Sandsynlighedsregning & Statistik

for matematikstuderende

Jørgen Larsen

2006

Roskilde Universitet

Teksten er sat med skriften Kp-Fonts ved hjælp af KOMA-Script og LaTeX. Tegningerne er fremstillet med METAPOST.

Oktober 2010.

Indhold

Forord

Sandsynlighedsregning og statistik er to emneområder der dels studeres på deres egne betingelser, dels optræder som støttefag eller hjælpefag i en række sammenhænge, og den måde man bør formidle fagenes indhold på, afhænger i høj grad af hvem der er målgruppen. Denne bog er ikke skrevet til personer der specialiserer sig i sandsynlighedsregning og/eller statistik, og heller ikke til personer der har brug for statistik som støttefag, men derimod til personer der som led i en generel matematikuddannelse skal vide noget om sandsynlighedsregning og statistik. – Forskellige udkast til bogen har i en årrække været anvendt på RUCs matematikuddannelse.

Ved tilrettelæggelsen af undervisningsforløb der introducerer til sandsynlighedsregning, er det et stadigt tilbagevendende spørgsmål hvor meget (eller måske snarere hvor lidt) vægt man skal lægge på en generel aksiomatisk fremstilling. Hvis der er tale om en almen introduktion der henvender sig til en ikke nødvendigvis matematisk interesseret eller kvalificeret målgruppe, er der ikke så meget at være i tvivl om – den målteoretiske aksiomatisering à la Kolmogorov skal ikke med, eller den kan måske blive nævnt i en diskret fodnote. Men når der er tale om en del af en matematikuddannelse, stiller sagen sig anderledes; her kan der være en god pointe i at studere hvordan den almindeligvis »genstandsløse« matematikformalisme fungerer når man ønsker at etablere et sæt byggesten til en helt bestemt slags modelleringsopgaver (modellering af tilfældighedsfænomener), og det kan derfor være på sin plads at beskæftige sig med fundamentet for den matematiske teoribygning.

Sandsynlighedsregning er afgjort en matematikdisciplin, og statistik må siges at være overordentlig matematik-involveret. Men begge emneområder er, hvad angår deres matematikindhold, organiseret væsentlig anderledes end »almindelige« matematiske emneområder (i hvert fald dem som normalt indgår i undervisningsprogrammer), fordi de i overvejende grad er styret/reguleret af at de skal kunne bestemte ting, f.eks. bevise Store tals Lov og Den centrale Grænseværdisætning, og kun i mindre grad af de gældende internt matematiske normer for hvordan teoriområder skal opbygges og præsenteres, og man vil eksempelvis gå fejl af mange pointer hvis man tror at sandsynlighedsregningen (den målteoribaserede sandsynlighedsregning) »bare« er et specialtilfælde af

emneområdet mål- og integralteori. Endvidere vil den der skal sætte sig ind i emneområderne sandsynlighedsregning og statistik, hurtigt opleve at man skal benytte begreber, metoder og resultater fra vidt forskellige »traditionelle« matematikområder, og dette er formentlig én grund til at sandsynlighedsregning og statistik opfattes som svært.

Fremstillingen er på traditionel vis delt op i en sandsynlighedsregningsdel og en statistikdel. De to dele er temmelig forskellige i stil og opbygning.

Del I præsenterer sandsynlighedsregningens grundlæggende begrebsdannelser og tankegange og de sædvanlige eksempler, men sådan at stofmængden holdes i meget stramme tøjler. Først vises hvordan sandsynlighedsregning i Kolmogorovs aksiomatisering ser ud når man holder sig til endelige udfaldsrum – derved holdes mængden af matematiske besværligheder på et minimum, uden at man behøver give afkald på at kunne bevise de formulerede sætninger ved hjælp af det givne teoriapparat. Derefter udvides teorien til tællelige udfaldsrum, eller i hvert fald til udfaldsrummet $\mathbb{N}_0$ forsynet med σ-algebraen af alle delmængder; her er det stadig muligt at bevise »alle« sætninger, selv med et beskedent matematisk apparatur (det forventes at læseren har kendskab til teorien for uendelige rækker), men man får dog indblik i nogle af vanskelighederne ved uendelige udfaldsrum. Den formalistiske aksiomatiske tilgang fortsættes imidlertid ikke i kapitlet om kontinuerte fordelinger på $\mathbb{R}$ og $\mathbb{R}^n$, dvs. fordelinger som har en tæthedsfunktion, og nu er der ikke længere tale om at alle påstande bevises (blandt andet ikke sætningen om transformation af tætheder, der snarere bør bevises i et analysekursus). Del I afsluttes med to lidt anderledes kapitler, dels et kapitel der behandler et mere afgrænset område, nemlig frembringende funktioner (inklusive lidt om forgreningsprocesser), dels et kapitel der kort giver nogle antydninger af hvordan og hvorfor man beskæftiger sig med sandsynlighedsmål på generelle udfaldsrum.

Del II præsenterer den klassiske matematiske statistik baseret på likelihoodfunktionen: de statistiske modeller er bygget op af almindelige standardfordelinger og et beskedent antal parametre, estimatorerne er som hovedregel maksimaliseringsestimatorer, og hypoteserne testes med likelihoodkvotienttests. Fremstillingen er bygget op med et kapitel om begrebet statistisk model, et kapitel om estimation og et kapitel om hypoteseprøvning; disse kapitler er forsynet med en række eksempler der viser hvordan teorien tager sig ud når den anvendes på bestemte modeltyper eller modeller, og eksemplerne fortsætter ofte fra det ene kapitel til det andet. Efter denne teoriorganiserede fremstilling følger et eksempelorienteret kapitel med tre større gennemregnede eksempler der illustrerer den generelle teori. Del II afsluttes med en introduktion til teorien for lineære normale modeller formuleret i lineær algebra-sprog, i god overensstemmelse med en henved 40-årig tradition inden for dansk matematisk statistik.

Roskilde i august 2006
Jørgen Larsen

Del I

Sandsynlighedsregning

Indledning

Sandsynlighedsregning er en disciplin der beskæftiger sig med en matematisk formalisering af dagligdagsbegreberne sandsynlighed og tilfældighed og dertil knyttede delbegreber. I første omgang kan man måske studse over at der overhovedet skulle kunne gives en matematisk formalisering af tilfældighed: hvis noget er tilfældigt, er det så netop ikke unddraget muligheden for en eksakt beskrivelse? Ikke ganske. Erfaringen viser at i hvert fald nogle typer af tilfældighedsfænomener og tilfældighedseksperimenter udviser betydelige grader af regelmæssighed når man gentager dem et stort antal gange, det gælder f.eks. kast med terninger og mønter, roulettespil og andre former for »lykkespil«. For at kunne tale nærmere om tingene er vi nødt til at indføre forskellige begreber og betegnelser; i første omgang er de lidt upræcise, men senere vil de få en præcis matematisk betydning (som forhåbentlig ikke er alt for fjern fra dagligsprogets).

