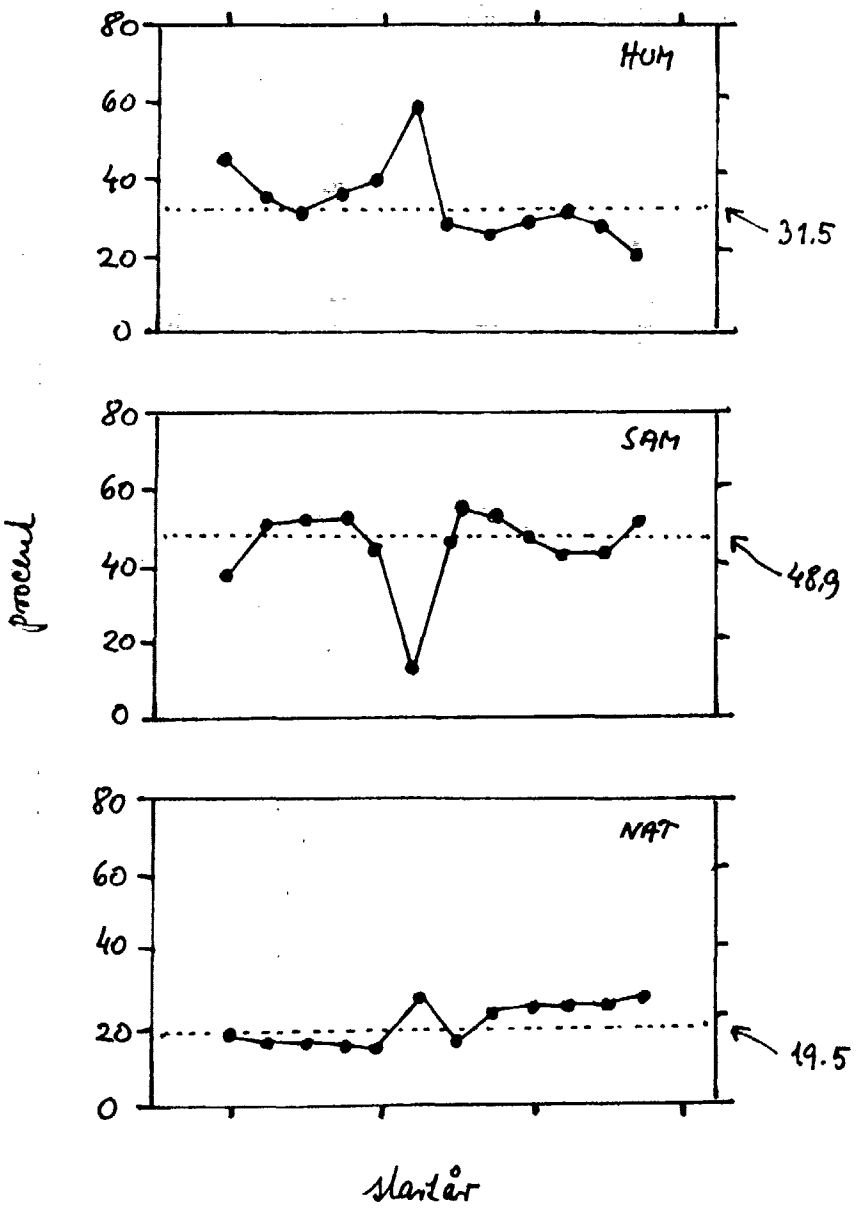


I 1974 fordeler de 413 optagne sig med 31% på HUM, 53% på SAM og 16% på NAT, og i 1982 fordeler de 537 optagne sig med andelene 29%, 46% og 25%. Kan man tillade sig at sige, at der egentlig består det samme "styrkeforhold" mellem basisuddannelserne de to år og at de observerede forskelle blot skyldes tilfældigheder?

Tilløb til en statistisk model

I forbindelse med præsentationen af (nogle træk ved) studiestatistiktallene fandt vi på at formulere forskellige statistiske spørgsmål, statistiske problemer. I de følgende kapitler skal vi beskæftige os med, hvordan man når frem til et svar på sådanne (og på mere sofistikerede) spørgsmål. Der findes "standardopskrifter" på hvad man stiller op med slige problemer. Disse opskrifter bygger imidlertid på visse grundlæggende forudsætninger, der sædvanligvis udmøntes i en *statistisk model* for det talmateriale der er tale om. Når man først har formuleret den statistiske model for talmaterialet er alle problemer på sin vis løst, fordi der findes almindeligt anerkendte statistiske principper for, hvordan man så bør udføre den videre analyse. Det svære er at finde en brugbar statistisk model.

Figur 1.11: Hver basisuddannelses årlige andel af den samlede tilgang, jf. Tabel 1.7. — De stiplede referencelinier angiver gennemsnitsværdien for alle årene excl. 1977.



Tabel 1.8: SAM-basister 1983 fordelt efter køn.

KØN	antal	andel
M	135	58%
K	98	42%
sum	233	100%

Hvad er dog en statistisk model da så, vil læseren spørge. Det kan der ikke gives et kortfattet forståeligt svar på, men i de følgende kapitler kommer der mange eksempler på statistiske modeller, og det skulle forhåbentlig give en idé om hvad det er. Som en begyndelse vil vi nu gøre visse tilløb til formuleringer af statistiske modeller for nogle af de situationer der blev nævnt i det foregående.

Der blev nævnt (side 13) det spørgsmål, om der på SAM-BAS årgang '83 var flere drenge end man egentlig skulle forvente, jf. Tabel 1.8, eller om det ikke er andet end hvad der kan forklares som tilfældigheder. Øjensynligt ligger den observerede andel af drenge noget over de godt 51% der gælder for den danske befolkning i de aldersklasser der normalt søger ind på et universitet. Men er det mere end hvad man kan forklare som tilfældigheder? For at kunne give et fornuftigt svar på det må vi præcisere, hvor tilfældigheder skal komme ind i billedet og hvordan de nærmere er beskafne. En sådan præcisering er i en vis forstand netop den statistiske model.

En og anden ville måske foreslå en naturalistisk model gående ud på, at man udvælger 233 personer tilfældigt fra populationen bestående af den danske befolkning og så spørger om sandsynligheden for at få netop 135 M (eller et andet bestemt antal M) og resten K. Men hvad er "den danske befolkning" i den forbindelse? Vi må i hvert fald fraregne personer under 17 år, og der må også ske en vis vægtning fordi det langt overvejende er unge mennesker det drejer sig om, og den viden skal man vel udnytte. Men skal man så også udnytte viden om, at drenge i højere grad end piger får længerevarende uddannelser? og viden om ... Sådan kunne man nok blive ved med at finde på indvendinger.

Vi vil overhovedet ikke bruge en naturalistisk model med 233 personer der udvælges som en stikprøve fra en virkelig population. I stedet må man tænke sig en *"hypotetisk uendelig population"* af SAM-basister årgang 1983; heraf udgør drenge en vis brøkdel p. Vores model skal nu

Tabel 1.9: Absolut og relativ fordeling efter køn for hver basisuddannelse, årgang 1983.

	HUM		SAM		NAT	
	antal	andel	antal	andel	antal	andel
M	36	41%	135	58%	78	63%
K	51	59%	98	42%	45	37%
i alt	87	100%	233	100%	123	100%

gå ud på, at der fra "populationen" vælges 233 personer tilfældigt og uafhængigt af hverandre; hver gang er der chancen p for at personen viser sig at være en dreng[4]. Størrelsen p er en såkaldt *parameter*; dens værdi er ukendt men kan *estimeres* ved $\hat{p}$ = den relative hyppighed = 135/233 = 58%. Det oprindelige spørgsmål, om det forholdsvis store observerede antal drenge kan forklares som en tilfældighed, konkretiserer man nu til: Er observationen 135 drenge ud af 233 noget man må regne med forekommer i praksis når personerne udvælges på den beskrevne måde og når p har værdien 0.51 ? Hvis ja, er der ingen grund til at antage at p ikke rent faktisk skulle have værdien 0.51; hvis nej, tyder observationerne på at p faktisk er forskellig fra 0.51.

Lad os tage endnu et eksempel. Der blev nævnt (side 13) det spørgsmål, om der i 1983 egentlig er den samme andel drenge på hver af de tre basisuddannelser, så at de observerede forskelle ikke er andet end hvad der kan skyldes tilfældigheder. Den talmæssige situation fremgår af Tabel 1.9. Igen vil vi ikke bruge en naturalistisk model; i stedet bruger vi en model der er en nærliggende udvidelse af den forrige: Man må denne gang forestille sig hele tre "hypotetiske uendelige populationer", en for hver af de tre basisuddannelser anno 1983. De tre "populationer" har hver deres egen brøkdel p_H, p_S og p_N af drenge; de tre p-er er ukendte parametre. Fra hver "population" udvælges et bestemt antal personer (hhv. 87, 233 og 123) tilfældigt og uafhængigt af hverandre. I denne "model" skal spørgsmålet, om der egentlig er den samme brøkdel drenge de tre steder, konkretiseres til spørgsmålet om $p_H = p_S = p_N$. Mere udførligt er det nu statistikkens opgave at vurdere, om de foreliggende observationer er noget man må regne med forekommer i praksis når

[4]Det vil fremgå af næste kapitel, at det samlede antal drenge derved bliver binomialfordelt.

Tabel 1.10: Absolut og relativ fordeling efter basisuddannelse for hver af årgangene 1974 og 1982.

	1974		1982	
	antal	andel	antal	andel
HUM	130	31%	158	29%
SAM	218	53%	247	46%
NAT	65	16%	132	25%
i alt	413	100%	537	100%

personerne udvælges på den beskrevne måde og $p_H = p_S = p_N$. I givet fald er der nemlig ikke grund til at antage at parametrene og dermed "populationerne" skulle være forskellige; i modsat fald tyder observationerne på, at de tre p-er og dermed de tre "populationer" faktisk ikke er ens.

Som et sidste eksempel i denne omgang vil vi se på det spørgsmål (fra side 17), om der er sket nogen ændring i styrkeforholdet mellem de tre basisuddannelser når man sammenligner årene 1974 og 1982. De fornødne tal er gengivet i Tabel 1.10. Denne gang skal man tænke sig to "hypotetiske uendelige populationer", en for hver årgang. Hver årgang skal have *sin* basisfordeling: i "populationen" for årgang '74 udgør HUM-er brøkdelen p_{1H}, SAM-er brøkdelen p_{1S} og NAT-er brøkdelen p_{1N} (hvor $p_{1H} + p_{1S} + p_{1N} = 1$), og i "populationen" for årgang '82 er de tilsvarende brøkdele p_{2H}, p_{2S} og p_{2N} (hvor $p_{2H} + p_{2S} + p_{2N} = 1$); alle p-erne er ukendte parametre. Fra hver "population" udvælges nu det fornødne antal personer (hhv. 413 og 537) tilfældigt og uafhængigt af hverandre, og hver gang er der en bestemt chance, givet ved det relevante af p-erne, for at den valgte person viser sig at høre til den og den basisuddannelse. (I dette eksempel er der altså *tre* forskellige klasser som hver person kan tilhøre.) At der ikke er nogen forskel i styrkeforholdet mellem de tre basisuddannelser kan nu formuleres som at

$$p_{1H} = p_{2H} \quad \text{og} \quad p_{1S} = p_{2S} \quad \text{og} \quad p_{1N} = p_{2N} \ ,$$

og opgaven består i at vurdere, om de foreliggende observationer er noget man må regne med forekommer i praksis når personerne udvælges på den beskrevne måde og $p_{1H} = p_{2H}$, $p_{1S} = p_{2S}$ og $p_{1N} = p_{2N}$. I givet fald er der nemlig ikke grund til at antage at "populationerne" skulle

være forskellige, dvs. at der skulle være forskel på styrkeforholdene de to år; i modsat fald tyder observationerne på at der faktisk er en forskel.

Afrunding

Afslutningsvis vil vi fremdrage nogle generelle træk ved statistiske modeller som man kan få øje på også i de just præsenterede "næsten-modeller". Udgangspunktet er et *statistisk problem*: man skal på baggrund af et talmateriale udtale sig om bestemte spørgsmål. Problemets strukturer og begreber udmøntes i forskellige modelingredienser:

1. Visse størrelser der anses for at være faste, givne konstanter.

2. Visse størrelser der anses for fremkommet ved et nærmere angivet tilfældigt valg. Disse størrelser kaldes gerne *observationer*.

3. Nogle "hypotetiske uendelige populationer" hvorfra observationerne er udvalgt tilfældigt.

4. Nogle principielt ukendte størrelser, *parametre*, der beskriver karakteristika ved de "hypotetiske uendelige populationer". Parametrene kan i visse situationer fortolkes som beskrivende forhold i virkeligheden.

Lad os som eksempel se på det senest omtalte spørgsmål, hvorvidt styrkeforholdet mellem de tre basisuddannelser er det samme i de to år 1974 og 1982. Her vil man anse de totale antal studerende hvert år (dvs. 413 og 537) for faste eller givne, fordi vi af de to tal i sig selv ikke kan uddrage information om styrkeforholdene. Det interessante er ikke at der kom 413 basister i 1974, men at de 413 fordelte sig med 130, 218 og 65. Nu vil/kan vi ikke beskrive de mekanismer og årsager osv. der faktisk har afstedkommet denne fordeling. Tværtimod sætter vi en (temmelig høj) grænse for beskrivelsesniveauet og siger, at hvad der ligger derunder *anses for* tilfældigt. Vi opfatter således hver af de 413 personer som tilfældigt udvalgt fra en "hypotetisk uendelig population" som består af uendelig mange "individer", hvoraf en vis brøkdel er HUM-er, en vis brøkdel SAM-er og resten NAT-er; disse brøkdele er den "hypotetiske uendelige populations" parametre. Derved anses det altså for tilfældigt at de 413 fordelte sig med netop 130, 218 og 65. —

På denne måde går vi ud fra, at vi ville kunne beskrive fænomenet fuldstændigt *på det valgte beskrivelsesniveau*, hvis blot vi kendte værdierne af de nævnte parametre.

Der gøres ganske tilsvarende overvejelser hvad årgang 1982 angår. De to årgange får hver sin "hypotetiske uendelige population" med hver sit sæt ukendte parametre. — Inden for disse rammer er det nu meningsfuldt at spørge, om det er rimeligt at mene at de to sæt ukendte parametre kan antages at være ens. I det omfang parametersættet hørende til en bestemt årgang kan siges at udtrykke styrkeforholdet mellem de tre basisuddannelser i den pågældende årgang, er det oprindelige spørgsmål nu fortolket ind i modellens rammer.

Vi ser, at den statistiske model *ikke* foregiver at give en beskrivelse af *faktiske* tilfældighedsmekanismer. Derimod påtager den statistiske model sig at beskrive, hvilke andre observationer end de faktisk foreliggende man også kunne have fået og med hvilke sandsynligheder, *når* man opererer med dét bestemte beskrivelsesniveau og siger at hvad der ligger derunder anses for tilfældigt.

Kapitel 2

Binomialfordelingen

I Kapitel 1 blev bl.a. fremsat det spørgsmål (side 13), om observationen "135 drenge og 98 piger på SAM-BAS i 1983" er forenelig med antagelsen om, at drenge og piger rekrutteres i forholdet 51 : 49, dvs. at den teoretiske brøkdel af drenge er 0.51. Den observerede brøkdel af drenge er $135/233 = 0.58$ hvilket jo ikke er 0.51, så når spørgsmålet overhovedet stilles, er det fordi der er en underliggende antagelse om, at det faktiske udfald (135 ud af 233) i et eller andet omfang har været bestemt af tilfældigheder og derfor sådan set lige så godt kunne være blevet noget andet. Vi vil nu søge at opstille en *statistisk model* der siger disse ting mere præcist. Det er (som allerede sagt i Kapitel 1) ikke meningen at det skal være en naturalistisk model der så vidt muligt inddrager alle tænkelige relevante faktorer (– der ville måske til sidst slet ikke være plads til/brug for tilfældigheder i dét projekt!). Tværtimod benytter vi en model der er helt uden ambitioner hvad angår detaljeringsgrad: Vi antager, at det ligger fast at der er netop 233 personer i alt, og at hver af de 233 er udvalgt tilfældigt (og uafhængigt af de andre) fra en "hypotetisk uendelig population"[1]. I denne "population" er brøkdelen p af personerne drenge og dermed brøkdelen $1 - p$ piger. Vi kender ikke værdien af p (fordi vi kender ikke hele den "hypotetiske uendelige population"), men på grundlag af de foreliggende data kan vi udregne et *skøn* over (et *estimat* over) værdien af p, nemlig $\widehat{p} = 135/233 = 0.58$. Det oprindelige spørgsmål kan nu formuleres som: er den foreliggende observation forenelig med en antagelse om at

[1] Man kan eventuelt tænke på "populationen" som en mængde af potentielle SAM-basister 1983.

parameteren p har værdien 0.51? I den statistiske jargon vil man sige, at vi skal *teste hypotesen* H_0 : $p = 0.51$. - Løsningen af dén opgave må vente til Kapitel 3. Her vil vi gå i gang med at formulere modellen i matematiksprog; vi vil i den forbindelse benytte forskellige begreber og resultater fra hæftet om grundbegreber i sandsynlighedsregningen.

Den simple binomialfordelingsmodel

Vores model går ud på, at der vælges $n = 233$ personer tilfældigt fra en "hypotetisk uendelig population" hvor brøkdelen p er drenge. Det betyder, at hver gang der vælges en person, er der sandsynligheden p for at det bliver en dreng og sandsynligheden $1 - p$ for at det bliver en pige. Vi indfører en *indikatorvariabel* I for hver person:

$$I_j = \begin{cases} 1 & \text{hvis person nr. } j \text{ er en dreng} \\ 0 & \text{hvis person nr. } j \text{ er en pige} \end{cases}, \quad j = 1, 2, \ldots, n .$$

Dermed er $Y = I_1 + I_2 + \ldots + I_n$ det samlede antal drenge i stikprøven på n personer. I vort eksempel kender vi ikke de enkelte I_j-er men kun Y, som har værdien $y = 135$; det gør imidlertid ikke noget her, hvor øvelsen går ud på at bestemme, hvilke (andre) værdier Y kan antage og med hvilke sandsynligheder.

I_j-erne er stokastiske variable om hvilke det antages at

1. $I_1, I_2, \ldots, I_n$ er stokastisk uafhængige.

2. $$\begin{cases} \mathrm{P}(I_j = 1) = p \\ \mathrm{P}(I_j = 0) = 1 - p \end{cases} \quad \text{for ethvert } j.$$

(Parameteren p er den ukendte brøkdel drenge i "populationen".) De to formler i 2. kan med fordel samles til én formel, der så angiver sandsynlighedsfunktionen for I_j:

$$\mathrm{P}(I_j = x) = p^x(1-p)^{1-x}, \; x = 0, 1 .$$

Da I_j-erne er stokastisk uafhængige, er den simultane sandsynlighedsfunktion f for $I_1, I_2, \ldots, I_n$ givet som

$$f(x_1, x_2, \ldots, x_n)$$

$$
\begin{aligned}
&= \mathrm{P}(I_1 = x_1) \times \mathrm{P}(I_2 = x_2) \times \ldots \times \mathrm{P}(I_n = x_n) \\
&= p^{x_1}(1-p)^{1-x_1} \times p^{x_2}(1-p)^{1-x_2} \times \ldots \times p^{x_n}(1-p)^{1-x_n} \\
&= p^{x_1+x_2+\ldots+x_n}\,(1-p)^{n-(x_1+x_2+\ldots+x_n)}
\end{aligned}
$$

når $(x_1, x_2, \ldots, x_n)$ er et talsæt bestående af 0-er og 1-er. Det ses at hvis $(x_1, x_2, \ldots, x_n)$ er et talsæt bestående af k 1-er og $n-k$ 0-er, så er

$$f(x_1, x_2, \ldots, x_n) \;=\; p^k(1-p)^{n-k}\,.$$

Da vi nu kender den simultane sandsynlighedsfunktion for I_j-erne, kan vi bestemme[2] sandsynlighedsfunktionen for $Y = I_1 + I_2 + \ldots + I_n$. Man får at

$$\mathrm{P}(Y = k) \;=\; \sum f(x_1, x_2, \ldots, x_n),$$

hvor der summeres over alle talsæt $(x_1, x_2, \ldots, x_n)$ bestående af 0-er og 1-er og for hvilke $x_1 + x_2 + \ldots + x_n = k$, dvs. alle talsæt $(x_1, x_2, \ldots, x_n)$ bestående af k 1-er og $n-k$ 0-er. Som vi netop er nået frem til, har ethvert af disse talsæt sandsynlighed $p^k(1-p)^{n-k}$, så derfor bliver

$$\mathrm{P}(Y = k) \;=\; A \times p^k(1-p)^{n-k}\,,$$

hvor A står for antallet af forskellige talsæt $(x_1, x_2, \ldots, x_n)$ bestående af k 1-er og $n-k$ 0-er. Antallet A afhænger af værdierne af n og k, og man plejer at betegne dette antal $\binom{n}{k}$ (udtales "n over k"). $\binom{n}{k}$ kaldes en *binomialkoefficient*. Sandsynlighedsfunktionen for Y får dermed udseendet

$$\mathrm{P}(Y = k) \;=\; \binom{n}{k} p^k(1-p)^{n-k},\; k = 0, 1, 2, \ldots, n. \qquad (2.1)$$

Den sandsynlighedsfordeling for Y som er fastlagt på denne måde hedder *binomialfordelingen med sandsynlighedsparameter p og antalsparameter n*, og man siger at Y er *binomialfordelt* med parametre n og p. – Antalsparameteren n er et kendt heltal, og sandsynlighedsparameteren p, som typisk er ukendt, er et tal mellem 0 og 1.

[2]Det generelle problem: at bestemme fordelingen af en sum af stokastiske variable, er behandlet i hæftet om grundbegreber i sandsynlighedsregningen.

Stokastiske variable der som I_j-erne kun kan antage værdierne 0 og 1, kaldes undertiden for *01-variable*. Med denne sprogbrug kan vi udtrykke hovedpointen i det hidtil udledte på den måde, at *hvis Y er en sum af et bestemt antal uafhængige identisk fordelte 01-variable, så er Y binomialfordelt.*

Den statistiske model for tallene vedrørende drenge og piger på SAM-BAS 1983 kan nu formuleres kort således:

> Observationen $y = 135$ drenge er en observeret værdi af en stokastisk variable Y som er binomialfordelt med antalsparameter $n = 233$ og ukendt sandsynlighedsparameter p $(0 \leq p \leq 1)$.

Inden for rammerne af denne model ønsker vi at teste den statistiske hypotese $H_0: \ p = 0.51$.

Før vi kan give os i kast med statistisk analyse af binomialfordelte observationer er det nødvendigt at lære forskelligt om binomialfordelingen og om binomialkoefficienter.

Binomialkoefficienter

Binomialkoefficienter vil vi definere på følgende måde:

Definition:
Binomialkoefficienten $\binom{n}{k}$ er antallet af forskellige måder hvorpå man kan placere to forskellige slags symboler (1 og 0) på n pladser, således at det første symbol (1) kommer på k af pladserne og det andet symbol (0) kommer på de resterende $n - k$ pladser.

Det følger uden videre af denne definition, at antallet af forskellige talsæt $(x_1, x_2, \ldots, x_n)$ bestående af k 1-er og $n - k$ 0-er netop er $\binom{n}{k}$, således som det blev påstået på side 27.

Ud fra definitionen kan man i princippet bestemme talværdier af enhver binomialkoefficient ved simpel optælling. Eksempelvis er $\binom{4}{3}$ lig

med fire, fordi der er de fire placeringer $(1,1,1,0)$, $(1,1,0,1)$, $(1,0,1,1)$ og $(0,1,1,1)$ af tre 1-er og et 0 på de fire pladser (, , ,).

I definitionen af $\binom{n}{k}$ skal man placere k 1-er og $n-k$ 0-er. Hvis man i en sådan placering kalder 1-erne for 0 og 0-erne for 1, så har vi i stedet en placering af $n-k$ 1-er og k 0-er. Heraf følger en nyttig relation:

$$\binom{n}{k} = \binom{n}{n-k}, \quad \begin{array}{rcl} n & = & 0,1,2,\ldots \\ k & = & 0,1,\ldots,n \end{array} \tag{2.2}$$

Det bliver let en uoverkommelig opgave at bestemme talværdier af binomialkoefficienter ved at tælle antal placeringer - tænk hvis man f.eks. skulle bestemme $\binom{37}{15}$ på den måde! Der kan derfor være grund til at vie beregningen af binomialkoefficienter lidt opmærksomhed.

Fra definitionen og fra (2.2) får man uden videre[3]

$$\binom{n}{0} = 1 \text{ og dermed } \binom{n}{n} = 1, \text{ for } n = 0,1,2,\ldots$$

$$\binom{n}{1} = n \text{ og dermed } \binom{n}{n-1} = n, \text{ for } n = 1,2,3,\ldots$$

Der findes en formel der fortæller hvordan man kan beregne en "svær" binomialkoefficient som en sum af to lidt mindre "svære" binomialkoefficienter. Her er først et eksempel der illustrerer ræsonnementet bag formlen: Vi søger værdien af den "svære" binomialkoefficient $\binom{5}{2}$, der jo er antallet af placeringer af to 1-er og tre 0-er på fem pladser. Vi

[3] At $\binom{0}{0}$ skal være 1 er nok til dels en konvention.

skriver alle disse placeringer op på en smart måde:

$\binom{5}{2}$ placeringer af to 1-er og tre 0-er.

- $\binom{4}{2}$ placeringer af to 1-er og to 0-er:
  - 0 0011
  - 0 0101
  - 0 0110
  - 0 1001
  - 0 1010
  - 0 1100
- $\binom{4}{1}$ placeringer af en 1-er og tre 0-er:
  - 1 0001
  - 1 0010
  - 1 0100
  - 1 1000

Det ses at $\binom{5}{2}$ kan findes som $\binom{4}{2} + \binom{4}{1}$, nemlig som "antal placeringer med et 0 på førstepladsen" + "antal placeringer med et 1 på førstepladsen". Man kan uden videre generalisere ræsonnementet, og man får derved den vigtige *rekursionsformel*

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}, \quad \begin{array}{l} k = 1,2,\ldots,n \\ n = 1,2,3,\ldots \end{array} \tag{2.3}$$

Vi kan nu bestemme talværdien af $\binom{5}{2}$: ifølge (2.3) er $\binom{5}{2} = \binom{4}{2} + \binom{4}{1}$, så hvis vi kender talværdierne af $\binom{4}{2}$ og $\binom{4}{1}$, kan vi løse opgaven. Men $\binom{4}{1}$ er 4 (fordi generelt er $\binom{n}{1}$ lig n), så det er kun $\binom{4}{2}$ der er et problem. Vi benytter så rekursionsformlen (2.3) en gang til: $\binom{4}{2} = \binom{3}{2} + \binom{3}{1}$; her er $\binom{3}{1}$ lig 3, og $\binom{3}{2}$ er også 3 (fordi $\binom{n}{n-1} = n$). Altså er $\binom{4}{2} = 3 + 3 = 6$, og dermed $\binom{5}{2} = \binom{4}{2} + \binom{4}{1} = 6 + 4 = 10$ — hvad man jo også kan se ved simpel optælling.

Figur 2.1: Pascals trekant.

n	binomialkoefficienterne $\binom{n}{k}$													
0						1								
1					1		1							
2				1		2		1						
3			1		3		3		1					
4		1		4		6		4		1				
5	1		5		10		10		5		1			
6		1		6		15		20		15		6		1
7	1		7		21		35		35		21		7		1
⋮						⋮								

Pascals trekant

Rekursionsformlen (2.3) er i og for sig ikke særlig velegnet når man ønsker at beregne en enkelt binomialkoefficient, men den er overordentlig praktisk når man ønsker at beregne alle binomialkoefficienter op til en eller anden øvre grænse for n. Vi kender på forhånd binomialkoefficienterne med $n = 0$ og $n = 1$ (de er $\binom{0}{0} = 1, \binom{1}{0} = \binom{1}{1} = 1$). Ved hjælp af formlen kan vi beregne alle koefficienter med $n = 2$, derefter alle med $n = 3$, derefter alle med $n = 4$, osv. Man plejer at stille resultaterne op i et skema der kaldes *Pascals trekant*, se Figur 2.1. Det ses at f.eks. er $\binom{7}{2}$ lig 21. Hvert tal i Pascals trekant fremkommer, ifølge (2.3), som summen af de to nærmeste tal i rækken lige ovenover, f.eks. er 21=6+15.

Ved brug af Pascals trekant vil det være muligt at bestemme talværdier af enhver binomialkoefficient; man skulle dog udføre en hel del additioner og have et temmelig stort stykke papir for at udregne f.eks. $\binom{37}{15}$. Heldigvis findes der også en anden og mindre pladskrævende metode, hvor man så til gengæld skal lave nogle multiplikationer og divisioner. Som forberedelse til denne metode skal vi bruge endnu en formel for binomialkoefficienter.

Lad os først se på et taleksempel. På fem pladser vil vi placere nogle symboler, men denne gang skal det være *tre* slags symboler, nemlig 0, 1 og 2. Der skal placeres tre 0-er, et 1 og et 2. Vi ønsker at bestemme antallet af forskellige placeringer. Det kan vi gøre på to måder. **Enten** kan vi starte med at placere tre 0-er og to "ikke-0"-er, hvilket kan ske på $\binom{5}{2}$ måder, og hver gang vi har en sådan placering, kan de to "ikke-0"-er på to måder ændres til et 1 og et 2:

$$\begin{aligned}
000xx &\rightarrow \begin{matrix}00012\\00021\end{matrix}\\
00x0x &\rightarrow \begin{matrix}00102\\00201\end{matrix}\\
00xx0 &\rightarrow \begin{matrix}00120\\00210\end{matrix}\\
0x00x &\rightarrow \begin{matrix}01002\\02001\end{matrix}\\
0x0x0 &\rightarrow \begin{matrix}01020\\02010\end{matrix}\\
0xx00 &\rightarrow \begin{matrix}01200\\02100\end{matrix}\\
x000x &\rightarrow \begin{matrix}10002\\20001\end{matrix}\\
x00x0 &\rightarrow \begin{matrix}10020\\20010\end{matrix}\\
x0x00 &\rightarrow \begin{matrix}10200\\20100\end{matrix}\\
xx000 &\rightarrow \begin{matrix}12000\\21000\end{matrix}
\end{aligned}$$

i alt $2 \times \binom{5}{2}$ forskellige placeringer. **Eller også** kan vi starte med at placere et 1 og fire "ikke-1"-er, hvilket kan ske på $\binom{5}{4}$ måder, og hver gang vi har en sådan placering, kan de fire "ikke-1-"-er på 4 måder

ændres til tre 0 og et 2:

$$\begin{array}{rcl} xxxx1 & \rightarrow & \begin{array}{l}00021\\00201\\02001\\20001\end{array} \\ xxx1x & \rightarrow & \begin{array}{l}00012\\00210\\02010\\20010\end{array} \\ xx1xx & \rightarrow & \begin{array}{l}00102\\00120\\02100\\20100\end{array} \\ x1xxx & \rightarrow & \begin{array}{l}01002\\01020\\01200\\21000\end{array} \\ 1xxxx & \rightarrow & \begin{array}{l}10002\\10020\\10200\\12000\end{array} \end{array},$$

i alt $4 \times \binom{5}{4}$ placeringer. Heraf kan man se, at $2 \times \binom{5}{2} = 4 \times \binom{5}{4}$.

Ræsonnementet kan uden videre generaliseres til den situation der handler om at placere $k-1$ 1-er, et 2 og $n-k$ 0-er på n pladser. Hvis man først fordeler efter 0 eller ikke-0, og derefter deler "ikke-0"-erne op i 1 og 2, så er der $k \times \binom{n}{k}$ forskellige placeringer. Hvis man derimod først deler op efter 1 eller ikke-1, og derefter deler "ikke-1"-erne op i 0 og 2, så er der $(n-k+1) \times \binom{n}{k-1}$ forskellige placeringer. Deraf følger at

$$k \times \binom{n}{k} = (n-k+1) \times \binom{n}{k-1},$$

og ved at flytte rundt på faktorerne fås

$$\binom{n}{k} = \frac{n-k+1}{k} \times \binom{n}{k-1}, \quad \begin{array}{l} k = 1,2,\ldots,n \\ n = 1,2,\ldots \end{array}$$

Denne rekursionsformel fortæller hvordan man finder $\binom{n}{k}$ hvis man kender $\binom{n}{k-1}$.

Ved gentagne anvendelser af rekursionsformlen fås i øvrigt

$$\begin{aligned} \binom{n}{k} &= \frac{n-k+1}{k} \times \binom{n}{k-1} \\ &= \frac{n-k+1}{k} \times \frac{n-k+2}{k-1} \times \binom{n}{k-2} \\ &= \frac{n-k+1}{k} \times \frac{n-k+2}{k-1} \times \frac{n-k+3}{k-2} \times \binom{n}{k-3} \\ &= \ldots \\ &= \frac{n-k+1}{k} \times \frac{n-k+2}{k-1} \times \ldots \times \frac{n-2}{3} \times \frac{n-1}{2} \times \frac{n}{1}, \end{aligned}$$

dvs.

$$\binom{n}{k} = \frac{n}{1} \times \frac{n-1}{2} \times \frac{n-2}{3} \times \ldots \times \frac{n-k+1}{n}, \quad k = 0,1,2,\ldots . \tag{2.4}$$

(Hvis k er 0, er højresiden "det tomme produkt", som er 1.)

Ved hjælp af denne formel (eller algoritmen, hvilket kommer ud på det samme) kan man med papir og blyant og lommeregner let finde at $\binom{37}{15} = 9\,364\,199\,760$.

Binomialformlen

Hvorfor hedder det en binomialkoefficient?

Et *bi*-nomium er en *to*-leddet størrelse som f.eks. $a + b$. En formodentlig velkendt formel fortæller hvad kvadratet på en toleddet størrelse er:

$$(a+b)^2 = a^2 + 2ab + b^2.$$

Denne formel kan generaliseres til at handle om n-te potensen af en toleddet størrelse. Hvis man i

$$(a+b)^n = \underbrace{(a+b)(a+b)\ldots(a+b)}_{n}$$

ganger parenteserne ud, får man 2^n led der hver især er et produkt af n faktorer, én fra hvert af de n binomier. Af disse 2^n led er der netop $\binom{n}{k}$ der består af k a-er og $n-k$ b-er. Følgelig er

$$\begin{aligned}(a+b)^n &= \binom{n}{0}a^0b^n + \binom{n}{1}a^1b^{n-1} + \binom{n}{2}a^2b^{n-2} + \ldots + \binom{n}{n}a^nb^0 \\ &= \sum_{k=0}^{n}\binom{n}{k}a^kb^{n-k}.\end{aligned}$$

Denne formel hedder *binomialformlen* fordi den handler om n-te potensen af et binomium. De koefficienter der indgår i binomialformlen må selvfølgelig hedde binomialkoefficienter.

Binomialfordelingen

Definition:
Binomialfordelingen med sandsynlighedsparameter p og antalsparameter n er den diskrete sandsynlighedsfordeling som er givet ved sandsynlighedsfunktionen

$$f(y) = \binom{n}{y}p^y(1-p)^{n-y},\ y = 0,1,2,\ldots,n.$$

Her er p et (som oftest ukendt) tal mellem 0 og 1, og n er et positivt heltal.

Som udledt tidligere i dette kapitel (side 28) kan en binomialfordelt stokastisk variabel Y fremkomme som en sum af uafhængige identisk fordelte 01-variable. Denne kendsgerning gør det uhyre let at bestemme middelværdi og varians af en binomialfordelt stokastisk variabel: Antag at $I_1, I_2, \ldots, I_n$ er uafhængige 01-variable med $\mathrm{P}(I_j = 1) = p$ for alle j. Den stokastiske variabel $Y = I_1 + I_2 + \ldots + I_n$ er da binomialfordelt med parametre n og p. Ifølge regnereglerne for middelværdi og varians[4] har Y middelværdi

$$\begin{aligned} \mathrm{E}Y &= \mathrm{E}\,I_1 + \mathrm{E}\,I_2 + \ldots + \mathrm{E}\,I_n \\ &= n\,\mathrm{E}\,I_1 \end{aligned}$$

og varians

$$\begin{aligned} \mathrm{Var}\,Y &= \mathrm{Var}\,I_1 + \mathrm{Var}\,I_2 + \ldots + \mathrm{Var}\,I_n \\ &= n\,\mathrm{Var}\,I_1\,, \end{aligned}$$

så problemet er nu reduceret til at bestemme middelværdi og varians af I_1. Men

$$\begin{aligned} \mathrm{E}\,I_1 &= 0 \times \mathrm{P}(I_1 = 0) + 1 \times \mathrm{P}(I_1 = 1) \\ &= 0 \times (1-p) + 1 \times p \\ &= p\,, \end{aligned}$$

og

$$\begin{aligned} \mathrm{Var}\,I_1 &= \mathrm{E}(I_1^2) - (\mathrm{E}\,I_1)^2 \\ &= \mathrm{E}\,I_1 - (\mathrm{E}\,I_1)^2 \\ &= p - p^2 \\ &= p(1-p)\,, \end{aligned}$$

[4] **M3** og **V8** i sandsynlighedsregningshæftet

så alt i alt har vi

Hvis den stokastiske variabel Y er binomialfordelt med parametre n og p, så er

$$\begin{aligned} \mathrm{E}\,Y &= np \\ \mathrm{Var}\,Y &= np(1-p)\,. \end{aligned}$$

Hvis man skal udregne binomialsandsynlighederne

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}$$

for $y = 0, 1, 2, \ldots, n$, er det som regel ikke hensigtsmæssigt bare uden videre at indsætte i formlen. Man kan med fordel benytte en rekursionsformel. Ved simple omskrivninger finder man

$$\frac{f(y)}{f(y-1)} = \frac{n-y+1}{y} \times \frac{p}{1-p}\,, \quad y = 1, 2, \ldots, n\,,$$

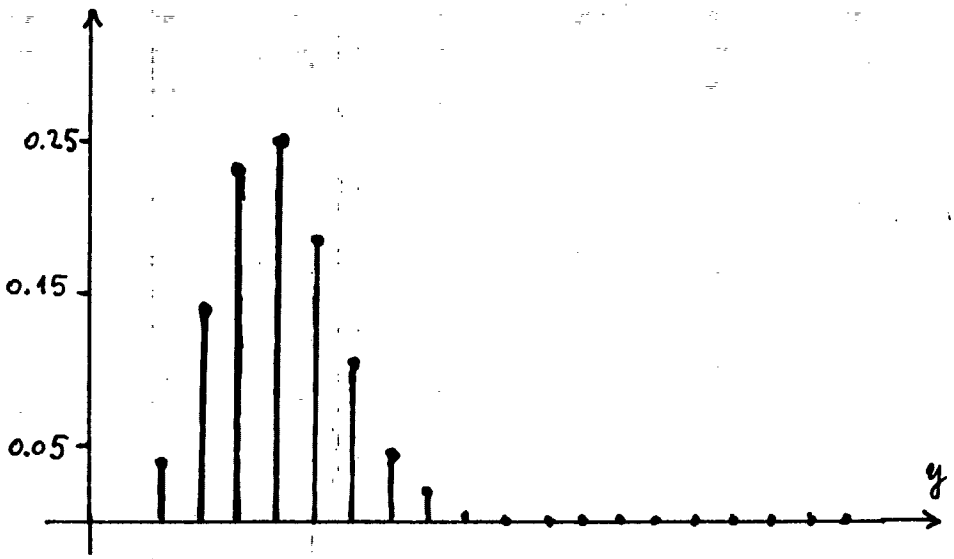
således at $f(y)$ let kan beregnes ud fra $f(y-1)$. Metoden bliver dermed

$$\begin{aligned} f(0) &= (1-p)^n \\ f(y) &= f(y-1) \times \frac{n-y+1}{y}\,\frac{p}{1-p}\,, \quad y = 1, 2, \ldots, n\,. \end{aligned}$$

Eksempel 2.1. ***En binomialfordeling***

Som eksempel vil vi beregne og tegne sandsynlighedsfunktionen for binomialfordelingen med $n = 18$ og $p = 1/6$. (Denne fordeling kunne f.eks. beskrive antallet af seksere ved 18 kast med en almindelig terning.) Fordelingen har i øvrigt middelværdi $18 \times \frac{1}{6} = 3$

Figur 2.2: Sandsynlighedsfunktionen $f(y)$ for binomialfordelingen med $n = 18$ og $p = 1/6$.



og varians $18 \times \frac{1}{6} \times \frac{5}{6} = 2.5$ *(svarende til standardafvigelsen 1.58).*

Man finder ved den beskrevne metode

y	$f(y) = \binom{18}{y} (\frac{1}{6})^y (\frac{5}{6})^{18-y}$
0	0.038
1	0.135
2	0.230
3	0.245
4	0.184
5	0.103
6	0.045
7	0.015
8	0.004
9	0.001
10	0.000
11	0.000
12	0.000
13	0.000
14	0.000
15	0.000
16	0.000
17	0.000
18	0.000
	1.000

Denne sandsynlighedsfunktion er vist i Figur 2.2. □

Kapitel 3

Statistisk analyse af den simple binomialfordelingsmodel

Dette kapitel introducerer nogle grundlæggende principper for hvordan man analyserer en statistisk model, og vi gennemgår det simpleste eksempel på statistisk analyse af binomialfordelte observationer.

Estimation af p

Det er i visse simple tilfælde ret klart, hvordan man "selvfølgelig" skal analysere sin statistiske model, idet der er en "umiddelbart indlysende" fremgangsmåde osv. I andre tilfælde (de fleste) er det knap så klart. Vi skal i dette kapitel introducere et sæt overordnede principper for, hvordan man bør analysere en statistisk model. Principperne (med visse tilføjelser) gælder for "enhver" model. Indførelsen af principperne betyder *ikke* at man slipper for overvejelser over hvad man "selvfølgelig" skal gøre og hvad der er "umiddelbart indlysende", *men* at man i stedet for at skulle gøre overvejelserne igen og igen i hvert enkelt tilfælde så at sige overstår dem alle på én gang ved at hæve dem fra enkelttilfældene op til et overordnet niveau, et meta-plan, hvor de udnævnes til generelle principper. — Et *princip* er i denne sammenhæng en norm, en retningslinie, som ikke bliver logisk-deduktivt bevist, men som retfærdiggøres dels gennem generelle betragtninger og overvejelser, dels ved at levere rimelige resultater i konkrete situationer.

Som gennemgående eksempel til at ledsage præsentationen af principperne bruger vi eksemplet fra begyndelsen af Kapitel 2; det handler om, at der i 1983 begyndte 135 drenge og 98 piger på SAM-BAS, og spørgsmålet er, om den faktiske overrepræsentation af drenge kan tilskrives tilfældigheder. Den statistiske model for dette talmateriale skulle være, at $y = 135$ opfattes som en observation af en stokastisk variabel Y der er binomialfordelt med antalsparameter $n = 135 + 98 = 233$ og ukendt sandsynlighedsparameter p ($0 \leq p \leq 1$). Inden for rammerne af denne model ønsker vi at teste den statistiske hypotese $H_0 : p = 0.51$.

Sandsynlighedsfunktionen for Y er

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y} \, , y = 0, 1, 2, \ldots, n \, .$$

For at fremhæve at udtrykket afhænger både af y og af p udskifter vi betegnelsen "$f(y)$" med "$f(y;p)$", dvs.

$$f(y;p) = \binom{n}{y} p^y (1-p)^{n-y} \, , \quad 0 \leq p \leq 1, \; y = 0, 1, 2, \ldots, n \, .$$

Funktionen f er nu en funktion af *to* variable, en observationsvariabel y og en parametervariabel p. Denne funktion kaldes *modelfunktionen* for den statistiske model, idet den fuldstændig specificerer modellen: for enhver kombination af en mulig observation y og en mulig parameterværdi p angiver den sandsynligheden for at observere netop det y hvis netop det p er den rigtige parameterværdi.

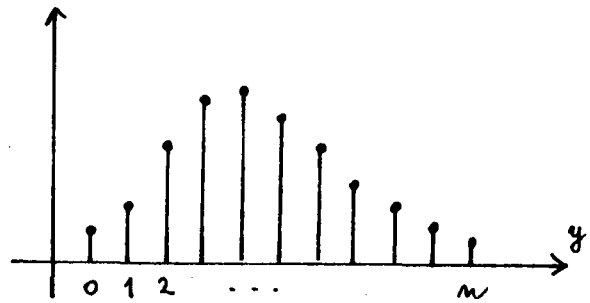
- Hvis vi i modelfunktionen fixerer p og kun opfatter funktionen som en funktion af y, så har vi *sandsynlighedsfunktionen* svarende til parameterværdien p. Figur 3.1 viser en "typisk" sandsynlighedsfunktion.

- Hvis vi i modelfunktionen fixerer y og kun opfatter funktionen som en funktion af p, så har vi en ny funktion som vi ikke tidligere har mødt. Denne nye funktion hedder *likelihoodfunktionen* svarende til observationen y og betegnes $L(\cdot)$ eller $L(\cdot\,; y)$:

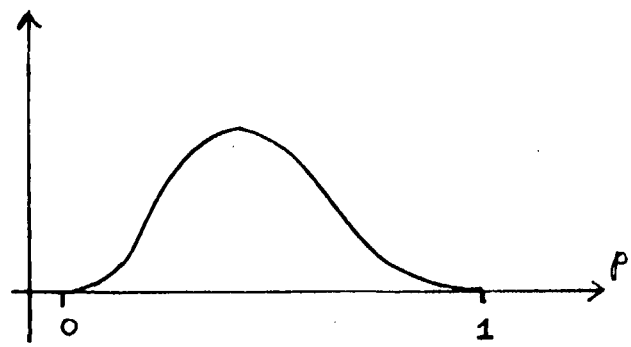
$$L(p) = L(p;y) = \binom{n}{y} p^y (1-p)^{n-y}, \; 0 \leq p \leq 1.$$

Figur 3.2 viser en "typisk" likelihoodfunktion.

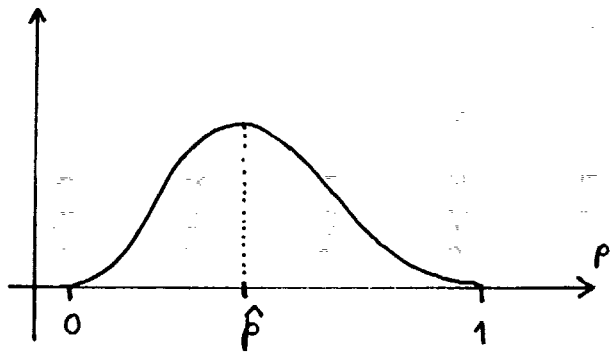
Figur 3.1: En sandsynlighedsfunktion $f(\,\cdot\,;p)$.



Figur 3.2: En likelihoodfunktion $L(\,\cdot\,;y) = f(y;\,\cdot\,)$.



Figur 3.3: En likelihoodfunktion $L(p)$, samt $\hat{p}$.



I vort eksempel er

$$L(p) = L(p;135) = \binom{233}{135} p^{135}(1-p)^{98}\ ,\ 0 \leq p \leq 1.$$

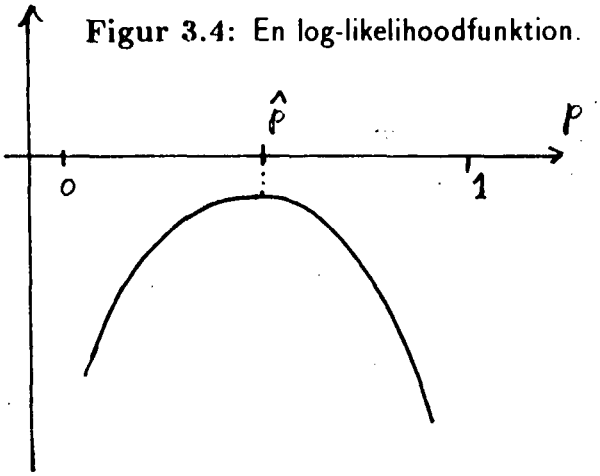
$L(p)$ er sandsynligheden for at observere det man faktisk har observeret, forudsat at den ukendte parameter har værdien p. Likelihoodfunktionen kan derfor anvendes til at sammenligne forskellige p-værdiers evne til at beskrive den faktiske observation y. For hvis f.eks. $L(p_1) < L(p_2)$, så er chancen for at observere netop dette y større når p er lig p_2 end når p er lig p_1, og det må betyde at p_2 giver en bedre beskrivelse af data end p_1 gør. Den p-værdi som giver den bedste beskrivelse efter disse retningslinier er den p-værdi som maksimaliserer likelihoodfunktionen. Denne p-værdi hedder *maksimaliseringsestimatet* (eller *maximum likelihood estimatet*) for p og betegnes $\hat{p}$ ("p hat"). $\hat{p}$ er altså bestemt ved at

$$L(\hat{p};y) \geq L(p;y) \text{ for alle } p.$$

Bemærk at $\hat{p}$ er en funktion af y.

Af bekvemmelighedsgrunde opererer man tit med *"log-likelihood-funktionen"*, dvs. funktionen $\ln L(p)$, og man bestemmer $\hat{p}$ som maksimumspunktet for $\ln L$ (resultatet bliver jo det samme). I vort eksempel er log-likelihoodfunktionen

$$\ln L(p) = \ln \binom{233}{135} + 135 \ln p + 98 \ln(1-p)\ .$$



Figur 3.4: En log-likelihoodfunktion.

Imidlertid vil talværdierne let gøre ræsonnementerne ugennemskuelige, så vi vender tilbage til den generelle binomialfordelingsmodel, hvor log-likelihoodfunktionen er

$$\ln L(p) = \ln \binom{n}{y} + y \ln p + (n-y) \ln(1-p) . \qquad (3.1)$$

Hvad er $\hat{p}$ i denne model? Svaret herpå får vi ved at løse den matematikopgave der hedder:

> Bestem maksimumspunkt(er) for funktionen $p \mapsto \ln L(p)$ når $0 \leq p \leq 1$.

(Denne funktion ser "typisk" ud som skitseret i Figur 3.4.) Fra matematikken ved vi, at kandidater til maksimumspunkter er dels intervalendepunkterne $p = 0$ og $p = 1$, dels de stationære punkter, dvs. punkter hvor $\frac{d}{dp} \ln L(p) = 0$. Nu er

$$\begin{aligned} \frac{d}{dp} \ln L(p) &= \frac{y}{p} - \frac{n-y}{1-p} \\ &= \frac{y - np}{p(1-p)} \end{aligned}$$

for $0 < p < 1$, og det betyder at hvis $0 < y < n$, så er punktet $p = y/n$ et stationært punkt for $\ln L(p)$. Hvis stadig $0 < y < n$, så er endepunkterne $p = 0$ og $p = 1$ i hvert fald ikke maksimumspunkter, for

dér er $\ln L$ lig $-\infty$. Altså er $\hat{p} = \hat{p}(y) = y/n$, foreløbig når $0 < y < n$. Hvis y er 0 er $\ln L(p) = n\ln(1-p)$; dette er et ikke-positivt tal, og det er lig 0 netop når $p = 0$, så derfor er $\hat{p} = 0 = y/n$ også i dette tilfælde. Hvis y er 1 er på tilsvarende måde $\hat{p} = 1 = y/n$. Vi er hermed nået frem til at

> I binomialmodellen med modelfunktion
>
> $$f(y;p) = \binom{n}{y} p^y (1-p)^{n-y}, \quad \begin{matrix} y = 0,1,2,\ldots,n \\ 0 \le p \le 1 \end{matrix},$$
>
> er maksimaliseringsestimatet $\hat{p}$ for p givet ved $\hat{p} = y/n$.

At p skal estimeres ved den relative hyppighed y/n kan næppe overraske nogen, det er næsten hvad man kan sige sig selv. Derimod kan det måske være en smule overraskende, at svaret $\hat{p} = y/n$ altså faktisk også er det svar man når frem til ved at benytte den generelle fremgangsmåde der lyder

- opstil modelfunktionen,
- dan derudfra likelihoodfunktionen,
- bestem $\hat{p}$ som maksimumspunktet for likelihoodfunktionen.

Det er vigtigt at have in mente, at der tænkes at være en bestemt sand parameterværdi, som er et bestemt, ukendt tal, et træk ved en "hypotetisk uendelig population". Vi kan principielt aldrig erfare den sande parameterværdi, men ud fra foreliggende observationer kan vi estimere den.

Maksimaliseringsestimatet $\hat{p} = y/n$ er det bedste bud vi kan give på den ukendte p-værdi, når vi har observeret antallet y ud af n. Den statistiske model fortæller, at y er at opfatte som en observation af en stokastisk variabel Y. Det medfører at vi også må opfatte estimatet y/n som en observation af den stokastiske variabel Y/n. Denne stokastiske variabel $\hat{p} = \hat{p}(Y) = Y/n$ kaldes *maksimaliseringsestimatoren* for p. Da Y er binomialfordelt med parametre n og p, er middelværdien $\mathrm{E}\,Y$ af Y lig np, og ifølge regnereglerne for middelværdi er så $\mathrm{E}\,\hat{p}(Y) = (\mathrm{E}\,Y)/n = p$, hvilket betyder at maksimaliseringsestimatoren $\hat{p}$ for p i middel[1] giver

[1] En estimator hvis middelværdi er lig den parameter der skal estimeres kaldes en *central* estimator (på engelsk: an *unbiased* estimator).

det rigtige svar p (– men deraf følger ikke noget om det konkrete enkelttilfælde 135/233). For at få en idé om størrelsen af maksimaliseringsestimatorens tilfældige variation omkring sin middelværdi p kan vi bestemme variansen af $\widehat{p}(Y)$ eller standardafvigelsen af $\widehat{p}(Y)$. Da Y har varians $np(1-p)$, har $\widehat{p}(Y) = Y/n$ varians $np(1-p)/n^2 = p(1-p)/n$, dvs. standardafvigelsen af $\widehat{p}(Y)$ er $\sqrt{p(1-p)/n}$. I vores konkrete eksempel er standardafvigelsen af $\widehat{p}$ lig $\sqrt{p(1-p)/233}$, der kan estimeres til $\sqrt{\widehat{p}(1-\widehat{p})/n} = \sqrt{0.58 \times 0.42/233} = 0.03$.

Sammenfattende kan vi sige, at binomialparameteren p estimeres ved $\widehat{p} = 0.58$ med en standardafvigelse på 0.03.

Kvotienttesstørrelsen

Det blev påstået, at man ved hjælp af likelihoodfunktionen kan sammenligne forskellige parameterværdiers evne til at beskrive det faktisk observerede y: hvis $L(p_1) < L(p_2)$, så giver parameterværdien p_2 en bedre beskrivelse end parameterværdien p_1 gør, inden for rammerne af den aktuelle statistiske model. I særdeleshed giver maksimaliseringsestimatet $\widehat{p} = \widehat{p}(y)$ den bedst mulige beskrivelse af observationen y. Parameterværdier der giver en værdi af likelihoodfunktionen som ligger tæt på den maksimale værdi $L(\widehat{p})$, må give en næsten lige så god beskrivelse af observationen y som $\widehat{p}(y)$ gør. Når vi derfor skal teste en statistisk hypotese $H_0 : p = p_0$ om, at den ukendte parameter p kan antages at have den kendte værdi p_0 (– i vort taleksempel er p_0 lig 0.51), så må det foregå ved at vi sammenligner likelihoodfunktionens værdi $L(p_0)$ i punktet p_0 med likelihoodfunktionens maksimale værdi $L(\widehat{p})$. Hvis $L(p_0)$ er næsten lige så stor som $L(\widehat{p})$, betyder det at p_0 beskriver observationen y næsten lige så godt som $\widehat{p}$ gør, og det betyder igen at man kan tillade sig at mene at p_0 er den sande værdi for p: man *akcepterer* eller *godkender* hypotesen H_0. Hvis derimod $L(p_0)$ er væsentlig mindre end $L(\widehat{p})$, betyder det at p_0 giver en væsentlig dårligere beskrivelse af observationen y end $\widehat{p}$ gør, og det er derfor ikke rimeligt at mene at p_0 skulle være den sande værdi af p_0: man *forkaster* hypotesen H_0.

Man skal sammenligne $L(p_0)$ og $L(\widehat{p})$ ved at dividere den mindste

med den største: man danner kvotienten

$$Q = Q(y) = \frac{L(p_0)}{L(\hat{p})} = \frac{L(p_0; y)}{L(\hat{p}; y)} .$$

Q bliver et tal mellem 0 og 1, og værdien benyttes så på den måde at

- En Q-værdi nær 1 betyder at p_0 er stort set lige så god som $\hat{p}$, dvs. man akcepterer H_0.
- En Q-værdi langt fra 1 betyder at p_0 er væsentligere dårligere end $\hat{p}$, dvs. man forkaster H_0.

Q hedder *kvotienttesstørrelsen* for den statistiske hypotese H_0.

I binomialfordelingsmodellen er $L(p) = \binom{n}{y} p^y(1-p)^{n-y}$, så at

$$\begin{aligned} Q = Q(y) &= \frac{p_0^y(1-p_0)^{n-y}}{\hat{p}^y(1-\hat{p})^{n-y}} \\ &= \left(\frac{np_0}{y}\right)^y \left(\frac{n(1-p_0)}{n-y}\right)^{n-y} \end{aligned}$$

idet $\hat{p} = y/n$. I eksemplet er $n = 233$, $y = 135$, $p_0 = 0.51$, så den observerede værdi Q_{obs} af Q er

$$Q_{\text{obs}} = \left(\frac{233 \times 0.51}{135}\right)^{135} \left(\frac{233 \times 0.49}{98}\right)^{98} = 0.105 .$$

Tallet $Q_{\text{obs}} = 0.105$ i sig selv kan vi ikke stille noget op med, — det har ingen mening at spørge om 0.105 er nær 1 eller langt fra 1, så længe vi ikke har nogen målestok, noget sammenligningsgrundlag[2]. Den statistiske model fortæller, at vi skal betragte y som en observation af en stokastisk variabel Y; dermed skal vi også betragte $Q_{\text{obs}} = Q(y)$ som en observation af den stokastiske variabel $Q = Q(Y)$. Fordelingen af Y beskriver hvilke y-værdier man også kunne have fået (i stedet for den faktisk observerede) og med hvilke sandsynligheder, og den tilsvarende fordeling af $Q = Q(Y)$ beskriver dermed hvilke Q-værdier man også kunne have fået (i stedet for 0.105) og med hvilke sandsynligheder.

[2]Det har heller ikke nogen mening at spørge, om en 185cm høj kvinde er høj; hvis hun færdes i Hamburg vil hun ikke virke udsædvanlig høj, men hvis hun færdes i en landsby i Syditalien vil hun afgjort være høj.

Takket være sandsynlighedsfordelingerne kan vi altså sammenholde den faktiske værdi $Q_{\text{obs}} = 0.105$ med alle de andre Q-værdier man også kunne have fået når p har værdien $p_0 = 0.51$.

- Hvis det er sådan at der, når $p = p_0$, er en pæn chance for at få Q-værdier som ligger længere væk fra 1 end Q_{obs} gør, dvs. for at få Q-værdier for hvilke $Q \leq Q_{\text{obs}}$, så vil man sige at Q_{obs} ikke ligger specielt langt fra 1, og man vil akceptere hypotesen $H_0 : p = p_0$.

- Hvis det derimod er sådan at der, når $p = p_0$, er meget lille chance for at få Q-værdier som ligger længere fra 1 end Q_{obs} gør, dvs. for at få Q-værdier for hvilke $Q \leq Q_{\text{obs}}$, så vil man fortolke det som at Q_{obs} i sig selv ligger usædvanlig langt fra 1, og man vil forkaste hypotesen $H_0 : p = p_0$.

Når man skal teste hypotesen H_0, skal man derfor bestemme *testsandsynligheden*

$$\varepsilon = \mathrm{P}_0(Q \leq Q_{\text{obs}}).$$

(Fodtegnet 0 på P-et betyder, at sandsynligheden skal udregnes under antagelse af at hypotesen H_0 er rigtig.) Testsandsynligheden er sandsynligheden under H_0 for at få en værre, dvs. mindre, Q-værdi end den faktisk observerede Q_{obs}.

Hvis testsandsynligheden ε er meget lille, så forkaster man H_0 på grund af følgende ræsonnement:

1. Vi har fået en Q_{obs}-værdi der er så langt fra 1 at der, forudsat at H_0 er rigtig, kun er den meget lille sandsynlighed ε for at få en værre Q-værdi.

2. I praksis plejer man ikke at få særlig ekstreme observationer, så der må være noget galt med grundlaget for beregningen af ε.

3. Da vi ikke kan lave om på observationerne, må det være hypotesen H_0 det er galt med.

Hvis testsandsynligheden ε har en rimelig størrelse, så kan man *ikke* forkaste H_0. Ræsonnementet er denne gang således:

1. Vi har fået en Q_{obs}-værdi der ikke ligger specielt langt fra 1, thi der er nemlig, forudsat at H_0 er rigtig, en rimelig chance ε for at få en værre Q-værdi.

2. Den faktiske værdi Q_{obs} er derfor udmærket forenelig med hypotesen H_0 og der er dermed *ikke* grundlag for at forkaste H_0.

Hvis testsandsynligheden ε er så lille at man forkaster hypotesen, så siger man at teststørrelsen Q_{obs} er *signifikant* eller at der er *signifikans*.

Bestemmelse af ε

Vi vil nu for en stund holde inde med generelle betragtninger over tests og i stedet vende tilbage til den konkrete binomialfordelingsmodel, hvor der viser sig et påtrængende problem, nemlig hvordan bestemmer man rent faktisk ε? Pr. definition er ε lig med sandsynligheden (når $p = p_0$) for at $Q(Y) \leq Q_{\text{obs}}$. Af forskellige grunde, hvoraf nogle er regnetekniske og andre vil fremgå lidt senere, udregner man ofte $-2\ln Q$ i stedet for Q, og testsandsynligheden er da sandsynligheden for at $-2\ln Q(Y) \geq -2\ln Q_{\text{obs}}$. Ud fra det tidligere fundne udtryk for Q får vi at

$$-2\ln Q(y) \;=\; 2\left(y\ln\frac{y}{np_0} + (n-y)\ln\frac{n-y}{n(1-p_0)}\right),$$

så i vores talcksempel er

$$\begin{aligned} -2\ln Q(y) &= 2\left(y\ln\frac{y}{233\times 0.51} + (233-y)\ln\frac{233-y}{233\times 0.49}\right) \\ &= 2\left(y\ln\frac{y}{118.83} + (233-y)\ln\frac{233-y}{114.17}\right) \end{aligned}$$

og dermed

$$-2\ln Q_{\text{obs}} \;=\; -2\ln Q(135) \;=\; 4.51\ .$$

Testsandsynligheden ε fremkommer nu ved at summere binomialsandsynlighederne $f(y;p_0) = \binom{n}{y} p_0^y(1-p_0)^{n-y}$ for alle de y for hvilke $-2\ln Q(Y) \geq -2\ln Q_{\text{obs}}$, dvs.

$$\varepsilon \;=\; \sum_{y:\ -2\ln Q(y)\geq -2\ln Q_{\text{obs}}} \binom{n}{y} p_0^y(1-p_0)^{n-y}\ . \qquad (3.2)$$

Her har vi ε udtrykt ved lutter kendte størrelse. I taleksemplet er således

$$\varepsilon = \sum_{y:\,-2\ln Q(y)\geq 4.51} \binom{233}{y} 0.51^y 0.49^{233-y} ,$$

hvor $-2\ln Q(y)$ er som ovenfor. Fremgangsmåden er derfor

1. udregn $-2\ln Q(y)$ for $y = 0, 1, 2, \ldots, 233$,
2. bestem de y-er for hvilke $-2\ln Q(y) \geq 4.51$,
3. bestem binomialsandsynlighederne for de således udpegede y-er,
4. ε er summen af disse sandsynligheder.

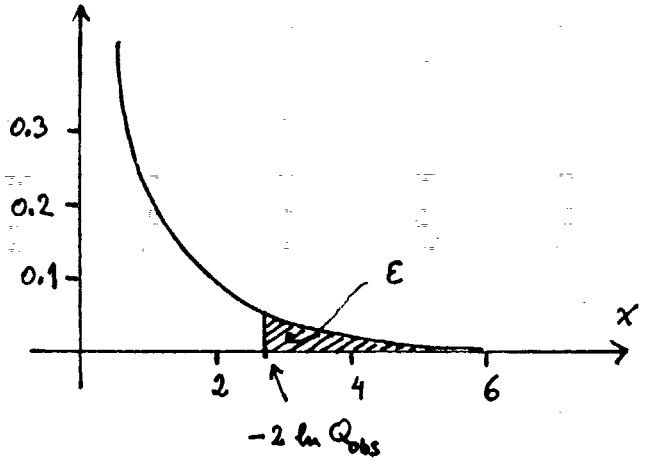
Man finder at $-2\ln Q(y) \geq 4.51$ for $y = 0, 1, 2, \ldots, 102$ og $y = 135, 136, \ldots, 233$. Videre finder man at $\mathrm{P}_0(Y \leq 102) = 0.0161$ og $\mathrm{P}_0(Y \geq 135) = 0.0198$, så at den *eksakte testsandsynlighed* er $\varepsilon = 0.0161 + 0.0198 = 0.0359 \approx 3.6\%$.

Ganske vist er der i Kapitel 2 vist en udmærket algoritme til beregning af binomialsandsynligheder, men alligevel må man nok sige, at ovennævnte regnestykke ikke er noget man lige klarer i en håndevending, medmindre man da har en datamat eller en programmerbar lommeregner til sin rådighed. Heldigvis kan matematikken komme os til hjælp, idet den kan fortælle hvordan man uden større besvær kan bestemme en god tilnærmet værdi af testsandsynligheden ε. Man kan nemlig bevise generelt, at for binomialmodellen (og for en lang række andre statistiske modeller) er den sandsynlighedsfordeling som kvotienttesstørrelsen $-2\ln Q$ følger når den testede hypotese er rigtig med god tilnærmelse af en ganske bestemt type: den er ca. en såkaldt χ^2-fordeling ("khi-i-anden fordeling") med et vist antal frihedsgrader, som i vores aktuelle tilfælde er 1. Da ε jo er sandsynligheden for at få en $-2\ln Q$-værdi som er større end $-2\ln Q_{\mathrm{obs}}$, betyder det, at ε med god tilnærmelse er lig med sandsynligheden for at få en værdi større end $-2\ln Q_{\mathrm{obs}}$ i en χ^2-fordeling med 1 frihedsgrad, og den sandsynlighed kan let bestemmes, f.eks. ved hjælp af tabeller over fraktiler i χ^2-fordelingen.

χ^2-fordelingen med 1 frihedsgrad er den kontinuerte sandsynlighedsfordeling som har tæthedsfunktion

$$f(x) = \frac{1}{\sqrt{2\pi}}\, x^{-1/2} \exp(-x/2), \quad x > 0 .$$

Figur 3.5: Tætheden for χ^2-fordelingen med én frihedsgrad.



Funktionen ser skitsemæssigt ud som vist i Figur 3.5. I en tabel over fraktiler i χ^2-fordelingen finder man, at svarende til 1 frihedsgrad er 95%-fraktilen 3.84 og 97.5%-fraktilen 5.02. Den aktuelle $-2\ln Q_{\text{obs}}$-værdi 4.51 ligger mellem disse to fraktiler, hvilket betyder at (det tilnærmede) ε ligger mellem 5% og 2.5%. (Dette harmonerer udmærket med at den eksakte testsandsynlighed er 3.6%.)

χ^2-fordelingen er som nævnt kun en approksimation til den rigtige fordeling af $-2\ln Q$ under H_0. Man er naturligvis nødt til at have nogle retningslinier for, hvornår approksimationen er god og hvornår ikke. Man plejer at gå ud fra, at hvis begge de *forventede antal* np_0 og $n(1-p_0)$ (det forventede antal drenge hhv. piger) er mindst fem, så kan man anvende χ^2-approksimationen. Ellers må man regne den eksakte testsandsynlighed ud efter "slavemetoden".

De mange udregninger må følges op af en konklusion: Vi fandt en testsandsynlighed på 3.6%, dvs. hvis hypotesen H_0 er rigtig, så er der kun en 3-4% chance for at få en større værdi end den faktisk observerede værdi $-2\ln Q = 4.51$. En så lille testsandsynlighed vil almindeligvis føre til at man forkaster hypotesen H_0. Vi må altså konkludere, at i 1983 er andelen af drenge på SAM-BAS signifikant større end andelen af drenge i den samme aldersgruppe i befolkningen som helhed.

Afrunding

Dette kapitel har behandlet to forskellige emner. På den ene side har det præsenteret en række generelle begreber og teknikker vedrørende den såkaldte likelihood-inferens; dette tema dukker op igen og igen i de følgende kapitler, og læseren skal derfor ikke fortvivle over det allerede her. På den anden side har kapitlet handlet om analyse af en bestemt statistisk model, og det kan der være grund til at give et resumé af.

Resumé 1. Statistisk analyse af den simple binomialfordelingsmodel.

Situation: n individer er klassificeret i to klasser:

klasse	obs. antal
"1"	y
"0"	$n-y$
i alt	n

Model: y er en observation fra en binomialfordeling med parametre n og p, hvoraf p er ukendt.

Estimation: p estimeres ved $\hat{p} = y/n$. Middelværdien af estimatoren $\hat{p}$ er p. Standardafvigelsen af $\hat{p}$ estimeres til $\sqrt{\hat{p}(1-\hat{p})/n}$.

Hypotese: Man ønsker at teste den statistiske hypotese $H_0 : p = p_0$, hvor p_0 er et på forhånd givet tal.

Teststørrelse: Under H_0 er situationen

klasse	obs. antal	"forventet" antal
"1"	y	$\hat{y} = np_0$
"0"	$n-y$	$n-\hat{y} = n(1-p_0)$
i alt	n	n

Kvotienteteststørrelsen er

$$-2\ln Q = 2\left(y\ln\frac{y}{\hat{y}} + (n-y)\ln\frac{n-y}{n-\hat{y}}\right).$$

Testsandsynlighed: Testsandsynligheden ε bestemmes således:

1. Hvis begge de "forventede" antal er mindst 5, kan ε med god tilnærmelse findes som sandsynligheden for at få en værdi større end $-2\ln Q_{\text{obs}}$ i χ^2-fordelingen med 1 frihedsgrad:

$$\varepsilon \;=\; \mathrm{P}\left(\chi_1^2 \geq -2\ln Q_{\text{obs}}\right) .$$

2. I modsat fald må man udregne den eksakte testsandsynlighed

$$\varepsilon \;=\; \sum_{y\,:\,-2\ln Q(y)\geq -2\ln Q_{\text{obs}}} \binom{n}{y} p_0^y (1-p_0)^{n-y}.$$

Konklusion: Hvis ε er meget lille, så er der en signifikant afvigelse mellem det observerede og det som H_0 foreskriver: man må da forkaste H_0.

Hvis ε ikke er meget lille, er H_0 forenelig med det observerede: man kan ikke forkaste H_0.

Kapitel 4

Sammenligning af binomialfordelinger

I Kapitel 1 blev bl.a. fremsat det spørgsmål (side 13), om der i 1983 egentlig er den samme andel drenge på hver af de tre basisuddannelser, således at de faktisk observerede forskelle ikke er andet end hvad der kan skyldes tilfældigheder, se Tabel 4.1.

For at vi skal kunne tale om at noget eventuelt kan skyldes tilfældigheder, må vi have en *statistisk model* der nærmere specificerer, på hvilke punkter der kommer tilfældigheder ind i billedet. Da formålet er at sammenligne de tre drenge-andele, må totalantallene 87, 233 og 123 anses for uinteressante, forstået på den måde at det er uinteressant at der er netop 87 HUM-basister (og ikke 86 eller 90 osv.). Derfor vil vi i den statistiske model *ikke* opfatte totalerne 87, 233 og 123 som observationer af stokastiske variable, men tværtimod betragte dem som givne

Tabel 4.1: (= Tabel 1.9) Absolut og relativ fordeling efter køn for hver basisuddannelse, årgang 1983.

	HUM		SAM		NAT	
	antal	andel	antal	andel	antal	andel
M	36	41%	135	58%	78	63%
K	51	59%	98	42%	45	37%
i alt	87	100%	233	100%	123	100%

konstanter. Det der er interessant er, at der ud af de hhv. 87, 233 og 123 er netop 36, 135 og 78 drenge. Derfor er det tallene 36, 135 og 78 der i den statistiske model skal opfattes som observationer af passende stokastiske variable.

I Kapitel 2 blev der formuleret en model for den del af det nu aktuelle talmateriale som handler om SAM-BAS. Der blev argumenteret for en model gående ud på, at man 233 gange havde udvalgt en person tilfældigt fra en vis "hypotetisk uendelig population" med en bestemt brøkdel drenge, og som resultat heraf havde man fået "udtrukket" 135 drenge. Den matematiske formalisering heraf var så, at det observerede antal (135) skulle betragtes som en observeret værdi af en binomialfordelt stokastisk variabel med antalsparameter 233 og med en ukendt sandsynlighedsparameter. Det er nærliggende at udvide denne model for SAM-BAS til en model for alle tre basisuddannelser på følgende måde.

- Der er forestille sig tre forskellige "hypotetiske uendelige populationer", kaldet H, S og N.
- Fra disse "populationer" udtrækkes hhv. $n_H = 87$, $n_S = 233$ og $n_N = 123$ personer tilfældigt og uafhængigt af hverandre.
- I de tre "populationer" er andelene af drenge p_H, p_S og p_N som er ukendte parametre.
- Vi har faktisk observeret $y_H = 36$, $y_S = 135$ og $y_N = 78$ drenge.

Ved at anvende ræsonnementet fra Kapitel 2 på hver af de tre tilfælde får vi, at vi må betragte de observerede antal y_H, y_S og y_N som observerede værdier af binomialfordelte stokastiske variable Y_H, Y_S og Y_N med antalsparametre n_H, n_S og n_N og med ukendte sandsynlighedsparametre p_H, p_S og p_N. Det faktum, at vi tænker os tre separate "hypotetiske uendelige populationer" og at alle udvælgelser sker uafhængigt af hverandre, medfører, at de tre stokastiske variable Y_H, Y_S og Y_N skal være stokastisk uafhængige. Vi har hermed specificeret fordelingen af de stokastiske variable Y_H, Y_S og Y_N fuldstændigt (nemlig at de er stokastisk uafhængige og hver især binomialfordelt på nærmere angivet måde), og vi er dermed i stand til at skrive modelfunktionen op, men først repeterer vi selve *modellen*:

De observerede antal drenge $y_H = 36$, $y_S = 135$ og $y_N = 78$ opfattes som observationer af stokastiske variable Y_H, Y_S og Y_N, som er stokastisk uafhængige og binomialfordelte med antalsparametre $n_H = 87$, $n_S = 233$ og $n_N = 123$ og med ukendte sandsynlighedsparametre p_H, p_S og p_N.

Modelfunktionen, dvs. den simultane sandsynlighedsfunktion opfattet som en funktion af både y-erne og p-erne, er derfor

$$\begin{aligned} & f(y_H, y_S, y_N; p_H, p_S, p_N) \\ &= \binom{87}{y_H} p_H^{y_H}(1-p_H)^{87-y_H} \\ &\quad \times \binom{233}{y_S} p_S^{y_S}(1-p_S)^{233-y_S} \\ &\quad \times \binom{123}{y_N} p_N^{y_N}(1-p_N)^{123-y_N} . \end{aligned}$$

Det oprindelige spørgsmål, om der egentlig er den samme brøkdel drenge på hver af de tre basisuddannelser, kan nu inden for rammerne af den opstillede model præciseres til spørgsmålet, om det kan antages at de tre ukendte parametre p_H, p_S og p_N er ens, dvs. til den statistiske hypotese $H_0 : p_H = p_S = p_N$.

Vi skal i dette kapitel vise, hvordan den statistiske analyse af denne model forløber når man benytter de principper der blev lanceret i Kapitel 3. Vi vil dog gøre det hele en anelse mere generelt ved at se på en situation med s binomialfordelinger der skal sammenlignes.