Tilfældighedseksperimentet giver når det udføres, et resultat af en slags, f.eks. resulterer terningkastet i at terningen viser et bestemt antal øjne; et sådant resultat kaldes et *udfald*. Mængden af mulige udfald kaldes *udfaldsrummet*.

Sandsynligheder er reelle tal der giver en kvantitativ beskrivelse af visse træk ved tilfældighedseksperimentet. Et simpelt eksempel på et sandsynlighedsudsagn kunne være »sandsynligheden for at terningkastet giver udfaldet fem øjne, er $1/6$«; et andet eksempel kunne være »sandsynligheden for at det bliver snevejr juleaften, er $1/20$«. Hvad betyder sådanne udsagn? Nogle mennesker hævder at sandsynlighedsudsagn skal fortolkes som udsagn der beskriver forudsigelser om udfaldet af et bestemt fremtidigt fænomen (f.eks. snevejr juleaften). Andre mener at sandsynlighedsudsagn beskriver den relative hyppighed hvormed det pågældende udfald indtræffer når tilfældighedseksperimentet (f.eks. terningkastet) gentages igen og igen. Den måde som sandsynlighedsregningen formaliseres/aksiomatiseres på, er i høj grad inspireret af at sandsynlighed skal kunne fortolkes som *relativ hyppighed i det lange løb*, men den er ikke bundet til denne bestemte fortolkning.

Sandsynlighedsregningen benytter sig af den simple mængdelæres notationer og begreber – dog med visse ændrede betegnelser, jf. oversigten på næste side. En *sandsynlighed* eller mere præcist et *sandsynlighedsmål* vil blive defineret som en afbildning fra en vis definitionsmængde ind i de reelle tal. Hvad definitions-

Oversigt over forskellige begreber fra mængdelæren og deres betegnelse inden for sandsynlighedsregningen.

Typisk notation	Sandsynlighedsregning	Mængdelære
Ω	udfaldsrum; den sikre hændelse	grundmængde, univers
$\varnothing$	den umulige hændelse	den tomme mængde
ω	udfald	element i Ω
A	hændelse	delmængde af Ω
$A \cap B$	både A og B	fællesmængden af A og B
$A \cup B$	enten A eller B	foreningsmængden af A og B
$A \setminus B$	A men ikke B	differensmængde
A^{c}	den modsatte hændelse til A	komplementærmængden til A, dvs. $\Omega \setminus A$

mængden skal være, er måske ikke ganske klart; eller rettere, i første omgang ville man jo nok tro at den ganske enkelt skulle være udfaldsrummet, men det giver problemer i situationer hvor udfaldsrummet er overtælleligt (f.eks. de reelle tal). Det har vist sig at den rigtige måde at gøre tingene på, er at tale om sandsynligheder for *hændelser*, dvs. visse nærmere fastlagte delmængder af udfaldsrummet. Et sandsynlighedsmål bliver derfor en afbildning der til visse delmængder af udfaldsrummet knytter et reelt tal.

1 Endelige udfaldsrum

I dette kapitel vil vi studere sandsynligheder på endelige udfaldsrum. Det vil foregå på den måde at vi præsenterer de generelle definitioner, men forsimplet til det endelige tilfælde. I forbindelse med mere generelle udfaldsrum dukker der forskellige matematiske besværligheder op som man i det endelige tilfælde helt slipper for.

1.1 Grundlæggende definitioner

Definition 1.1: Sandsynlighedsrum over en endelig mængde
Et sandsynlighedsrum over en endelig mængde er et tripel $(\Omega, \mathscr{F}, \mathrm{P})$ *bestående af*

1. *et udfaldsrum* Ω*, som er en ikke-tom, endelig mængde,*
2. *mængden* $\mathscr{F}$ *af alle delmængder af* Ω*,*
3. *et sandsynlighedsmål på* $(\Omega, \mathscr{F})$*, dvs. en afbildning* $\mathrm{P} : \mathscr{F} \to \mathbb{R}$ *som er*
   - *positiv:* $\mathrm{P}(A) \geq 0$ *for alle* $A \in \mathscr{F}$*,*
   - *normeret:* $\mathrm{P}(\Omega) = 1$*, og*
   - *additiv: hvis* $A_1, A_2, \ldots, A_n \in \mathscr{F}$ *er parvis disjunkte hændelser, så er*

$$\mathrm{P}\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} \mathrm{P}(A_i).$$

Her er to simple eksempler, der i øvrigt også kan bruges til at demonstrere at der faktisk findes matematiske objekter der opfylder definitionen:

Eksempel 1.1: Ligefordeling
Lad Ω være en endelig mængde med n elementer, $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, lad $\mathscr{F}$ være mængden af delmængder af Ω, og lad P være givet ved $\mathrm{P}(A) = \frac{1}{n}\#A$, hvor $\#A$ står for »antal elementer i A«. Så opfylder $(\Omega, \mathscr{F}, \mathrm{P})$ betingelserne for at være et sandsynlighedsrum (additiviteten følger af additiviteten af antalsfunktionen). Sandsynlighedsmålet P hedder *ligefordelingen* på Ω (fordi det fordeler »sandsynlighedsmassen« ligeligt ud over udfaldsrummet).

Eksempel 1.2: Etpunktsfordeling
Lad Ω være en endelig mængde, og lad $\omega_0 \in \Omega$ være et udvalgt punkt. Lad $\mathscr{F}$ være mængden af delmængder af Ω, og sæt $\mathrm{P}(A) = 1$ hvis $\omega_0 \in A$ og $\mathrm{P}(A) = 0$ ellers. Så opfylder $(\Omega, \mathscr{F}, \mathrm{P})$ betingelserne for at være et sandsynlighedsrum. Sandsynlighedsmålet P

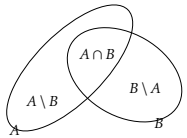
hedder *etpunktsfordelingen* i ω_0 (fordi det placerer al sandsynlighedsmassen i dette ene punkt).

Vi går straks i gang med at vise nogle resultater:

LEMMA 1.1
For vilkårlige hændelser A og B i sandsynlighedsrummet $(\Omega, \mathscr{F}, \mathrm{P})$ gælder:

1. $\mathrm{P}(A) + \mathrm{P}(A^{\mathrm{c}}) = 1$.
2. $\mathrm{P}(\varnothing) = 0$.
3. *Hvis $A \subseteq B$, så er $\mathrm{P}(B \setminus A) = \mathrm{P}(B) - \mathrm{P}(A)$ og dermed $\mathrm{P}(A) \leq \mathrm{P}(B)$.*
   (Dette udtrykkes undertiden på den måde at man siger at P er voksende.)
4. $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$.

BEVIS
Ad 1: De to hændelser A og A^{c} er disjunkte, og deres forening er Ω; derfor er ifølge additivitetsaksiomet $\mathrm{P}(A) + \mathrm{P}(A^{\mathrm{c}}) = \mathrm{P}(\Omega)$, og $\mathrm{P}(\Omega)$ er lig 1.
Ad 2: Da $\varnothing = \Omega^{\mathrm{c}}$, er ifølge det netop viste $\mathrm{P}(\varnothing) = 1 - \mathrm{P}(\Omega) = 1 - 1 = 0$.
Ad 3: Hændelserne A og $B \setminus A$ er disjunkte og deres forening er B; derfor er $\mathrm{P}(A) + \mathrm{P}(B \setminus A) = \mathrm{P}(B)$; da $\mathrm{P}(B \setminus A) \geq 0$, fås at $\mathrm{P}(A) \leq \mathrm{P}(B)$.
Ad 4: De tre hændelser $A \setminus B$, $B \setminus A$ og $A \cap B$ er parvis disjunkte og deres forening er $A \cup B$. Derfor er



$$\begin{aligned}\mathrm{P}(A \cup B) &= \mathrm{P}(A \setminus B) + \mathrm{P}(B \setminus A) + \mathrm{P}(A \cap B) \\ &= \big(\mathrm{P}(A \setminus B) + \mathrm{P}(A \cap B)\big) + \big(\mathrm{P}(B \setminus A) + \mathrm{P}(A \cap B)\big) - \mathrm{P}(A \cap B) \\ &= \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B).\end{aligned}$$

□

I fremstillinger af sandsynlighedsregningen inddrager man altid møntkast og terningkast som eksempler på tilfældighedsfænomener, så det gør vi også her.

Eksempel 1.3: Møntkast
Antag at vi kaster én gang med en mønt og ser efter om den viser Plat eller Krone.

Udfaldsrummet er topunktsmængden $\Omega = \{\text{Plat}, \text{Krone}\}$. Mængden af hændelser er $\mathscr{F} = \{\Omega, \{\text{Krone}\}, \{\text{Plat}\}, \varnothing\}$. Det sandsynlighedsmål P der svarer til at mønten er symmetrisk, altså har lige stor sandsynlighed for at falde på enhver af de to sider, er ligefordelingen på Ω:

$$\mathrm{P}(\Omega) = 1, \quad \mathrm{P}(\{\text{Krone}\}) = 1/2, \quad \mathrm{P}(\{\text{Plat}\}) = 1/2, \quad \mathrm{P}(\varnothing) = 0.$$

Eksempel 1.4: Terningkast
Antag at vi kaster én gang med en almindelig terning og ser efter hvor mange øjne den viser.

Udfaldsrummet er mængden $\Omega = \{1, 2, 3, 4, 5, 6\}$. Mængden $\mathscr{F}$ af hændelser er mængden af delmængder af Ω (så der er $2^6 = 64$ forskellige hændelser). Det sandsynlighedsmål P der svarer til at terningen er symmetrisk, er ligefordelingen på Ω.

Dermed er eksempelvis sandsynligheden for hændelsen $\{3, 6\}$ (antallet af øjne er deleligt med 3) givet som $\mathrm{P}(\{3, 6\}) = 2/6$, fordi hændelsen består af to udfald, og der er seks mulige udfald i alt.

Eksempel 1.5: Simpel stikprøveudtagning
Man har en kasse (eller urne) med s sorte og h hvide kugler, og herfra udtager man en k-stikprøve, dvs. en delmængde med k elementer (det forudsættes at $k \leq s + h$). Kuglerne tænkes udtaget ved *simpel stikprøveudtagning*, dvs. alle $\binom{s+h}{k}$ forskellige delmængder med k elementer (jf. definitionen af binomialkoefficienter side 32) har samme sandsynlighed for at blive udtaget. – Her er altså tale om en ligefordeling på mængden Ω bestående af alle disse delmængder.

Sandsynligheden for hændelsen »netop x sorte kugler« er derfor lig antal k-stikprøver med x sorte kugler og $k - x$ hvide kugler divideret med det samlede antal stikprøver, altså $\binom{s}{x}\binom{h}{k-x} \Big/ \binom{s+h}{k}$. Se også side 34, herunder sætning 1.13.

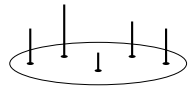
Punktsandsynligheder

Læseren kan med nogen ret undre sig over den noget kringlede måde at matematificere sandsynligheder på, hvorfor kan man ikke bare have en funktion der til hvert udfald knytter sandsynligheden for at det indtræffer? Så længe man opererer med endelige (og tællelige) udfaldsrum kunne man faktisk godt gribe sagen an på den måde, men med overtællelige udfaldsrum går det helt galt (fordi overtælleligt mange positive tal ikke kan summere til noget endeligt). Men da vi nu er i det endelige tilfælde, er følgende definition og sætning af interesse.

DEFINITION 1.2: PUNKTSANDSYNLIGHEDER
Lad $(\Omega, \mathscr{F}, \mathrm{P})$ være et sandsynlighedsrum over en endelig mængde Ω. Funktionen

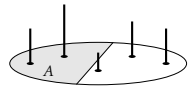
$$\begin{aligned} p : \Omega &\longrightarrow [0\,;1] \\ \omega &\longmapsto \mathrm{P}(\{\omega\}) \end{aligned}$$

kaldes punktsandsynlighederne for P.



Punktsandsynligheder anskueliggøres ofte som sandsynlighedspinde.

SÆTNING 1.2
Hvis p er punktsandsynlighederne for sandsynlighedsmålet P, *så gælder for en vilkårlig hændelse A at* $\mathrm{P}(A) = \sum_{\omega \in A} p(\omega)$.



Bevis
Vi skriver A som disjunkt forening af sine etpunkts-delmængder og bruger additiviteten: $\mathrm{P}(A) = \mathrm{P}\big(\bigcup_{\omega\in A}\{\omega\}\big) = \sum_{\omega\in A}\mathrm{P}(\{\omega\}) = \sum_{\omega\in A} p(\omega)$. □

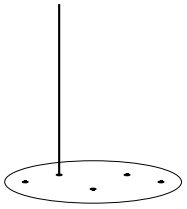
Bemærkninger: En konsekvens af sætningen er at to forskellige sandsynlighedsmål ikke kan have samme punktsandsynlighedsfunktion. En anden konsekvens er at p summerer til 1, dvs. $\sum_{\omega\in\Omega} p(\omega) = 1$; det ser man ved at sætte $A = \Omega$.