Modellen

Antag at vi har klassificeret nogle individer i to forskellige klasser "1" og "0". Individerne er på forhånd delt op i grupper, idet der er s forskellige grupper med hhv. $n_1, n_2, \ldots, n_s$ individer. Det har vist sig, at i gruppe j hører y_j af individerne til klassen "1" og de resterende $n_j - y_j$ af individerne til klassen "0", $j = 1, 2, \ldots, s$. Skematisk ser

situationen sådan ud:

	gruppe nr.				
klasse	1	2	3	...	s
"1"	y_1	y_2	y_3	...	y_s
"0"	$n_1 - y_1$	$n_2 - y_2$	$n_3 - y_3$	...	$n_s - y_s$
i alt	n_1	n_2	n_3	...	n_s

I eksemplet svarer grupperne til de tre basisuddannelser så $s = 3$, og klasserne er M og K.

Den statistiske model der benyttes til at beskrive denne situation er, at $y_1, y_2, \ldots, y_s$ betragtes som observerede værdier af stokastiske variable $Y_1, Y_2, \ldots, Y_s$, der er indbyrdes uafhængige binomialfordelte, således at Y_j har antalsparameter n_j og ukendt sandsynlighedsparameter p_j, $j = 1, 2, \ldots, s$. — Modellen tager altså udgangspunkt i at grupperne er forskellige (mht. den aktuelle klassificering), hvilket giver sig udtryk i at der er en sandsynlighedsparameter for hver gruppe. Opgaven er at undersøge om grupperne kan anses for ens, dvs. den er at teste den statistiske hypotese $H_0 : p_1 = p_2 = \ldots = p_s$.

De generelle retningslinier for hvordan man analyserer en given statistisk model siger, at vi nu først skal opskrive modelfunktionen og likelihoodfunktionen. *Modelfunktionen* er den simultane sandsynlighedsfunktion for Y-erne, opfattet som en funktion af både observationer og parametre, altså

$$f(y_1, y_2, \ldots, y_s;\ p_1, p_2, \ldots, p_s) \;=\; \prod_{j=1}^{s} \binom{n_j}{y_j} p_j^{y_j} (1 - p_j)^{n_j - y_j} \,.$$

Ved her at holde y-erne fast og kun opfatte udtrykket som en funktion af p-erne får vi *likelihoodfunktionen* svarende til observationen $(y_1, y_2, \ldots, y_s)$:

$$L(p_1, p_2, \ldots, p_s) \;=\; \prod_{j=1}^{s} \binom{n_j}{y_j} p_j^{y_j} (1 - p_j)^{n_j - y_j}$$

og dermed log-likelihoodfunktionen

$$\ln L(p_1, p_2, \ldots, p_s) =$$
$$\sum_{j=1}^{s} \ln \binom{n_j}{y_j} + \sum_{j=1}^{s} \left(y_j \ln p_j + (n_j - y_j) \ln(1 - p_j) \right) \,. \qquad (4.1)$$

I taleksemplet er således log-likelihoodfunktionen

$$\begin{aligned}\ln L(p_H, p_S, p_N) = \quad & \ln\binom{87}{36} + \ln\binom{233}{135} + \ln\binom{123}{78} \\ + \quad & 36\ln p_H + 51\ln(1-p_H) \\ + \quad & 135\ln p_S + 98\ln(1-p_S) \\ + \quad & 78\ln p_N + 123\ln(1-p_N)\,.\end{aligned}$$

Likelihoodfunktionen er sandsynligheden for at observere det faktisk observerede, som funktion af det ukendte sæt parametre. Det bedste estimat over de ukendte parametres værdier er derfor det talsæt $(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_s)$ som maksimaliserer likelihoodfunktionen eller log-likelihoodfunktionen. Log-likelihoodfunktionen er en funktion af s variable, men heldigvis en meget skikkelig funktion, idet den (bortset fra et konstantled) er en sum af s led der hver især kun er en funktion af én variabel. Det j-te led hedder $y_j \ln p_j + (n_j - y_j)\ln(1-p_j)$, og vi ved allerede fra Kapitel 3 (side 46) at dette udtryk antager sit maksimum når $p_j = y_j/n_j$. Vi har hermed fundet at *maksimaliseringsestimatet* for $(p_1, p_2, \ldots, p_s)$ er $(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_s) = (\frac{y_1}{n_1}, \frac{y_2}{n_2}, \ldots, \frac{y_s}{n_s})$. I eksemplet er specielt $(\hat{p}_H, \hat{p}_S, \hat{p}_N) = (0.41, 0.58, 0.63)$.

Hypoteseprøvning

Vi skal herefter undersøge om det er rimeligt at antage at hypotesen $H_0 : p_1 = p_2 = \ldots = p_s$ om ens sandsynlighedsparametre holder. Under H_0 er der ingen forskel på de s grupper, og i så fald kan vi lige så godt slå dem sammen til én stor gruppe bestående af $n_\bullet = n_1 + n_2 + \ldots + n_s$ individer, der fordeler sig med $y_\bullet = y_1 + y_2 + \ldots + y_s$ individer i klassen "1" og resten, dvs. $n_\bullet - y_\bullet$, i klassen "0". Derfor må man formode at den fælles p-værdi skal estimeres ved $y_\bullet/n_\bullet$, men lad os benytte likelihood-metoden og se hvad den siger om estimation af den fælles p-værdi.

Vi kalder den fælles værdi (under H_0) af $p_1, p_2, \ldots, p_s$ for p. I den gamle log-likelihoodfunktion (4.1) erstatter vi alle p_j-erne med p og får derved *log-likelihoodfunktionen under* H_0 svarende til observationen $(y_1, y_2, \ldots, y_s)$:

$$\ln L(p, p, \ldots, p)$$

$$= \sum_{j=1}^{s} \ln \binom{n_j}{y_j} + \sum_{j=1}^{s} \Big(y_j \ln p + (n_j - y_j) \ln(1-p) \Big)$$

$$= \sum_{j=1}^{s} \ln \binom{n_j}{y_j} + y_\cdot \ln p + (n_\cdot - y_\cdot) \ln(1-p) \,. \qquad (4.2)$$

Maksimaliseringestimatet $\hat{p}$ for p er den p-værdi der maksimaliserer (4.2), dvs. den p-værdi der maksimaliserer

$$y_\cdot \ln p + (n_\cdot - y_\cdot) \ln(1-p)$$

mht. p. Vi ved fra Kapitel 3 (side 46) at svaret herpå er $\hat{p} = y_\cdot / n_\cdot$. Likelihoodmetoden giver altså det svar som vi formodede måtte være det rigtige. - I vort eksempel bliver $\hat{p} = 249/443 = 0.56$.

Likelihoodfunktionen bruges til at vurdere et sæt parameterværdiers evne til at beskrive det faktisk observerede. Det bedste sæt parameterværdier overhovedet er $(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_s)$. Under H_0 er det bedste sæt værdier $(\hat{p}, \hat{p}, \ldots, \hat{p})$. Vi sammenligner disse to parametersæts beskrivelsesevne ved hjælp af *kvotienttesstørrelsen*

$$Q = \frac{L(\hat{p}, \hat{p}, \ldots, \hat{p})}{L(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_s)} \,.$$

Q antager værdier mellem 0 og 1; en Q-værdi tæt på 1 betyder, at sættet $(\hat{p}, \hat{p}, \ldots, \hat{p})$ beskriver det observerede næsten lige så godt som $(p_1, p_2, \ldots, p_s)$ gør, dvs. vi kan godtage hypotesen H_0, hvorimod en Q-værdi langt fra 1 betyder, at H_0 giver en væsentlig dårligere beskrivelse af det observerede end grundmodellen gør. Som oftest udregner man dog ikke Q men $-2 \ln Q$, som er

$$-2 \ln Q = 2 \Big(\ln L(\hat{p}, \hat{p}, \ldots, \hat{p}) - \ln L(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_s) \Big)$$

$$= 2 \sum_{j=1}^{s} \left(y_j \ln \frac{\hat{p}_j}{\hat{p}} + (n_j - y_j) \ln \frac{1-\hat{p}_j}{1-\hat{p}} \right) .$$

Hvis vi indfører betegnelsen $\hat{y}_j = n_j \hat{p}$, så kan $-2 \ln Q$ omskrives til

$$-2 \ln Q = 2 \sum_{j=1}^{s} \left(y_j \ln \frac{y_j}{\hat{y}_j} + (n_j - y_j) \ln \frac{n_j - y_j}{n_j - \hat{y}_j} \right) ; \qquad (4.3)$$

man kan tænke på $\hat{y}_j$ som *det "forventede" antal* individer fra gruppe j der klassificeres som "1" og på $n_j - \hat{y}_j$ som det "forventede" antal individer fra gruppe j der klassificeres som "0".

I eksemplet er de "forventede" antal[1]

$$\begin{aligned}
\hat{y}_H &= 87 \times 249/443 &&= 48.9 \\
n_H - \hat{y}_H &= 87 - 48.9 &&= 38.1 \\
\hat{y}_S &= 233 \times 249/443 &&= 131.0 \\
n_S - \hat{y}_S &= 233 - 131.0 &&= 102.0 \\
\hat{y}_N &= 123 \times 249/443 &&= 69.1 \\
n_H - \hat{y}_N &= 123 - 69.1 &&= 53.9
\end{aligned}$$

så at

$$\begin{aligned}
-2\ln Q_{\text{obs}} &= 2\Big(36\ln\frac{36}{48.9} + 51\ln\frac{51}{38.1} + \\
&\quad 135\ln\frac{135}{131.0} + 98\ln\frac{98}{102.0} + \\
&\quad 78\ln\frac{78}{69.1} + 45\ln\frac{45}{53.9}\Big) \\
&= 10.6\,.
\end{aligned}$$

Q-værdier tæt på 1 svarer til $-2\ln Q$-værdier tæt på 0. Det vil sige, at hvis $-2\ln Q_{\text{obs}}$ er tæt på 0 så kan vi godtage H_0, hvorimod en stor værdi af $-2\ln Q_{\text{obs}}$ tyder på en signifikant afvigelse mellem det observerede og det som H_0 foreskriver, dvs. vi må forkaste H_0. For at afgøre om tallet $-2\ln Q_{\text{obs}}$ er stort eller lille er vi nødt til at sammenligne det med alle de andre værdier man også kunne have fået ifølge den aktuelle model når H_0 er rigtig. Derfor skal vi bestemme *testsandsynligheden* ε som er sandsynligheden for at få noget værre end det faktisk observerede, dvs. en større $-2\ln Q$-værdi end den observerede, under forudsætning af at H_0 er rigtig:

$$\varepsilon = \mathrm{P}_0\big(-2\ln Q \geq -2\ln Q_{\text{obs}}\big)\,. \tag{4.4}$$

Mere udførligt er ε defineret på følgende måde: Den statistiske model siger, at observationerne $y_1, y_2, \ldots, y_s$ er observerede værdier af stokastiske variable $Y_1, Y_2, \ldots, Y_s$ der er binomialfordelte med antalsparametre

[1] der passende kan udregnes med én decimal men *uden* at bruge den afrundede værdi 0.56 af $\hat{p}$

$n_1, n_2, \ldots, n_s$ og, da H_0 antages rigtig, med samme sandsynlighedsparameter p. Testsandsynligheden ε er da sandsynligheden for, at disse stokastiske variable antager værdier som giver anledning til en $-2\ln Q$-værdi der er større end den faktisk observerede $-2\ln Q_{\text{obs}}$[2].

Det lyder til at være en omstændelig opgave at beregne ε, men takket være matematikken får statistikeren nu alligevel mulighed for at pleje sin dovenskab. Der er nemlig en generel sætning der fortæller, at når H_0 er rigtig, så er $-2\ln Q$ med god tilnærmelse χ^2-fordelt med et antal frihedsgrader som er $s-1$. Det betyder, at ε med god tilnærmelse kan bestemmes som sandsynligheden for at få en værdi større end $-2\ln Q_{\text{obs}}$ i en χ^2-fordeling med $s-1$ frihedsgrader, kort

$$\varepsilon \;=\; \mathrm{P}\left(\chi^2_{s-1} \geq -2\ln Q_{\text{obs}}\right) \;,$$

og den sandsynlighed er let at bestemme, f.eks. ved hjælp af tabeller over fraktiler i χ^2-fordelingen.

Antallet af frihedsgrader for $-2\ln Q$ findes som ændringen i antallet af frie parametre: i grundmodellen er der s frie parametre $p_1, p_2, \ldots, p_s$, under H_0 er der én fri parameter p, derfor bliver der $s-1$ frihedsgrader til teststørrelsen.

I eksemplet er $-2\ln Q_{\text{obs}} = 10.6$ og der er tre grupper, dvs. teststørrelsen har to frihedsgrader. I en tabel over fraktiler i χ^2-fordelingen finder man, at 10.6 netop er 99.5%-fraktilen i χ^2-fordelingen med to frihedsgrader, og det vil sige at testsandsynligheden ε er 0.5%. Værdien 10.6 er altså så stor at der, under forudsætning af at hypotesen er rigtig, kun er 0.5% chance for at få en endnu større værdi, dvs. 10.6 er en særdeles stor værdi. Vi må derfor forkaste hypotesen H_0, eller sagt på en anden måde: Der er en signifikant forskel på de tre basisuddannelsers brøkdele af drenge.

Som nævnt er χ^2-fordelingen kun en approksimation til den rigtige fordeling af $-2\ln Q$. *For at approksimationen skal kunne bruges, skal alle de "forventede" antal $\hat{y}_j$ og $n_j - \hat{y}_j$, $j = 1, 2, \ldots, s$ være mindst fem.* Hvis denne betingelse ikke er opfyldt kan man eventuelt udelade de problematiske grupper eller slå nogle af grupperne sammen på forhånd. Hvis der kun er to grupper i det hele taget, så duer det ikke rigtig at udelade en gruppe eller at slå nogle grupper sammen; man kan da i

[2]En pikant detalje , at for at kunne beregne talværdien af ε skal man formodentlig kende talværdien af p, som er en ukendt parameter! Ganske vist har vi et estimat over værdien af p, men alligevel ...

stedet lave et såkaldte Fisher's eksakte test. Før vi går i gang med det, kommer der dog et kort resumé af de generelle ideer og principper og definitioner der er præsenteret i det foregående.

Resumé 2. Nogle begreber og principper for statistisk inferens

En statistisk model for nogle observationer er et udsagn om, at observationerne opfattes som observerede værdier af en bestemt slags stokastiske variable (dvs. at observationerne opfattes som værende fremkommet som tilfældige tal fra en bestemt slags sandsynlighedsfordelinger).

Parametre. I den fuldstændige angivelse af de stokastiske variables fordelinger indgår også nogle **ukendte parametre**.

Modelfunktionen er sandsynlighedsfunktionen opfattet som en funktion af observationer såvel som parametre. Modelfunktionen angiver sandsynligheden for at få et bestemt udfald når de ukendte parametre har en bestemt værdi.

Likelihoodfunktionen svarende til et bestemt sæt observationer fremkommer ved at man i modelfunktionen indsætter dette sæt observationer og derved får en funktion af de ukendte parametre.

En statistisk hypotese er et udsagn om at de ukendte parametre opfylder visse betingelser (f.eks. at nogle af dem er ens, eller lig 0, etc.).

Maksimaliseringsestimatet (under en vis hypotese) for de ukendte parametre er den værdi af parametrene som maksimaliserer likelihoodfunktionen eller log-likelihoodfunktionen (og som opfylder de betingelser som hypotesen angiver).

Kvotienttesteststørrelsen Q for en bestemt hypotese er kvotienten mellem den maksimale likelihoodfunktion under hypotesen og den maksimale likelihoodfunktion under den aktuelle grundmodel.

Testsandsynligheden ε er sandsynligheden for at $Q \leq Q_{\text{obs}}$, dvs. $-2\ln Q \geq -2\ln Q_{\text{obs}}$, under antagelse af den hypotese der testes.

Man kan bevise en matematisk sætning der fortæller, at i visse nærmere angivne situationer er $-2\ln Q$ med god tilnærmelse χ^2-fordelt med et antal frihedsgrader der findes som "antal frie parametre i den aktuelle grundmodel" minus "antal parametre under den hypotese der testes".

Det var nu nogle matematiske definitioner og en enkelt sætning.

Disse henter deres interesse fra nogle *statistiske principper* (som *ikke* er matematiske definitioner eller sætninger):

Likelihood-metodens grundprincip går ud på, at når man har lagt sig fast på sin statistiske model, så gælder, at al den information vedrørende de ukendte parametre som observationerne kan give kan man hente ud af likelihoodfunktionen (dvs. man kan smide observationerne væk og nøjes med at gemme likelihoodfunktionen). — Dette princip har indtil nu ikke i særlig grad været inddraget i diskussionen.

Dernæst nogle principper om, hvordan man benytter den information der således er i likelihoodfunktionen.

Om sammenligning: At et sæt parameterværdier er *bedre end* et andet til at beskrive de foreliggende observationer betyder, at det første sæt giver en højere værdi af likelihoodfunktionen end det andet.

Maximum likelihood princippet: Det *bedste* estimat over de ukendte parametre er det sæt parameterværdier der maksimaliserer likelihoodfunktionen, altså maksimaliseringsestimatet.

Princippet om kvotienttest: Når man ønsker at teste en bestemt statistisk hypotese skal man som teststørrelse bruge kvotientteststørrelsen Q. En Q-værdi nær 1 betyder, at hypotesen giver en næsten lige så god beskrivelse af det observerede som grundmodellen gør. En Q-værdi langt fra 1 betyder, at hypotesen ikke er særlig forenelig med det observerede.

Om vurdering af Q: En Q-værdi Q_{obs} *ligger langt fra 1* hvis det er usandsynligt at få en værdi der ligger endnu længere fra 1 end Q_{obs} gør. Det man egentlig har brug for at vide er derfor ikke værdien af Q_{obs}, men den tilsvarende værdi af testsandsynligheden.

Om signifikansgrænser. Hvis man skal have nogen fornøjelse af det sidste princip, så må man have nogle *retningslinier* for, hvornår man skal sige at testsandsynligheden ε er så lille at det er usandsynligt at få en værre Q-værdi, altså at Q er signifikant. Man vil ofte sige, at hvis ε er væsentlig mindre end 5%, f.eks. 2.5%, så er Q *signifikant*, dvs. hypotesen er uforenelig med det observerede; hvis omvendt ε er pænt større end 5%, f.eks. 10%, så er Q *ikke-signifikant*, dvs. hypotesen er pænt forenelig med

det obseverede; hvis endelig ε er tæt på 5%, så er vi havnet i et "gråt område" hvor de ikke er noget klart svar. — Det må dog siges at være god tone altid at angive værdien af ε og ikke blot skrive om teststørrelsen er signifikant eller ikke-signifikant.

Det eksakte test

Som nævnt tidligere i kapitlet (side 62) kan der opstå tvivl om anvendeligheden af χ^2-approksimationen til $-2\ln Q$ når nogle af de "forventede" antal er små. Vi skal nu se hvordan man kan sammenligne *to* binomialfordelinger hvor nogle af antallene er for små. Tag som eksempel den situation der er skitseret i Tabel 4.2. Ved at efterligne ræsonnementerne i begyndelsen af dette kapitel kan man nå frem til følgende (forslag til den) statistiske model for observationerne i Tabel 4.2:

> De observerede antal drenge $y_1 = 2$ og $y_2 = 6$ opfattes som observationer af stokastiske variable Y_1 og Y_2, som er stokastisk uafhængige og binomialfordelte med antalsparametre $n_1 = 6$ og $n_2 = 9$ og med ukendte sandsynlighedsparametre p_1 hhv. p_2. Den tilsvarende *modelfunktion* er
>
> $$f(y_1, y_2;\, p_1, p_2) \;=\; \binom{6}{y_1} p_1^{y_1}(1-p_1)^{6-y_1} \times \binom{9}{y_2} p_2^{y_2}(1-p_2)^{9-y_2}.$$
>
> Maksimaliseringsestimaterne for p_1 og p_2 er $\widehat{p}_1 = 2/6 = 1/3$ og $\widehat{p}_2 = 6/9 = 2/3$.

Lad os sætte, at opgaven er at undersøge, om der er en signifikant forskel på kønsfordelingen i de to grupper, eller om det tværtimod er sådan at de observerede forskelle ikke er andet end hvad man kan komme ud for på grund af tilfældigheder. Vi vil derfor teste den statistiske hypotese $H_0 : p_1 = p_2$.

Tabel 4.2: Fordeling efter køn i to projektgrupper.

	gr. 1	gr. 2	SUM
M	2	6	8
K	4	3	7
i alt	6	9	15

Tabel 4.3: Forventet kønsfordeling under H_0 i de to projektgrupper.

	gr. 1	gr. 2	SUM
M	3.2	4.8	8
K	2.8	4.2	7
i alt	6	9	15

Problemet

Da vi har at gøre med et specialtilfælde af det generelle problem "sammenligning af binomialfordelinger" der blev behandlet tidligere i kapitlet, kan vi nu blot gå frem efter opskriften. Under H_0 er maksimaliseringsestimatet for den fælles værdi af p_1 og p_2 givet som $\hat{p} = 8/15 = 0.53$ og de "forventede" antal $\hat{y}_1 = n_1\hat{p}$, osv. er derfor som i Tabel 4.3. Kvotienttesstørrelsen $-2\ln Q$ er dermed

$$\begin{aligned} -2\ln Q &= 2 \times \sum \left(\text{obs. antal} \times \ln \frac{\text{obs antal}}{\text{forv. antal}} \right) \\ &= 2\left(2\ln\frac{2}{3.2} + 4\ln\frac{4}{2.8} + 6\ln\frac{6}{4.8} + 3\ln\frac{3}{4.2} \right) \\ &= 1.63 \, . \end{aligned}$$

Store værdier af $-2\ln Q$ tyder på at hypotesen H_0 ikke holder; for at afgøre om 1.63 er en "stor" værdi, skal vi bestemme testsandsynligheden ε, dvs. sandsynligheden for at få en $-2\ln Q$-værdi som er større end 1.63 under forudsætning af at H_0 er rigtig:

$$\varepsilon = \mathrm{P}_0(-2\ln Q \geq 1.63) \, .$$

Der gælder, at hvis de "forventede antal" alle er mindst fem, så kan ε findes med god tilnærmelse som sandsynligheden for at få en værdi på mindst 1.63 i en χ^2-fordeling med én frihedsgrad. *Men* i vores tilfælde er ingen af de "forventede" antal (Tabel 4.3) over fem, så vi kan *ikke* gå ud fra at χ^2-approksimationen er anvendelig.

Et betinget test

Derfor må man prøve at udregne ε fra 'first principles'. Hvis man udtrykker $-2\ln Q$ ved y_1 og y_2 får man (jf. (4.3))

$$-2\ln Q(y_1,y_2) = 2\left(y_1\ln\frac{y_1}{n_1\frac{y_\cdot}{n_\cdot}}+(n_1-y_1)\ln\frac{n_1-y_1}{n_1(1-\frac{y_\cdot}{n_\cdot})}+\right.$$
$$\left.y_2\ln\frac{y_2}{n_2\frac{y_\cdot}{n_\cdot}}+(n_2-y_2)\ln\frac{n_2-y_2}{n_2(1-\frac{y_\cdot}{n_\cdot})}\right),$$

hvor $y_\cdot = y_1+y_2$ og $n_\cdot = n_1+n_2$. Her kan parret (y_1,y_2) antage 70 forskellige sæt værdier svarende til at $y_1 = 0,1,2,\ldots,6$ og $y_2 = 0,1,2,\ldots,9$. Man kunne så udregne $-2\ln Q$ for hvert af de 70 mulige udfald og derved bestemme de udfald (y_1,y_2) for hvilke $-2\ln Q(y_1,y_2)$ er mindst 1.63 . Man finder at det er de kombinationer (y_1,y_2) som er markeret med $\star$ i nedenstående skema:

Kombinationer af (y_1,y_2) for hvilke $-2\ln Q(y_1,y_2)\geq 1.63$

		y_1						
		0	1	2	3	4	5	6
	0	·	·	★	★	★	★	★
	1	·	·	·	★	★	★	★
	2	★	·	·	·	★	★	★
	3	★	·	·	·	★	★	★
y_2	4	★	·	·	·	·	★	★
	5	★	★	·	·	·	·	★
	6	★	★	★	·	·	·	★
	7	★	★	★	·	·	·	★
	8	★	★	★	★	·	·	·
	9	★	★	★	★	★	·	·

Testsandsynligheden ε kan så findes som summen af sandsynlighederne $f(y_1,y_2;p,p)$ for alle udfald (y_1,y_2) for hvilke $-2\ln Q(y_1,y_2)\geq 1.63$. Denne fremgangsmåde indebærer, som man hurtigt vil erfare, en hel del regnearbejde[3], men der er også en komplikation af mere fundamental

[3]men alle de moderne regnetekniske hjælpemidler taget i betragtning kan *det* nu ikke være nogen alvorlig endsige principiel hindring for at benytte fremgangsmåden.

karakter.

I Kapitel 3 testede vi hypoteser gående ud på at den eneste ukendte parameter havde en bestemt, på forhånd givet, værdi. Når en sådan hypotese var rigtig, var der ikke flere ukendte parametre inde i billedet — den slags hypoteser plejer man at kalde *simple hypoteser*. De hypoteser vi tester i indeværende kapitel er af en anden slags: Der er tale om modeller med mere end en ukendt parameter, og hypoteserne går ud på at nogle af disse parametre er ens. Når en sådan hypotese er rigtig, er der stadigvæk ukendte parametre i modellen — den slags hypoteser plejer man at kalde *sammensatte hypoteser*.

I det aktuelle hypoteseprøvnings-problem, der altså handler om en sammensat hypotese, nåede vi ovenfor frem til at testsandsynligheden ε måtte skulle bestemmes som en sum af nogle sandsynligheder $f(y_1, y_2; p, p)$ hvor der summeres over en vis mængde (y_1, y_2)-er, og hvor der indgår den fælles men *ukendte* parameter p. For at beregne ε skal vi altså kende (den rigtige værdi af) den ukendte parameter p ! Nu ville læseren måske nok uden at blegne indsætte værdien af $\widehat{p}$ (som er 8/15) og så udregne ε på det grundlag (hvorved man får ε til 27%), men det ændrer ikke ved det principielle problem. Der findes imidlertid en fremgangsmåde ved hjælp af hvilken man helt kan eliminere det famøse p:

Parameteren p er sandsynligheden for at en tilfældigt valgt person er en dreng, når gr.1-populationen og gr.2-populationen er ens. Den del af observationsmaterialet der indeholder information om p må være, at der ud af de i alt 15 personer viste sig at være netop 8 drenge. Nu kan man sige, at det er uinteressant at der netop er 8 (og ikke 7 eller 10) drenge; det interessante er at de 8 fordeler sig med 2 i gr.1 og 6 i gr.2 . Derfor skal man (sådan siger et statistisk princip) se på *den betingede fordeling* givet at der netop var 8 drenge. I denne betingede fordeling vil det vise sig, at den oprindelige sammensatte hypotese H_0 bliver til en simpel hypotese. For at se hvordan det går til må vi oversætte det netop sagte til matematik.

Modelfunktionen i grundmodellen er som allerede nævnt

$$f(y_1, y_2; p_1, p_2) \;=\; \binom{6}{y_1} p_1^{y_1}(1-p_1)^{6-y_1} \binom{9}{y_2} p_2^{y_2}(1-p_2)^{9-y_2} \;.$$

Når H_0 er rigtig har p_1 og p_2 den fælles værdi p, og modelfunktionen

kommer så til at se sådan ud:

$$f(y_1, y_2; p, p) = \binom{6}{y_1} p^{y_1}(1-p)^{6-y_1} \times \binom{9}{y_2} p^{y_2}(1-p)^{9-y_2}$$

$$= \binom{6}{y_1}\binom{9}{y_2} \times p^{y_1+y_2}(1-p)^{15-(y_1+y_2)} .$$

Heraf fremgår, at likelihoodfunktionen under H_0 er

$$L(p) = \text{konstant} \times p^{y_1+y_2}(1-p)^{15-(y_1+y_2)} ,$$

dvs. man kan bestemme likelihoodfunktionen (pånær en konstant faktor) blot man kender det totale antal drenge $y_1 + y_2$ — man behøver ikke at kende y_1 og y_2 hver for sig[4].

Påstanden er, at i den betingede fordeling givet det samlede antal drenge $Y_1 + Y_2$ bliver hypotesen H_0 til en simpel hypotese[5]. For at indse det vil vi bestemme den betingede fordeling af Y_1 og Y_2 givet at $Y_1 + Y_2 = 8$. Ifølge de sædvanlige formler for betingede sandsynligheder er

$$\mathrm{P}(Y_1 = y_1, Y_2 = y_2 \,\big|\, Y_1 + Y_2 = 8)$$

$$= \begin{cases} \dfrac{\mathrm{P}(Y_1 = y_1) \times \mathrm{P}(Y_2 = 8 - y_1)}{\mathrm{P}(Y_1 + Y_2 = 8)} & \text{hvis } y_1 + y_2 = 8 \\ 0 & \text{hvis } y_1 + y_2 \neq 8 , \end{cases}$$

og udtrykket svarende til tilfældet $y_1 + y_2 = 8$ kan videre omskrives således (hvor y_1 erstattes af y):

$$\frac{\mathrm{P}(Y_1 = y_1) \times \mathrm{P}(Y_2 = 8 - y_1)}{\mathrm{P}(Y_1 + Y_2 = 8)}$$

$$= \frac{f(y, 8-y; p_1, p_2)}{\sum_{z=0}^{8} f(z, 8-z; p_1, p_2)}$$

$$= \frac{\binom{6}{y} p_1^{y}(1-p_1)^{6-y} \times \binom{9}{8-y} p_2^{8-y}(1-p_2)^{9-(8-y)}}{\sum_{z=0}^{8} \binom{6}{z} p_1^{z}(1-p_1)^{6-z} \times \binom{9}{8-z} p_2^{8-z}(1-p_2)^{9-(8-z)}}$$

[4]Statistikeren udtrykker dette på den måde, at $y_1 + y_2$ er en *minimalsufficient reduktion* af data.

[5]Det hænger i øvrigt sammen med at under H_0 er $Y_1 + Y_2$ minimalsufficient.

$$= \frac{\binom{6}{y}\binom{9}{8-y}\theta^y}{\sum_{z=0}^{8}\binom{6}{z}\binom{9}{8-z}\theta^z},$$

hvor

$$\theta = \frac{p_1}{1-p_1}/\frac{p_2}{1-p_2} = \frac{p_1(1-p_2)}{p_2(1-p_1)}.$$

Det ses, at hvor grundmodellen har to ukendte parametre p_1 og p_2, har den betingede model kun én parameter, nemlig θ. Modelfunktionen for den betingede model er

$$\overline{f}(y;\theta) = \frac{\binom{6}{y}\binom{9}{8-y}\theta^y}{\sum_{z=0}^{8}\binom{6}{z}\binom{9}{8-z}\theta^z}.$$

Da parameteren θ afhænger af grundmodellens parametre på den måde at

$$\theta = \frac{p_1(1-p_2)}{p_2(1-p_1)},$$

er det klart at grundmodellens hypotese

$$H_0 : \ p_1 = p_2$$

er ensbetydende med hypotesen

$$\overline{H}_0 : \ \theta = 1$$

i den betingede model. Den sammensatte hypotese i grundmodellen er blevet til en simpel hypotese i den betingede model.

Vi kan nu teste hypotesen $\overline{H}_0$ ved brug af de sædvanlige principper, og da $\overline{H}_0$ er en simpel hypotese opstår der ikke nogen principielle problemer[6]. Der foreligger observationen $y = 2$. Det tilsvarende bedste

[6]Til gengæld er der regnetekniske problemer, så vi er stadig ikke nået til den metode man plejer at benytte.

skøn $\hat{\theta}$ over θ er den θ-værdi der maksimaliserer den betingede likelihoodfunktion

$$\overline{L}(\theta) = \overline{f}(2;\theta) ,$$

dvs. den θ-værdi som er løsning til

$$\frac{d}{d\theta}\overline{L}(\theta) = 0 .$$

Man finder at $\hat{\theta} = \hat{\theta}(2) = 0.276$. Kvotienttesstørrelsen for $\overline{H}_0$ er

$$Q = Q(2) = \frac{\overline{L}(1)}{\overline{L}(\hat{\theta}(2))} = \frac{\overline{L}(1)}{\overline{L}(0.276)} = 0.468.$$

Hvis Q er langt fra 1 er det tegn på at $\overline{H}_0$ skal forkastes. For at vurdere om 0.468 ligger langt fra 1 skal vi udregne testsandsynligheden ε, som er sandsynligheden (under $\overline{H}_0$) for at få et y således at $Q(y)$ er mindre end eller lig med 0.468:

$$\varepsilon = \sum_{y\,:\,Q(y)\leq 0.468} \overline{f}(y;1) .$$

Bestemmelsen af ε er ukompliceret men noget besværlig. Man finder følgende resultater:

	y	$Q(y)$	$\overline{f}(y;1)$
	0	0.002	0.001
	1	0.069	0.034
→	2	0.468	0.196
	3	0.979	0.392
	4	0.713	0.294
	5	0.166	0.078
	6	0.006	0.006
	7	-	0
	8	-	0
			1.001

Det ses heraf, at de y-er som giver en Q-værdi der er $\leq Q(2) = 0.468$, dvs. de y-er der er mindst lige så uforenelige med $\overline{H}_0$ som $y = 2$ er, er y-erne 0,1,2,5,6 , således at testsandsynligheden er

$$\begin{aligned} \varepsilon &= \overline{f}(0;1) + \overline{f}(1;1) + \overline{f}(2;1) + \overline{f}(5;1) + \overline{f}(6;1) \\ &= 0.315. \end{aligned}$$

Der er altså ca. 31% chance for at få et "lige så slemt eller værre" y end det observerede $y = 2$, når $\overline{H}_0$ er rigtig. Man vil derfor sige at der *ikke* er nogen signifikant uoverensstemmelse mellem hypotesen $\overline{H}_0$ og det observerede $y = 2$. Sagt på en anden måde: vi kan ikke forkaste $\overline{H}_0$.

Vi er gået meget let hen over, hvordan man egentlig skal finde talværdien af $\hat{\theta}$ og hvordan man egentlig beregner værdier af funktionerne $\overline{L}$ og $\overline{f}$. Grunden hertil er, at den just beskrevne metode, som er den principielt rigtigste, faktisk sædvanligvis ikke bruges. Den er nemlig besværlig rent regnemæssigt, såfremt man skal regne med håndkraft. Det er ganske vist ingen sag at skrive et lille computer-program der kan udføre beregningerne, men man bruger alligevel (endnu) for det meste en regnemæssigt simplere metode, som vi nu vil beskrive i detaljer.

Det eksakte test

Det man gør når man tester en statistisk hypotese er, at man udregner værdien af en vis teststørrelse, sædvanligvis kvotienttesteststørrelsen Q eller $-2\ln Q$, der er et udtryk for hvor godt hypotesen er forenelig med de foreliggende data; dernæst bestemmer man testsandsynligheden, dvs. sandsynligheden for at få et sæt observationer som er mindst lige så "uforenelige" med hypotesen som de faktiske observationer er. I den simplere metode til løsning af det aktuelle testproblem benytter man ikke Q som teststørrelse, men derimod sandsynlighedsfunktionen $\overline{f}(\cdot;1)$ svarende til at hypotesen $\overline{H}_0$ er rigtig. (Derved slipper vi bl.a. for at skulle bestemme $\hat{\theta}$.) Denne funktion er forholdsvis simpel:

$$\overline{f}(y;1) = \frac{\binom{6}{y}\binom{9}{8-y}}{\sum_{z=0}^{8}\binom{6}{z}\binom{9}{8-z}} . \tag{4.5}$$

(Man kan i øvrigt vise, at nævneren er lig $\binom{15}{8}$.) Det såkaldte *eksakte test* for $\overline{H}_0$ forløber nu på følgende måde.

Vi har observeret $y = 2$. Vi skal bestemme de y-er for hvilke $\overline{f}(y;1) \leq \overline{f}(2;1)$. For at gøre det udregner vi tælleren i (4.5) for alle de

mulige y-er, hvilket er hurtigt gjort, f.eks. ved brug af Pascal's trekant (side 31). Man får

y	$\binom{6}{y}$	$\times$	$\binom{9}{8-y}$	$=$	$\overline{f}(y;1)$
0	1	$\times$	9	$=$	9
1	6	$\times$	36	$=$	216
2	15	$\times$	84	$=$	1260
3	20	$\times$	126	$=$	2520
4	15	$\times$	126	$=$	1890
5	6	$\times$	84	$=$	504
6	1	$\times$	36	$=$	36
7	0	$\times$	9	$=$	0
8	0	$\times$	1	$=$	0
					6435

Det ses at de y-er som er mere ekstreme end $y = 2$ (i den forstand at $\overline{f}(y;1) \leq \overline{f}(2;1) = 1260/6435$) er alle y-erne undtagen $y = 3$ og $y = 4$. Testsandsynligheden, dvs. sandsynligheden for at få en mindst lige så ekstrem observation som $y = 2$, er derfor

$$\begin{aligned} \varepsilon &= 1 - \left(\overline{f}(3;1) + \overline{f}(4;1)\right) \\ &= 1 - \frac{2520 + 1890}{6435} \\ &= 31\%. \end{aligned}$$

Det eksakte test giver således (i dette eksempel) præcis samme resultat som det rigtige betingede test.

Hvad angår det oprindelige praktiske problem kan vi i første omgang konkludere, at $\overline{H}_0$ må akcepteres, dvs. der er ikke nogen signifikant forskel på kønsfordelingen i de to grupper set fra den betingede models synspunkt. Da man kan sige (jf. side 70), at det der adskiller den betingede model og den oprindelige (ubetingede) model er noget som er uinteressant for spørgsmålet om ens kønsfordeling i de to grupper, kan vi videre konkludere, at også H_0 må akcepteres, dvs. heller ikke fra grundmodellens synspunkt er der nogen signifikant forskel på kønsfordelingen i de to grupper.