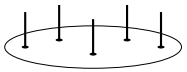
Sætning 1.3
Hvis $p : \Omega \to [0\,;1]$ summerer til 1, dvs. $\sum_{\omega\in\Omega} p(\omega) = 1$, så findes netop et sandsynlighedsmål P *på* Ω *der har p som sine punktsandsynligheder.*

Bevis
Vi kan definere en funktion $\mathrm{P} : \mathscr{F} \to [0\,;+\infty[$ ved $\mathrm{P}(A) = \sum_{\omega\in A} p(\omega)$. Denne funktion er positiv fordi $p \geq 0$, og normeret fordi p summerer til 1. Den er desuden additiv: hvis $A_1, A_2, \ldots, A_n$ er parvis disjunkte hændelser, så er

$$\begin{aligned}\mathrm{P}(A_1 \cup A_2 \cup \ldots \cup A_n) &= \sum_{\omega\in A_1\cup A_2\cup\ldots\cup A_n} p(\omega)\\ &= \sum_{\omega\in A_1} p(\omega) + \sum_{\omega\in A_2} p(\omega) + \ldots + \sum_{\omega\in A_n} p(\omega)\\ &= \mathrm{P}(A_1) + \mathrm{P}(A_2) + \ldots + \mathrm{P}(A_n),\end{aligned}$$

hvor det andet lighedstegn følger af den associative lov for regneoperationen +. Altså opfylder P betingelserne for at være et sandsynlighedsmål. Pr. konstruktion er P's punktsandsynligheder p, og som nævnt i bemærkningen til sætning 1.2 er der kun ét sandsynlighedsmål der kan have p som punktsandsynligheder. □


En etpunktsfordeling


En ligefordeling

Eksempel 1.6
Punktsandsynlighederne for etpunktsfordelingen i ω_0 (jf. eksempel 1.2) er givet ved $p(\omega_0) = 1$, og $p(\omega) = 0$ når $\omega \neq \omega_0$.

Eksempel 1.7
Punktsandsynlighederne for ligefordelingen på $\{\omega_1, \omega_2, \ldots, \omega_n\}$ (jf. eksempel 1.1) er givet ved $p(\omega_i) = 1/n$, $i = 1, 2, \ldots, n$.

Betingede sandsynligheder og fordelinger

Man er ofte interesseret i sandsynligheden for at en hændelse A indtræffer, *givet* at en anden hændelse B vides at indtræffe (eller at være indtruffet) – man taler om den *betingede sandsynlighed* for A givet B.

Definition 1.3: Betinget sandsynlighed
Lad $(\Omega, \mathscr{F}, \mathrm{P})$ være et sandsynlighedsrum, lad A og B være hændelser, og antag at $\mathrm{P}(B) > 0$. Tallet

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

kaldes den betingede sandsynlighed for A givet B.

Eksempel 1.8
Man slår Plat eller Krone med to mønter, en 10-krone og en 20-krone, på én gang. Hvad er sandsynligheden for at 10-kronen viser Krone, givet at mindst en af de to mønter viser Krone?

Der er (iflg. standardmodellen) fire mulige udfald, og udfaldsrummet er

$$\Omega = \{(\text{Plat}, \text{Plat}), (\text{Plat}, \text{Krone}), (\text{Krone}, \text{Plat}), (\text{Krone}, \text{Krone})\},$$

hvor vi skriver 10-kronens resultat først; hvert af disse fire udfald antages at have sandsynlighed $1/4$. Den betingende hændelse B (mindst en Krone) og den omspurgte hændelse A (at 10-kronen viser Krone) er hhv.

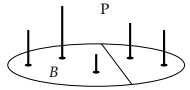
$$\begin{aligned}B &= \{(\text{Plat}, \text{Krone}), (\text{Krone}, \text{Plat}), (\text{Krone}, \text{Krone})\} \quad \text{og}\\ A &= \{(\text{Krone}, \text{Plat}), (\text{Krone}, \text{Krone})\},\end{aligned}$$

så den betingede sandsynlighed for A givet B er

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{2/4}{3/4} = 2/3.$$

Af definition 1.3 følger umiddelbart

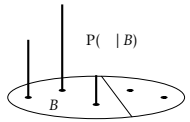
Sætning 1.4
Hvis A og B er hændelser, og hvis $\mathrm{P}(B) > 0$, så er $\mathrm{P}(A \cap B) = \mathrm{P}(A \mid B)\,\mathrm{P}(B)$.



Definition 1.4: Betinget fordeling
Lad $(\Omega, \mathscr{F}, \mathrm{P})$ være et sandsynlighedsrum, og lad B være en hændelse med den egenskab at $\mathrm{P}(B) > 0$. Funktionen

$$\begin{aligned}\mathrm{P}(\quad \mid B) : \mathscr{F} &\longrightarrow [0\,;1]\\ A &\longmapsto \mathrm{P}(A \mid B)\end{aligned}$$

kaldes den betingede fordeling givet B.



Bayes' formel

Antag at Ω kan skrives som en disjunkt forening af de k hændelser $B_1, B_2, \ldots, B_k$ (eller som man også siger: $B_1, B_2, \ldots, B_k$ er en *klassedeling* af Ω). Desuden er der en hændelse A. Det antages endvidere at vi forlods (eller *a priori*) kender

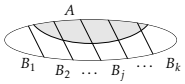
sandsynlighederne $\mathrm{P}(B_1), \mathrm{P}(B_2), \ldots, \mathrm{P}(B_k)$ for de enkelte klasser i klassedelingen, og desuden kendes også alle de betingede sandsynligheder $\mathrm{P}(A \mid B_j)$ for A givet B_j. Opgaven er nu at bestemme sandsynlighederne $\mathrm{P}(B_j \mid A)$ for de enkelte B_j-er, givet at hændelsen A vides at være indtruffet; disse sandsynligheder kaldes *a posteriori* sandsynligheder.

Thomas Bayes
engelsk matematiker og teolog (1702-61). »Bayes' formel« (der hidrører fra Bayes (1763)) spiller i vore dage en altafgørende rolle i den såkaldte bayesianske statistik og i bayesianske netværk.

(Som illustration kan man eksempelvis tænke på en medicinsk diagnosticeringssituation: A er det sæt af symptomer man observerer på patienten, og B-erne er forskellige (hinanden udelukkende) sygdomme der kunne forklare symptomerne. Lægerne har bud på de hyppigheder hvormed sygdommene forekommer, og på sandsynlighederne for at en patient udviser netop symptombilledet A, givet at patienten har sygdommen B_i, $i = 1, 2, \ldots, k$. Lægerne er interesserede i de betingede sandsynligheder for at den patient som har symptomerne A, fejler sygdommen B_i.)

Da $A = \bigcup_{i=1}^{k} A \cap B_i$ hvor der er tale om en disjunkt forening, er

$$\mathrm{P}(A) = \sum_{i=1}^{k} \mathrm{P}(A \cap B_i) = \sum_{i=1}^{k} \mathrm{P}(A \mid B_i)\,\mathrm{P}(B_i),$$



og da $\mathrm{P}(B_j \mid A) = \mathrm{P}(A \cap B_j)/\mathrm{P}(A) = \mathrm{P}(A \mid B_j)\,\mathrm{P}(B_j)/\mathrm{P}(A)$, er dermed

$$\mathrm{P}(B_j \mid A) = \frac{\mathrm{P}(A \mid B_j)\,\mathrm{P}(B_j)}{\sum_{i=1}^{k} \mathrm{P}(A \mid B_i)\,\mathrm{P}(B_i)}. \tag{1.1}$$

Formel (1.1) kaldes *Bayes' formel*; den fortæller hvordan man udregner a posteriori sandsynlighederne $\mathrm{P}(B_j \mid A)$ ud fra a priori sandsynlighederne $\mathrm{P}(B_j)$ og de betingede sandsynligheder $\mathrm{P}(A \mid B_j)$.

Uafhængighed

Hændelser kaldes uafhængige hvis det er sådan at sandsynlighedsudsagn om nogle af dem ikke ændres af kendskabet til hvorvidt andre af dem er indtruffet eller ej.

Man kunne overveje at definere uafhængighed af hændelserne A og B til at betyde at $\mathrm{P}(A \mid B) = \mathrm{P}(A)$, hvilket ved anvendelse af definitionen på betinget sandsynlighed bliver til $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$; den sidste formel har den fordel at den er meningsfuld også når $\mathrm{P}(B) = 0$, samt at A og B indgår symmetrisk. Man definerer derfor uafhængighed af *to* hændelser A og B til at betyde at $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$. – Hvis man vil gøre tingene ordentligt, skal man imidlertid

kunne tale om uafhængighed af k hændelser. Den generelle definition ser sådan ud:

Definition 1.5: Uafhængighed af hændelser
Hændelserne $A_1, A_2, \ldots, A_k$ siges at være uafhængige hvis der for enhver delmængde $\{A_{i_1}, A_{i_2}, \ldots, A_{i_m}\}$ af disse hændelser gælder at $\mathrm{P}\Big(\bigcap_{j=1}^{m} A_{i_j}\Big) = \prod_{j=1}^{m} \mathrm{P}(A_{i_j})$.

Uafhængighed
Termen *uafhængig* bruges i forskellige betydninger i forskellige delområder af matematikken, så undertiden kan det være nødvendigt med en præcisere sprogbrug. Den her præsenterede form for uafhængighed er *stokastisk* uafhængighed.

Bemærk: Når man skal undersøge uafhængighed af hændelser, er det ikke tilstrækkeligt at tjekke at de er parvis uafhængige, jf. eksempel 1.9. Det er heller ikke tilstrækkeligt at kontrollere at sandsynligheden for fællesmængden af *alle* hændelserne er lig produktet af sandsynlighederne for de enkelte hændelser, jf. eksempel 1.10.

Eksempel 1.9
Lad $\Omega = \{a, b, c, d\}$ og lad P være ligefordelingen på Ω. De tre hændelser $A = \{a, b\}$, $B = \{a, c\}$ og $C = \{a, d\}$ har hver især sandsynlighed $1/2$. Hændelserne A, B og C er parvis uafhængige, f.eks. er $\mathrm{P}(B \cap C) = \mathrm{P}(B)\mathrm{P}(C)$, idet $\mathrm{P}(B \cap C) = \mathrm{P}(\{a\}) = 1/4$ og $\mathrm{P}(B)\mathrm{P}(C) = 1/2 \cdot 1/2 = 1/4$, og tilsvarende er $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$ og $\mathrm{P}(A \cap C) = \mathrm{P}(A)\mathrm{P}(C)$.

Derimod er de tre hændelser *ikke* uafhængige, eksempelvis gælder at $\mathrm{P}(A \cap B \cap C) \neq \mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C)$, idet $\mathrm{P}(A \cap B \cap C) = \mathrm{P}(\{a\}) = 1/4$ og $\mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C) = 1/8$.

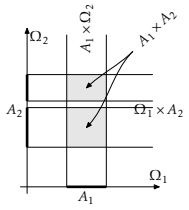
Eksempel 1.10
Lad $\Omega = \{a, b, c, d, e, f, g, h\}$ og lad P være ligefordelingen på Ω. De tre hændelser $A = \{a, b, c, d\}$, $B = \{a, e, f, g\}$ og $C = \{a, b, c, e\}$ har hver især sandsynlighed $1/2$. Da $A \cap B \cap C = \{a\}$, er $\mathrm{P}(A \cap B \cap C) = \mathrm{P}(\{a\}) = 1/8 = \mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C)$, men hændelserne A, B og C er *ikke* uafhængige; eksempelvis er $\mathrm{P}(A \cap B) \neq \mathrm{P}(A)\mathrm{P}(B)$ (fordi $\mathrm{P}(A \cap B) = \mathrm{P}(\{a\}) = 1/8$ og $\mathrm{P}(A)\mathrm{P}(B) = 1/2 \cdot 1/2 = 1/4$).

Uafhængige delforsøg; produktrum

De tilfældighedsfænomener der skal modelleres, består meget ofte af et antal separate delfænomener (f.eks. kan ét kast med fem terninger opfattes som sammensat af fem udgaver af »et kast med én terning«). Hvis delfænomenerne antages uafhængige af hverandre, kan man let sammensætte modeller for delfænomenerne til en stor model for det samlede fænomen.

I de indledende overvejelser vil vi for nemheds skyld antage at det sammensatte fænomen består af *to* delfænomener I og II. Lad os sige at de to delfænomener kan modelleres med sandsynlighedsrummene $(\Omega_1, \mathscr{F}_1, \mathrm{P}_1)$ hhv. $(\Omega_2, \mathscr{F}_2, \mathrm{P}_2)$.

Vi søger et sandsynlighedsrum $(\Omega, \mathscr{F}, \mathrm{P})$ der kan modellere det sammensatte fænomen bestående af I og II. Det er nærliggende at sige at udfaldene i det sammensatte forsøg skal skrives på formen (ω_1, ω_2) hvor $\omega_1 \in \Omega_1$ og $\omega_2 \in \Omega_2$, altså at Ω skal være produktmængden $\Omega_1 \times \Omega_2$, og $\mathscr{F}$ kan så være mængden af alle delmængder af Ω. Men hvad skal P være?



Tag en I-hændelse $A_1 \in \mathscr{F}_1$ og dan hændelsen $A_1 \times \Omega_2 \in \mathscr{F}$ svarende til at i det sammensatte fænomen giver I-delen et udfald i A_1 og II-delen hvad som helst, det vil sige at i det sammensatte fænomen interesserer man sig kun for hvad første delfænomen giver. De to hændelser $A_1 \times \Omega_2$ og A_1 svarer til det samme fænomen, blot i to forskellige sandsynlighedsrum, og derfor skulle det sandsynlighedsmål P som vi er på jagt efter, gerne være indrettet sådan at $\mathrm{P}(A_1 \times \Omega_2) = \mathrm{P}_1(A_1)$. På samme måde må man forlange at hvis $A_2 \in \mathscr{F}_2$, så er $\mathrm{P}(\Omega_1 \times A_2) = \mathrm{P}_2(A_2)$.