Resumé 3. Sammenligning af binomialfordelinger

Situation: n_j individer fra gruppe j er klassificeret i to klasser, $j = 1, 2, \ldots, s$:

	gruppe nr.				
klasse	1	2	...	s	sum
"1"	y_1	y_2	...	y_s	$y_\bullet$
"0"	$n_1 - y_1$	$n_2 - y_2$	...	$n_s - y_s$	$n_\bullet - y_\bullet$
i alt	n_1	n_2	...	n_s	$n_\bullet$

Model: $y_1, y_2, \ldots, y_s$ er uafhængige observationer fra binomialfordelinger, således at y_j stammer fra en binomialfordeling med parametre n_j og p_j.

$p_1, p_2, \ldots, p_s$ er ukendte parametre.

Estimation: p_j estimeres ved $\hat{p}_j = y_j/n_j$, $j = 1, 2, \ldots, s$.

Hypotese: Man ønsker at teste den statistiske hypotese

$$H_0 : \; p_1 = p_2 = \ldots = p_s \;\; (= p) \; .$$

Teststørrelse: Under H_0 er den "forventede" situation

	gruppe nr.				
klasse	1	2	...	s	sum
"1"	$\hat{y}_1$	$\hat{y}_2$	...	$\hat{y}_s$	$y_\bullet$
"0"	$n_1 - \hat{y}_1$	$n_2 - \hat{y}_2$	...	$n_s - \hat{y}_s$	$n_\bullet - y_\bullet$
i alt	n_1	n_2	...	n_s	$n_\bullet$

hvor $\hat{y}_j = n_j \hat{p}$ og $\hat{p} = y_\bullet/n_\bullet$.

Kvotienttesstørrelsen er

$$-2 \ln Q \;\; = \;\; 2 \sum_{j=1}^{s} \left(y_j \ln \frac{y_j}{\hat{y}_j} + (n_j - y_j) \ln \frac{n_j - y_j}{n_j - \hat{y}_j} \right) .$$

Testsandsynlighed:

1. Hvis alle de $2s$ "forventede" antal er mindst 5, kan testsandsynligheden ε med god tilnærmelse findes som sandsynligheden for at få en værdi større end $-2 \ln Q_{\text{obs}}$ i χ^2-fordelingen med $s - 1$ frihedsgrader:

$$\varepsilon \;\; = \;\; \mathrm{P}\left(\chi^2_{s-1} \geq -2 \ln Q_{\text{obs}} \right) \; .$$

2. I modsat fald må man prøve at slå nogle smågrupper sammen, således at 1) bliver opfyldt. Eller hvis $s = 2$ kan man benytte det eksakte test.

Det eksakte test når $s = 2$**:** Udregn størrelsen

$$g(z) = \binom{n_1}{z}\binom{n_2}{y_\bullet - z}$$

for $z = 0, 1, 2, \ldots, y_\bullet$, samt $S = g(0) + g(1) + \ldots + g(y_\bullet)$ (der i øvrigt er lig $\binom{n_1 + n_2}{y_\bullet}$). Testsandsynligheden er da

$$\varepsilon = \frac{1}{S} \sum_{z\,:\,g(z) \le g(y_1)} g(z) \,.$$

Konklusion: Hvis ε er meget lille, så er der en signifikant afvigelse mellem det observerede og det som H_0 foreskriver, og man vil da forkaste H_0. I modsat fald er H_0 forenelig med det observerede, og man kan ikke forkaste H_0.

Kapitel 5

Multinomialfordelingen

Dette kapitel beskæftiger sig med multinomialfordelingen og med sammenligning af multinomialfordelinger. Endnu en gang tager vi udgangspunkt i et af spørgsmålene fra Kapitel 1, nemlig spørgsmålet (side 17) om der er nogen reel forskel på styrkeforholdene mellem de tre basisuddannelser i de to år 1974 og 1982. Det talmateriale det drejer sig om er vist i Tabel 5.1.

Hvis vi skal kunne tale om at de observerede forskelle mellem de to år eventuelt kan skyldes tilfældigheder, må vi have en *statistisk model* der nærmere specificerer, hvor der kommer tilfældigheder ind i billedet. Vi er ude på at sammenligne *fordelingerne* på de tre basisuddannelser, og derfor må vi anse totalantallene 413 og 537 for uinteressante, således forstået at det er uinteressant, at der i 1974 netop er 413 (og ikke 414 eller 410 osv.) og at der i 1982 netop er 537 optagne; i den statistiske

Tabel 5.1: (= Tabel 1.10) Absolut og relativ fordeling efter basisuddannelse for hver af årgangene 1974 og 1982.

	1974		1982	
	antal	andel	antal	andel
HUM	130	31%	158	29%
SAM	218	53%	247	46%
NAT	65	16%	132	25%
i alt	413	100%	537	100%

model skal totalantallene derfor indgå som givne konstanter. Det der er interessant er, at de 413 fordelte sig med 130, 218 og 65, og at de 537 fordelte sig med 158, 247 og 132 på de tre basisuddannelser; i den statistiske model er det derfor talsættene (130, 218, 65) og (158, 247, 132) der skal opfattes som observationer af stokastiske variable.

Lad os i første omgang se på året 1974. Talsættet (130, 218, 65) er fremkommet på den måde, at man har klassificeret $n = 413$ personer i tre klasser (HUM, SAM, NAT), hvorved det har vist sig, at der kom $y_1 = 130$ personer i klasse nr. 1, $y_2 = 218$ personer i klasse nr. 2 og $y_3 = 65$ personer i klasse nr. 3. Vi vil gøre diskussionen lidt mere generel ved at se på en situation hvor n individer klassificeres i r klasser $A_1, A_2, \ldots, A_r$.

5.1 Den simple multinomialfordelingsmodel

Antag at vi har klassificeret n individer i r klasser $A_1, A_2, \ldots, A_r$. Skematisk ser situationen således ud:

niveau	klasse	observeret antal
1	A_1	y_1
2	A_2	y_2
3	A_3	y_3
$\vdots$	$\vdots$	$\vdots$
r	A_r	y_r
i alt		n

Vi går ud fra, at de n individer stammer fra en og samme "hypotetiske uendelige population", som er indrettet på den måde, at hver gang man tilfældigt udvælger et individ fra "populationen", så er der sandsynligheden p_1 for at individet tilhører A_1, sandsynligheden p_2 for at individet tilhører A_2, osv. Sandsynlighederne $p_1, p_2, \ldots, p_r$ (der summerer til 1) er *ukendte parametre* der er karakteristiske for "populationen".

Hermed har vi sådan set beskrevet den statistiske model for ét individ. Når der er et større antal individer angiver man ikke hvert enkelt

individs klasse, men man angiver kun de observerede værdier af de stokastiske variable $Y_1, Y_2, \ldots, Y_r$ defineret ved

$$Y_i = \begin{array}{l}\text{antal individer der ved den}\\ \text{tilfældige udvælgelse viser}\\ \text{sig at tilhøre klassen } A_i\end{array} \quad (i = 1, 2, \ldots, r) .$$

Den statistiske model vi skal nå frem til skal derfor specificere sandsynlighedsfordelingen for $(Y_1, Y_2, \ldots, Y_r)$, dvs. vi skal bestemme $\mathrm{P}(Y_1 = y_1,\ Y_2 = y_2,\ \ldots,\ Y_r = y_r)$.

Hvis der kun er *to* klasser, så står vi med et problem der allerede er løst i Kapitel 2, hvor vi lærte, at når man klassificerer i *to* klasser "1" og "0", så er antal individer der klassificeres som "1" binomialfordelt. Ved at generalisere ræsonnementet fra begyndelsen af Kapitel 2 kan vi finde den søgte fordeling også når der er mere end to klasser. Det kommer til at forløbe på følgende måde:

Vi indfører nogle hjælpevariable $X_1, X_2, \ldots, X_n$ således at X_d er navnet på den klasse som individ nr. d tilhører, dvs. $X_d = A_i$ hvis og kun hvis individ nr. d tilhører klassen A_i. Der gælder så, at $\mathrm{P}(X_d = A_i) = p_i$. Da individerne tænkes valgt uafhængigt af hverandre, må de forskellige X_d-er være stokastisk uafhængige, således at f.eks.

$$\mathrm{P}(X_{d_1} = A_{i_1},\ X_{d_2} = A_{i_2}) \ = \ p_{i_1} p_{i_2} .$$

Hvis vi har en stribe klassenavne $x_1, x_2, \ldots, x_n$, hvor y_1 af x-erne er et A_1, y_2 af x-erne er et A_2 osv., så er

$$\begin{aligned}&\mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)\\ &= \mathrm{P}(X_1 = x_1) \times \mathrm{P}(X_2 = x_2) \times \ldots \times \mathrm{P}(X_n = x_n)\\ &= p_1^{y_1} p_2^{y_2} \ldots p_r^{y_r} .\end{aligned}$$

Ved at summere disse sandsynligheder over alle mulige sæt $(x_1, x_2, \ldots, x_n)$ bestående af y_1 A_1-er, y_2 A_2-er, $\ldots$, y_r A_r-er får vi den søgte sandsynlighed:

$$\begin{aligned}&\mathrm{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_r = y_r)\\ &= \sum \mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)\\ &= \sum p_1^{y_1} p_2^{y_2} \ldots p_r^{y_r}\\ &= \Big(\sum 1\Big) \times p_1^{y_1} p_2^{y_2} \ldots p_r^{y_r} ,\end{aligned}$$

hvor summationstegnet hver gang betyder "summation over de $(x_1, x_2, \ldots, x_n)$ som består af y_1 A_1-er, y_2 A_2-er osv.". Symbolet $\sum 1$ betyder på den måde *antallet* af forskellige sådanne n-tupler $(x_1, x_2, \ldots, x_n)$. Når vi får bestemt dette antal, som man plejer at betegne

$$\binom{n}{y_1\ y_2\ \cdots\ y_r}, \tag{5.1}$$

har vi fundet det fuldstændige udtryk for den søgte sandsynlighed. Antallet hedder i øvrigt en *multinomialkoefficient* (eller *polynomialkoefficient*), og den fundne sandsynlighedsfunktion

$$\begin{aligned} &\mathrm{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_r = y_r) \\ &= \binom{n}{y_1\ y_2\ \cdots\ y_r} p_1^{y_1} p_2^{y_2} \cdots p_r^{y_r} \end{aligned} \tag{5.2}$$

er en sandsynlighedsfunktion for en *multinomialfordeling* (eller *polynomialfordeling*) *med parametre* n *og* $\boldsymbol{p}$, *hvor*

$$\boldsymbol{p} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_r \end{pmatrix}.$$

Hvad angår den statistiske model for 1974-tallene kan vi altså sige, at de observerede antal $y_1 = 130$, $y_2 = 218$ og $y_3 = 65$ er observationer fra en multinomialfordeling[1] med antalsparameter $n = 413$ og med ukendte sandsynlighedsparametre p_1, p_2 og p_3 ($p_1 + p_2 + p_3 = 1$). *Modelfunktionen* er

$$f(y_1, y_2, y_3;\ p_1, p_2, p_3) = \binom{413}{y_1\ y_2\ y_3} p_1^{y_1} p_2^{y_2} p_3^{y_3}.$$

Hvis man her holder y-erne fast og kun betragter udtrykket som en funktion af p-erne, har man *likelihoodfunktionen*. I særdeleshed er likelihoodfunktionen svarende til $y_1 = 130$, $y_2 = 218$ og $y_3 = 65$

$$L(p_1, p_2, p_3) = \binom{413}{130\ 218\ 65} p_1^{130} p_2^{218} p_3^{65}.$$

[1] som i dette tilfælde er en *tri*-nomialfordeling, da der er tre klasser.

(Man skal ganske afgjort ikke give sig i kast med at udregne multinomialkoefficienten $\binom{413}{130\ 218\ 65}$; resultatet bliver et heltal med 176 cifre, og vi skal ikke bruge det til noget.)

Multinomialkoefficienter

Multinomialkoefficienten (5.1) er hurtigt bestemt. Vi illustrerer først fremgangsmåden med et eksempel: Bestem talværdien af $\binom{7}{2\ 3\ 2}$. Det søgte tal er antallet af forskellige måder hvorpå man kan placere tre forskellige symboler A_1, A_2 og A_3 på syv pladser, således at der er to A_1-er, tre A_2-er og to A_3-er. Een mulig placering er

$$A_1 \quad A_3 \quad A_1 \quad A_2 \quad A_2 \quad A_2 \quad A_3.$$

Vi kan bestemme en placering ved først at bestemme hvilke to pladser der skal have et A_1, dernæst at bestemme hvilke tre pladser der skal have et A_2, og så endelig placere et A_3 på de to tiloversblevne pladser.

Der er $\binom{7}{2} = 21$ forskellige placeringer af de to A_1-er, jf. definitionen af binomialkoefficienter; her er nogle af dem:

$$\begin{array}{ccccccc}
A_1 & A_1 & - & - & - & - & - \\
A_1 & - & A_1 & - & - & - & - \\
A_1 & - & - & A_1 & - & - & - \\
 & & \vdots & & & & \\
- & A_1 & A_1 & - & - & - & - \\
- & A_1 & - & A_1 & - & - & - \\
 & & \vdots & & & & \\
- & - & - & - & - & A_1 & A_1\ .
\end{array}$$

Hver gang vi har placerer de to A_1-er er der fem pladser tilbage, og på de fem pladser skal vi fordele tre A_2-er og to A_3-er; dette kan gøres på $\binom{5}{3} = 10$ forskellige måder. Hver gang vi har en af de $\binom{7}{2}$ placeringer af A_1 er der $\binom{5}{3}$ placeringer af A_2 og A_3; derfor er der i alt

$\binom{7}{2} \times \binom{5}{3}$ forskellige placeringer af A-erne. Alt i alt er

$$\begin{aligned}\binom{7}{2\,3\,2} &= \binom{7}{2} \times \binom{5}{3} \\ &= \binom{7}{2} \times \binom{5}{3} \\ &= 21 \times 10 \\ &= 210\,.\end{aligned}$$

Vi kan benytte formel (2.4) fra side 34 og få

$$\begin{aligned}\binom{7}{2\,3\,2} &= \binom{7}{2} \times \binom{5}{3} \\ &= \frac{7 \times 6}{1 \times 2} \times \frac{5 \times 4 \times 3}{1 \times 2 \times 3} \\ &= \frac{7 \times 6 \times 5 \times 4 \times 3}{(1 \times 2) \times (1 \times 2 \times 3)} \\ &= \frac{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(1 \times 2) \times (1 \times 2 \times 3) \times (1 \times 2)}\,.\end{aligned}$$

Man benytter tit betegnelsen $k!$ for $1 \times 2 \times 3 \times \ldots \times k$ [2]. Udtrykket for vores multinomialkoefficient kan så skrives

$$\binom{7}{2\,3\,2} = \frac{7!}{2!\,3!\,2!}\,.$$

Et generelt udtryk for koefficienten (5.1) fås på ganske tilsvarende måde. Man skal placere y_1 A_1-er, y_2 A_2-er, ..., y_r A_r-er på n pladser ($n = y_1 + y_2 + \ldots + y_r$). Først er der $\binom{n}{y_1}$ forskellige placeringer af A_1-erne. På de resterende $n - y_1 = y_2 + y_3 + \ldots + y_r$ pladser er der $\binom{n-y_1}{y_2}$ forskellige placeringer af A_2; dernæst er der $n - y_1 - y_2 = y_3 + y_4 + \ldots + y_r$

[2] $k!$ udtales 'k fakultet' eller 'k udråbstegn'. Funktionen $k \mapsto k!$ hedder *fakultetsfunktionen*.

pladser hvor der skal placeres y_3 A_3-er, osv. Slutresultatet er at

$$\binom{n}{y_1\ y_2\ \cdots\ y_r} = \frac{n!}{y_1!y_2!\ldots y_r!}$$

når $y_1 + y_2 + \ldots + y_r = n$.

Estimation af parametrene

I den generelle situation er modelfunktionen givet ved (5.2) og likelihoodfunktionen er dermed

$$L(p_1, p_2, \ldots, p_r) = \text{konstant} \times p_1^{y_1} p_2^{y_2} \ldots p_r^{y_r} .$$

Spørgsmålet er nu, hvordan man estimerer parametersættet

$$\boldsymbol{p} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_r \end{pmatrix}.$$

De almene principper for analyse af statistiske modeller (jf. Resumé 2) påbyder, at vi skal estimere parametersættet $(p_1, p_2, \ldots, p_r)$ ved det talsæt $(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_r)$ der maksimaliserer likelihoodfunktionen. Likelihoodfunktionen er en funktion af r variable $p_1, p_2, \ldots, p_r$ der ikke kan variere frit, men opfylder "bibetingelserne"

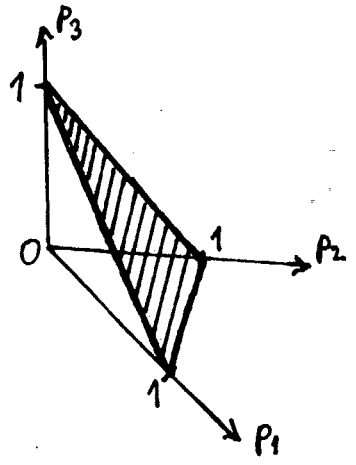
$$p_1 \geq 0,\ p_2 \geq 0,\ \ldots\ , p_r \geq 0,\ \sum_{i=1}^{r} p_r = 1 .$$

I specialtilfældet $r = 3$ kan vi anskueliggøre mulighedsområdet, dvs. mængden af $\boldsymbol{p}$-er der opfylder bibetingelserne, som et trekantet område, det såkaldte sandsynlighedssimplex, i det tredimensionale rum, se Figur 5.1.

Opgaven er at bestemme det punkt

$$\hat{\boldsymbol{p}} = \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \vdots \\ \hat{p}_r \end{pmatrix}$$

Figur 5.1: Sandsynlighedssimplexet i det tredimensionale rum.



i mulighedsområdet hvor likelihoodfunktionen L antager sin største værdi. I matematikken diskuteres generelle metode til bestemmelse af maksimumspunkter for funktioner af mange variable, men disse metoder skal vi ikke komme ind på her. Derimod vil vi løse det specielle problem der vedrører multinomialfordelingen. Dertil skal vi bruge følgende

Sætning 5.1 *Antag at $a_1, a_2, \ldots, a_r$ er givne ikke-negative tal, og betragt funktionen*

$$f : (x_1, x_2, \ldots, x_r) \mapsto x_1^{a_1} x_2^{a_2} \ldots x_r^{a_r}$$

defineret på mængden af ikke-negative talsæt $(x_1, x_2, \ldots, x_r)$ der summerer til 1.

Da gælder

1. *f antager sin maksimumsværdi i punktet $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_r)$ hvor $\hat{x}_i = a_i/a_\bullet$, og $a_\bullet = a_1 + a_2 + \ldots + a_r$,*

2. *hvis $(x_1, x_2, \ldots, x_r) \neq (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_r)$ så er $f(x_1, x_2, \ldots, x_r) < f(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_r)$.*

Bevis: Beviset for sætningen bygger (måske lidt overraskende) på at logaritmefunktionen har den egenskab at

$$\ln t \;\leq\; t-1 \text{ for alle } t>0, \tag{5.3}$$

jf. Figur 5.2. Endvidere gælder der lighedstegn i (5.3) hvis og kun hvis $t=1$. Nu indsætter vi i (5.3) $t=x_i/\widehat{x}_i$, hvor $\widehat{x}_i = a_i/a_\bullet$ og får

$$\ln\frac{x_i}{\widehat{x}_i} \;\leq\; \frac{x_i}{\widehat{x}_i}-1\,. \tag{5.4}$$

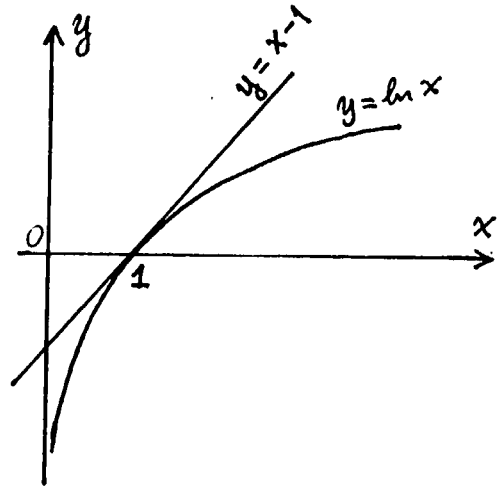
Gang så med $\widehat{x}_i$ på begge sider og summér:

$$\begin{aligned}\sum_{i=1}^{r}\widehat{x}_i\ln\frac{x_i}{\widehat{x}_i} \;&\leq\; \sum_{i=1}^{r}\widehat{x}_i\left(\frac{x_i}{\widehat{x}_i}-1\right) \\ &=\; \left(\sum_{i=1}^{r}x_i\right)-\left(\sum_{i=1}^{r}\widehat{x}_i\right) \\ &=\; 1-1 \\ &=\; 0\,.\end{aligned} \tag{5.5}$$

Endvidere kan venstresiden i (5.5) omskrives således:

$$\begin{aligned}\sum_{i=1}^{r}\widehat{x}_i\ln\frac{x_i}{\widehat{x}_i} \;&=\; \sum_{i=1}^{r}\widehat{x}_i\left(\ln x_i-\ln\widehat{x}_i\right) \\ &=\; \sum_{i=1}^{r}\widehat{x}_i\ln x_i-\sum_{i=1}^{r}\widehat{x}_i\ln\widehat{x}_i\end{aligned}$$

Figur 5.2: Graferne for $y=\ln x$ og $y=x-1$.



$$= \frac{1}{a_\cdot}\left(\sum_{i=1}^{r} a_i \ln x_i - \sum_{i=1}^{r} a_i \ln \hat{x}_i\right)$$

$$= \frac{1}{a_\cdot}\left(\sum_{i=1}^{r} \ln x_i^{a_i} - \sum_{i=1}^{r} \ln \hat{x}_i^{a_i}\right)$$

$$= \frac{1}{a_\cdot}\left(\ln f(x_1, x_2, \ldots, x_r) - \ln f(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_r)\right) .$$

Da vi netop har vist at dette altid er ≤ 0, er påstand 1 hermed vist.

Antag så at $(x_1, x_2, \ldots, x_r) \neq (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_r)$. Eftersom $\sum x_i = 1 = \sum \hat{x}_i$, må der være et indeksnummer i for hvilket $\hat{x}_i \neq 0$ og $x_i \neq \hat{x}_i$. For dette i er ulighedstegnet i (5.4) og dermed også i (5.5) skarpt, dvs. $f(x_1, x_2, \ldots, x_r) < f(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_r)$. Dermed er sætningen bevist.

Anvendt på funktionen

$$(p_1, p_2, \ldots, p_r) \mapsto p_1^{y_1} p_2^{y_2} \ldots p_r^{y_r}$$

fortæller sætningen, at likelihoodfunktionen L antager sit maksimum i et entydigt bestemt punkt, nemlig $(y_1/n,\ y_2/n,\ \ldots,\ y_r/n)$. Derfor:

Maksimaliseringsestimatet $\hat{\boldsymbol{p}}$ for $\boldsymbol{p}$ er givet ved

$$\hat{\boldsymbol{p}} = \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \vdots \\ \hat{p}_r \end{pmatrix} = \begin{pmatrix} y_1/n \\ y_2/n \\ \vdots \\ y_r/n \end{pmatrix} .$$

Parameteren p_i, der jo er sandsynligheden for at et individ tilhører klassen A_j, skal altså estimeres ved den relative hyppighed y_i/n af A_i-individer i stikprøven.

I taleksemplet vedrørende fordelingen på de tre basisuddannelser i 1974 bliver estimaterne over de tre sandsynlighedsparametre

$$\hat{p}_1 = 130/413 = 0.31 ,$$

$$\hat{p}_2 = 218/413 = 0.53 ,$$

$$\hat{p}_3 = 65/413 = 0.16 .$$

5.2 Sammenligning af multinomialfordelinger

Dette kapitel indledtes med at der blev præsenteret et spørgsmål der handler om at sammenligne fordelingerne i to forskellige år. Hidtil har vi set på, hvordan man kan opstille og analysere en model for fordelingen i et enkelt år. Hvis vi skal kunne sammenligne de to år, må det ske med udgangspunkt i én enkelt model for hele materialet (dvs. for de to år).

Når vi overhovedet stiller det spørgsmål, om der en forskel på fordelingen på basisuddannelser i de to år, er det ud fra en forestilling om, at det faktiske udfald Tabel 5.1 i et eller andet omfang er bestemt af tilfældigheder, således at resultatet lige så godt kunne have været lidt anderledes. Den statistiske model skal beskrive disse tilfældigheder. Det må være rimeligt at mene,

- at vi for hvert af årene kan benytte en multinomialfordelingsmodel, jf. diskussionen tidligere i kapitlet, og
- at hvad der sker i 1974 er stokastisk uafhængigt af det der sker i 1982.

Det betyder, at den samlede modelfunktion bliver et produkt af modelfunktionerne for hvert af de to år:

$$
\begin{aligned}
& f(y_{1,74}, y_{2,74}, y_{3,74}, y_{1,82}, y_{2,82}, y_{3,82};\ p_{1,74}, p_{2,74}, p_{3,74}, p_{1,82}, p_{2,82}, p_{3,82}) \\
&= f_{74}(y_{1,74}, y_{2,74}, y_{3,74}; p_{1,74}, p_{2,74}, p_{3,74}) \times \\
&\quad f_{82}(y_{1,82}, y_{2,82}, y_{3,82}; p_{1,82}, p_{2,82}, p_{3,82}) \\
&= \binom{413}{y_{1,74}\ y_{2,74}\ y_{3,74}} \binom{537}{y_{1,82}\ y_{2,82}\ y_{3,82}} \times \\
&\quad p_{1,74}^{y_{1,74}} p_{2,74}^{y_{2,74}} p_{3,74}^{y_{3,74}} \times p_{1,82}^{y_{1,82}} p_{2,82}^{y_{2,82}} p_{3,82}^{y_{3,82}} .
\end{aligned}
$$

Spørgsmålet om der er nogen forskel på styrkeforholdene mellem de tre basisuddannelser bliver i modellens sprog til den statistiske hypotese

$$H_0 : (p_{1,74}, p_{2,74}, p_{3,74}) = (p_{1,82}, p_{2,82}, p_{3,82}) .$$

Som vanligt behandler vi situationen generelt.

Modellen

Antag at vi har klassificeret nogle individer i r forskellige klasser $A_1, A_2, \ldots, A_r$. Individerne er på forhånd delt op i grupper, idet der er s forskellige grupper med hhv. $n_1, n_2, \ldots, n_s$ individer. Det har vist sig, at i gruppe j hører $y_{1,j}$ af individerne til gruppen A_1, $y_{2,j}$ af individerne til gruppen A_2, $y_{3,j}$ af individerne til gruppen A_3, osv. Skematisk ser situationen sådan ud:

	gruppe nr.				
klasse	1	2	3	...	s
A_1	y_{11}	y_{12}	y_{13}	$\cdots$	y_{1s}
A_2	y_{21}	y_{22}	y_{23}	$\cdots$	y_{2s}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
A_r	y_{r1}	y_{r2}	y_{r3}	$\cdots$	y_{rs}
i alt	n_1	n_2	n_3	...	n_s

Med andre ord, y_{ij} betegner antal individer fra gruppe j der tilhører klassen A_i. — I vort eksempel er der $s = 2$ grupper svarende til de to år og $r = 3$ klasser svarende til de tre forskellige basisuddannelser.

Den statistiske model der benyttes til at beskrive denne situation er:

- For hvert j (dvs. for hver gruppe) opfattes det r-dimensionale talsæt

$$\boldsymbol{y}_j = \begin{pmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{rj} \end{pmatrix}$$

som en observeret værdi af en r-dimensional stokastisk variabel

$$\boldsymbol{Y}_j = \begin{pmatrix} Y_{1j} \\ Y_{2j} \\ \vdots \\ Y_{rj} \end{pmatrix} ;$$

- De stokastiske variable $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_s$ er stokastisk uafhængige (dvs. de forskellige grupper er stokastisk uafhængige).

- $\boldsymbol{Y}_j$ er multinomialfordelt med antalsparameter n_j og ukendt sandsynlighedsparameter

$$\boldsymbol{p}_j = \begin{pmatrix} p_{1j} \\ p_{2j} \\ \vdots \\ p_{rj} \end{pmatrix}$$

hvor p_{ij}-erne er ikke-negative tal med $p_{1j} + p_{2j} + \ldots + p_{rj} = 1$ for hvert j.

Modellen tager altså udgangspunkt i at grupperne er systematisk forskellige (mht. den aktuelle klassificering), og den beskriver *den systematiske variation mellem grupperne* ved hjælp af de s sandsynlighedsparametre $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_s$. *Den tilfældige variation inden for grupper* beskrives ved sandsynlighedsfordelingerne (multinomialfordelingerne).

Opgaven er nu at undersøge om grupperne kan anses for ens, dvs. den er at teste den statistiske hypotese

$$H_0 : \boldsymbol{p}_1 = \boldsymbol{p}_2 = \ldots = \boldsymbol{p}_s ,$$

eller mere udførligt

$$H_0 : \begin{pmatrix} p_{11} \\ p_{21} \\ \vdots \\ p_{r1} \end{pmatrix} = \begin{pmatrix} p_{12} \\ p_{22} \\ \vdots \\ p_{r2} \end{pmatrix} = \ldots = \begin{pmatrix} p_{1s} \\ p_{2s} \\ \vdots \\ p_{rs} \end{pmatrix} .$$

De generelle retningslinier for hvordan man analyserer en given statistisk model siger, at vi skal begynde med at opskrive modelfunktionen og derudaf få likelihoodfunktionen. Da de enkelte grupper er stokastisk uafhængige, er den samlede modelfunktion lig med et produkt af delmodelfunktionerne for de enkelte grupper, dvs. *den samlede modelfunktion* er

$$f = \prod_{j=1}^{s} \binom{n_j}{y_{1j}\ y_{2j}\ \cdots\ y_{rj}} p_{1j}^{y_{1j}} p_{2j}^{y_{2j}} \ldots p_{rj}^{y_{rj}} .$$

Likelihoodfunktionen er dermed

$$L(\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_s) = \text{konstant} \times \prod_{j=1}^{s} p_{1j}^{y_{1j}} p_{2j}^{y_{2j}} \ldots p_{rj}^{y_{rj}} , \qquad (5.6)$$

hvor konstanten er produktet af de s multinomialkoefficienter. I taleksemplet er likelihoodfunktionen altså

$$L(\boldsymbol{p}_{74}, \boldsymbol{p}_{82}) \;=\; \text{konstant} \;\times p_{11}^{130} p_{21}^{218} p_{31}^{65} p_{12}^{158} p_{22}^{247} p_{32}^{132} \,.$$

Likelihoodfunktionen er sandsynligheden for at observere det faktisk observerede, betragtet som funktion af det ukendte sæt parametre. Som sædvanlig er det bedste estimat over de ukendte parametres værdier de værdier der maksimaliserer likelihoodfunktionen (eller log-likelihoodfunktionen). Nu er likelihoodfunktionen et produkt af s del-likelihoodfunktioner der hver især vedrører én enkelt gruppe og ét enkelt $\boldsymbol{p}_j$. Når vi skal maksimalisere L mht. $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_s$ kan det derfor ske ved at maksimalisere hver del-likelihoodfunktion for sig. Det j-te delproblem er en simpel multinomialfordelingsmodel, så derfor følger det uden videre af resultatet på side 88 at

$$\widehat{p}_{ij} \;=\; \frac{y_{ij}}{n_j} \,.$$

I taleksemplet er specielt

$$\widehat{\boldsymbol{p}}_{74} = \begin{pmatrix} \widehat{p}_{1,74} \\ \widehat{p}_{2,74} \\ \widehat{p}_{3,74} \end{pmatrix} \;=\; \begin{pmatrix} 130/413 \\ 218/413 \\ 65/413 \end{pmatrix} = \begin{pmatrix} 0.31 \\ 0.53 \\ 0.16 \end{pmatrix} ,$$

$$\widehat{\boldsymbol{p}}_{82} = \begin{pmatrix} \widehat{p}_{1,82} \\ \widehat{p}_{2,82} \\ \widehat{p}_{3,82} \end{pmatrix} \;=\; \begin{pmatrix} 158/537 \\ 247/537 \\ 132/537 \end{pmatrix} = \begin{pmatrix} 0.29 \\ 0.46 \\ 0.25 \end{pmatrix} .$$

Hypoteseprøvning

Vi skal herefter undersøge, om det er rimeligt at antage at hypotesen $H_0 \;:\; \boldsymbol{p}_1 = \boldsymbol{p}_2 = \ldots = \boldsymbol{p}_s$ om ens sandsynlighedsparametre holder. Under H_0 er der ingen forskel på de s grupper, så da kan vi lige så godt slå dem sammen til én stor gruppe bestående af $n_\bullet = n_1 + n_2 + \ldots + n_s$

individer, der fordeler sig med

$$\begin{aligned}
y_{1\cdot} &= y_{11} + y_{12} + \ldots + y_{1s} &&= \textstyle\sum_{j=1}^{s} y_{1j} && \text{i klassen } A_1 ,\\
y_{2\cdot} &= y_{21} + y_{22} + \ldots + y_{2s} &&= \textstyle\sum_{j=1}^{s} y_{2j} && \text{i klassen } A_2 ,\\
&\vdots &&\vdots && \\
y_{i\cdot} &= y_{i1} + y_{i2} + \ldots + y_{is} &&= \textstyle\sum_{j=1}^{s} y_{ij} && \text{i klassen } A_i ,\\
&\vdots &&\vdots && \\
y_{r\cdot} &= y_{r1} + y_{r2} + \ldots + y_{rs} &&= \textstyle\sum_{j=1}^{s} y_{rj} && \text{i klassen } A_r .
\end{aligned}$$

Man må derfor formode, at den fælles værdi p_i af sandsynligheden for at tilhøre klassen A_i skal estimeres ved $y_{i\cdot}/n_{\cdot}$, men lad os prøve at gå frem efter likelihoodmetoden.

Vi kalder den fælles værdi (under H_0) af $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_s$ for $\boldsymbol{p}$,

$$\boldsymbol{p} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_r \end{pmatrix} .$$

I likelihoodfunktionen (5.6) erstatter vi alle $\boldsymbol{p}_j$-erne med $\boldsymbol{p}$ og får derved *likelihoodfunktionen under* H_0:

$$\begin{aligned}
L(\boldsymbol{p}, \boldsymbol{p}, \ldots, \boldsymbol{p}) &= \text{konstant} \times \prod_{j=1}^{s} p_1^{y_{1j}} p_2^{y_{2j}} \ldots p_r^{y_{rj}} \\
&= \text{konstant} \times p_1^{y_{1\cdot}} p_2^{y_{2\cdot}} \ldots p_r^{y_{r\cdot}} .
\end{aligned}$$

Det valg af $p_1, p_2, \ldots, p_r$ der maksimaliserer denne likelihoodfunktion er ifølge sætningen på side 86 netop $\hat{p}_i = y_{i\cdot}/n_{\cdot}$ som formodet.

I taleksemplet bliver

$$\hat{\boldsymbol{p}} = \begin{pmatrix} 288/950 \\ 465/950 \\ 197/950 \end{pmatrix} = \begin{pmatrix} 0.30 \\ 0.49 \\ 0.21 \end{pmatrix} .$$

Når vi vil vurdere hvor godt det faktisk observerede kan beskrives under H_0 i forhold til den bedst mulige beskrivelse vi kan få med den

foreliggende grundmodel, skal vi udregne *kvotienttesstørrelsen*

$$Q = \frac{L(\hat{\boldsymbol{p}}, \hat{\boldsymbol{p}}, \ldots, \hat{\boldsymbol{p}})}{L(\hat{\boldsymbol{p}}_1, \hat{\boldsymbol{p}}_2, \ldots, \hat{\boldsymbol{p}}_s)}$$

eller $-2\ln Q$. En Q-værdi tæt på 1, dvs. en $-2\ln Q$-værdi tæt på 0, betyder at H_0 beskriver data næsten lige så godt som grundmodellen gør, hvorimod en Q-værdi nær 0, dvs. en stor $-2\ln Q$-værdi, betyder at H_0 giver en væsentlig dårligere beskrivelse end grundmodellen gør. Man plejer at udregne $-2\ln Q$ (og ikke Q).

Når man indsætter udtrykkene for L i Q får man let at

$$\begin{aligned} -2\ln Q &= 2\sum_{j=1}^{s}\left(y_{1j}\ln\frac{y_{1j}}{\hat{y}_{1j}} + y_{2j}\ln\frac{y_{2j}}{\hat{y}_{2j}} + \ldots + y_{rj}\ln\frac{y_{rj}}{\hat{y}_{rj}}\right) \\ &= 2\sum_{j=1}^{s}\sum_{i=1}^{r} y_{ij}\ln\frac{y_{ij}}{\hat{y}_{ij}} \end{aligned} \tag{5.7}$$

hvor

$$\begin{aligned} \hat{y}_{ij} &= \hat{p}_i n_j \\ &= y_{i\cdot} n_j / n_{\cdot} \end{aligned} \tag{5.8}$$

er det "forventede" antal individer fra gruppe j der klassificeres som A_i.

For i taleksemplet at bestemme $-2\ln Q$ udregnes først de "forventede" antal ved brug af (5.8). Man får

	1974 antal	1982 antal
HUM	125.2	162.8
SAM	202.2	262.8
NAT	85.6	111.4
i alt	413.0	537.0

og dermed

$$\begin{aligned} -2\ln Q_{\text{obs}} &= 2\Big(130\ln\frac{130}{125.2} + 218\ln\frac{218}{202.2} + 65\ln\frac{65}{85.6} \\ &\quad +158\ln\frac{158}{162.8} + 247\ln\frac{247}{262.8} + 132\ln\frac{132}{111.4}\Big) \\ &= 11.5\,. \end{aligned}$$

For at afgøre om en opnået $-2\ln Q_{\text{obs}}$-værdi (som f.eks. 11.5) nu er tæt på 0 eller ej skal vi sammenligne den med alle de andre $-2\ln Q$-værdier man også kunne have fået ifølge den aktuelle model når H_0 er rigtig. Vi skal derfor finde *testsandsynligheden* ε, dvs. sandsynligheden for at få en værre (større) $-2\ln Q$-værdi end den observerede, under forudsætning af at H_0 er rigtig:

$$\varepsilon \;=\; \mathrm{P}_0\left(-2\ln Q \geq -2\ln Q_{\text{obs}}\right) .$$

Når man skal bestemme ε, kan man udnytte, at der findes en generel matematisk sætning der fortæller, at når H_0 er rigtig, så er $-2\ln Q$ med god tilnærmelse χ^2-fordelt med $(r-1)(s-1)$ frihedsgrader, således at ε med god tilnærmelse kan bestemmes som sandsynligheden for at få en værdi større end $-2\ln Q_{\text{obs}}$ i en χ^2-fordeling med $(r-1)(s-1)$ frihedsgrader, kort

$$\varepsilon \;=\; \mathrm{P}\left(\chi^2_{(r-1)(s-1)} \geq -2\ln Q_{\text{obs}}\right) ,$$

og denne sandsynlighed er let at bestemme ved hjælp af tabeller over fraktiler i χ^2-fordelingen.