Hvis det sammensatte fænomen skal have den egenskab at delfænomenerne er uafhængige af hinanden, så må det betyde at de to hændelser $A_1 \times \Omega_2$ og $\Omega_1 \times A_2$ (der jo vedrører hver sit delfænomen) skal være uafhængige hændelser, og da deres fællesmængde er $A_1 \times A_2$, skal der derfor gælde at

$$\begin{aligned}\mathrm{P}(A_1 \times A_2) &= \mathrm{P}\big((A_1 \times \Omega_2) \cap (\Omega_1 \times A_2)\big) \\ &= \mathrm{P}(A_1 \times \Omega_2)\,\mathrm{P}(\Omega_1 \times A_2) = \mathrm{P}_1(A_1)\,\mathrm{P}_2(A_2),\end{aligned}$$

dvs. vi har et krav til P som vedrører alle produktmængder i $\mathscr{F}$. Da etpunktsmængder er produktmængder (fordi $\{(\omega_1,\omega_2)\} = \{\omega_1\} \times \{\omega_2\}$), har vi specielt et krav til punktsandsynlighederne for P; med nærliggende betegnelser er kravet at $p(\omega_1,\omega_2) = p_1(\omega_1)p_2(\omega_2)$ for alle $(\omega_1,\omega_2) \in \Omega$.

Inspireret af denne analyse af problemstillingen kan man nu gå frem som følger:

1. Lad p_1 og p_2 være punktsandsynlighederne for hhv. P_1 og P_2.
2. Definér så en funktion $p : \Omega \to [0\,;1]$ ved $p(\omega_1,\omega_2) = p_1(\omega_1)p_2(\omega_2)$ for $(\omega_1,\omega_2) \in \Omega$.
3. Der gælder at p summerer til 1:
$$\begin{aligned}\sum_{\omega\in\Omega} p(\omega) &= \sum_{(\omega_1,\omega_2)\in\Omega_1\times\Omega_2} p_1(\omega_1)p_2(\omega_2) \\ &= \sum_{\omega_1\in\Omega_1} p_1(\omega_1) \sum_{\omega_2\in\Omega_2} p_2(\omega_2) \\ &= 1 \cdot 1 = 1.\end{aligned}$$
4. Ifølge sætning 1.3 findes derfor et entydigt bestemt sandsynlighedsmål P på Ω der har p som sine punktsandsynligheder.
5. Dette sandsynlighedsmål opfylder det stillede krav om at $\mathrm{P}(A_1 \times A_2) = \mathrm{P}_1(A_1)\mathrm{P}_2(A_2)$ for alle $A_1 \in \mathscr{F}_1$ og $A_2 \in \mathscr{F}_2$, idet
$$\begin{aligned}\mathrm{P}(A_1 \times A_2) &= \sum_{(\omega_1,\omega_2)\in A_1\times A_2} p_1(\omega_1)p_2(\omega_2) \\ &= \sum_{\omega_1\in A_1} p_1(\omega_1) \sum_{\omega_2\in A_2} p_2(\omega_2) = \mathrm{P}_1(A_1)\mathrm{P}_2(A_2).\end{aligned}$$

Hermed har vi løst det stillede problem. – Det fundne sandsynlighedsmål P kaldes i øvrigt *produktet* af P_1 og P_2.

Man kan udvide ovenstående betragtninger til situationer med n delfænomener og derved nå frem til at hvis et tilfældighedsfænomen er sammensat af n uafhængige delfænomener med punktsandsynligheder $p_1, p_2, \ldots, p_n$, så er den samlede punktsandsynlighedsfunktion givet ved

$$p(\omega_1,\omega_2,\ldots,\omega_n) = p_1(\omega_1)p_2(\omega_2)\ldots p_n(\omega_n),$$

og om de tilsvarende sandsynlighedsmål gælder

$$\mathrm{P}(A_1 \times A_2 \times \ldots \times A_n) = \mathrm{P}_1(A_1)\,\mathrm{P}_2(A_2)\ldots\mathrm{P}_n(A_n).$$

Man kalder i slige forbindelser p og P for den *simultane* punktsandsynlighedsfunktion hhv. fordeling, og p_i-ene og P_i-erne for de *marginale* punktsandsynlighedsfunktioner hhv. fordelinger.

Eksempel 1.11
Hvis man kaster én gang med en mønt og én gang med en terning, så har udfaldet (Krone, 5 øjne) sandsynlighed $\frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$, og hændelsen »Krone og mindst fem øjne« sandsynlighed $\frac{1}{2} \cdot \frac{2}{6} = \frac{1}{6}$. – Hvis man kaster 100 gange med en mønt, så er sandsynligheden for at de 10 sidste kast alle giver Krone, lig $(\frac{1}{2})^{10} \approx 0.001$.

1.2 Stokastiske variable

En af grundene til matematikkens store succes er utvivlsomt at den i meget vid udstrækning betjener sig af symboler. Når man vil sætte talen om sandsynlighed og tilfældighed på matematiksprog, handler det blandt meget andet om at vælge en hensigtsmæssig notation. Det er uhyre praktisk at kunne operere med symboler der står for »det tilfældige numeriske udfald som tilfældighedseksperimentet nu leverer når vi udfører det«. Sådanne symboler kaldes *stokastiske variable*; stokastiske variable betegnes oftest med store bogstaver (især X, Y, Z). Vi vil benytte stokastiske variable som om de var almindelige reelle tal og altså lade dem indgå i udtryk som $X + Y = 5$ eller $Z \in B$.

Eksempel: I forbindelse med kast med to terninger kan man indføre stokastiske variable X_1 og X_2 som skal stå for antal øjne som terning nr. 1 hhv. 2 viser. At terningerne viser samme antal øjne, kan da kort skrives som $X_1 = X_2$, og at summen af øjnene er mindst 10, kan skrives som $X_1 + X_2 \geq 10$, osv.

Selv om ovenstående måske antyder hvad *meningen* med en stokastisk variabel skal være, så er det jo afgjort ikke nogen klar definition af hvad det er for et matematisk objekt. Omvendt fortæller nedenstående definition ikke meget om hvad meningen er:

Definition 1.6: Stokastisk variabel
Lad $(\Omega, \mathscr{F}, \mathrm{P})$ *være et sandsynlighedsrum over en endelig mængde. En stokastisk variabel på* $(\Omega, \mathscr{F}, \mathrm{P})$ *er en afbildning* X *af* Ω *ind i de reelle tal* $\mathbb{R}$.

Mere generelt er en n-dimensional stokastisk variabel på $(\Omega, \mathscr{F}, \mathrm{P})$ *en afbildning* $\boldsymbol{X}$ *af* Ω *ind i* $\mathbb{R}^n$.