Antallet af frihedsgrader for $-2\ln Q$ findes som *ændringen i antallet af frie parametre*: i grundmodellen er der for hver af de s grupper $r-1$ parametre (fordi der er r klasser og dermed r sandsynligheder der skal summere til 1), altså i alt $s(r-1)$ parametre; under H_0 er der i realiteten kun én gruppe og dermed $r-1$ frie parametre; antallet af frihedsgrader for teststørrelsen er dermed $s(r-1)-(r-1) = (r-1)(s-1)$.

Bemærk at χ^2-fordelingen kun er en approksimation; for at man skal kunne bruge den *skal alle de "forventede" antal* (5.8) *være mindst fem.* Hvis denne betingelse ikke er opfyldt, kan man måske opnå at den bliver opfyldt ved at man udelader nogle grupper eller klasser eller slår nogle grupper eller klasser sammen.

I vores gennemgående taleksempel er der ingen problemer med at nogle af de "forventede" antal er for små. Vi kan derfor uden videre sammenligne $-2\ln Q_{\text{obs}} = 11.5$ med χ^2-fordelingen med $(3-1)(2-1) = 2$ frihedsgrader. Da 99.5%-fraktilen i denne fordeling er 10.6, er testsandsynligheden ε mindre end 0.5%. Da det således er temmelig usandsynligt at få en værre værdi af teststørrelsen $-2\ln Q$ end 11.5, er teststørrelsen *signifikant* og vi forkaster H_0. Man må altså sige, at hvad angår fordelingen på de tre basisuddannelser er der en signifikant forskel på de to årgange. Hvis man vil have en idé om, *hvad*

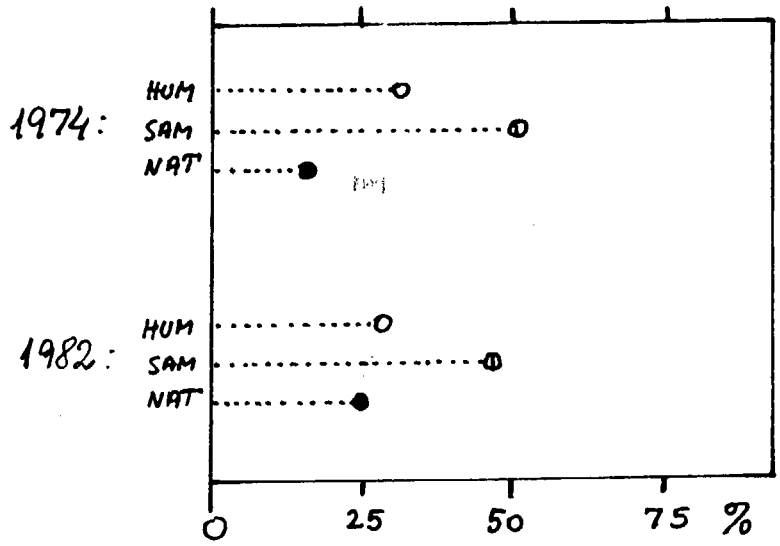
Tabel 5.2: Fordeling efter basisuddannelse for årgangene 1974 og 1982, **absolutte tal**. De observerede værdier sammenholdt med de værdier der måtte forventes under antagelsen om, at forholdet mellem basisuddannelserne er det samme i de to år.

	1974		1982	
	obs.	forv.	obs.	forv.
HUM	130	125.2	158	162.8
SAM	218	202.2	247	262.8
NAT	65	85.6	132	111.4
i alt	413	413.0	537	537.0

Tabel 5.3: Fordeling efter basisuddannelse for årgangene 1974 og 1982, **relative tal**. De observerede værdier sammenholdt med de værdier der måtte forventes under antagelsen om, at forholdet mellem basisuddannelserne er det samme i de to år.

	1974		1982	
	obs.%	forv.%	obs.%	forv.%
HUM	31	30	29	30
SAM	53	49	46	49
NAT	16	21	25	21
i alt	100	100	100	100

Figur 5.3: Fordeling efter basisuddannelse for årgangene 1974 og 1982, relative tal, jf. Tabel 5.3.



forskellen består i, kan man se på tallene i Tabel 5.2 og 5.3 eller man kan se på en tegning som f.eks. Figur 5.3.

Resumé 4. Sammenligning af multinomialfordelinger

Situation: n_j individer fra gruppe j er klassificeret i r klasser:

		gruppe nr.				sum
	A_1	y_{11}	y_{12}	$\dots$	y_{1s}	$y_{1\bullet}$
	A_2	y_{21}	y_{22}	$\dots$	y_{2s}	$y_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
klasse	A_i	y_{i1}	y_{i2}	$\dots$	y_{is}	$y_{i\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	A_r	y_{r1}	y_{r2}	$\dots$	y_{rs}	$y_{r\bullet}$
	i alt	n_1	n_2	$\dots$	n_s	$n_\bullet$

Model: De enkelte søjler af y-er er uafhængige observationer fra multinomialfordelinger, således at multinomialfordelingen hørende til søjle nr. j har antalsparameter n_j og sandsynlighedsparametre $p_{1j}, p_{2j}, \dots, p_{rj}$.

p_{ij}-erne er ukendte parametre således at $p_{1j}+p_{2j}+\dots+p_{rj} = 1$ for alle j.

Estimation: p_{ij} estimeres ved den observerede relative hyppighed af A_i i gruppe j, dvs.

$$\hat{p}_{ij} = \frac{y_{ij}}{n_j}.$$

Hypotese: Man ønsker at teste hypotesen

$$H_0 : \begin{pmatrix} p_{11} \\ p_{21} \\ \vdots \\ p_{r1} \end{pmatrix} = \begin{pmatrix} p_{12} \\ p_{22} \\ \vdots \\ p_{r2} \end{pmatrix} = \dots = \begin{pmatrix} p_{1s} \\ p_{2s} \\ \vdots \\ p_{rs} \end{pmatrix}$$

om at der ikke er forskel på søjlerne.

Teststørrelse: Under H_0 er den "forventede" situation

		gruppe nr.				sum
	A_1	$\hat{y}_{11}$	$\hat{y}_{12}$	$\dots$	$\hat{y}_{1s}$	$y_{1\bullet}$
	A_2	$\hat{y}_{21}$	$\hat{y}_{22}$	$\dots$	$\hat{y}_{2s}$	$y_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
klasse	A_i	$\hat{y}_{i1}$	$\hat{y}_{i2}$	$\dots$	$\hat{y}_{is}$	$y_{i\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	A_r	$\hat{y}_{r1}$	$\hat{y}_{r2}$	$\dots$	$\hat{y}_{rs}$	$y_{r\bullet}$
	i alt	n_1	n_2	$\dots$	n_s	$n_\bullet$

hvor $\widehat{y}_{ij} = y_{i\bullet} n_j / n_{\bullet\bullet}$.
Kvotienttesteststørrelsen er

$$-2 \ln Q \;=\; 2 \sum_{j=1}^{s} \sum_{i=1}^{r} y_{ij} \ln \frac{y_{ij}}{\widehat{y}_{ij}} \; .$$

Testsandsynlighed:

1. Hvis alle de rs "forventede" antal er mindst fem, kan testsandsynligheden ε med god tilnærmelse findes som sandsynligheden for at få en større værdi end $-2 \ln Q_{\text{obs}}$ i χ^2-fordelingen med $(r-1)(s-1)$ frihedsgrader:

$$\varepsilon \;=\; \mathrm{P}\left(\chi^2_{(r-1)(s-1)} \geq -2 \ln Q_{\text{obs}}\right) \; .$$

2. I modsat fald må man prøve at slå nogle smågrupper sammen eller at slå nogle klasser sammen el.lgn. for at opnå at 1 bliver opfyldt.

Konklusion: Hvis ε er meget lille, så er der en signifikant afvigelse mellem det observerede og det som H_0 foreskriver, og man vil da forkaste H_0. I modsat fald er H_0 forenelig med det observerede, og man kan ikke forkaste H_0.

Kapitel 6

Tosidede kontingenstabeller

En af pointerne i kapitlet om multinomialfordelingen (Kapitel 5) er, at når man klassificerer et antal individer (fra en eller anden population) efter ét kriterium med r klasser $A_1, A_2, \ldots, A_r$, så kan det være fornuftigt at forsøge sig med en model der siger, at hvis Y_i betegner antallet af A_i-individer i stikprøven, $i = 1, 2, \ldots, r$, så er $(Y_1, Y_2, \ldots, Y_r)$ multinomialfordelt.

I dette kapitel skal vi se, hvorledes en bestemt art *struktur i inddelingskriteriet* kan afspejle sig i den statistiske model. Den struktur der er tale om, er, at inddelingskriteriet rent faktisk består i, at man inddeler efter *to* kriterier på én gang. Inden vi går i gang, kommer en præsentation af det talmateriale der benyttes som gennemgående eksempel i dette kapitel.

Eksempel 6.1. ***Hjernesvulstpatienter***

141 hjernesvulstpatienter er blevet klassificeret efter svulstens art *(*GODARTET, ONDARTET, ANDET*) og* placering *i hjernevævet (*VED PANDEN, VED TINDINGEN, ANDRE STEDER*). Resultaterne heraf fremgår af Tabel 6.1. Man er interesseret i at finde ud af, om disse tal tyder på, at der er en sammenhæng mellem svulstens art og placering.* □

Tabel 6.1: 141 hjernesvulstpatienter fordelt efter svulstens art og placering.

		placering			
		pande	tinding	andet	sum
art	godartet	23	21	34	78
	ondartet	9	4	24	37
	andet	6	3	17	26
	sum	38	28	75	141

Eksemplet går ud på, at man har klassificeret $n = 141$ patienter[1] som hørende til én af ni forskellige klasser. Ifølge overvejelserne i Kapitel 5 kan man da betragte det observerede talsæt (23, 21, ..., 17) som en observation af en multinomialfordelt stokastisk variabel. Imidlertid kan man også tænke på situationen på den måde, at patienterne er klassificeret efter to kriterier på én gang, hvor hvert kriterium har tre niveauer. Den generelle beskrivelse kommer derfor til at se ud på følgende måde.

Grundmodellen

Antag at vi har klassificeret n individer efter *to* kriterier. Det første kriterium har r niveauer og klasserne $A_1, A_2, \ldots, A_r$, det andet har s niveauer og klasserne $B_1, B_2, \ldots, B_s$. Skematisk ser situationen sådan ud:

		kriterium 2				
	klasse	B_1	B_2	...	B_s	sum
kriterium 1	A_1	y_{11}	y_{12}	...	y_{1s}	$y_{1\bullet}$
	A_2	y_{21}	y_{22}	...	y_{2s}	$y_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	A_r	y_{r1}	y_{r2}	...	y_{rs}	$y_{r\bullet}$
	sum	$y_{\bullet 1}$	$y_{\bullet 2}$	...	$y_{\bullet s}$	n

hvor

$$y_{ij} = \text{antal individer i klassen } A_iB_j \ (= A_i \cap B_j),$$

[1] fra en forhåbentlig ensartet gruppe.

$$y_{i\bullet} = \sum_{j=1}^{s} y_{ij} = \text{ antal individer i klassen } A_i,$$

$$y_{\bullet j} = \sum_{i=1}^{r} y_{ij} = \text{ antal individer i klassen } B_j.$$

Da der er tale om, at nogle individer er klassificeret i et antal grupper, bruger vi som grundmodel en multinomialfordelingsmodel: Den rs-dimensionale observation

$$\boldsymbol{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{rs} \end{pmatrix}$$

er en observeret værdi af en rs-dimensional stokastisk variabel

$$\boldsymbol{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{rs} \end{pmatrix}$$

som er multinomialfordelt med antalsparameter n og sandsynlighedsparameter

$$\boldsymbol{p} = \begin{pmatrix} p_{11} \\ p_{12} \\ \vdots \\ p_{rs} \end{pmatrix}.$$

Derved betegner p_{ij} sandsynligheden for at et individ udvalgt tilfældigt fra "populationen" vil tilhøre klassen A_iB_j. Størrelsen p_{ij} estimeres ved

$$\widehat{p}_{ij} = y_{ij}/n\,. \tag{6.1}$$

Uafhængighedshypotesen

Den struktur der er i inddelingskriteriet (nemlig at der inddeles efter to kriterier på en gang) har foreløbig kun givet sig udslag i den måde

de variable og parametrene er navngivet på (med index ij). Vi skal nu udlede en model der går ud på, at der ikke er nogen sammenhæng mellem de to inddelingskriterier.

Den "sammenhæng" der skal være tale om er ikke en årsags-sammenhæng, men en statistisk sammenhæng. At der ikke er nogen sammenhæng mellem kriterium A og kriterium B skal betyde, at A og B i en vis forstand "virker" uafhængigt af hinanden, således at forstå at en oplysning om, hvilken B-klasse et individ tilhører ikke indeholder nogen information om, hvilken A-klasse individet tilhører, og omvendt. Det må vi oversætte til matematik for at kunne se hvad det betyder. Vi indfører[2] nogle hjælpevariable $\boldsymbol{X}_d = (X_{dA}, X_{dB})$, således at X_{dA} er navnet på den A-klasse som individ nr. d tilhører, og tilsvarende for X_{dB} , dvs.

> $\boldsymbol{X}_d = (A_i, B_j)$ hvis og kun hvis individ nr. d tilhører A-klassen A_i *og* B-klassen B_j.

At der ikke er nogen sammenhæng mellem A og B betyder hermed, at en oplysning om værdien af X_{dB} ikke indeholder nogen information om værdien af X_{dA} (og omvendt), og det betyder, at *de stokastiske variable X_{dA} og X_{dB} er stokastisk uafhængige.* At X_{dA} og X_{dB} er stokastisk uafhængige betyder at

$$\begin{aligned} &\mathrm{P}(X_{dA} = A_i,\ X_{dB} = B_j) \\ &\quad = \mathrm{P}(X_{dA} = A_i) \times \mathrm{P}(X_{dB} = B_j)\ . \end{aligned}$$

Nu er pr. definition $\mathrm{P}(X_{dA} = A_i,\ X_{dB} = B_j) = p_{ij}$, så at der ikke er nogen sammenhæng mellem A og B betyder altså, at $p_{ij} = \alpha_i \times \beta_j$ hvor $\alpha_i = \mathrm{P}(X_{dA} = A_i)$ og $\beta_j = \mathrm{P}(X_{dB} = B_j)$. Sammenfattende kan vi derfor sige, at antagelsen om, at der ikke er nogen (statistisk) sammenhæng mellem kriterierne A og B oversat til matematik beløber sig til at

$$p_{ij} = \alpha_i\beta_j \quad \text{for alle } i \text{ og } j,$$

hvor $\sum_{i=1}^{r} \alpha_i = \sum_{j=1}^{s} \beta_j = 1$.

[2] i lighed med side 81

I stedet for at tale om, at der ikke er nogen sammenhæng mellem A og B, taler man ofte om, at der er *uafhængighed* mellem A og B, og den statistiske hypotese

$$H_0 : p_{ij} = \alpha_i \beta_j \quad \text{for alle } i \text{ og } j,$$

hvor de ukendte parametre $(\alpha_1, \alpha_2, \ldots, \alpha_r)$ og $(\beta_1, \beta_2, \ldots, \beta_s)$ er ikkenegative talsæt der hver især summerer til 1, hedder da *uafhængighedshypotesen.*

At der er uafhængighed mellem A og B udtrykker man undertiden på den måde, at der ikke er nogen (signifikant) *vekselvirkning* mellem A og B. Når der ikke er nogen vekselvirkning mellem A og B, beskrives hele den *systematiske variation* i talmaterialet ved hjælp af *række-virkningerne* (A-virkningerne) $\alpha_1, \alpha_2, \ldots, \alpha_r$, der beskriver den systematiske forskel mellem rækker, og ved hjælp af *søjlevirkningerne* (B-virkningerne) $\beta_1, \beta_2, \ldots, \beta_s$, der beskriver den systematiske forskel mellem søjler.

Estimation af parametrene

Likelihoodfunktionen i grundmodellen er en almindelig multinomial-likelihoodfunktion:

$$L(\boldsymbol{p}) \;=\; \text{konstant} \times \prod_{i=1}^{r} \prod_{j=1}^{s} p_{ij}^{y_{ij}} \,, \tag{6.2}$$

hvor konstanten er en multinomialkoefficient.

De bedste skøn over parametrene $\alpha_1, \alpha_2, \ldots, \alpha_r$ og $\beta_1, \beta_2, \ldots, \beta_s$ i uafhængighedsmodellen er de værdier der maksimaliserer $L(\boldsymbol{p})$ hvor man for p_{ij} indsætter $p_{ij} = \alpha_i \beta_j$, dvs. de værdier der maksimaliserer

$$\begin{aligned} & L_0(\alpha_1, \alpha_2, \ldots, \alpha_r, \beta_1, \beta_2, \ldots, \beta_s) \\ &= \text{konstant} \times \prod_{i=1}^{r} \prod_{j=1}^{s} (\alpha_i \beta_j)^{y_{ij}} \\ &= \text{konstant} \times \prod_{i=1}^{r} \prod_{j=1}^{s} \alpha_i^{y_{ij}} \prod_{i=1}^{r} \prod_{j=1}^{s} \beta_j^{y_{ij}} \\ &= \text{konstant} \times \prod_{i=1}^{r} \alpha_i^{y_{i\bullet}} \times \prod_{j=1}^{s} \beta_j^{y_{\bullet j}} \,. \end{aligned}$$

Tabel 6.2: Skønnene over grundmodellens parametre p_{ij} og uafhængighedsmodellens parametre α_i og β_j i hjernesvulsteksemplet. Tallene er sandsynligheder i procent.

		placering			sum =
		pande	tinding	andet	$\widehat{\alpha}_i$
art	godartet	14.9	11.0	29.4	55.3
	ondartet	7.1	5.2	14.0	26.2
	andet	5.0	3.7	9.8	18.4
sum = $\widehat{\beta}_j$		27.0	19.9	53.2	100.0

Det ses at L_0 er et produkt af en funktion af α-erne og en funktion af β-erne. Ifølge Sætning 5.1 antager disse to funktioner deres maksimumsværdier i hhv.

$$(\widehat{\alpha}_1, \widehat{\alpha}_2, \ldots, \widehat{\alpha}_r) = \left(\frac{y_{1\bullet}}{n}, \frac{y_{2\bullet}}{n}, \ldots, \frac{y_{r\bullet}}{n}\right) \tag{6.3}$$

og

$$(\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_s) = \left(\frac{y_{\bullet 1}}{n}, \frac{y_{\bullet 2}}{n}, \ldots, \frac{y_{\bullet s}}{n}\right) . \tag{6.4}$$

Hermed har vi bestemt maksimaliseringestimaterne over parametrene $\alpha_1, \alpha_2, \ldots, \alpha_r, \beta_1, \beta_2, \ldots, \beta_s$. Resultatet er i øvrigt hvad man umiddelbart skulle forvente, idet f.eks. sandsynligheden α_i for at tilhøre A-klassen A_i estimeres ved den observerede relative hyppighed $y_{i\bullet}/n$ af A_i.

I taleksemplet bliver

$$L = \text{konstant} \times p_{11}^{23} p_{12}^{21} p_{13}^{34} p_{21}^{9} p_{22}^{4} p_{23}^{24} p_{31}^{6} p_{32}^{3} p_{33}^{17} .$$

Ved at indsætte de aktuelle talværdier i (6.1), (6.3) og (6.4) fås estimaterne over de ukendte parametre, se Tabel 6.2.

Test for uafhængighed

Teststørrelsen for uafhængighedshypotesen H_0 er likelihoodkvotientstørrelsen Q eller $-2\ln Q$. Når man indsætter de fundne estimater i

udtrykket for Q får man

$$\begin{aligned} Q &= \frac{L_0(\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_r, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_s)}{L(\hat{p}_{11}, \hat{p}_{12}, \ldots, \hat{p}_{rs})} \\ &= \frac{\prod_{i=1}^{r}\prod_{j=1}^{s}\left(\hat{\alpha}_i\hat{\beta}_j\right)^{y_{ij}}}{\prod_{i=1}^{r}\prod_{j=1}^{s}(\hat{p}_{ij})^{y_{ij}}} \\ &= \prod_{i=1}^{r}\prod_{j=1}^{s}\left(\frac{\hat{\alpha}_i\hat{\beta}_j}{\hat{p}_{ij}}\right)^{y_{ij}} \\ &= \prod_{i=1}^{r}\prod_{j=1}^{s}\left(\frac{\hat{y}_{ij}}{y_{ij}}\right)^{y_{ij}}, \end{aligned}$$

hvor

$$\hat{y}_{ij} = n\hat{\alpha}_i\hat{\beta}_j = \frac{y_{i\cdot}y_{\cdot j}}{n} \tag{6.5}$$

er det "forventede" antal individer i klassen A_iB_j under uafhængighedshypotesen. Dermed bliver

$$-2\ln Q = 2\sum_{i=1}^{r}\sum_{j=1}^{s} y_{ij}\ln\frac{y_{ij}}{\hat{y}_{ij}}.$$

$-2\ln Q$-værdier tæt på 0 tyder på, at H_0 giver en næsten lige så god beskrivelse af data som grundmodellen gør, hvorimod store $-2\ln Q$-værdier betyder, at H_0 giver en væsentlig dårligere beskrivelse end grundmodellen gør, og i så fald vil man forkaste hypotesen om uafhængighed mellem rækker og søjler.

De "forventede" antal i hjernesvulsteksemplet er vist i Tabel 6.3; herudfra får man

$$\begin{aligned} -2\ln Q_{\text{obs}} &= 2\Big(23\ln\frac{23}{21.0} + 21\ln\frac{21}{15.5} + 34\ln\frac{34}{41.5} \\ &\quad +9\ln\frac{9}{10.0} + 4\ln\frac{4}{7.3} + 24\ln\frac{24}{19.7} \\ &\quad +6\ln\frac{6}{7.0} + 3\ln\frac{3}{5.2} + 17\ln\frac{17}{13.8}\Big) \\ &= 8.1\,. \end{aligned}$$

Tabel 6.3: Den "forventede" fordeling af 141 hjernesvulstpatienter under forudsætning af uafhængighed mellem svulstens art og placering.

		placering			
		pande	tinding	andet	sum
art	godartet	21.0	15.5	41.5	78
	ondartet	10.0	7.3	19.7	37
	andet	7.0	5.2	13.8	26
	sum	38.0	28.0	75.0	141

Når vi skal afgøre om en opnået $-2\ln Q_{\text{obs}}$-værdi (som f.eks. 8.1) er signifikant stor, skal vi sammenligne den med alle de andre $-2\ln Q$-værdier man også kunne have fået såfremt uafhængighedshypotesen H_0 var rigtig. Vi skal derfor bestemme *testsandsynligheden* ε, dvs. sandsynligheden for at få en værre (større) $-2\ln Q$-værdi end den observerede, under forudsætning af at H_0 er rigtig:

$$\varepsilon \;=\; \mathrm{P}_0\big(-2\ln Q \geq -2\ln Q_{\text{obs}}\big)\ .$$

Når man skal bestemme ε, kan man udnytte, at der findes en generel matematisk sætning der fortæller, at når H_0 er rigtig så er $-2\ln Q$ med god tilnærmelse χ^2-fordelt med $(r-1)(s-1)$ frihedsgrader, således at ε med god tilnærmelse kan bestemmes som sandsynligheden for at få en værdi større end $-2\ln Q_{\text{obs}}$ i en χ^2-fordeling med $(r-1)(s-1)$ frihedsgrader, kort

$$\varepsilon \;=\; \mathrm{P}\left(\chi^2_{(r-1)(s-1)} \geq -2\ln Q_{\text{obs}}\right)\ .$$

Denne sandsynlighed er let at bestemme ved hjælp af tabeller over fraktiler i χ^2-fordelingen.

Antallet af frihedsgrader for $-2\ln Q$ findes som *ændringen i antallet af frie parametre*: i grundmodellen er der rs sandsynlighedsparametre der summerer til 1, dvs. der er $rs-1$ frie parametre; under H_0 er der r rækkeparametre der summerer til 1 plus s søjleparametre der summerer til 1, dvs. $(r-1)+(s-1)$ frie parametre; antallet af frihedsgrader for teststørrelsen er dermed

$$\big(rs-1\big) - \big((r-1)+(s-1)\big) \;=\; (r-1)(s-1)\ .$$

Bemærk at χ^2-fordelingen kun er en approksimation; for at man skal kunne bruge den *skal alle de "forventede" antal* (6.5) *være mindst fem*. Hvis denne betingelse ikke er opfyldt, kan man eventuelt slå nogle rækker eller nogle søjler sammen.

I hjernesvulsteksemplet er de "forventede" antal over fem, så vi kan roligt anvende χ^2-approksimationen. I χ^2-fordelingen med $(3-1)(3-1) = 4$ frihedsgrader er 90%-fraktilen 7.78 og 95%-fraktilen 9.49, således at teststørrelsen $-2\ln Q_{\text{obs}} \doteq 8.1$ svarer til en testsandsynlighed på mellem 5% og 10%. På det grundlag vil man sædvanligvis *ikke* forkaste H_0. Det kan altså konkluderes at der tilsyneladende ikke er nogen sammenhæng mellem svulstens *art* og dens *placering*. Det vil (bl.a.) sige, at man ikke ud fra kendskab til *placeringen* af en svulst kan sige noget om, hvorvidt den vil være godartet eller ej.

Jævnføring med andre tilsvarende modeller

Den læser der har studeret Kapitel 5 vil måske have bemærket, at metoderne i det kapitel har store ligheder med dem i indeværende kapitel. Vi kan opregne nogle ligheder:

1. Der foreligger nogle observerede *antal* y_{ij} anbragt i et tosidet skema.

2. Man udregner nogle "forventede" antal $\hat{y}_{ij}$ efter opskriften *rækkesum* gange *søjlesum* divideret med *totalsum*.

3. Man udregner en teststørrelse $-2\ln Q_{\text{obs}} = \sum y \ln(y/\hat{y})$.

4. Man sammenligner $-2\ln Q_{\text{obs}}$ med χ^2-fordelingen med $(r-1)(s-1)$ frihedsgrader.

Selv om man *gør* det samme i de to tilfælde, er det dog på grundlag af to forskellige modeller[3].

[3]De to modeller er dog nært beslægtede; hvis man i dette kapitels model betinger med søjlesummerne, dvs. betinger med at $Y_{\cdot 1} = n_1$, $Y_{\cdot 2} = n_2$, ..., $Y_{\cdot s} = n_s$, så får man modellen i Kapitel 5, og uafhængighedshypotesen overføres til Kapitel 5s H_0.

- I det ene tilfælde (dette kapitel) klassificerer man nogle individer efter *to* kriterier, og opgaven er da at undersøge om der er en sammenhæng mellem disse to kriterier.

- I det andet tilfælde (Kapitel 5) er individerne på forhånd delt ind i nogle grupper inden de klassificeres efter *et* kriterium. Opgaven er da at undersøge om der er forskel på grupperne (med hensyn til hvordan gruppernes individer fordeles på klasserne).

Om man skal benytte den ene eller den anden model er således et spørgsmål om, hvorledes man har designet selve det forsøg der har leveret talmaterialet. I eksemplet i dette kapitel sagde vi, at det handlede om at man havde taget 141 hjernesvulstpatienter og klassificeret dem efter *to* kriterier; derved blev det et eksempel der illustrerede dette kapitels model og metoder. Hvis det derimod havde handlet om, at man havde taget 38 patienter med svulst i panden, 28 med svulst i tindingen og 75 hvor svulsten ikke var lokaliseret til pande eller tinding, og dernæst klassificeret disse patienter efter svulstens art, så havde det været et Kapitel 5-eksempel.

Resumé 5. Uafhængighedstest i en $r \times s$-tabel

Situation: n individer er klassificeret efter to kriterier:

		kriterium 2				
	klasse	B_1	B_2	...	B_s	sum
kriterium 1	A_1	y_{11}	y_{12}	...	y_{1s}	$y_{1\bullet}$
	A_2	y_{21}	y_{22}	...	y_{2s}	$y_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	A_r	y_{r1}	y_{r2}	...	y_{rs}	$y_{r\bullet}$
	sum	$y_{\bullet 1}$	$y_{\bullet 2}$	...	$y_{\bullet s}$	n

hvor

$$y_{ij} \quad = \quad \text{antal individer i klassen } A_iB_j (= A_i \cap B_j).$$

Model: De rs værdier y_{ij} udgør tilsammen en observation fra en multinomialfordeling med rs klasser, med antalsparameter n og med sandsynlighedsparametre p_{ij}, $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, s$.

p_{ij}-erne er ukendte parametre der summerer til 1: $p_{11} + p_{12} + \ldots + p_{rs} = 1$.

Estimation: p_{ij} estimeres ved den observerede relative hyppighed af A_iB_j, dvs.

$$\widehat{p}_{ij} \quad = \quad \frac{y_{ij}}{n}.$$

Hypotese: Man ønsker at teste uafhængighedshypotesen (hypotesen om forsvindende vekselvirkning)

$$H_0 : \; p_{ij} = \alpha_i \times \beta_j \quad \text{for alle } i \text{ og } j,$$

hvor de ukendte parametre $(\alpha_1, \alpha_2, \ldots, \alpha_r)$ og $(\beta_1, \beta_2, \ldots, \beta_s)$ er ikke-negative talsæt der hver især summerer til 1.

Teststørrelse: Under H_0 er den "forventede" situation

		kriterium 2				
	klasse	B_1	B_2	...	B_s	sum
kriterium 1	A_1	$\widehat{y}_{11}$	$\widehat{y}_{12}$	...	$\widehat{y}_{1s}$	$y_{1\bullet}$
	A_2	$\widehat{y}_{21}$	$\widehat{y}_{22}$	...	$\widehat{y}_{2s}$	$y_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	A_r	$\widehat{y}_{r1}$	$\widehat{y}_{r2}$	...	$\widehat{y}_{rs}$	$y_{r\bullet}$
	sum	$y_{\bullet 1}$	$y_{\bullet 2}$	...	$y_{\bullet s}$	n

hvor $\widehat{y}_{ij} = y_{i\bullet}y_{\bullet j}/y_{\bullet\bullet}$.
Kvotienttesteststørrelsen er

$$-2\ln Q \;=\; 2\sum_{j=1}^{s}\sum_{i=1}^{r} y_{ij}\ln\frac{y_{ij}}{\widehat{y}_{ij}} \,.$$

Testsandsynlighed:

1. Hvis alle de rs "forventede" antal er mindst fem, kan testsandsynligheden ε med god tilnærmelse findes som sandsynligheden for at få en større værdi end $-2\ln Q_{\text{obs}}$ i χ^2-fordelingen med $(r-1)(s-1)$ frihedsgrader:

$$\varepsilon \;=\; \mathrm{P}\left(\chi^2_{(r-1)(s-1)} \geq -2\ln Q_{\text{obs}}\right) \,.$$

2. I modsat fald må man prøve at slå nogle rækker eller søjler sammen for at opnå at 1. bliver opfyldt.

Konklusion: Hvis ε er meget lille, så er der en signifikant afvigelse mellem det observerede og det som H_0 foreskriver, og man vil da forkaste H_0. I modsat fald er H_0 forenelig med det observerede, og man kan ikke forkaste H_0.

Kapitel 7

Poissonfordelingen

Der er i de forrige kapitler givet eksempler på, hvorledes man kan beskrive antals-observationer ved hjælp af binomial- og multinomialfordelingsmodeller. I nogle situationer er det imidlertid mere hensigtsmæssigt at benytte Poissonfordelingsmodeller.

Dette kapitel introducerer Poissonfordelingen[1]. Talmaterialet til det gennemgående eksempel er meget berømt[2] og kan måske i første omgang forekomme lidt kuriøst.

Eksempel 7.1. *Hestespark*

For hvert af de 20 år fra 1875 til 1894 har man for hvert af den preussiske armés 10 regimenter registreret, hvor mange soldater der døde fordi de blev sparket af en hest. Det vil sige, at man for hvert af de 200 "regiment-år" kender antal dødsfald som følge af hestespark.

Man kan give en oversigt over disse tal ved at fortælle, i hvor mange regiment-år der var 0 dødsfald, i hvor mange der var 1 dødsfald, i hvor mange der var 2, osv., dvs. man klassificerer regiment-årene efter antal dødsfald. Man kunne jo ikke på forhånd vide hvor mange klasser der skal være, man det viste sig, at der rent faktisk

[1] opkaldt efter franskmanden S.-D. Poisson (1781-1840).

[2] idet det optræder i næsten alle lærebøger i statistik,

ikke i noget regiment og i noget år var mere end fire dødsfald. Oversigten over de faktiske tal ser nemlig således ud:

antal dødsfald y	antal regiment-år med y dødsfald
0	109
1	65
2	22
3	3
4	1
	200

Man må formode, at det i høj grad var tilfældigheder der bestemte, om en given soldat blev sparket til døde af en hest eller ej. Derfor er det også i høj grad tilfældigheder der har afgjort, om et givet regiment i et givet år nu fik 0 eller 1 eller 2 osv. døde som følge af hestespark. Det kan være anledningen til at man at formulerer sig denne modelbygningsopgave:

> *Find et fornuftigt forslag til en statistisk model der kan levere sandsynligheder for at have netop y døde i et bestemt regiment, $y = 0, 1, 2, \ldots$.*

Det er denne opgave der skal løses i dette kapitel. □

En del af problemløsningsprocessen består i at oversætte problemet til matematik i en passende generel formulering. Vi går frem i en række skridt, der dels leder frem til en passende formulering, dels leverer en løsning på problemet.

1. Hestespark-eksemplet handler om, at man 200 gange har foretaget sig noget bestemt, nemlig fulgt et regiment igennem et år og set hvor mange dødsfald der var.

2. "Grund-eksperimentet" består i, at man i et vist tidsinterval (af længde 1 år) holder øje med hvor mange gange en bestemt art begivenhed (dødsfald ved hestespark) indtræffer.

3. Grund-eksperimentet består i at der i tidsintervallet fra t_0 til t_1 registreres antal forekomster af en bestemt art begivenhed.

4. Vi kan dele intervallet fra t_0 til t_1 op i et antal lige store delintervaller, som hver især har længden Δt. På den måde bliver der

$$n = n(\Delta t) = \frac{t_1 - t_0}{\Delta t}$$

delintervaller. (I hestesparkeksemplet kan man f.eks. dele intervallet $]t_0, t_1]$ af længde 1 år op i 365 delintervaller af længde $\Delta t = 1$ dag.)

Antallet af begivenheder i det store interval er (selvfølgelig) lig med summen af antal begivenheder i de enkelte delintervaller.

5. Fidusen ved at dele op i delintervaller er, at hvis Δt er tilstrækkelig lille, så er det meget usandsynligt at der indtræffer to eller flere begivenheder i *samme* delinterval. Sagt på en anden måde, hvis Δt er meget lille, så er det samlede antal begivenheder i intervallet $]t_0, t_1]$ stort set altid lig med antallet af delintervaller hvori der forekommer mindst én begivenhed.

6. Vi har nu fået lavet problemet om til noget der handler om 01-variable, nemlig om variablene

$$I_j = \begin{cases} 1 & \text{hvis mindst én begivenhed i delinterval nr. } j \\ 0 & \text{hvis ingen begivenhed i delinterval nr. } j \end{cases},$$

$j = 1, 2, \ldots, n$. Hvis Δt er meget lille, så er det samlede antal Y af begivenheder i $]t_0, t_1]$ ca. lig med $I_1 + I_2 + \ldots + I_n$.

7. Antag så at der i alle $n = n(\Delta t)$ delintervaller er den samme sandsynlighed $p = p(\Delta t)$ for at der sker begivenheder. (Der bliver altså ikke i løbet af perioden indført nye sikkerhedsforanstaltninger der nedsætter chancen for at blive sparket af en hest og dø af det. Og antallet af soldater og af heste i regimentet er stort set konstant året igennem.) Antag også at det der sker i ét interval er *stokastisk uafhængigt* af det der sker i andre intervaller. (Hvis der *tilfældigvis* var to soldater der i begyndelsen af året blev sparket til døde af heste, så tager de øvrige soldater i regimentet *ikke* i den anledning ekstra forholdsregler i resten af året.)

8. Det følger nu af Kapitel 2 (jf. side 28) at $\sum_{j=1}^{n} I_j$ er binomialfordelt med parametre $n = n(\Delta t)$ og $p = p(\Delta t)$, og da totalantallet Y af begivenheder i $]t_0, t_1]$ er cirka lig med $\sum_{j=1}^{n} I_j$, er Y således

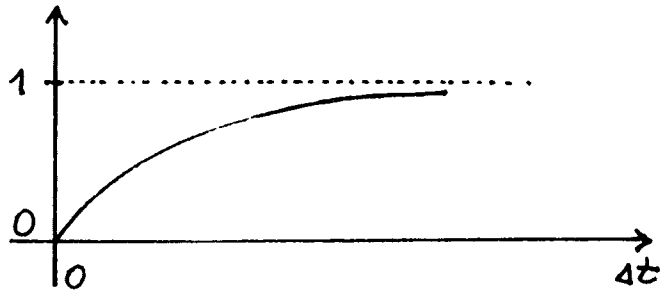
cirka binomialfordelt med parametre n og p. Forbeholdet "cirka" bortfalder når Δt bliver mindre og mindre, dvs. vi skal på et senere stadium lade $\Delta t \to 0$.

9. Den måde hvorpå $n = n(\Delta t)$ afhænger af Δt er simpel, idet som tidligere anført

$$n = n(\Delta t) = \frac{t_1 - t_0}{\Delta t} .$$

Derimod kender vi ikke p's afhængighed af Δt.