Vi skal i det følgende studere det matematiske begreb en stokastisk variabel. Allerførst må vi præcisere hvordan det kan indgå i sproget: Lad X være en stokastisk variabel på det aktuelle udfaldsrum. Hvis u er et udsagn om reelle tal sådan at for hvert $x \in \mathbb{R}$ er $u(x)$ enten sandt eller falsk, så skal $u(X)$ forstås som et meningsfuldt udsagn om X, og dette er udsagn er sandt hvis og kun hvis hændelsen $\{\omega \in \Omega : u(X(\omega))\}$ indtræffer. Derved bliver vi i stand til at tale om sandsynligheden $\mathrm{P}(u(X))$ for at $u(X)$ er sandt; denne sandsynlighed er pr. definition $\mathrm{P}(u(X)) = \mathrm{P}(\{\omega \in \Omega : u(X(\omega))\})$.

Skrivemåden $\{\omega \in \Omega : u(X(\omega))\}$ er præcis, men temmelig omstændelig og ofte unødigt detaljeret. Derfor skriver man næsten altid den pågældende hændelse på den kortere form $\{u(X)\}$.

Eksempelvis svarer udsagnet $X \geq 3$ til hændelsen $\{X \geq 3\}$ der mere udførligt er lig med $\{\omega \in \Omega : X(\omega) \geq 3\}$, og man skriver $\mathrm{P}(X \geq 3)$ hvilket pr. definition er $\mathrm{P}(\{X \geq 3\})$ eller mere udførligt $\mathrm{P}(\{\omega \in \Omega : X(\omega) \geq 3\})$. – Dette udvides på oplagt måde til situationer med flere stokastiske variable.

Hvis B er en delmængde af $\mathbb{R}$, svarer udsagnet $X \in B$ til hændelsen $\{X \in B\} = \{\omega \in \Omega : X(\omega) \in B\}$. Sandsynligheden for dette udsagn (eller denne hændelse) er $\mathrm{P}(X \in B)$. Hvis vi ser på $\mathrm{P}(X \in B)$ som funktion af B, så opfylder den betingelserne for at være et sandsynlighedsmål på $\mathbb{R}$; dette sandsynlighedsmål vil matematikere kalde det *transformerede mål* og sandsynlighedsteoretikere og statistikere vil kalde det *fordelingen af* X. – Det foregående skal lige korrigeres en smule: Faktisk kan vi på dette sted ikke tale om et sandsynlighedsmål på $\mathbb{R}$, al den stund vi endnu kun er nået til sandsynligheder på endelige mængder. Derfor må vi »nøjes med« at opfatte sandsynlighedsmålet med de to navne som et sandsynlighedsmål på den endelige mængde $X(\Omega) \subset \mathbb{R}$.

Lad os for en kort bemærkning betegne det transformerede sandsynlighedsmål $X(\mathrm{P})$; så er $X(\mathrm{P})(B) = \mathrm{P}(X \in B)$ når B er en delmængde af $\mathbb{R}$. Denne tilsyneladende ganske uskyldige formel fortæller at alle sandsynlighedsudsagn vedrørende X kan omskrives til sandsynlighedsudsagn der alene involverer (delmængder af) $\mathbb{R}$ og sandsynlighedsmålet $X(\mathrm{P})$ på (den endelige delmængde $X(\Omega)$ af) $\mathbb{R}$. Vi kan altså helt se bort fra det oprindelige sandsynlighedsrum Ω.

Eksempel 1.12
Lad $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$ og lad P være det sandsynlighedsmål på Ω som har punktsandsynlighederne $p(\omega_1) = 0.3$, $p(\omega_2) = 0.1$, $p(\omega_3) = 0.4$, $p(\omega_4) = 0$, og $p(\omega_5) = 0.2$. Vi definerer en stokastisk variabel $X : \Omega \to \mathbb{R}$ ved $X(\omega_1) = 3$, $X(\omega_2) = -0.8$, $X(\omega_3) = 4.1$, $X(\omega_4) = -0.8$, og $X(\omega_5) = -0.8$.

Værdimængden for X er $X(\Omega) = \{-0.8, 3, 4.1\}$, og fordelingen af X er det sandsynlighedsmål på $X(\Omega)$ der har punktsandsynlighederne 0.3, 0.3 og 0.4, fordi

$$\begin{aligned}
\mathrm{P}(X = -0.8) &= \mathrm{P}(\{\omega_2, \omega_4, \omega_5\}) = p(\omega_2) + p(\omega_4) + p(\omega_5) = 0.3,\\
\mathrm{P}(X = 3) &= \mathrm{P}(\{\omega_1\}) = p(\omega_1) = 0.3,\\
\mathrm{P}(X = 4.1) &= \mathrm{P}(\{\omega_3\}) = p(\omega_3) = 0.4.
\end{aligned}$$

Eksempel 1.13: Fortsættelse af eksempel 1.12
Vi indfører nu to yderligere stokastiske variable Y og Z, sådan at situationen alt i alt er som følger

ω	$p(\omega)$	$X(\omega)$	$Y(\omega)$	$Z(\omega)$
ω_1	0.3	3	−0.8	3
ω_2	0.1	−0.8	3	−0.8
ω_3	0.4	4.1	4.1	4.1
ω_4	0	−0.8	4.1	4.1
ω_5	0.2	−0.8	3	−0.8

Det ses at X, Y og Z er forskellige, eksempelvis er $X(\omega_1) \neq Y(\omega_1)$, $X(\omega_4) \neq Z(\omega_4)$ og $Y(\omega_1) \neq Z(\omega_1)$, men de har samme den værdimængde, nemlig $\{-0.8, 3, 4.1\}$. Almindelig udregning viser at

$$\begin{aligned}
\mathrm{P}(X = -0.8) &= \mathrm{P}(Y = -0.8) = \mathrm{P}(Z = -0.8) = 0.3,\\
\mathrm{P}(X = 3) &= \mathrm{P}(Y = 3) = \mathrm{P}(Z = 3) = 0.3,\\
\mathrm{P}(X = 4.1) &= \mathrm{P}(Y = 4.1) = \mathrm{P}(Z = 4.1) = 0.4,
\end{aligned}$$

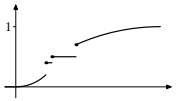
dvs. X, Y og Z har *samme fordeling*. Vi ser at $\mathrm{P}(X \neq Z) = \mathrm{P}(\{\omega_4\}) = 0$, dvs. $X = Z$ med sandsynlighed 1, og at $\mathrm{P}(X = Y) = \mathrm{P}(\{\omega_3\}) = 0.4$.

Fordelingen af en stokastisk variabel X kan – da der er tale om en fordeling på en endelig mængde – beskrives (og anskueliggøres) ved sine punktsandsynligheder. Da fordelingen »lever« på en delmængde af de reelle tal, kan vi imidlertid også beskrive og anskueliggøre den på en anden måde, nemlig ved hjælp af dens fordelingsfunktion.