Det må være rimeligt at formode, at p er en forholdsvis pæn funktion af Δt, bl.a. med den egenskab at $p(\Delta t) \to 0$ når $t \to 0$ og at $p(\Delta t) \to 1$ når $t \to +\infty$. $p(\Delta t)$ må have et udseende i retning af



Vi vil gå ud fra, at $p(\Delta t)$ er differentiabel fra højre i $\Delta t = 0$, dvs. at der eksisterer et tal $\lambda \geq 0$ således at

$$\lim_{\Delta t \to 0} \frac{p(\Delta t)}{\Delta t} = \lambda . \tag{7.1}$$

Der gælder altså at $p(\Delta t) \approx \lambda \Delta t$ for små værdier af Δt.

10. I punkt 8 nåede vi frem til at Y er cirka binomialfordelt, dvs. at

$$\mathrm{P}(Y = y) \approx \binom{n}{y} p^y (1-p)^{n-y} , \tag{7.2}$$

hvor "$\approx$" bliver til "$=$" når $\Delta t \to 0$. Derfor må det næste skridt være at bestemme

$$\lim \binom{n}{y} p^y (1-p)^{n-y}$$

under den grænseovergang hvor $\Delta t \to 0$ og dermed $n = (t_1 - t_0)/\Delta t \to \infty$. Som følge af (7.1) vil der under denne grænseovergang gælde $p/\Delta t = p(\Delta t)/\Delta t \to \lambda$ og dermed også $np = (t_1 - t_0)p/\Delta t \to \lambda \times (t_1 - t_0)$.

Vi omskriver binomialsandsynligheden på følgende måde (hvorved (2.4) på side 34 udnyttes)

$$\begin{aligned}
&\binom{n}{y} p^y (1-p)^{n-y} \\
&= \frac{n}{1} \times \frac{n-1}{2} \times \ldots \times \frac{n-y+1}{y} \times p^y \times (1-p)^{-y} \times (1-p)^n \\
&= 1 \times (1 - \tfrac{1}{n}) \times \ldots \times (1 - \tfrac{y-1}{n}) \times \frac{(np)^y}{y!} \times (1-p)^{-y} \times (1-p)^n .
\end{aligned}$$

Under grænseovergangen vil

(a) $\underbrace{1 \times (1 - \frac{1}{n}) \times \ldots \times (1 - \frac{y-1}{n})}_{y \text{ faktorer}} \to \underbrace{1 \times 1 \times \ldots \times 1}_{y \text{ faktorer}} = 1.$

(b) $\dfrac{(np)^y}{y!} \to \dfrac{\left(\lambda \times (t_1 - t_0)\right)^y}{y!}.$

(c) $(1-p)^{-y} \to (1-0)^{-y} = 1\ .$

(d) $(1-p)^n \to \exp(-\lambda(t_1 - t_0))$, hvilket indses på følgende måde:

i. Da $x \mapsto \ln x$ er differentiabel i $x = 1$ med differentialkvotient 1, vil for $p \to 0$

$$\begin{aligned}
\frac{\ln(1-p)}{p} &= \frac{\ln(1-p) - \ln 1}{p} \\
&\to -1\ .
\end{aligned}$$

ii. Derfor vil

$$\begin{aligned}
n \ln(1-p) &= \lambda(t_1 - t_0) \times \frac{np}{\lambda(t_1 - t_0)} \times \frac{n \ln(1-p)}{np} \\
&\to \lambda(t_1 - t_0) \times 1 \times (-1) \\
&= -\lambda \times (t_1 - t_0)\ .
\end{aligned}$$

iii. Ved at tage exp på begge sider heraf fås, at

$$(1-p)^n \;\to\; \exp\Big(-\lambda(t_1-t_0)\Big)$$

som ønsket.

Alt i alt vil binomialsandsynligheden (7.2) derfor konvergere mod

$$\frac{\big(\lambda(t_1-t_0)\big)^y}{y!}\exp\Big(-\lambda(t_1-t_0)\Big).$$

Vi er hermed nået frem til følgende forslag til en statistisk model: Sandsynligheden for at der i et bestemt regiment er netop y dødsfald i perioden $]t_0,t_1]$ må være

$$\mathrm{P}(Y=y) \;=\; \frac{\big(\lambda(t_1-t_0)\big)^y}{y!}\exp\big(-\lambda(t_1-t_0)\big), \qquad (7.3)$$

hvor λ er en positiv konstant og $y = 0,1,2,3,\ldots$. - Bemærk at de hjælpestørrelser n og Δt som vi indførte i 4. helt er forsvundet.

I (7.3) optræder der den ukendte parameter λ, der i (7.1) blev indført som værende ca. "sandsynligheden for en begivenhed i et meget kort tidsinterval divideret med tidsintervallets længde". λ har derfor dimensionen tid^{-1}, dvs. λ angives i f.eks. dag^{-1} eller år^{-1}.

Jo større λ er, jo tilbøjeligere er begivenhederne til at indtræffe; λ er en såkaldt *intensitet* (der i hestespark-eksemplet specielt kunne kaldes for en *ulykkesintensitet* eller en *dødsintensitet*)[3].

Man definerer Poissonfordelingen således:

Definition:

Poissonfordelingen med parameter $\mu \geq 0$ er den sandsynlighedsfordeling på udfaldsrummet $\mathcal{X} = \{0,1,2,\ldots\}$ som har sandsynlighedsfunktion

$$f(y;\mu) \;=\; \frac{\mu^y}{y!}\exp(-\mu y)\,.$$

[3]Antagelsen 4 går ud på, at begivenhederne indtræffer med *samme* tilbøjelighed, med samme intensitet, overalt på (den betragtede del af) tidsaksen. Der er imidlertid ikke noget i vejen for at konstruere mere indviklede modeller hvor intensiteten er tidsafhængig, dvs. $\lambda = \lambda(t)$.

Det fundne resultat kan derefter udtrykkes på den måde, at antallet af dødsfald i et bestemt regiment i perioden fra t_0 til t_1 er Poissonfordelt med parameter $\lambda \times (t_1 - t_0)$, hvor λ betegner dødsintensiteten.

Strengt taget bør definitionen følges op af en redegørelse for, at $f(y;\mu)$ faktisk *er* en sandsynlighedsfunktion, dvs. at $f(y;\mu) \geq 0$ og $\sum_{y \in \mathcal{X}} f(y;\mu) = 1$. Det er klart at f-erne er ikke-negative; at de summerer til 1 følger af eksponentialfunktionens rækkeudvikling

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

som ikke vil blive bevist her.

En anden ting som heller ikke bliver bevist er, at

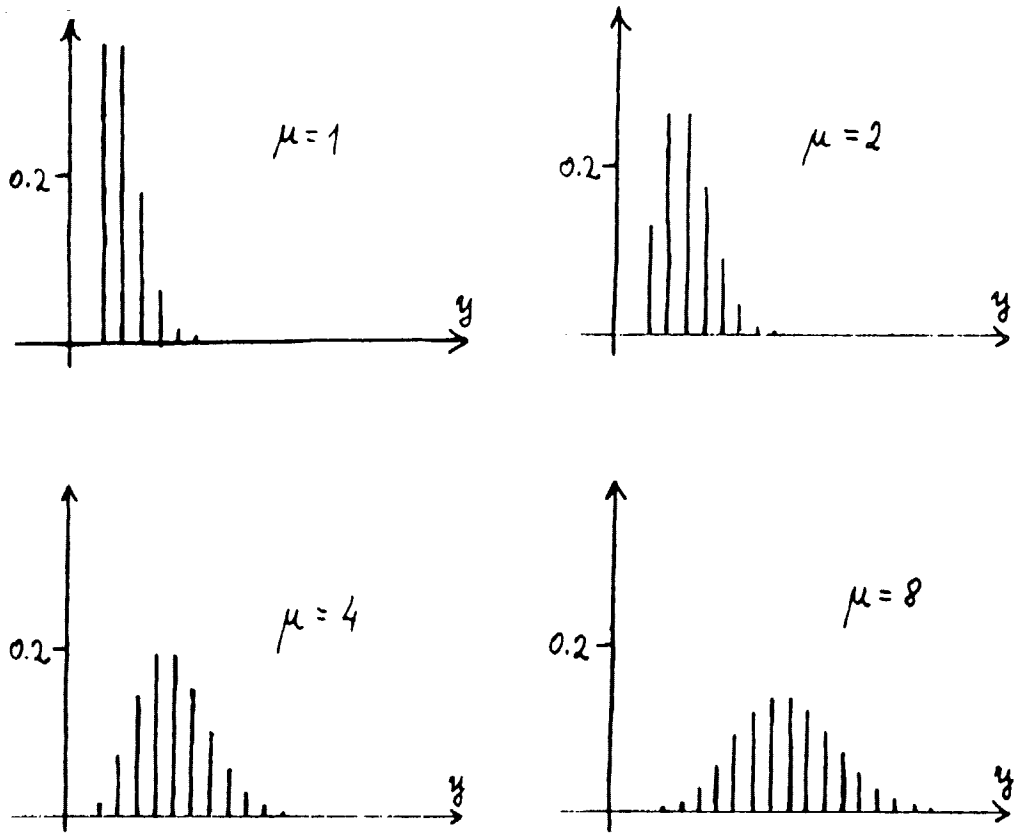
> Hvis den stokastiske variabel Y er Poissonfordelt med parameter μ så er
>
> $$\begin{aligned} \mathrm{E}\, Y &= \mu \\ \mathrm{Var}\, Y &= \mu\,, \end{aligned}$$
>
> dvs. både middelværdien og variansen af Y er lig μ.

Figur 7.1 viser nogle Poissonfordelinger.

Enstikprøveproblemet i Poisson-fordelingen

Ved ganske teoretiske overvejelser nåede vi frem til, at antallet af dødsfald i et regiment i et år måtte være Poissonfordelt med parameter $\mu = \lambda \times 1$ år, men passer det overhovedet med virkeligheden?

Figur 7.1: Poissonfordelinger.



Estimation af parameteren

Vi vil først estimere den ukendte intensitetsparameter λ og dernæst undersøge hvor god en beskrivelse af talmaterialet vi får ved dens hjælp.

Da der er 200 regimentår, er situationen altså den, at der er $n = 200$ uafhængige observationer $y_1, y_2, \ldots, y_n$ fra Poissonfordelingen med parameter $\mu = \lambda \times 1$ år. Den generelle situation er således, at der foreligger uafhængige observationer $y_1, y_2, \ldots, y_n$ fra en Poissonfordeling med parameter μ, svarende til at *modelfunktionen* er

$$\begin{aligned} f(y_1, y_2, \ldots, y_n; \mu) &= \prod_{j=1}^{n} \frac{\mu^{y_j}}{y_j!} \exp(-\mu) \\ &= \frac{\mu^{\sum_{j=1}^{n} y_j}}{\prod_{j=1}^{n} y_j!} \exp(-n\mu) \, . \end{aligned}$$

Likelihoodfunktionen er dermed

$$L(\mu) = \frac{\mu^{\sum_{j=1}^{n} y_j}}{\text{konstant}} \exp(-n\mu) \, ,$$

så at

$$\ln L(\mu) = \left(\sum_{j=1}^{n} y_j \right) \ln \mu - n\mu + \text{konstant.}$$

Ifølge de sædvanlige principper er det bedste skøn over μ den værdi $\hat{\mu} = \hat{\mu}(y_1, y_2, \ldots, y_n)$ der maksimaliserer L eller $\ln L$. For at bestemme denne værdi finder vi den afledede af $\ln L$ og løser ligningen $\frac{d}{d\mu} \ln L = 0$. Man får at $\frac{d}{d\mu} \ln L(\mu)$ er lig med

$$\frac{\sum_{j=1}^{n} y_j}{\mu} - n \, ,$$

som er lig 0 netop når μ er lig $\overline{y} = \sum_{j=1}^{n} y_j / n$. Funktionen $\ln L$ har altså stationært punkt i $\mu = \overline{y}$, og da dens anden afledede

$$\frac{d^2}{d\mu^2} \ln L(\mu) = -\frac{\sum_{j=1}^{n} y_j}{\mu^2}$$

altid er negativ er $\overline{y}$ et maksimumspunkt. Dermed er vist[4] at maksimaliseringsestimatet er

$$\widehat{\mu} \;=\; \overline{y} \;=\; \frac{1}{n}\sum_{j=1}^{n} y_j \;. \tag{7.4}$$

I taleksemplet får man

$$\sum_{j=1}^{200} y_j = 0 \times 109 + 1 \times 65 + 2 \times 22 + 3 \times 3 + 4 \times 1 = 122 \;,$$

så at $\widehat{\mu} = 122/200 = 0.61$ og dermed

$$\widehat{\lambda} \;=\; \frac{\widehat{\mu}}{1 \text{ år}} \;=\; 0.61 \text{ år}^{-1} \;,$$

dvs. dødsintensiteten er 0.61 dødsfald pr. år for hvert regiment. Det ses at $\widehat{\lambda}$ fremkommer som antal dødsfald divideret med antal regiment-år.

Kontrol af modellen

Når vi holder os inden for klassen af Poissonfordelingsmodeller får vi den bedste beskrivelse ved at bruge intensiteten $\widehat{\lambda} = 0.61$ dødsfald pr. år for hvert regiment.

For at få et fingerpeg om, hvor god denne "bedste beskrivelse" er, udregner vi nogle "forventede" antal under forudsætning af at modellen er rigtig. Modellens beskrivelse går ud på, at sandsynligheden for, at der i et bestemt regiment-år er netop y dødsfald, er

$$f(y;\widehat{\lambda}) \;=\; \frac{(\widehat{\lambda} \times 1 \text{ år})^y}{y!} \exp(-\widehat{\lambda} \times 1 \text{ år}) \;.$$

Ud af i alt 200 regiment-år skulle man derfor forvente ca. $200 \times f(0;\widehat{\lambda})$ tilfælde med 0 dødsfald, ca. $200 \times f(1;\widehat{\lambda})$ tilfælde med 1 dødsfald, ca. $200 \times f(2;\widehat{\lambda})$ tilfælde med 2 dødsfald, osv. Disse forventede tal udregnes[5] og man får Tabel 7.1. Det ses at de "forventede" antal stemmer fint overens med de observerede, og det må vi tage som tegn på, at Poissonmodellen faktisk ikke er helt hen i vejret.

[4]Det er i udledningen forudsat at $\sum y_j \neq 0$. Hvis $\sum y_j = 0$ er log-likelihoodfunktionen lig

$$-n\mu + \text{konstant} \;,$$

og den antager sit maksimum når $\mu = 0$.

[5]Ved udregning af Poissonsandsynligheder kan man med fordel udregne dem successivt, startende med $y = 0$: $f(0;\mu) = \exp(-\mu)$ og $f(y+1;\mu) = \frac{\mu}{y+1} f(y;\mu)$.

Tabel 7.1: Hestespark-eksemplet. De observerede antal år med y dødsfald sammenlignet med de "forventede" antal år med y dødsfald beregnet ud fra Poissonmodellen.

antal dødsfald y	observeret antal år	"forventet" antal år
0	109	108.7
1	65	66.3
2	22	20.2
3	3	4.1
4	1	0.6
5+	0	0.1
	200	200.0

Afrunding

Vi afrunder kapitlet med et resumé over, hvad det var for omstændigheder der førte til en Poissonmodel for antal dødsfald pr. regiment-år. Se i øvrigt også side 160-163.

Resumé 6. Betingelser for en Poissonmodel

1. Det der observeres er *antallet* af en bestemt slags begivenheder i et bestemt tidsinterval.
2. Der indtræffer aldrig to (eller flere) begivenheder samtidig.
3. Begivenhederne indtræffer uafhængigt af hinanden, således at forstå at det der indtræffer f.eks. i tidsintervallet $]a, b]$ er stokastisk uafhængigt af hvad der sker uden for dette interval.
4. Sandsynligheden for at der indtræffer en begivenhed i et tidsinterval $]t, t + \Delta t]$ af længde Δt afhænger *ikke* af, hvor på tidsaksen intervallet er beliggende, dvs. den afhænger ikke af t (sålænge vi befinder os inden for det overordnede tidsinterval der i det hele taget er tale om).
5. Hvis $p(\Delta t)$ betegner sandsynligheden for at der indtræffer mindst en begivenhed i et tidsinterval af længde Δt, så vil

$$\lim_{\Delta t \searrow 0} \frac{p(\Delta t)}{\Delta t} = \lambda ,$$

hvor λ er en positiv konstant kaldet *intensiteten*.

Under disse omstændigheder vil antallet af begivenheder i tidsintervallet $]a, b]$ være Poissonfordelt med parameter $\mu = \lambda \times (b - a)$.

Her er nogle flere eksempler på situationer der kan give Poissonfordelte antal:

- Antal tilfælde af en betemt (ikke-smittende) sygdom i et bestemt tidsrum.
- Antal ulykkestilfælde af en bestemt art i et bestemt tidsrum.
- Antal omdannelser af atomer i et radio-aktivt stof i et bestemt tidsrum (der er forsvindende i forhold til stoffets halveringstid).
- Antal trykfejl i en bog. – Her er "tidsaksen" simpelthen teksten forstået som en tegnsekvens. Denne "tidsakse" er en diskret tidsakse, og ræsonnementerne der førte frem til Poissonfordelingen beror i høj grad på at tidsaksen er kontinuert. Men hvis der kun er få trykfejl i forhold til antallet af bogstaver og tegn, så kan man "næsten ikke se" at tidsaksen faktisk er diskret. Derfor finder man på at anvende Poissonfordelingen.
- Antal bombenedfald i London under det tyske bombardement under Anden Verdenskrig – her er "tidsaksen" det (todimensionale) geografiske område London.

Kapitel 8

Statistisk analyse af Poissonfordelte observationer

Dette kapitel indeholder to eksempler på statistisk analyse af Poissonfordelte observationer. Det ene eksempel viser, hvorledes man sammenligner to Poissonfordelinger, det andet er et eksempel på en såkaldt multiplikativ Poissonmodel.

8.1 Sammenligning af to Poissonfordelinger

Eksempel 8.1. ***Ultralydsscanning***

Det er meget udbredt at foretage ultralydsscanning af gravide kvinder. Det menes/frygtes imidlertid, at fostrene kan lide skade derved, idet der måske kan ske kromosomforandringer. For at undersøge dette nærmere har man udført en række laboratorieforsøg med mus[1].

[1] L. Meillier og I. Toldbod: *På skærmen står et lille hjerte og banker... Ultralyd og biologiske skadevirkninger - afprøvet for kromosombrud i mikrokernetesten.* Biologispecialerapport, RUC, 1985.

Et antal drægtige mus udsættes for ultralydsbestråling i et vist stykke tid, hvorefter man undersøger leverceller fra fostrene for at se, om der er dannet såkaldte mikrokærne-celler. Mikrokærner i en celle opstår som følge af kromosomforandringer/ødelæggelser.

I dette eksempel (der kun behandler en del af forsøgets talmateriale) optræder to grupper à tre mus: en behandlingsgruppe og en kontrolgruppe. Behandlingsgruppen har fået ultralyd, hvorefter man har ladet gå 18 timer inden musen er dræbt og prøverne udtaget. Kontrolgruppen er behandlet på samme måde, på nær at der denne gang ikke blev tændt for ultralydapparatet. Fra hver mus udtog man otte prøver; i alt undersøgte man for hver mus ca. 2000 celler og afgjorde, om det var en mikrokærnecelle eller ej. Derved fremkom resultaterne i Tabel 8.1.

Spørgsmålet er nu, om disse tal tyder på, at ultralyd har en skadelig virkning. □

Modelopstilling

For hver mus er der øjensynlig *to* størrelser der er uforudsigelige, nemlig antal optalte celler n og antal mikrokærneceller y. Når vi skal formulere den statistiske model, skal vi tage stilling til, om både n og y eller kun den ene af dem skal opfattes som observation af en stokastisk variabel. De størrelser der opfattes som udfald af stokastiske variable er de størrelser, for hvilke den statistiske model påtager sig at beskrive hvilke andre udfald man også kunne have fået.

I den foreliggende problemstilling er det der er genstand for den grundlæggende interesse formentlig chancen for at en celle omdannes til en mikrokærnecelle. I den forbindelse er det uinteressant at søge at opstille en model der kan påtage sig at beskrive variationen i antal optalte celler pr. mus. Derimod er det interessant at formulere en model der kan beskrive variationen i antallet af mikrokærneceller i en prøve af en given størrelse. I modellen skal n-erne derfor indgå som givne konstanter og y-erne som udfald af stokastiske variable.

Da der for en enkelt mus optælles et meget stort antal celler der hver især har en meget lille chance for at være blevet omdannet til en mikrokærnecelle, kan vi (jf. trykfejlseksemplet side 124) antage, at antal mikrokærneceller i en prøve med n celler er Poissonfordelt med

Tabel 8.1: Resultater af mikrokærne-tællinger.

1. Behandlingsgruppen

Mus nr.	Antal optalte celler n	Antal mikrokærne-celler y
1	2096	1
2	2138	10
3	2086	7
sum	6320	18

2. Kontrolgruppen

Mus nr.	Antal optalte celler n	Antal mikrokærne-celler y
1	2077	2
2	2181	6
3	2030	2
sum	6288	10

parameter $\mu = \lambda \times n$, hvor λ er en "omdannelsesintensitet", nemlig sandsynligheden pr. optalt celle for at registrere en mikrokærnecelle. Den systematiske forskel mellem behandlingsgrupperne skal beskrives ved hjælp af intensitetsparametre, så derfor skal mus med samme behandling have samme intensitet λ, hvorimod behandlingsgruppen og kontrolgruppen skal have hver sit λ.

Vi indfører lidt notation for at kunne formulere modellen præcist: Lad

$$n_{ij} = \text{antal optalte celler fra mus nr. } j \text{ i gruppe } i,$$

$$y_{ij} = \text{antal mikrokærneceller fra mus nr. } j \text{ i gruppe } i,$$

hvor $i = 1$ svarer til behandlingsgruppen og $i = 2$ til kontrolgruppen. Det vil sige, at Tabel 8.1 skematisk ser således ud:

	$i = 1$	
1	n_{11}	y_{11}
2	n_{12}	y_{12}
3	n_{13}	y_{13}
	$n_{1\bullet}$	$y_{1\bullet}$

	$i = 2$	
1	n_{21}	y_{21}
2	n_{22}	y_{22}
3	n_{23}	y_{23}
	$n_{2\bullet}$	$y_{2\bullet}$

Modellen er da, at tallene $y_{11}, y_{12}, y_{13}, y_{21}, y_{22}, y_{23}$ opfattes som observerede værdier af stokastisk uafhængige Poissonfordelte stokastiske variable $Y_{11}, Y_{12}, Y_{13}, Y_{21}, Y_{22}, Y_{23}$, hvor Y_{ij} har parameter $\mu_{ij} = \lambda_i n_{ij}$. Her er λ_1 og λ_2 ukendte parametre der beskriver den systematiske forskel mellem behandlingsgruppen og kontrolgruppen. *Modelfunktionen* er

$$\prod_{i=1}^{2} \prod_{j=1}^{3} \frac{(\lambda_i n_{ij})^{y_{ij}}}{y_{ij}!} \exp(-\lambda_i n_{ij}) \,. \tag{8.1}$$

Det oprindelige spørgsmål om tallene tyder på at ultralyd er skadeligt kan nu oversættes til modellens sprog. Da den systematiske forskel mellem grupperne beskrives ved hjælp af parametrene λ_1 og λ_2, kommer spørgsmålet til at gå ud på, om tallene tyder på at λ_1 og λ_2 er signifikant forskellige, dvs. vi skal teste den statistiske hypotese $H_0 : \lambda_1 = \lambda_2$.

Estimation af parametre

Maksimaliseringsestimaterne over λ_1 og λ_2 skal bestemmes på grundlag af *likelihoodfunktionen*. Ud fra modelfunktionen (8.1) får vi

$$\begin{aligned} L(\lambda_1, \lambda_2) &= \prod_{i=1}^{2} \prod_{j=1}^{3} \frac{(\lambda_i n_{ij})^{y_{ij}}}{y_{ij}!} \exp(-\lambda_i n_{ij}) \\ &= \text{konstant} \times \prod_{i=1}^{2} \prod_{j=1}^{3} \lambda_i^{y_{ij}} \exp(-\lambda_i n_{ij}) \\ &= \text{konstant} \times \prod_{i=1}^{2} \lambda_i^{y_{i\bullet}} \exp(-\lambda_i n_{i\bullet}) \ , \end{aligned}$$

hvor konstanten afhænger af n-erne og y-erne, men ikke af λ_1 og λ_2. Vi ser, at likelihoodfunktionen er den samme som man ville have fået hvis man udelukkende havde set på totalantallene $y_{1\bullet}$ og $y_{2\bullet}$ for hver mus og havde sagt, at det var $Y_{1\bullet}$ og $Y_{2\bullet}$ der var Poissonfordelte med parametre $\lambda_1 n_{1\bullet}$ hhv. $\lambda_2 n_{2\bullet}$. Derfor bliver skønnet over λ_i

$$\widehat{\lambda}_i = \frac{y_{i\bullet}}{n_{i\bullet}} \ ,$$

dvs. det totale observerede antal mikrokærneceller i gruppe i divideret med det totale antal optalte celler i gruppe i, hvilket også er det estimat der umiddelbart tilbyder sig.

For at estimere det fælles λ under H_0 betragtes likelihoodfunktionen $L_0(\lambda) = L(\lambda, \lambda)$:

$$\begin{aligned} L_0(\lambda) &= \text{konstant} \times \prod_{i=1}^{2} \lambda^{y_{i\bullet}} \exp(-\lambda n_{i\bullet}) \\ &= \text{konstant} \times \lambda^{y_{\bullet\bullet}} \exp(-\lambda n_{\bullet\bullet}) \ , \end{aligned}$$

således at

$$\widehat{\lambda} = \frac{y_{\bullet\bullet}}{n_{\bullet\bullet}} \ .$$

Det er også hvad man umiddelbart skulle vente, thi når H_0 er rigtig er der ingen forskel på de to grupper, dvs. der er i realiteten kun tale om én gruppe, bestående af $n_{\bullet\bullet}$ celler hvoraf $y_{\bullet\bullet}$ er mikrokærneceller.

I eksemplet bliver estimaterne

$$\hat{\lambda}_{\text{behandl}} = \frac{18}{6320} = 2.8 \times 10^{-3}$$
$$\sim \text{knap 3 mikrokærneceller pr. 1000 celler}$$

$$\hat{\lambda}_{\text{kontrol}} = \frac{10}{6288} = 1.6 \times 10^{-3}$$
$$\sim \text{godt 1.5 mikrokærneceller pr. 1000 celler}$$

$$\hat{\lambda}_{\text{fælles}} = \frac{28}{12608} = 2.2 \times 10^{-3}$$
$$\sim \text{godt 2 mikrokærneceller pr. 1000 celler}$$

Man kan spørge om, hvor stor tiltro man nu kan have til disse tal. Det er ikke i statistikerens magt at udtale noget fornuftigt om diverse eksterne fejlkilder der eventuelt måtte have været i spil (det ved eksperimentator bedre). Statistikeren kan udtale sig om dén tilfældige variation der beskrives af den statistiske model. I den konkrete situation kan det siges, at når f.eks. $Y_{1\cdot}$ er Poissonfordelt med parameter $\lambda_1 n_{1\cdot}$, så er[2] variansen på $Y_{1\cdot}$ givet ved $\text{Var}\, Y_{1\cdot} = \lambda_1 n_{1\cdot}$. Da $\hat{\lambda}_1 = Y_{1\cdot}/n_{1\cdot}$, er

$$\begin{aligned} \text{Var}\, \hat{\lambda}_1 &= (\text{Var}\, Y_{1\cdot})/n_{1\cdot}^2 \\ &= \lambda_1/n_{1\cdot}\,, \end{aligned}$$

således at et *estimat* over $\text{Var}\, \hat{\lambda}_1$ er

$$\hat{\lambda}_1/n_{1\cdot} = y_{1\cdot}/n_{1\cdot}^2\,;$$

Imidlertid er vi mere interesserede i et skøn over den såkaldte *middelfejl* på $\hat{\lambda}$, dvs. *standardafvigelsen* af $\hat{\lambda}_1$, eftersom standardafvigelsen af $\hat{\lambda}_1$ er på samme skala som $\hat{\lambda}_1$. Skønnet over middelfejlen på $\hat{\lambda}_1$ er

$$\sqrt{y_{1\cdot}/n_{1\cdot}^2} = \sqrt{y_{1\cdot}}/n_{1\cdot}\,,$$

dvs. ca. 0.67×10^{-3}; på samme måde får man skønnene over standardafvigelserne på de to øvrige parameterskøn til ca. 0.50×10^{-3} og ca. 0.42×10^{-3}.

Som læseren vi have bemærket benytter vi ved beregningen af de forskellige parameterskøn slet ikke de individuelle værdier af n og y for de enkelte mus, vi benytter kun totalerne for hver gruppe. Er det da

[2]jf. side 119

lige meget hvad værdierne for de enkelte mus er? Ja, det er det faktisk, *sålænge der ikke er tvivl om Poissonmodellens brugbarhed.* Men hvis vi er på udkig efter indicier for (eller imod) anvendeligheden af Poissonmodellen, så er det i høj grad påkrævet at kende de enkelte værdier. For den statistiske model skal jo beskrive enkeltobservationernes tilfældige variation omkring et bestemt niveau, så hvis man vil vurdere antagelsen om at den tilfældige variation kan beskrives ved netop en Poissonfordeling, så skal man se på enkeltobservationernes faktiske variation og vurdere, om den ligner den fittede Poissonfordeling.

Hypoteseprøvning

Som nævnt skal vi teste den statistiske hypotese $H_0 : \lambda_1 = \lambda_2$. Det gøres på traditionel vis med et kvotienttest. Vi udregner kvotientteststørrelsen

$$\begin{aligned} Q &= \frac{L(\hat{\lambda}, \hat{\lambda})}{L(\hat{\lambda}_1, \hat{\lambda}_2)} \\ &= \frac{L_0(\hat{\lambda})}{L(\hat{\lambda}_1, \hat{\lambda}_2)} \\ &= \frac{\hat{\lambda}^{y_{1\cdot}+y_{2\cdot}} \exp(-\hat{\lambda} n_{1\cdot} - \hat{\lambda} n_{2\cdot})}{\hat{\lambda}_1^{y_{1\cdot}} \hat{\lambda}_2^{y_{2\cdot}} \exp(-\hat{\lambda}_1 n_{1\cdot} - \hat{\lambda}_2 n_{2\cdot})} \\ &= \frac{(\hat{\lambda} n_{1\cdot})^{y_{1\cdot}} (\hat{\lambda} n_{2\cdot})^{y_{2\cdot}} \exp(-y_{1\cdot} - y_{2\cdot})}{(\hat{\lambda}_1 n_{1\cdot})^{y_{1\cdot}} (\hat{\lambda}_2 n_{2\cdot})^{y_{2\cdot}} \exp(-y_{1\cdot} - y_{2\cdot})} \\ &= \left(\frac{\hat{\lambda} n_{1\cdot}}{y_{1\cdot}}\right)^{y_{1\cdot}} \left(\frac{\hat{\lambda} n_{2\cdot}}{y_{2\cdot}}\right)^{y_{2\cdot}} \\ &= \left(\frac{\hat{y}_{1\cdot}}{y_{1\cdot}}\right)^{y_{1\cdot}} \left(\frac{\hat{y}_{2\cdot}}{y_{2\cdot}}\right)^{y_{2\cdot}} , \end{aligned}$$

hvor $\widehat{y_{i\cdot}} = \hat{\lambda} n_{i\cdot}$ er det "forventede" antal mikrokærneceller i gruppe i, forudsat at H_0 er rigtig.

Derfor er

$$-2 \ln Q = 2 \left(y_{1\cdot} \ln \frac{y_{1\cdot}}{\hat{y}_{1\cdot}} + y_{2\cdot} \ln \frac{y_{2\cdot}}{\hat{y}_{2\cdot}} \right) .$$

Små værdier af Q, dvs. store værdier af $-2\ln Q$, er *signifikante*, dvs. de er tegn på at hypotesen H_0 *ikke* er forenelig med de foreliggende data.

For at vurdere om $-2\ln Q_{\text{obs}}$ er signifikant stor, skal man bestemme testsandsynligheden

$$\varepsilon \;=\; \mathrm{P}_0\left(-2\ln Q \geq -2\ln Q_{\text{obs}}\right) \;,$$

altså sandsynligheden under H_0 for at få et mindst lige så afvigende observationssæt som det foreliggende. Når man skal beregne ε plejer man at udnytte, at når H_0 er rigtig, så er $-2\ln Q$ med god tilnærmelse[3] χ^2-fordelt med $f = 2-1$ frihedsgrader[4], således at ε med god tilnærmelse kan udregnes som sandsynligheden for at få en værdi større end eller lig med $-2\ln Q_{\text{obs}}$ i χ^2-fordelingen med 1 frihedsgrad:

$$\varepsilon \;=\; \mathrm{P}\left(\chi^2_1 \geq -2\ln Q_{\text{obs}}\right) \;.$$

I taleksemplet er $\hat{y}_{1\bullet} = 14.0$ og $\hat{y}_{2\bullet} = 14.0$, så

$$\begin{aligned} -2\ln Q \;&=\; 2\left(18\,\ln\frac{18}{14.0} + 10\,\ln\frac{10}{14.0}\right) \\ &=\; 2.32 \;. \end{aligned}$$

I χ^2-fordelingen med 1 frihedsgrad er 80%-fraktilen 1.64 og 90%-fraktilen 2.71, således at den fundne $-2\ln Q$-værdi svarer til et ε på mellem 10% og 20%. Man vil almindeligvis sige, at en sådan ε-værdi ikke er lille nok til at man vil forkaste H_0. Vi kan dermed konkludere, at de foreliggende tal ikke giver statistisk belæg for at man kan mene at ultralyd er skadeligt.

---

[3] χ^2-approksimationen kan anvendes når de forventede antal $\hat{y}_{i\bullet}$ er mindst fem.

[4] antal parametre i grundmodellen er 2; antal parametre under H_0 er 1; antal frihedsgrader er derfor $f = 2-1$.

Resumé 7. Sammenligning af Poissonfordelinger

Situation: Man har optalt antal begivenheder af en bestemt slags i r forskellige grupper. I gruppe i er

$$\begin{aligned} y_i &= \text{observeret antal begivenheder} \\ t_i &= \text{observations-perioden} \\ n_i &= \text{størrelsen af "risikogruppen"} \end{aligned}$$

Model: $y_1, y_2, \ldots, y_r$ er observerede værdier af uafhængige Poissonfordelte stokastiske variable $Y_1, Y_2, \ldots, Y_r$, hvor Y_i har parameter $\lambda_i n_i t_i$ hvor λ_i er en ukendt parameter der beskriver intensiteten af begivenhederne pr. tid og pr. individ for gruppe i.

Estimation: λ_i estimeres ved antal begivenheder divideret med den samlede 'gennemlevede' tid i gruppe i:

$$\widehat{\lambda}_i = \frac{y_i}{n_i t_i} .$$

Hypotese: Man ønsker at teste hypotesen

$$H_0 : \lambda_1 = \lambda_2 = \ldots = \lambda_r$$

om at intensiteterne er ens i de r grupper.

Teststørrelse: Den eventuelle fælles værdi af λ estimeres ved $\widehat{\lambda} = y_\bullet / \sum_{i=1}^{r} n_i t_i$, og det "forventede" antal begivenheder i gruppe i under H_0 er

$$\begin{aligned} \widehat{y}_i &= \widehat{\lambda} \times n_i t_i \\ &= \frac{n_i t_i}{\sum_{i=1}^{r} n_i t_i} y_\bullet . \end{aligned}$$

Kvotienttesstørrelsen er

$$-2 \ln Q = 2 \sum_{i=1}^{r} y_i \ln \frac{y_i}{\widehat{y}_i} .$$

Testsandsynlighed: Hvis de "forventede" antal $\widehat{y}_i$ er mindst fem, kan testsandsynligheden ε med god tilnærmelse bestemmes som sandsynligheden for at få en større værdi end $-2\ln Q_{\text{obs}}$ i χ^2-fordelingen med $f = r - 1$ frihedsgrader:

$$\varepsilon \;=\; \mathrm{P}\left(\chi^2_{r-1} \geq -2\ln Q_{\text{obs}}\right).$$

Konklusion: Hvis ε er meget lille, så er der en signifikant afvigelse mellem det observerede og det som modellen foreskriver, og man vil da forkaste H_0. I modsat fald er H_0 forenelig med det observerede, og man kan da ikke forkaste H_0.

8.2 Multiplikative Poissonmodeller: et eksempel

I dette afsnit skal vi gennemgå et eksempel på en såkaldt multiplikativ Poissonmodel. Modellen (og især analysen af den) er ganske vist en smule mere indviklet end hvad der hidtil er blevet præsenteret, men på den anden side er det en type modeller der benyttes en del.

Først præsenteres det gennemgående eksempel.

Eksempel 8.2. ***Lungekræft i Fredericia***

I midten af 1970-erne var der en større debat om, hvorvidt der var særlig stor risiko for at få lungekræft når man boede i byen Fredericia. Grunden til at der kunne være en større risiko var, at der i Fredericia var en betydelig mængde luftforurenende industri, som tilmed lå midt inde i byen. For at kunne afgøre spørgsmålet indsamlede man numeriske data om lungekræfthyppigheden i perioden 1968-71, dels i Fredericia, dels i byerne Horsens, Kolding og Vejle. De tre sidste byer skulle tjene som sammenligningsgrundlag, idet det var byer af nogenlunde samme art som Fredericia, pånær den mistænkte industri.

Lungekræft opstår tit som et resultat af mange års daglige påvirkninger af skadelige stoffer. En eventuel større risiko i Fredericia kunne måske derfor vise sig ved at lungekræftpatienterne fra Fredericia var yngre end dem fra kontrolbyerne. Og det er under alle omstændigheder tilfældet, at lungekræft optræder med meget forskellig hyppighed i forskellige aldersklasser. Det er derfor ikke nok at se på totalantallene af lungekræfttilfælde, man skal se på antallene af tilfælde i forskellige aldersklasser. De foreliggende tal[5] *er vist som Tabel 8.2. Da antallene af lungekræfttilfælde i sig selv ikke siger noget sålænge man ikke kender risikogruppernes størrelse, må man også rapportere antal indbyggere i de forskellige aldersklasser og byer, se Tabel 8.3.*

Det der nu er statistikerens opgave er at søge at beskrive tallene i Tabel 8.2 ved hjælp af en statistisk model hvori der indgår nogle

[5] citeret efter E. B. Andersen (1977): Multiplicative Poisson models with unequal cell rates. *Scand. J. Statist.* **4**, 153–158.

Tabel 8.2: Lungekræfttilfælde i fire byer fordelt på aldersklasser.

aldersklasse	Fredericia	Horsens	Kolding	Vejle	i alt
40-54	11	13	4	5	33
55-59	11	6	8	7	32
60-64	11	15	7	10	43
65-69	10	10	11	14	45
70-74	11	12	9	8	40
75+	10	2	12	7	31
i alt	64	58	51	51	224

Tabel 8.3: Antal indbyggere i de forskellige aldersklasser i de fire byer.

aldersklasse	Fredericia	Horsens	Kolding	Vejle	i alt
40-54	3059	2879	3142	2520	11600
55-59	800	1083	1050	878	3811
60-64	710	923	895	839	3367
65-69	581	834	702	631	2748
70-74	509	634	535	539	2217
75+	605	782	659	619	2665
i alt	6264	7135	6983	6026	26408

parametre der i en passende forstand beskriver risikoen for at få lungekræft når man tilhører en bestemt aldersgruppe og bor i en bestemt by. Endvidere ville det være formålstjenligt, hvis man kunne udskille nogle parametre der beskrev "by-virkninger" (dvs. forskelle mellem byer) efter at man på en eller anden måde havde taget højde for forskellene mellem aldersgrupperne. □

Opstilling af model for lungekræft-eksemplet

Den statistiske model skal ikke modellere variationen i antallet af indbyggere i de forskellige byer og aldersklasser, så derfor vil vi anse disse antal for givne konstanter. Det er antallene af lungekræfttilfælde der skal modelleres.

Vi indfører lidt notation:

$$y_{ij} = \text{antal tilfælde i aldersgruppe } i \text{ i by } j,$$

$$n_{ij} = \text{antal personer i aldersgruppe } i \text{ i by } j,$$

hvor $i = 1,2,3,4,5,6$ nummererer aldersgrupperne og $j = 1,2,3,4$ nummererer byerne. y_{ij}-erne opfattes som observerede værdier af stokastiske variable Y_{ij}.

Inspireret af Kapitel 7 kunne man foreslå, at Y_{ij} skulle være Poissonfordelt med en parameter der er lig med en intensitet (som afhænger af i og j) ganget med periodelængden. For nemheds skyld sætter vi tidsenheden til 1 (dvs. 1 tidsenhed = 4 år). Da vi ikke er interesserede i en intensitet pr. by men i en intensitet der kan fortolkes som en risiko pr. person, skriver vi intensiteten som risikoen pr. person ganget med antallet n_{ij} af personer i den pågældende by og aldersklasse. Alt i alt skal Y_{ij} være Poissonfordelt med parameter $\lambda_{ij}n_{ij}$, hvor λ_{ij} er sandsynligheden pr. tid for at en person i aldersgruppe i i by j udvikler lungekræft, dvs. λ_{ij} er den alders- og by-specifikke cancer-incidens. Endvidere vil vi gå ud fra, at de enkelte Y_{ij}-er er stokastisk uafhængige. Grundmodellen er således

> Y_{ij}-erne er stokastisk uafhængige og Poissonfordelte; Y_{ij} har parameter $\lambda_{ij}n_{ij}$, hvor λ_{ij}-erne er ukendte positive parametre.

Det er let nok at estimere parametrene i grundmodellen. Eksempelvis estimeres intensiteten λ_{21} for 55-59 -årige i Fredericia til $11/800 = 0.014$ (dvs. 0.014 tilfælde pr. person pr. 4 år).

Nu var det jo tanken, at vi skulle kunne sammenligne byerne efter at vi på en eller anden måde havde taget hensyn til eller elimineret aldersforskellene, og det kan ikke uden videre lade sig gøre i grundmodellen. Derfor vil vi undersøge, om det lader sig gøre at beskrive data med en anden model hvori λ_{ij} er spaltet op i et produkt af en *alders-virkning* α_i og en *by-virkning* β_j: $\lambda_{ij} = \alpha_i \times \beta_j$. *Hvis* dette lader sig gøre er vi heldigt stillede, for så kan vi sammenligne byerne ved at sammenligne by-parametrene β_j. Vi vil derfor teste den statistiske hypotese

$$H_0 : \lambda_{ij} = \alpha_i \times \beta_j$$

hvor $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_1, \beta_2, \beta_3, \beta_4$ er ukendte parametre[6].

En detalje vedrørende parametriseringen

Der er det særlige ved parametriseringen af modellen under H_0, at den ikke er injektiv. At en parametrisering er *injektiv* betyder, at forskellige parametersæt giver forskellige udgaver af modellen.

I det foreliggende tilfælde indgår de 10 parametre α_1, α_2, α_3, α_4, α_5, α_6, β_1, β_2, β_3, β_4 i modellen udelukkende gennem produkterne $\alpha_i\beta_j$. Antag at to parametersæt

$$\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_1, \beta_2, \beta_3, \beta_4$$

og

$$\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*, \alpha_5^*, \alpha_6^*, \beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*$$

giver anledning til de samme produkter, dvs. antag at

$$\alpha_i\beta_j = \alpha_i^*\beta_j^* \tag{8.2}$$

for alle i og j. Så gælder også

$$\alpha_i/\alpha_i^* = \beta_i^*/\beta_i \tag{8.3}$$

[6]Sådan formulerer man det i den sædvanlige statistiske jargon. Mere udførligt kan man sige, at vi skal teste den statistiske hypotese H_0 der siger: Der findes parametre $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_1, \beta_2, \beta_3, \beta_4$, således at der for hver by j og hver aldersgruppe i gælder, at lungekræftrisikoen λ_{ij} fås som $\lambda_{ij} = \alpha_i \times \beta_j$.

for alle i og j. Da højresiden af (8.3) ikke involverer i kan venstresiden heller ikke afhænge af i, og tilsvarende kan højresiden ikke afhænge af j, dvs. der er en konstant c således at $\alpha_i/\alpha_i^* = c$ og dermed

$$\alpha_i^* = \tfrac{1}{c}\alpha_i$$

for alle i, og således at $\beta_j^*/\beta_j = c$ og dermed

$$\beta_j^* = c\beta_j$$

for alle j. Parametersættet $\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*, \alpha_5^*, \alpha_6^*, \beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*$ er altså af formen

$$\begin{aligned}&(\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*, \alpha_5^*, \alpha_6^*, \beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*)\\ &= (\tfrac{1}{c}\alpha_1, \tfrac{1}{c}\alpha_2, \tfrac{1}{c}\alpha_3, \tfrac{1}{c}\alpha_4, \tfrac{1}{c}\alpha_5, \tfrac{1}{c}\alpha_6, c\beta_1, c\beta_2, c\beta_3, c\beta_4) \qquad (8.4)\end{aligned}$$

hvor c er en positiv konstant. Omvendt gælder også, at hvis det stjernede parametersæt er defineret ved (8.4), så er (8.2) opfyldt. Dermed har vi fået klarlagt dels at parametriseringen ikke er injektiv, dels hvilke parametersæt der giver den samme model.

De 10 parametre skal pålægges ét bånd for at få en injektiv parametrisering. Et sådant bånd kan være at $\alpha_1 = 1$ eller at $\sum_{i=1}^{6} \alpha_i = 1$ eller at $\prod_{i=1}^{6} \alpha_i = 1$ (eller det tilsvarende for β) osv.

I det aktuelle eksempel vil vi bruge betingelsen $\beta_1 = 1$, dvs. vi definerer at by-parameteren for Fredericia skal være lig 1. Med denne betingelse er parametriseringen injektiv, for hvis både β_1 og $\beta_1^* = c\beta_1$ skal være 1, så må c nødvendigvis være lig 1.

Samtidig noterer vi, at der er $10 - 1 = 9$ forskellige parametre at estimere.

Estimation af parametrene under H_0

Dette eksempel adskiller sig fra dem vi hidtil har mødt derved, at man *ikke* kan opskrive simple udtryk for estimaterne; man er nødt til at benytte en særlig fremgangsmåde for at udregne talværdierne af estimaterne i det enkelte konkrete tilfælde. En datamat med noget ordentligt statistik-programmel vil uden videre kunne levere estimaterne, men hvis man foretrækker at gå ud fra de grundlæggende principper og at foretage udregningerne med håndkraft/lommeregner er det faktisk heller særlig besværligt. Det skal vi se på de næste sider.

Parametrene $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_1, \beta_2, \beta_3, \beta_4$ (med den just indførte bibetingelse $\beta_1 = 1$) skal ifølge de sædvanlige principper bestemmes så de maksimaliserer likelihoodfunktionen. I grundmodellen er likelihoodfunktionen

$$\begin{aligned} L &= \prod_{i=1}^{6}\prod_{j=1}^{4} \frac{(\lambda_{ij}n_{ij})^{y_{ij}}}{y_{ij}!} \exp(-\lambda_{ij}n_{ij}) \\ &= \text{konstant} \times \prod_{i=1}^{6}\prod_{j=1}^{4} \lambda_{ij}^{y_{ij}} \exp(-\lambda_{ij}n_{ij}) \ . \end{aligned}$$

Når vi her erstatter λ_{ij} med $\alpha_i\beta_j$ får vi likelihoodfunktionen under H_0:

$$\begin{aligned} L_0 &= \text{konstant} \times \prod_{i=1}^{6}\prod_{j=1}^{4} \alpha_i^{y_{ij}} \beta_j^{y_{ij}} \exp(-\alpha_i\beta_j n_{ij}) \\ &= \text{konstant} \times \left(\prod_{i=1}^{6} \alpha_i^{y_{i\bullet}}\right) \left(\prod_{j=1}^{4} \beta_j^{y_{\bullet j}}\right) \exp\left(-\sum_{i=1}^{6}\sum_{j=1}^{4} \alpha_i\beta_j n_{ij}\right) \ . \end{aligned}$$

Opgaven består nu i at bestemme det parametersæt $\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\alpha}_5, \hat{\alpha}_6, 1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$ der maksimaliserer L_0. Vi vil dog benytte *log-likelihoodfunktionen* $\ln L_0$ i stedet:

$$\begin{aligned} &\ln L_0 \\ &= \text{konstant} + \left(\sum_{i=1}^{6} y_{i\bullet} \ln \alpha_i\right) + \left(\sum_{j=1}^{4} y_{\bullet j} \ln \beta_j\right) - \sum_{i=1}^{6}\sum_{j=1}^{4} \alpha_i\beta_j n_{ij} \ . \end{aligned}$$

Opgaven at maksimalisere $\ln L_0$ lader sig ikke løse sådan lige uden videre, og man må derfor inddrage den generelle matematiske teori for, hvordan man maksimaliserer en funktion af mange variable. Hvis $\ln L_0$ havde været en funktion af én variabel θ, så kunne man (under visse omstændigheder) bestemme maksimumspunktet $\hat{\theta}$ som løsningen til "likelihood-ligningen"

$$\frac{d}{d\theta} \ln L_0 = 0 \ .$$

Tilsvarende[7] kan maksimumspunktet $\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\alpha}_5, \hat{\alpha}_6, 1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$ for den aktuelle log-likelihoodfunktion bestemmes som løsning til

[7] Se f.eks. afsnit 4.6 i: Mogens Brun Heefelt (1987): *Kursusmateriale til Matematik på NAT-BAS*. IMFUFA-tekst 143.

"likelihood-ligningerne"

$$\begin{aligned}
\tfrac{\partial}{\partial\alpha_1}\ln L_0 &= 0\,,\\
\tfrac{\partial}{\partial\alpha_2}\ln L_0 &= 0\,,\\
\tfrac{\partial}{\partial\alpha_3}\ln L_0 &= 0\,,\\
\tfrac{\partial}{\partial\alpha_4}\ln L_0 &= 0\,,\\
\tfrac{\partial}{\partial\alpha_5}\ln L_0 &= 0\,,\\
\tfrac{\partial}{\partial\alpha_6}\ln L_0 &= 0\,,\\
\tfrac{\partial}{\partial\beta_1}\ln L_0 &= 0\,,\\
\tfrac{\partial}{\partial\beta_2}\ln L_0 &= 0\,,\\
\tfrac{\partial}{\partial\beta_3}\ln L_0 &= 0\,,\\
\tfrac{\partial}{\partial\beta_4}\ln L_0 &= 0\,,
\end{aligned} \tag{8.5}$$

Her er f.eks. $\frac{\partial}{\partial\alpha_2}\ln L_0$ den såkaldte *partielle afledede af* $\ln L_0$ *med hensyn til* α_2, dvs. den afledede af $\ln L_0$ med hensyn til α_2 når de øvrige variable[8] holdes fast, - man finder den til

$$\frac{\partial}{\partial\alpha_2}\ln L_0 = \frac{y_{2\bullet}}{\alpha_2} - \sum_{j=1}^{4}\beta_j n_{2j}\,.$$

Deraf ses at ligningen $\frac{\partial}{\partial\alpha_2}\ln L_0 = 0$ er ensbetydende med at

$$\alpha_2 = \frac{y_{2\bullet}}{\sum_{j=1}^{4}\beta_j n_{2j}}\,.$$

Hvis man på samme måde løser alle de øvrige ligninger i (8.5), får man at følgende relationer skal være opfyldt:

$$\alpha_i = \frac{y_{i\bullet}}{\sum_{j=1}^{4}\beta_j n_{ij}}\,, \quad i = 1,2,3,4,5,6 \tag{8.6}$$

$$\beta_j = \frac{y_{\bullet j}}{\sum_{i=1}^{6}\alpha_i n_{ij}}\,, \quad j = 1,2,3,4 \tag{8.7}$$

[8] dvs. $\alpha_1,\alpha_3,\alpha_4,\alpha_5,\alpha_6,\beta_1,\beta_2,\beta_3,\beta_4$

og stadig

$$\beta_1 = 1 .$$

I ligningerne (8.6) og (8.7) er y-erne og n-erne kendte tal, α-erne og β-erne er de ubekendte. Man kan ikke løse ligningerne eksplicit, dvs. man kan ikke få en løsning af formen "skønnene over α og β = en kendt funktion af y-erne og n-erne". I stedet er man henvist til at bestemme en numerisk løsning iterativt ved brug af følgende algoritme:

1. Vælg startværdier for $\beta_2, \beta_3, \beta_4$ ($\beta_1 = 1$).
2. Indsæt β-værdierne i (8.6) og få derved $\alpha_1, \alpha_2, \ldots, \alpha_6$.
3. Indsæt α-værdierne i (8.7) og få derved $\beta_1, \beta_2, \beta_3, \beta_4$.
4. Juster β-værdierne så $\beta_1 = 1$, dvs. erstat de beregnede værdier med
$$(\frac{\beta_1}{\beta_1} = 1, \frac{\beta_2}{\beta_1}, \frac{\beta_3}{\beta_1}, \frac{\beta_4}{\beta_1}) .$$
5. Du har nu gennemført et iterations-trin. Du kan så enten gå tilbage til 2 eller du kan slutte.

Man vil almindeligvis vælge at slutte hvis enten det samlede antal iterations-trin er blevet for stort eller parameterskønnene næsten ikke har ændret sig siden forrige trin.

Denne metode prøver vi.

Som startværdi vælges $(\beta_1, \beta_2, \beta_3, \beta_4) = (1, 1, 1, 1)$ svarende til at byerne er ens. Skridt 2 leverer værdierne

$$\begin{aligned} \alpha_1 &= 0.0028 \\ \alpha_2 &= 0.0084 \\ \alpha_3 &= 0.0128 \\ \alpha_4 &= 0.0164 \\ \alpha_5 &= 0.0180 \\ \alpha_6 &= 0.0116 . \end{aligned}$$

Skridt 3 leverer et sæt nye β-er:

$$\begin{aligned}\beta_1 &= 1.2779\\ \beta_2 &= 0.9187\\ \beta_3 &= 0.8814\\ \beta_4 &= 0.9733\,.\end{aligned}$$

Skridt 4 består i at dividere hver af de fundne β-værdier med 1.2779 hvorved fås

$$\begin{aligned}\beta_1 &= 1\\ \beta_2 &= 0.7189\\ \beta_3 &= 0.6897\\ \beta_4 &= 0.7616\,.\end{aligned}$$

Således fortsættes et par gange indtil værdierne (med ca. tre betydende cifre) har stabiliseret sig. De på denne måde bestemte estimater er

$$\begin{aligned}\hat{\alpha}_1 &= 0.0036\\ \hat{\alpha}_2 &= 0.0108\\ \hat{\alpha}_3 &= 0.0164\\ \hat{\alpha}_4 &= 0.0210\\ \hat{\alpha}_5 &= 0.0229\\ \hat{\alpha}_6 &= 0.0148\\ \beta_1 &= 1\\ \hat{\beta}_2 &= 0.719\\ \hat{\beta}_3 &= 0.690\\ \hat{\beta}_4 &= 0.762\,.\end{aligned}$$

Den multiplikative models beskrivelse af data

Efter at have bestemt de bedste skøn over α-erne og β-erne skal vi nu beskæftige os med, hvor god en beskrivelse de faktisk giver af datama-

terialet.

Formelt består opgaven i at *teste* multiplikativitetshypotesen H_0, og dette gøres som sædvanlig med et kvotienttest: Man udregner

$$Q = \frac{L_0(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\alpha}_5, \hat{\alpha}_6, 1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)}{L(\hat{\lambda}_{11}, \hat{\lambda}_{12}, \ldots, \hat{\lambda}_{63}, \hat{\lambda}_{64})}$$

eller $-2\ln Q$. Små værdier af Q eller store værdier af $-2\ln Q$ er signifikante, dvs. de tyder på at H_0 *ikke* giver en tilstrækkelig god beskrivelse af data. For at afgøre om $-2\ln Q_{\text{obs}}$ er signifikant stor skal vi se på testsandsynligheden ε som er sandsynligheden for at få en værre $-2\ln Q$-værdi forudsat at H_0 er rigtig:

$$\varepsilon = \mathrm{P}_0\left(-2\ln Q \geq -2\ln Q_{\text{obs}}\right) .$$

Når H_0 er rigtig, er $-2\ln Q$ med god tilnærmelse χ^2-fordelt med $f = 24 - 9 = 15$ frihedsgrader[9]. Det betyder at testsandsynligheden kan bestemmes som

$$\varepsilon = \mathrm{P}\left(\chi^2_{15} \geq -2\ln Q_{\text{obs}}\right) .$$

Udtrykket for kvotienttesstørrelsen Q kan omformes således:

$$\begin{aligned} Q &= \frac{\prod_{i=1}^{6}\prod_{j=1}^{4}\left(\hat{\alpha}_i\hat{\beta}_j\right)^{y_{ij}}\exp(-\hat{\alpha}_i\hat{\beta}_j n_{ij})}{\prod_{i=1}^{6}\prod_{j=1}^{4}\hat{\lambda}_{ij}^{y_{ij}}\exp(-\hat{\lambda}_{ij}n_{ij})} \\ &= \prod_{i=1}^{6}\prod_{j=1}^{4}\left(\frac{\hat{\alpha}_i\hat{\beta}_j}{\hat{\lambda}_{ij}}\right)^{y_{ij}} \times \exp\left(-\sum_{i=1}^{6}\sum_{j=1}^{4}\hat{\alpha}_i\hat{\beta}_j n_{ij} + \sum_{i=1}^{6}\sum_{j=1}^{4}\hat{\lambda}_{ij}n_{ij}\right) \\ &= \prod_{i=1}^{6}\prod_{j=1}^{4}\left(\frac{\hat{\alpha}_i\hat{\beta}_j n_{ij}}{y_{ij}}\right)^{y_{ij}} \times \exp\left(-\sum_{i=1}^{6}\sum_{j=1}^{4}\hat{\alpha}_i\hat{\beta}_j n_{ij} + \sum_{i=1}^{6}\sum_{j=1}^{4}y_{ij}\right) \\ &= \prod_{i=1}^{6}\prod_{j=1}^{4}\left(\frac{\hat{y}_{ij}}{y_{ij}}\right)^{y_{ij}} \times \exp\left(-\sum_{i=1}^{6}\sum_{j=1}^{4}\hat{y}_{ij} + \sum_{i=1}^{6}\sum_{j=1}^{4}y_{ij}\right) , \end{aligned}$$

[9]Tilnærmelsen er som sædvanlig god nok når de forventede antal (se senere) alle er mindst fem.

hvor

$$\hat{y}_{ij} = \hat{\alpha}_i \hat{\beta}_j n_{ij} \tag{8.8}$$

er det forventede antal lungekræfttilfælde i aldersklasse i i by j. Nu gælder at det totale forventede antal $\sum_{i=1}^{6} \sum_{j=1}^{4} \hat{y}_{ij}$ er lig med det totale observerede antal $y_{\cdot\cdot} = \sum_{i=1}^{6} \sum_{j=1}^{4} y_{ij}$, fordi da parameterskønnene opfylder ligningerne (8.6) og (8.7) er

$$\begin{aligned} \hat{y}_{ij} &= \hat{\alpha}_i \hat{\beta}_j n_{ij} \\ &= \frac{y_{i\cdot}}{\sum_{j=1}^{4} \hat{\beta}_j n_{ij}} \hat{\beta}_j n_{ij} \\ &= y_{i\cdot} \frac{\hat{\beta}_j n_{ij}}{\sum_{j=1}^{4} \hat{\beta}_j n_{ij}} , \end{aligned}$$

så at

$$\begin{aligned} \sum_{j=1}^{4} \hat{\alpha}_i \hat{\beta}_j n_{ij} &= y_{i\cdot} \sum_{j=1}^{4} \frac{\hat{\beta}_j n_{ij}}{\sum_{j=1}^{4} \hat{\beta}_j n_{ij}} \\ &= y_{i\cdot} , \end{aligned}$$

og dermed

$$\begin{aligned} \sum_{i=1}^{6} \sum_{j=1}^{4} \hat{\alpha}_i \hat{\beta}_j n_{ij} &= \sum_{i=1}^{6} y_{i\cdot} \\ &= y_{\cdot\cdot} . \end{aligned}$$

Det betyder at det fundne udtryk for Q reduceres til

$$Q = \prod_{i=1}^{6} \prod_{j=1}^{4} \left(\frac{\hat{y}_{ij}}{y_{ij}} \right)^{y_{ij}} ,$$

og dermed

$$-2 \ln Q = 2 \sum_{i=1}^{6} \sum_{j=1}^{4} y_{ij} \ln \frac{y_{ij}}{\hat{y}_{ij}} . \tag{8.9}$$

Tabel 8.4: Estimerede alders- og byspecifikke lungekræftintensiteter i perioden 1986-71 under forudsætning af den multiplikative Poissonmodel. Værdierne er antal pr. 1000 indbyggere pr. 4 år.

aldersklasse	Fredericia	Horsens	Kolding	Vejle
40-54	3.6	2.6	2.5	2.7
55-59	10.8	7.8	7.5	8.2
60-64	16.4	11.8	11.3	12.5
65-69	21.0	15.1	14.5	16.0
70-74	22.9	16.5	15.8	17.4
75+	14.8	10.6	10.2	11.3

Tabel 8.5: De forventede antal $\widehat{y}_{ij}$ af lungekræfttilfælde under den multiplikative Poissonmodel.

aldersklasse	Fredericia	Horsens	Kolding	Vejle	i alt
40-54	11.01	7.45	7.80	6.91	33.17
55-59	8.64	8.41	7.82	7.23	32.10
60-64	11.64	10.88	10.13	10.48	43.13
65-69	12.20	12.59	10.17	10.10	45.06
70-74	11.66	10.44	8.45	9.41	39.96
75+	8.95	8.32	6.73	6.98	30.98
i alt	64.10	58.09	51.10	51.11	224.40

Som et led i beregningerne af $\widehat{y}_{ij}$-erne udregnes de estimerede alders- og by-specifikke lungekræftintensitetet $\widehat{\alpha}_i\widehat{\beta}_j$; i Tabel 8.4 ses værdierne af $1000\widehat{\alpha}_i\widehat{\beta}_j$, dvs. de forventede antal tilfælde pr. 1000 indbyggere. Selve de forventede antal[10] $\widehat{y}_{ij}$ i de forskellige byer og aldersklasser er gengivet i Tabel 8.5.

Indsættes tallene fra Tabel 8.2 og Tabel 8.5 i udtrykket (8.9) for $-2\ln Q$ får man

$$-2\ln Q_{\text{obs}} \;=\; 22.6\ .$$

[10]Bemærk at man er nødt til at udregne de forventede antal med (mindst) to decimaler, ellers bliver talværdien af $-2\ln Q$-størrelsen alt for forkert på grund af afrundingsfejl.

I χ^2-fordelingen med $f = 24-9 = 15$ frihedsgrader er 90%-fraktilen 22.3 og 95%-fraktilen 25.0. Den opnåede værdi $-2\ln Q_{\text{obs}} = 22.6$ svarer altså til en testsandsynlighed ε på godt 5%, og der er dermed ikke alvorlig evidens imod modellens brugbarhed. Vi tillader os at gå ud fra, at modellen faktisk *er* anvendelig, dvs. at ***lungekræftrisikoen afhænger multiplikativt af by og alder.*** Hermed er vi nået frem til en statistisk model der beskriver data ved hjælp af nogle by-parametre og nogle alders-parametre, men ingen parametre der svarer til en vekselvirkning mellem by og alder. Det betyder, at den forskel der er mellem byerne er den samme for alle aldersklasser (og omvendt at forskellen mellem aldersklasserne er den samme i alle byer). Når vi skal sammenligne byerne kan vi derfor gøre det ved udelukkende at betragte β-erne.

Ens byer?

Det hele går ud på at vurdere om der er nogen signifikant forskel på byerne. Hvis der ikke er nogen forskel, så må by-parametrene have den samme værdi, dvs. $\beta_1 = \beta_2 = \beta_3 = \beta_4$, og da $\beta_1 = 1$ må den fælles værdi være 1. Derfor vil vi teste den statistiske hypotese

$$H_1 : \ \beta_1 = \beta_2 = \beta_3 = \beta_4 = 1 \ .$$

Hypotesen skal testes i forhold til den aktuelle grundmodel H_0, så teststørrelsen bliver

$$Q \ = \ \frac{L_1(\hat{\hat{\alpha}}_1, \hat{\hat{\alpha}}_2, \hat{\hat{\alpha}}_3, \hat{\hat{\alpha}}_4, \hat{\hat{\alpha}}_5, \hat{\hat{\alpha}}_6)}{L_0(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\alpha}_5, \hat{\alpha}_6, 1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)} \ ,$$

hvor

$$L_1(\alpha_1, \alpha_2, \ldots, \alpha_6) \ = \ L_0(\alpha_1, \alpha_2, \ldots, \alpha_6, 1, 1, 1, 1) \qquad (8.10)$$

er likelihoodfunktionen under H_1, og hvor $\hat{\hat{\alpha}}_1, \hat{\hat{\alpha}}_2, \ldots, \hat{\hat{\alpha}}_6$ er de bedste skøn over $\alpha_1, \alpha_2, \ldots, \alpha_6$ under H_1, dvs. $\hat{\hat{\alpha}}_1, \hat{\hat{\alpha}}_2, \ldots, \hat{\hat{\alpha}}_6$ maksimaliserer (8.10).

Likelihoodfunktionen L_1 kan omskrives til et produkt af seks funktioner[11], hver med sit α:

$$L_1(\alpha_1, \alpha_2, \ldots, \alpha_6) \ = \ \text{konstant} \times \prod_{i=1}^{6} \prod_{j=1}^{4} \alpha_i^{y_{ij}} \exp(-\alpha_i n_{ij})$$

[11] nemlig seks Poisson-likelihoodfunktioner

$$= \text{konstant} \times \prod_{i=1}^{6} \alpha_i^{y_{i\bullet}} \exp(-\alpha_i n_{i\bullet}) .$$

Maksimaliseringsestimatet findes derfor til

$$\hat{\hat{\alpha}}_i = \frac{y_{i\bullet}}{n_{i\bullet}} .$$

Talværdierne bliver

$$\begin{aligned}
\hat{\hat{\alpha}}_1 &= 33/11600 = 0.002845 \\
\hat{\hat{\alpha}}_2 &= 32/3811 = 0.00840 \\
\hat{\hat{\alpha}}_3 &= 43/3367 = 0.0128 \\
\hat{\hat{\alpha}}_4 &= 45/2748 = 0.0164 \\
\hat{\hat{\alpha}}_5 &= 40/2217 = 0.0180 \\
\hat{\hat{\alpha}}_6 &= 31/2665 = 0.0116 .
\end{aligned}$$

Kvotienttesstørrelsen omskrives således:

$$\begin{aligned}
Q &= \frac{L_1(\hat{\hat{\alpha}}_1, \hat{\hat{\alpha}}_2, \hat{\hat{\alpha}}_3, \hat{\hat{\alpha}}_4, \hat{\hat{\alpha}}_5, \hat{\hat{\alpha}}_6)}{L_0(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\alpha}_5, \hat{\alpha}_6, \hat{\beta}_1 = 1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)} \\
&= \frac{\prod_{i=1}^{6}\prod_{j=1}^{4} \hat{\hat{\alpha}}_i^{y_{ij}} \exp(-\hat{\hat{\alpha}}_i n_{ij})}{\prod_{i=1}^{6}\prod_{j=1}^{4} \left(\hat{\alpha}_i \hat{\beta}_j\right)^{y_{ij}} \exp(-\hat{\alpha}_i \hat{\beta}_j n_{ij})} \\
&= \prod_{i=1}^{6}\prod_{j=1}^{4} \left(\frac{\hat{\hat{y}}_{ij}}{\hat{y}_{ij}}\right)^{y_{ij}} \times \exp\left(-\sum_{i=1}^{6}\sum_{j=1}^{4} \hat{\hat{y}}_{ij} + \sum_{i=1}^{6}\sum_{j=1}^{4} \hat{y}_{ij}\right) ,
\end{aligned}$$

hvor $\hat{y}_{ij} = \hat{\alpha}_i \hat{\beta}_j n_{ij}$ (som hidtil) og

$$\hat{\hat{y}}_{ij} = \hat{\hat{\alpha}}_i n_{ij} .$$

Da $\sum_{i=1}^{6}\sum_{j=1}^{4} \hat{\hat{y}}_{ij} = \sum_{i=1}^{6}\sum_{j=1}^{4} \hat{y}_{ij} = \sum_{i=1}^{6}\sum_{j=1}^{4} y_{ij}$, er

$$Q = \prod_{i=1}^{6}\prod_{j=1}^{4} \left(\frac{\hat{\hat{y}}_{ij}}{\hat{y}_{ij}}\right)^{y_{ij}}$$

Tabel 8.6: De forventede antal $\widehat{\widehat{y}}_{ij}$ af lungekræfttilfælde under antagelsen om at der ikke er forskel på byerne.

aldersklasse	Fredericia	Horsens	Kolding	Vejle	i alt
40-54	8.70	8.19	8.94	7.17	33.00
55-59	6.72	9.10	8.82	7.38	32.02
60-64	9.09	11.81	11.46	10.74	43.10
65-69	9.53	13.68	11.51	10.35	45.07
70-74	9.16	11.41	9.63	9.70	39.90
75+	7.02	9.07	7.64	7.18	30.91
i alt	50.22	63.26	58.00	52.52	224.00

og dermed

$$-2\ln Q \;=\; 2\sum_{i=1}^{6}\sum_{j=1}^{4} y_{ij}\ln\frac{\widehat{y}_{ij}}{\widehat{\widehat{y}}_{ij}}\,. \tag{8.11}$$

Store værdier af $-2\ln Q$ er signifikante. $-2\ln Q$ skal sammenholdes med χ^2-fordelingen med $f = 9-6 = 3$ frihedsgrader.

De forventede tal[12] er vist som Tabel 8.6. Indsættes værdierne fra Tabel 8.2, Tabel 8.5 og Tabel 8.6 i (8.11) fås

$$-2\ln Q_{\text{obs}} \;=\; 5.67\,.$$

I χ^2-fordelingen med $f = 9-6 = 3$ frihedsgrader er 80%-fraktilen 4.64 og 90%-fraktilen 6.25, således at testsandsynligheden ε er næsten 20%. De foreliggende observationer er altså fint forenelige med hypotesen H_1 om at der ikke er nogen forskel på byerne. Sagt på en anden måde, *der er ikke nogen signifikant forskel på byerne.*

En anden mulighed

Det er sjældent tilfældet at der er én bestemt måde at undersøge en praktisk problemstilling på ved hjælp af en statistisk model og en statistisk hypotese. Det aktuelle spørgsmål om der er en øget risiko for lungekræft ved at bo i Fredericia blev i forrige afsnit belyst ved at vi

[12] der stadig skal beregnes med mindst to decimaler

testede hypotesen H_1 om ens byparametre. Det viste sig at H_1 kunne akcepteres, og man kan således sige at der ikke er nogen signifikant forskel på de fire byer.

Nu kan man også angribe det aktuelle problem på en anden måde. Man kan sige at det hele drejer sig om at vurdere, om det er farligere at bo i Fredericia end i en af de tre øvrige byer; dermed er det indirekte forudsat, at de tre øvrige byer er rimelig ens, hvilket man bør teste! Man kunne derfor anlægge følgende strategi for formulering og test af hypoteser:

1. Vi går stadig ud fra den multiplikative Poissonmodel H_0 som grundmodel.

2. Først undersøges om det kan antages at de tre byer Horsens, Kolding og Vejle er ens, dvs. vi skal teste hypotesen

$$H_2 : \beta_2 = \beta_3 = \beta_4 .$$

3. Hvis H_2 bliver akcepteret er der et fælles niveau β_0 for de tre "kontrol-byer". Vi kan derefter sammenligne Fredericia med dette fælles niveau ved at teste om $\beta_1 = \beta_0$. Da β_1 pr. definition er lig 1, er den hypotese der skal testes

$$H_3 : \beta_0 = 1 .$$

Sammenligning af de tre kontrolbyer

Vi skal teste hypotesen $H_2 : \beta_2 = \beta_3 = \beta_4$ om ens kontrolbyer i forhold til den multiplikative model H_0. Det gøres med et kvotienttest

$$Q = \frac{L_2(\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_6, \tilde{\beta})}{L_0(\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_6, 1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)} ,$$

hvor

$$L_2(\alpha_1, \alpha_2, \ldots, \alpha_6, \beta) = L_0(\alpha_1, \alpha_2, \ldots, \alpha_6, 1, \beta, \beta, \beta)$$

er likelihoodfunktionen under H_2 og $\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_6, \tilde{\beta}$ er maksimaliseringsestimaterne under H_2.

Når H_2 er rigtig, er $-2 \ln Q$ med god tilnærmelse χ^2-fordelt med $f = 9 - 7 = 2$ frihedsgrader.

Modellen H_2 svarer til en multiplikativ Poissonmodel med *to* byer (nemlig Fredericia og resten) og seks aldersklasser, og der er derfor ikke nogen principielt nye problemer forbundet med at estimere parametrene under H_2. Man finder

$$\begin{aligned}
\tilde{\alpha}_1 &= 0.00358 \\
\tilde{\alpha}_2 &= 0.0108 \\
\tilde{\alpha}_3 &= 0.0164 \\
\tilde{\alpha}_4 &= 0.0210 \\
\tilde{\alpha}_5 &= 0.0230 \\
\tilde{\alpha}_6 &= 0.0148 \\
\tilde{\beta}_1 &= 1 \\
\tilde{\beta}_0 &= 0.7220
\end{aligned}$$

Endvidere bliver

$$-2\ln Q = -2\sum_{i=1}^{6}\sum_{j=1}^{4} y_{ij} \ln \frac{\hat{y}_{ij}}{\tilde{y}_{ij}} , \tag{8.12}$$

hvor $\hat{y}_{ij}$ er givet ved (8.8) og Tabel 8.5, og

$$\begin{aligned}
\tilde{y}_{i1} &= \tilde{\alpha}_i n_{i1} \\
\tilde{y}_{ij} &= \tilde{\alpha}_i \tilde{\beta}_0 , \quad j = 2,3,4.
\end{aligned}$$

De forventede antal $\tilde{y}_{ij}$ er vist i Tabel 8.7.

Når man indsætter værdierne fra Tabel 8.2, Tabel 8.5 og Tabel 8.7 i (8.12) får man

$$-2\ln Q_{\text{obs}} = 0.40 ,$$

hvilket skal sammenholdes med χ^2-fordelingen med $f = 9 - 7 = 2$ frihedsgrader. I χ^2-fordelingen med $f = 2$ frihedsgrader er 20%-fraktilen 0.446, så testsandsynligheden er altså godt 80%, og det betyder, at H_2 er udmærket forenelig med de foreliggende data. Vi kan altså udmærket tillade os at gå ud fra, at der ikke er nogen signifikant forskel mellem de tre byer.

Tabel 8.7: De forventede antal $\tilde{y}_{ij}$ af lungekræfttilfælde under H_2.

aldersklasse	Fredericia	Horsens	Kolding	Vejle	i alt
40-54	10.95	7.44	8.12	6.51	33.02
55-59	8.64	8.44	8.19	6.85	32.12
60-64	11.64	10.93	10.60	9.93	43.10
65-69	12.20	12.65	10.64	9.57	45.06
70-74	11.71	10.53	8.88	8.95	40.07
75+	8.95	8.36	7.04	6.61	30.96
i alt	64.09	58.35	53.47	48.42	224.33

Herefter kan vi gå over til at teste H_3, som går ud på, at alle fire byer er ens, og at der er de seks forskellige aldersgrupper med hver sin parameter α_i. H_3 er under forudsætning af H_2 identisk med hypotesen H_1 fra tidligere, så skønnene over aldersparametrene er $\hat{\tilde{\alpha}}_1, \hat{\tilde{\alpha}}_2, \ldots, \hat{\tilde{\alpha}}_6$.

I denne omgang skal vi teste H_3 $(= H_1)$ i forhold til den nu gældende grundmodel H_2. Teststørrelsen er

$$\begin{aligned} Q &= \frac{L_1(\hat{\tilde{\alpha}}_1, \hat{\tilde{\alpha}}_2, \ldots, \hat{\tilde{\alpha}}_6)}{L_2(\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_6, \tilde{\beta}_0)} \\ &= \frac{L_0(\hat{\tilde{\alpha}}_1, \hat{\tilde{\alpha}}_2, \ldots, \hat{\tilde{\alpha}}_6, 1, 1, 1, 1)}{L_0(\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_6, 1, \tilde{\beta}_0, \tilde{\beta}_0, \tilde{\beta}_0)} \end{aligned}$$

der let omformes til

$$Q = \prod_{i=1}^{6} \prod_{j=1}^{4} \left(\frac{\hat{\tilde{y}}_{ij}}{\tilde{y}_{ij}} \right)^{y_{ij}} ,$$

så at

$$-2 \ln Q = 2 \sum_{i=1}^{6} \sum_{j=1}^{4} y_{ij} \ln \frac{\tilde{y}_{ij}}{\hat{\tilde{y}}_{ij}} . \tag{8.13}$$

Store værdier af $-2 \ln Q$ er signifikante. Når H_3 er rigtig, er $-2 \ln Q$ med god tilnærmelse[13] χ^2-fordelt med $f = 7 - 6 = 1$ frihedsgrad.

[13]forudsat at de indgående forventede antal er mindst fem

Ved at indsætte værdierne fra Tabel 8.2, Tabel 8.6 og Tabel 8.7 i (8.13) fås

$$-2\ln Q_{\text{obs}} = 5.27 .$$

I χ^2-fordelingen med 1 frihedsgrad er 97.5%-fraktilen 5.02 og 99%-fraktilen 6.63, så testsandsynligheden er omkring 2%. På det grundlag vil man almindeligvis *forkaste* hypotesen H_3 ($= H_1$). Konklusionen bliver altså, at *der ikke er signifikant forskel på lungekræfthyppigheden i de tre byer Horsens, Kolding og Vejle, hvorimod Fredericia har en signifikant anderledes lungekræfthyppighed.*

Den relative lungekræfthyppighed i de tre ens byer i forhold til Fredericia estimeres til $\widetilde{\beta}_0 = 0.7$, så lungekræfthyppigheden i Fredericia er altså signifikant *større*.

Se det var jo en pæn og klar konklusion, der blot er stik modsat den vi nåede frem til på side 149!

Sammenligning af de to fremgangsmåder

De to fremgangsmåder er begge opbygget over følgende skema:

1. Find en passende grundmodel.
2. Formuler en hypotese der giver en forsimpling af den aktuelle grundmodel.
3. Test hypotesen i forhold til den aktuelle grundmodel.
4. (a) Hvis hypotesen accepteres, så har vi derved fået en ny aktuel grundmodel (nemlig den gamle med de simplifikationer som den accepterede hypotese giver).

   Fortsæt ved pkt. 2

   (b) Hvis hypotesen forkastes, så slut. Data beskrives da ved den senest anvendte grundmodel.

Begge de anvendte fremgangsmåder tog udgangspunkt i den samme Poissonmodel, de adskiller sig kun ved valgene af hypoteser i pkt. 2. Tabel 8.8 og Tabel 8.9 giver oversigter over de to fremgangsmåder.

I den første fremgangsmåde tages skridtet fra den multiplikative model til "fire ens" på én gang, hvilket giver en teststørrelse på 5.67,

Tabel 8.8: Oversigt over den første fremgangsmåde.

Model/Hypotese	$-2\ln Q$	f	ε
M: vilkårlige parametre H: multiplikativitet	22.65	24-9=15	godt 5%
M: multiplikativitet H: fire ens byer	5.67	9-6=3	ca. 20%

Tabel 8.9: Oversigt over den anden fremgangsmåde.

Model/Hypotese	$-2\ln Q$	f	ε
M: vilkårlige parametre H: multiplikativitet	22.65	24-9=15	godt 5%
M: multiplikativitet H: de tre byer ens	0.40	9-7=2	godt 80%
M: de tre byer ens H: de fire byer ens	5.27	7-6=1	ca. 2%

som, da den kan fordeles på 3 frihedsgrader, ikke er signifikant. I den anden fremgangsmåde spalter vi op i

1. multiplikativitet → "tre ens", og

2. "tre ens" → "fire ens",

og det viser sig så, at de 5.67 med 3 frihedsgrader spaltes op i 0.40 med 2 frihedsgrader og 5.27 med 1 frihedsgrad, hvoraf den sidste er temmelig signifikant.

Det kan undertiden være hensigtsmæssigt at foretage en sådan trinvis testning. Man bør dog ikke stræbe efter at spalte op i så mange tests som muligt, men kun teste hypoteser der er *rimelige* i den foreliggende faglige sammenhæng.

Om teststørrelser

Læseren vil måske have bemærket visse fælles træk ved de $-2\ln Q$-udtryk der forekommer i dette kapitel. De er alle af formen

$$-2\ln Q \;=\; 2\sum \text{obs.antal} \times \ln \frac{\text{Modellens forventede antal}}{\text{Hypotesens forventede antal}}$$

og er (tilnærmelsesvis) χ^2-fordelt med et antal frihedsgrader som er "det reelle antal parametre under Modellen" minus "det reelle antal parametre under Hypotesen". Dette gælder faktisk helt generelt[14] når man tester hypoteser om Poissonfordelte observationer.

[14]under forudsætning af at summen af de forventede antal er lig summen af de observerede antal

1/78 "TANKER OM EN PRAKSIS" - et matematikprojekt.
Projektrapport af: Anne Jensen, Lena Lindenskov, Marianne Kesselhahn og Nicolai Lomholt.
Vejleder: Anders Madsen

2/78 "OPTIMERING" - Menneskets forøgede beherskelsesmuligheder af natur og samfund.
Projektrapport af: Tom J. Andersen, Tommy R. Andersen, Gert Krenøe og Peter H. Lassen
Vejleder: Bernhelm Boss.

3/78 "OPGAVESAMLING", breddekursus i fysik.
Af: Lasse Rasmussen, Aage Bonde Kræmmer og Jens Højgaard Jensen.

4/78 "TRE ESSAYS" - om matematikundervisning, matematiklæreruddannelsen og videnskabsrindalismen.
Af: Mogens Niss
Nr. 4 er p.t. udgået.

5/78 "BIBLIOGRAFISK VEJLEDNING til studiet af DEN MODERNE FYSIKS HISTORIE".
Af: Helge Kragh.
Nr. 5 er p.t. udgået.

6/78 "NOGLE ARTIKLER OG DEBATINDLÆG OM - læreruddannelse og undervisning i fysik, og - de naturvidenskabelige fags situation efter studenteroprøret".
Af: Karin Beyer, Jens Højgaard Jensen og Bent C. Jørgensen.

7/78 "MATEMATIKKENS FORHOLD TIL SAMFUNDSØKONOMIEN".
Af: B.V. Gnedenko.
Nr. 7 er udgået.

8/78 "DYNAMIK OG DIAGRAMMER". Introduktion til energy-bond-graph formalismen.
Af: Peder Voetmann Christiansen.

9/78 "OM PRAKSIS' INDFLYDELSE PÅ MATEMATIKKENS UDVIKLING". - Motiver til Kepler's: "Nova Stereometria Doliorum Vinariom".
Projektrapport af: Lasse Rasmussen.
Vejleder: Anders Madsen.

10/79 "TERMODYNAMIK I GYMNASIET".
Projektrapport af: Jan Christensen og Jeanne Mortensen.
Vejledere: Karin Beyer og Peder Voetmann Christiansen.

11/79 "STATISTISKE MATERIALER".
Af: Jørgen Larsen.

12/79 "LINEÆRE DIFFERENTIALLIGNINGER OG DIFFERENTIALLIGNINGSSYSTEMER".
Af: Mogens Brun Heefelt.
Nr. 12 er udgået.

13/79 "CAVENDISH'S FORSØG I GYMNASIET".
Projektrapport af: Gert Kreinøe.
Vejleder: Albert Chr. Paulsen.

14/79 "BOOKS ABOUT MATHEMATICS: History, Philosophy, Education, Models, System Theory, and Works of".
Af: Else Høyrup.
Nr. 14 er p.t. udgået.

15/79 "STRUKTUREL STABILITET OG KATASTROFER i systemer i og udenfor termodynamisk ligevægt".
Specialeopgave af: Leif S. Striegler.
Vejleder: Peder Voetmann Christiansen.

16/79 "STATISTIK I KRÆFTFORSKNINGEN".
Projektrapport af: Michael Olsen og Jørn Jensen.
Vejleder: Jørgen Larsen.

17/79 "AT SPØRGE OG AT SVARE i fysikundervisningen".
Af: Albert Christian Paulsen.

18/79 "MATHEMATICS AND THE REAL WORLD", Proceedings af an International Workshop, Roskilde University Centre, Denmark, 1978.
Preprint.
Af: Bernhelm Booss og Mogens Niss (eds.)

19/79 "GEOMETRI, SKOLE OG VIRKELIGHED".
Projektrapport af: Tom J. Andersen, Tommy R. Andersen og Per H.H. Larsen.
Vejleder: Mogens Niss.

20/79 "STATISTISKE MODELLER TIL BESTEMMELSE AF SIKRE DOSER FOR CARCINOGENE STOFFER".
Projektrapport af: Michael Olsen og Jørn Jensen.
Vejleder: Jørgen Larsen

21/79 "KONTROL I GYMNASIET-FORMÅL OG KONSEKVENSER".
Projektrapport af: Crilles Bacher, Per S.Jensen, Preben Jensen og Torben Nysteen.

22/79 "SEMIOTIK OG SYSTEMEGENSKABER (1)".
1-port lineært response og støj i fysikken.
Af: Peder Voetmann Christiansen.

23/79 "ON THE HISTORY AF EARLY WAVE MECHANICS - with special emphasis on the role af realitivity".
Af: Helge Kragh.

24/80 a+b "MATEMATIKOPFATTELSER HOS 2.G'ERE".
1. En analyse. 2. Interviewmateriale.
Projektrapport af: Jan Christensen og Knud Lindhardt Rasmussen.
Vejleder: Mogens Niss.

25/80 "EKSAMENSOPGAVER", Dybdemodulet/fysik 1974-79.

26/80 "OM MATEMATISKE MODELLER".
En projektrapport og to artikler.
Af: Jens Højgaard Jensen m.fl.

27/80 "METHODOLOGY AND PHILOSOPHY AF SCIENCE IN PAUL DIRAC's PHYSICS".
Af: Helge Kragh.

28/80 "DIELEKTRISK RELAXATION - et forslag til en ny model bygget på væskernes viscoelastiske egenskaber".
Projektrapport af: Gert Kreinøe.
Vejleder: Niels Boye Olsen.

29/80 "ODIN - undervisningsmateriale til et kursus i differentialligningsmodeller".
Projektrapport af: Tommy R. Andersen, Per H.H. Larsen og Peter H. Lassen.
Vejleder: Mogens Brun Heefelt.

30/80 "FUSIONSENERGIEN - - - ATOMSAMFUNDETS ENDESTATION".
Af: Oluf Danielsen.
Nr. 30 er udgået.

31/80 "VIDENSKABSTEORETISKE PROBLEMER VED UNDERVISNINGSSYSTEMER BASERET PÅ MÆNGDELÆRE".
Projektrapport af: Troels Lange og Jørgen Karrebæk.
Vejleder: Stig Andur Pedersen.
Nr. 31 er p.t. udgået.

32/80 "POLYMERE STOFFERS VISCOELASTISKE EGENSKABER - BELYST VED HJÆLP AF MEKANISKE IMPEDANSMÅLINGER OG MÖSSBAUEREFFEKTMÅLINGER".
Projektrapport af: Crilles Bacher og Preben Jensen.
Vejledere: Niels Boye Olsen og Peder Voetmann Christiansen.

33/80 "KONSTITUERING AF FAG INDEN FOR TEKNISK - NATURVIDENSKABELIGE UDDANNELSER. I-II".
Af: Arne Jakobsen.

34/80 "ENVIRONMENTAL IMPACT AF WIND ENERGY UTILIZATION".
ENERGY SERIES NO. I.
Af: Bent Sørensen
Nr. 34 er udgået.

35/80 "HISTORISKE STUDIER I DEN NYERE ATOMFYSIKS UDVIKLING".
Af: Helge Kragh.

36/80 "HVAD ER MENINGEN MED MATEMATIKUNDERVISNINGEN?".
Fire artikler.
Af: Mogens Niss.

37/80 "RENEWABLE ENERGY AND ENERGY STORAGE".
ENERGY SERIES NO. 2.
Af: Bent Sørensen.

38/81 "TIL EN HISTORIETEORI OM NATURERKENDELSE, TEKNOLOGI OG SAMFUND".
Projektrapport af: Erik Gade, Hans Hedal, Henrik Lau og Finn Physant.
Vejledere: Stig Andur Pedersen, Helge Kragh og Ib Thiersen.
Nr. 38 er p.t. udgået.

39/81 "TIL KRITIKKEN AF VÆKSTØKONOMIEN".
Af: Jens Højgaard Jensen.

40/81 "TELEKOMMUNIKATION I DANMARK - oplæg til en teknologivurdering".
Projektrapport af: Arne Jørgensen, Bruno Petersen og Jan Vedde.
Vejleder: Per Nørgaard.

41/81 "PLANNING AND POLICY CONSIDERATIONS RELATED TO THE INTRODUCTION OF RENEWABLE ENERGY SOURCES INTO ENERGY SUPPLY SYSTEMS".
ENERGY SERIES NO. 3.
Af: Bent Sørensen.

42/81 "VIDENSKAB TEORI SAMFUND - En introduktion til materialistiske videnskabsopfattelser".
Af: Helge Kragh og Stig Andur Pedersen.

43/81 1."COMPARATIVE RISK ASSESSMENT OF TOTAL ENERGY SYSTEMS".
2."ADVANTAGES AND DISADVANTAGES OF DECENTRALIZATION".
ENERGY SERIES NO. 4.
Af: Bent Sørensen.

44/81 "HISTORISKE UNDERSØGELSER AF DE EKSPERIMENTELLE FORUDSÆTNINGER FOR RUTHERFORDS ATOMMODEL".
Projektrapport af: Niels Thor Nielsen.
Vejleder: Bent C. Jørgensen.

45/82 Er aldrig udkommet.

46/82 "EKSEMPLARISK UNDERVISNING OG FYSISK ERKENDESE-
1+11 ILLUSTRERET VED TO EKSEMPLER".
Projektrapport af: Torben O.Olsen, Lasse Rasmussen og Niels Dreyer Sørensen.
Vejleder: Bent C. Jørgensen.

47/82 "BARSEBÄCK OG DET VÆRST OFFICIELT-TÆNKELIGE UHELD".
ENERGY SERIES NO. 5.
Af: Bent Sørensen.

48/82 "EN UNDERSØGELSE AF MATEMATIKUNDERVISNINGEN PÅ ADGANGSKURSUS TIL KØBENHAVNS TEKNIKUM".
Projektrapport af: Lis Eilertzen, Jørgen Karrebæk, Troels Lange, Preben Nørregaard, Lissi Pedesen, Laust Rishøj, Lill Røn og Isac Showiki.
Vejleder: Mogens Niss.

49/82 "ANALYSE AF MULTISPEKTRALE SATELLITBILLEDER".
Projektrapport af: Preben Nørregaard.
Vejledere: Jørgen Larsen og Rasmus Ole Rasmussen.

50/82 "HERSLEV - MULIGHEDER FOR VEDVARENDE ENERGI I EN LANDSBY".
ENERGY SERIES NO. 6.
Rapport af: Bent Christensen, Bent Hove Jensen, Dennis B. Møller, Bjarne Laursen, Bjarne Lillethorup og Jacob Mørch Pedersen.
Vejleder: Bent Sørensen.

51/82 "HVAD KAN DER GØRES FOR AT AFHJÆLPE PIGERS BLOKERING OVERFOR MATEMATIK ?"
Projektrapport af: Lis Eilertzen, Lissi Pedersen, Lill Røn og Susanne Stender.

52/82 "DESUSPENSION OF SPLITTING ELLIPTIC SYMBOLS".
Af: Bernhelm Booss og Krzysztof Wojciechowski.

53/82 "THE CONSTITUTION OF SUBJECTS IN ENGINEERING EDUCATION".
Af: Arne Jacobsen og Stig Andur Pedersen.

54/82 "FUTURES RESEARCH" - A Philosophical Analysis of Its Subject-Matter and Methods.
Af: Stig Andur Pedersen og Johannes Witt-Hansen.

55/82 "MATEMATISKE MODELLER" - Litteratur på Roskilde Universitetsbibliotek.
En biografi.
Af: Else Høyrup.

Vedr. tekst nr. 55/82 se også tekst nr. 62/83.

56/82 "EN - TO - MANGE" -
En undersøgelse af matematisk økologi.
Projektrapport af: Troels Lange.
Vejleder: Anders Madsen.

57/83 "ASPECT EKSPERIMENTET"-
Skjulte variable i kvantemekanikken?
Projektrapport af: Tom Juul Andersen.
Vejleder: Peder Voetmann Christiansen.
Nr. 57 er udgået.

58/83 "MATEMATISKE VANDRINGER" - Modelbetragtninger over spredning af dyr mellem småbiotoper i agerlandet.
Projektrapport af: Per Hammershøj Jensen og Lene Vagn Rasmussen.
Vejleder: Jørgen Larsen.

59/83 "THE METHODOLOGY OF ENERGY PLANNING".
ENERGY SERIES NO. 7.
Af: Bent Sørensen.

60/83 "MATEMATISK MODEKSPERTISE"- et eksempel.
Projektrapport af: Erik O. Gade, Jørgen Karrebæk og Preben Nørregaard.
Vejleder: Anders Madsen.

61/83 "FYSIKS IDEOLOGISKE FUNKTION, SOM ET EKSEMPEL PÅ EN NATURVIDENSKAB - HISTORISK SET".
Projektrapport af: Annette Post Nielsen.
Vejledere: Jens Høyrup, Jens Højgaard Jensen og Jørgen Vogelius.

62/83 "MATEMATISKE MODELLER" - Litteratur på Roskilde Universitetsbibliotek.
En biografi 2. rev. udgave.
Af: Else Høyrup.

63/83 "CREATING ENERGY FUTURES:A SHORT GUIDE TO ENERGY PLANNING".
ENERGY SERIES No. 8.
Af: David Crossley og Bent Sørensen.

64/83 "VON MATEMATIK UND KRIEG".
Af: Berhelm Booss og Jens Høyrup.

65/83 "ANVENDT MATEMATIK - TEORI ELLER PRAKSIS".
Projektrapport af: Per Hedegård Andersen, Kirsten Habekost, Carsten Holst-Jensen, Annelise von Moos, Else Marie Pedersen og Erling Møller Pedersen.
Vejledere: Bernhelm Booss og Klaus Grünbaum.

66/83 "MATEMATISKE MODELLER FOR PERIODISK SELEKTION I ESCHERICHIA COLI".
Projektrapport af: Hanne Lisbet Andersen, Ole Richard Jensen og Klavs Frisdahl.
Vejledere: Jørgen Larsen og Anders Hede Madsen.

67/83 "ELEPSOIDE METODEN - EN NY METODE TIL LINEÆR PROGRAMMERING?"
Projektrapport af: Lone Biilmann og Lars Boye.
Vejleder: Mogens Brun Heefelt.

68/83 "STOKASTISKE MODELLER I POPULATIONSGENETIK"
- til kritikken af teoriladede modeller.
Projektrapport af: Lise Odgård Gade, Susanne Hansen, Michael Hviid og Frank Mølgård Olsen.
Vejleder: Jørgen Larsen.

69/83 "ELEVFORUDSÆTNINGER I FYSIK"
- en test i 1.g med kommentarer.

Af: Albert C. Paulsen.

70/83 "INDLÆRINGS - OG FORMIDLINGSPROBLEMER I MATEMATIK PÅ VOKSENUNDERVISNINGSNIVEAU".
Projektrapport af: Hanne Lisbet Andersen, Torben J. Andreasen, Svend Åge Houmann, Helle Glerup Jensen, Keld Fl. Nielsen, Lene Vagn Rasmussen.
Vejleder: Klaus Grünbaum og Anders Hede Madsen.

71/83 "PIGER OG FYSIK"
- et problem og en udfordring for skolen?
Af: Karin Beyer, Sussanne Blegaa, Birthe Olsen, Jette Reich og Mette Vedelsby.

72/83 "VERDEN IFØLGE PEIRCE" - to metafysiske essays, om og af C.S Peirce.
Af: Peder Voetmann Christiansen.

73/83 ""EN ENERGIANALYSE AF LANDBRUG"
- økologisk contra traditionelt.
ENERGY SERIES NO. 9
Specialeopgave i fysik af: Bent Hove Jensen.
Vejleder: Bent Sørensen.

74/84 "MINIATURISERING AF MIKROELEKTRONIK" - om videnskabeliggjort teknologi og nytten af at lære fysik.
Projektrapport af: Bodil Harder og Linda Szkotak Jensen.
Vejledere: Jens Højgaard Jensen og Bent C. Jørgensen.

75/84 "MATEMATIKUNDERVISNINGEN I FREMTIDENS GYMNASIUM"
- Case: Lineær programmering.
Projektrapport af: Morten Blomhøj, Klavs Frisdahl og Frank Mølgaard Olsen.
Vejledere: Mogens Brun Heefelt og Jens Bjørneboe.

76/84 "KERNEKRAFT I DANMARK?" - Et høringssvar indkaldt af miljøministeriet, med kritik af miljøstyrelsens rapporter af 15. marts 1984.
ENERGY SERIES No. 10
Af: Niels Boye Olsen og Bent Sørensen.

77/84 "POLITISKE INDEKS - FUP ELLER FAKTA?"
Opinionsundersøgelser belyst ved statistiske modeller.
Projektrapport af: Svend Åge Houmann, Keld Nielsen og Susanne Stender.
Vejledere: Jørgen Larsen og Jens Bjørneboe.

78/84 "JÆVNSTRØMSLEDNINGSEVNE OG GITTERSTRUKTUT I AMORFT GERMANIUM".
Specialrapport af: Hans Hedal, Frank C. Ludvigsen og Finn C. Physant.
Vejleder: Niels Boye Olsen.

79/84 "MATEMATIK OG ALMENDANNELSE".
Projektrapport af: Henrik Coster, Mikael Wennerberg Johansen, Povl Kattler, Birgitte Lydholm og Morten Overgaard Nielsen.
Vejleder: Bernhelm Booss.

80/84 "KURSUSMATERIALE TIL MATEMATIK B".
Af: Mogens Brun Heefelt.

81/84 "FREKVENSAFHÆNGIG LEDNINGSEVNE I AMORFT GERMANIUM".
Specialerapport af: Jørgen Wind Petersen og Jan Christensen.
Vejleder: Niels Boye Olsen.

82/84 "MATEMATIK - OG FYSIKUNDERVISNINGEN I DET AUTOMATISEREDE SAMFUND".
Rapport fra et seminar afholdt i Hvidovre 25-27 april 1983.
Red.: Jens Højgaard Jensen, Bent C. Jørgensen og Mogens Niss.

83/84 "ON THE QUANTIFICATION OF SECURITY":
PEACE RESEARCH SERIES NO. 1
Af: Bent Sørensen
nr. 83 er p.t. udgået

84/84 "NOGLE ARTIKLER OM MATEMATIK, FYSIK OG ALMENDANNELSE".
Af: Jens Højgaard Jensen, Mogens Niss m. fl.

85/84 "CENTRIFUGALREGULATORER OG MATEMATIK".
Specialerapport af: Per Hedegård Andersen, Carsten Holst-Jensen, Else Marie Pedersen og Erling Møller Pedersen.
Vejleder: Stig Andur Pedersen.

86/84 "SECURITY IMPLICATIONS OF ALTERNATIVE DEFENSE OPTIONS FOR WESTERN EUROPE".
PEACE RESEARCH SERIES NO. 2
Af: Bent Sørensen.

87/84 "A SIMPLE MODEL OF AC HOPPING CONDUCTIVITY IN DISORDERED SOLIDS".
Af: Jeppe C. Dyre.

88/84 "RISE, FALL AND RESURRECTION OF INFINITESIMALS".
Af: Detlef Laugwitz.

89/84 "FJERNVARMEOPTIMERING".
Af: Bjarne Lillethorup og Jacob Mørch Pedersen.

90/84 "ENERGI I 1.G - EN TEORI FOR TILRETTELÆGGELSE".
Af: Albert Chr. Paulsen.

91/85 "KVANTETEORI FOR GYMNASIET".
1. Lærervejledning
Projektrapport af: Biger Lundgren, Henning Sten Hansen og John Johansson.
Vejleder: Torsten Meyer.

92/85 "KVANTETEORI FOR GYMNASIET".
2. Materiale
Projektrapport af: Biger Lundgren, Henning Sten Hansen og John Johansson.
Vejleder: Torsten Meyer.

93/85 "THE SEMIOTICS OF QUANTUM - NON - LOCALITY".
Af: Peder Voetmann Christiansen.

94/85 "TREENIGHEDEN BOURBAKI - generalen, matematikeren og ånden".
Projektrapport af: Morten Blomhøj, Klavs Frisdahl og Frank M. Olsen.
Vejleder: Mogens Niss.

95/85 "AN ALTERNATIV DEFENSE PLAN FOR WESTERN EUROPE".
PEACE RESEARCH SERIES NO. 3
Af: Bent Sørensen

96/85 "ASPEKTER VED KRAFTVARMEFORSYNING".
Af: Bjarne Lilletorup.
Vejleder: Bent Sørensen.

97/85 "ON THE PHYSICS OF A.C. HOPPING CONDUCTIVITY".
Af: Jeppe C. Dyre.

98/85 "VALGMULIGHEDER I INFORMATIONSALDEREN".
Af: Bent Sørensen.

99/85 "Der er langt fra Q til R".
Projektrapport af: Niels Jørgensen og Mikael Klintorp.
Vejleder: Stig Andur Pedersen.

100/85 "TALSYSTEMETS OPBYGNING".
Af: Mogens Niss.

101/85 "EXTENDED MOMENTUM THEORY FOR WINDMILLS IN PERTURBATIVE FORM".
Af: Ganesh Sengupta.

102/85 OPSTILLING OG ANALYSE AF MATEMATISKE MODELLER, BELYST VED MODELLER OVER KØERS FODEROPTAGELSE OG - OMSÆTNING".
Projektrapport af: Lis Eilertzen, Kirsten Habekost, Lill Røn og Susanne Stender.
Vejleder: Klaus Grünbaum.

103/85 "ØDSLE KOLDKRIGERE OG VIDENSKABENS LYSE IDEER".
Projektrapport af: Niels Ole Dam og Kurt Jensen.
Vejleder: Bent Sørensen.

104/85 "ANALOGREGNEMASKINEN OG LORENZLIGNINGER".
Af: Jens Jæger.

105/85 "THE FREQUENCY DEPENDENCE OF THE SPECIFIC HEAT AF THE GLASS REANSITION".
Af: Tage Christensen.

"A SIMPLE MODEL AF AC HOPPING CONDUCTIVITY".
Af: Jeppe C. Dyre.
Contributions to the Third International Conference on the Structure of Non - Crystalline Materials held in Grenoble July 1985.

106/85 "QUANTUM THEORY OF EXTENDED PARTICLES".
Af: Bent Sørensen.

107/85 "EN MYG GØR INGEN EPIDEMI",
- flodblindhed som eksempel på matematisk modellering af et epidemiologisk problem.
Projektrapport af: Per Hedegård Andersen, Lars Boye, CarstenHolst Jensen, Else Marie Pedersen og Erling Møller Pedersen.
Vejleder: Jesper Larsen.

108/85 "APPLICATIONS AND MODELLING IN THE MATEMATICS CURRICULUM" - state and trends -
Af: Mogens Niss.

109/85 "COX I STUDIETIDEN" - Cox's regressionsmodel anvendt på studenteroplysninger fra RUC.
Projektrapport af: Mikael Wennerberg Johansen, Poul Katler og Torben J. Andreasen.
Vejleder: Jørgen Larsen.

110/85 "PLANNING FOR SECURITY".
Af: Bent Sørensen

111/85 JORDEN RUNDT PÅ FLADE KORT".
Projektrapport af: Birgit Andresen, Beatriz Quinones og Jimmy Staal.
Vejleder: Mogens Niss.

112/85 "VIDENSKABELIGGØRELSE AF DANSK TEKNOLOGISK INNOVATION FREM TIL 1950 - BELYST VED EKSEMPLER".
Projektrapport af: Erik Odgaard Gade, Hans Hedal, Frank C. Ludvigsen, Annette Post Nielsen og Finn Physant.
Vejleder: Claus Bryld og Bent C. Jørgensen.

113/85 "DESUSPENSION OF SPLITTING ELLIPTIC SYMBOLS II".
Af: Bernhelm Booss og Krzysztof Wojciechowski.

114/85 "ANVENDELSE AF GRAFISKE METODER TIL ANALYSE AF KONTIGENSTABELLER".
Projektrapport af: Lone Biilmann, Ole R. Jensen og Arne-Lise von Moos.
Vejleder: Jørgen Larsen.

115/85 "MATEMATIKKENS UDVIKLING OP TIL RENÆSSANCEN".
Af: Mogens Niss.

116/85 "A PHENOMENOLOGICAL MODEL FOR THE MEYER-NELDEL RULE".
Af: Jeppe C. Dyre.

117/85 "KRAFT & FJERNVARMEOPTIMERING"
Af: Jacob Mørch Pedersen.
Vejleder: Bent Sørensen

118/85 "TILFÆLDIGHEDEN OG NØDVENDIGHEDEN IFØLGE PEIRCE OG FYSIKKEN".
Af: Peder Voetmann Christiansen

119/86 "DET ER GANSKE VIST - - EUKLIDS FEMTE POSTULAT KUNNE NOK SKABE RØRE I ANDEDAMMEN".
Af: Iben Maj Christiansen
Vejleder: Mogens Niss.

120/86 "ET ANTAL STATISTISKE STANDARDMODELLER".
Af: Jørgen Larsen

121/86 "SIMULATION I KONTINUERT TID".
Af: Peder Voetmann Christiansen.

122/86 "ON THE MECHANISM OF GLASS IONIC CONDUCTIVITY".
Af: Jeppe C. Dyre.

123/86 "GYMNASIEFYSIKKEN OG DEN STORE VERDEN".
Fysiklærerforeningen, IMFUFA, RUC.

124/86 "OPGAVESAMLING I MATEMATIK".
Samtlige opgaver stillet i tiden 1974-jan. 1986.

125/86 "UVBY,β - systemet - en effektiv fotometrisk spektralklassifikation af B-,A- og F-stjerner".
Projektrapport af: Birger Lundgren.

126/86 "OM UDVIKLINGEN AF DEN SPECIELLE RELATIVITETSTEORI".
Projektrapport af: Lise Odgaard & Linda Szkotak Jensen
Vejledere: Karin Beyer & Stig Andur Pedersen.

127/86 "GALOIS' BIDRAG TIL UDVIKLINGEN AF DEN ABSTRAKTE ALGEBRA".
Projektrapport af: Pernille Sand, Heine Larsen & Lars Frandsen.
Vejleder: Mogens Niss.

128/86 "SMÅKRYB" - om ikke-standard analyse.
Projektrapport af: Niels Jørgensen & Mikael Klintorp.
Vejleder: Jeppe Dyre.

129/86 "PHYSICS IN SOCIETY"
Lecture Notes 1983 (1986)
Af: Bent Sørensen

130/86 "Studies in Wind Power"
Af: Bent Sørensen

131/86 "FYSIK OG SAMFUND" - Et integreret fysik/historieprojekt om naturanskuelsens historiske udvikling og dens samfundsmæssige betingethed.
Projektrapport af: Jakob Heckscher, Søren Brønd, Andy Wierød.
Vejledere: Jens Høyrup, Jørgen Vogelius, Jens Højgaard Jensen.

132/86 "FYSIK OG DANNELSE"
Projektrapport af: Søren Brønd, Andy Wierød.
Vejledere: Karin Beyer, Jørgen Vogelius.

133/86 "CHERNOBYL ACCIDENT: ASSESSING THE DATA.
ENERGY SERIES NO. 15.
AF: Bent Sørensen.

134/87 "THE D.C. AND THE A.C. ELECTRICAL TRANSPORT IN AsSeTe SYSTEM"
Authors: M.B.El-Den, N.B.Olsen, Ib Høst Pedersen, Petr Viščor

135/87 "INTUITIONISTISK MATEMATIKS METODER OG ERKENDELSESTEORETISKE FORUDSÆTNINGER"
MASTEMATIKSPECIALE: Claus Larsen
Vejledere: Anton Jensen og Stig Andur Pedersen

136/87 "Mystisk og naturlig filosofi: En skitse af kristendommens første og andet møde med græsk filosofi"
Projektrapport af Frank Colding Ludvigsen
Vejledere: Historie: Ib Thiersen
Fysik: Jens Højgaard Jensen

137/87 "HOPMODELLER FOR ELEKTRISK LEDNING I UORDNEDE FASTE STOFFER" - Resume af licentiatafhandling
Af: Jeppe Dyre
Vejledere: Niels Boye Olsen og Peder Voetmann Christiansen.

138/87 "JOSEPHSON EFFECT AND CIRCLE MAP."

Paper presented at The International Workshop on Teaching Nonlinear Phenomena at Universities and Schools, "Chaos in Education". Balaton, Hungary, 26 April-2 May 1987.

By: Peder Voetmann Christiansen

139/87 "Machbarkeit nichtbeherrschbarer Technik durch Fortschritte in der Erkennbarkeit der Natur"

Af: Bernhelm Booss-Bavnbek
Martin Bohle-Carbonell

140/87 "ON THE TOPOLOGY OF SPACES OF HOLOMORPHIC MAPS"

By: Jens Gravesen

141/87 "RADIOMETERS UDVIKLING AF BLODGASAPPARATUR - ET TEKNOLOGIHISTORISK PROJEKT"

Projektrapport af Finn C. Physant

Vejleder: Ib Thiersen

142/87 "The Calderón Projektor for Operators With Splitting Elliptic Symbols"

by: Bernhelm Booss-Bavnbek og
Krzysztof P. Wojciechowski

143/87 "Kursusmateriale til Matematik på NAT-BAS"

af: Mogens Brun Heefelt

144/87 "Context and Non-Locality - A Peircan Approach

Paper presented at the Symposium on the Foundations of Modern Physics The Copenhagen Interpretation 60 Years after the Como Lecture. Joensuu, Finland, 6 - 8 august 1987.

By: Peder Voetmann Christiansen

145/87 "AIMS AND SCOPE OF APPLICATIONS AND MODELLING IN MATHEMATICS CURRICULA"

Manuscript of a plenary lecture delivered at ICMTA 3, Kassel, FRG 8.-11.9.1987

By: Mogens Niss

146/87 "BESTEMMELSE AF BULKRESISTIVITETEN I SILICIUM"

- en ny frekvensbaseret målemetode.

Fysikspeciale af Jan Vedde

Vejledere: Niels Boye Olsen & Petr Viščor

147/87 "Rapport om BIS på NAT-BAS"

redigeret af: Mogens Brun Heefelt

148/87 "Naturvidenskabsundervisning med Samfundsperspektiv"

af: Peter Colding-Jørgensen DLH
Albert Chr. Paulsen

149/87 "In-Situ Measurements of the density of amorphous germanium prepared in ultra high vacuum"

by: Petr Viščor

150/87 "Structure and the Existence of the first sharp diffraction peak in amorphous germanium prepared in UHV and measured in-situ"

by: Petr Viščor

151/87 "DYNAMISK PROGRAMMERING"

Matematikprojekt af:
Birgit Andresen, Keld Nielsen og Jimmy Staal

Vejleder: Mogens Niss

152/87 "PSEUDO-DIFFERENTIAL PROJECTIONS AND THE TOPOLOGY OF CERTAIN SPACES OF ELLIPTIC BOUNDARY VALUE PROBLEMS"

by: Bernhelm Booss-Bavnbek
Krzysztof P. Wojciechowski

153/88 "HALVLEDERTEKNOLOGIENS UDVIKLING MELLEM MILITÆRE OG CIVILE KRÆFTER"

Et eksempel på humanistisk teknologihistorie
Historiespeciale

Af: Hans Hedal

Vejleder: Ib Thiersen

154/88 "MASTER EQUATION APPROACH TO VISCOUS LIQUIDS AND THE GLASS TRANSITION"

By: Jeppe Dyre

155/88 "A NOTE ON THE ACTION OF THE POISSON SOLUTION OPERATOR TO THE DIRICHLET PROBLEM FOR A FORMALLY SELFADJOINT DIFFERENTIAL OPERATOR"

by: Michael Pedersen

156/88 "THE RANDOM FREE ENERGY BARRIER MODEL FOR AC CONDUCTION IN DISORDERED SOLIDS"

by: Jeppe C. Dyre

157/88 " STABILIZATION OF PARTIAL DIFFERENTIAL EQUATIONS BY FINITE DIMENSIONAL BOUNDARY FEEDBACK CONTROL: A pseudo-differential approach."

by: Michael Pedersen

158/88 "UNIFIED FORMALISM FOR EXCESS CURRENT NOISE IN RANDOM WALK MODELS"

by: Jeppe Dyre

159/88 "STUDIES IN SOLAR ENERGY"

by: Bent Sørensen

160/88 "LOOP GROUPS AND INSTANTONS IN DIMENSION TWO"

by: Jens Gravesen

161/88 "PSEUDO-DIFFERENTIAL PERTURBATIONS AND STABILIZATION OF DISTRIBUTED PARAMETER SYSTEMS:

Dirichlet feedback control problems"

by: Michael Pedersen

162/88 "PIGER & FYSIK - OG MEGET MERE"

AF: Karin Beyer, Sussanne Blegaa, Birthe Olsen,
Jette Reich , Mette Vedelsby

163/88 "EN MATEMATISK MODEL TIL BESTEMMELSE AF PERMEABILITETEN FOR BLOD-NETHINDE-BARRIEREN"

Af: Finn Langberg, Michael Jarden, Lars Frellesen

Vejleder: Jesper Larsen