Definition 1.7: Fordelingsfunktion
Fordelingsfunktionen for en stokastisk variabel X *er funktionen*

$$\begin{aligned}
F : \mathbb{R} &\longrightarrow [0\,;1]\\
x &\longmapsto \mathrm{P}(X \leq x).
\end{aligned}$$

Fordelingsfunktioner har altid bestemte egenskaber:



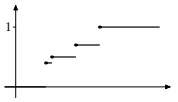
en »vilkårlig« fordelingsfunktion

Lemma 1.5
Hvis den stokastiske variabel X har fordelingsfunktion F, så er

$$\begin{aligned} \mathrm{P}(X \le x) &= F(x), \\ \mathrm{P}(X > x) &= 1 - F(x), \\ \mathrm{P}(a < X \le b) &= F(b) - F(a), \end{aligned}$$

for vilkårlige reelle tal x og $a < b$.

Bevis
Den første ligning er en gentagelse af definitionen af fordelingsfunktion. De to andre ligninger følger af punkterne 1 og 3 i lemma 1.1 (side 16). □



»vilkårlig« fordelingsfunktion, endeligt udfaldsrum

Sætning 1.6
Fordelingsfunktionen F for en stokastisk variabel X har følgende egenskaber:

1. *Den er ikke-aftagende, dvs. hvis $x \le y$, så er $F(x) \le F(y)$.*
2. $\lim_{x\to-\infty} F(x) = 0$ *og* $\lim_{x\to+\infty} F(x) = 1$.
3. *Den er højrekontinuert, dvs. $F(x+) = F(x)$ for alle x.*
4. *I ethvert punkt x gælder $\mathrm{P}(X = x) = F(x) - F(x-)$.*
5. *Et punkt x er et diskontinuitetspunkt for F hvis og kun hvis $\mathrm{P}(X = x) > 0$.*

Bevis
Ad 1: Hvis $x \le y$, så er $F(y) - F(x) = \mathrm{P}(x < X \le y)$ ifølge lemma 1.5, og da sandsynligheder er ikke-negative, er dermed $F(x) \le F(y)$.
Ad 2: Da X kun antager endeligt mange forskellige værdier, findes to tal $x_{\min}$ og $x_{\max}$ således at $x_{\min} < X(\omega) < x_{\max}$ for alle ω. Da er $F(x) = 0$ for alle $x < x_{\min}$, og $F(x) = 1$ for alle $x > x_{\max}$.
Ad 3: Da X kun antager endeligt mange forskellige værdier, gælder for et givet x at for alle tilstrækkeligt små tal $h > 0$ kan X ikke kan antage nogen værdi i intervallet $]x\,;x+h]$, dvs. hændelserne $\{X \le x\}$ og $\{X \le x+h\}$ er identiske, altså er $F(x) = F(x+h)$. Heraf følger det ønskede.
Ad 4: For et givet x ser vi på den del af X's værdimængde som ligger til venstre for x, altså mængden $X(\Omega) \cap]-\infty\,;x[$. Hvis denne mængde er tom, sætter vi $a = -\infty$, og ellers sætter vi a lig det største element i $X(\Omega) \cap]-\infty\,;x[$.

Da $X(\omega)$ ikke ligger i $]a\,;x[$ for noget ω, er det sådan at for ethvert tal $x^* \in]a\,;x[$ er hændelserne $\{X = x\}$ og $\{x^* < X \le x\}$ identiske, dvs. $\mathrm{P}(X = x) = \mathrm{P}(x^* < X \le x) = \mathrm{P}(X \le x) - \mathrm{P}(X \le x^*) = F(x) - F(x^*)$. For $x^* \nearrow x$ fås det ønskede.
Ad 5: Det følger af 4 og 3. □

Om voksende funktioner
Lad $F : \mathbb{R} \to \mathbb{R}$ være en funktion som er voksende, dvs. for alle x og y gælder at $x \le y$ medfører $F(x) \le F(y)$. Da gælder at F i ethvert punkt x har en *grænseværdi fra venstre*

$$F(x-) = \lim_{h\searrow 0} F(x-h)$$

og en *grænseværdi fra højre*

$$F(x+) = \lim_{h\searrow 0} F(x+h),$$

og der gælder at

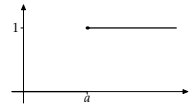
$$F(x-) \le F(x) \le F(x+).$$

Hvis $F(x-) \ne F(x+)$, er x et *springpunkt* for F (og ellers er x et kontinuitetspunkt for F).

Stokastiske variable vil komme til at optræde igen og igen, og læseren vil hurtigt nå til at jonglere aldeles ubesværet med selv avancerede eksemplarer af slagsen. Men på dette sted må vi hellere præsentere nogle simple (men ikke ligegyldige) eksempler på stokastiske variable.

Eksempel 1.14: Konstant stokastisk variabel
Den simpleste type stokastiske variable er dem som altid har den samme værdi, altså stokastiske variable af formen $X(\omega) = a$ for alle $\omega \in \Omega$; her er a et reelt tal. Dens fordelingsfunktion er

$$F(x) = \begin{cases} 1 & \text{når } a \le x \\ 0 & \text{når } x < a \end{cases}.$$



Faktisk kunne vi her erstatte betingelsen »$X(\omega) = a$ for alle $\omega \in \Omega$« med betingelsen »$X = a$ med sandsynlighed 1«, altså $\mathrm{P}(X = a) = 1$; fordelingsfunktionen ville være uændret.

Eksempel 1.15: 01-variabel
Den næstsimpleste type stokastiske variable må være dem der kun antager *to* forskellige værdier, som man ofte kalder for 0 og 1. Sådanne variable benyttes blandt andet i forbindelse med (modeller for) binære forsøg, dvs. forsøg med to mulige udfald (Plat/Krone, Succes/Fiasko, Gunstig/Ikke-gunstig, ...), og de kaldes 01-*variable* eller *Bernoulli-variable*.

Fordelingen af en 01-variabel kan specificeres ved hjælp af en parameter p der angiver sandsynligheden for værdien 1, dvs.

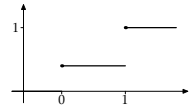
$$\mathrm{P}(X = x) = \begin{cases} p & \text{for } x = 1 \\ 1-p & \text{for } x = 0 \end{cases}$$

hvilket også kan skrives som

$$\mathrm{P}(X = x) = p^x(1-p)^{1-x}, \quad x = 0, 1. \tag{1.2}$$

Fordelingsfunktionen for en 01-variabel med parameter p er

$$F(x) = \begin{cases} 1 & \text{når } 1 \le x \\ 1-p & \text{når } 0 \le x < 1 \\ 0 & \text{når } x < 0. \end{cases}$$



Eksempel 1.16: Indikatorfunktion
Hvis A er en hændelse (dvs. en delmængde af Ω), så er dens *indikatorfunktion* funktionen

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{når } \omega \in A \\ 0 & \text{når } \omega \in A^{\mathrm{c}}. \end{cases}$$

En sådan indikatorfunktion er en 01-variabel med parameter $p = \mathrm{P}(A)$.

Hvis omvendt X er en 01-variabel, så er X indikatorfunktionen for hændelsen $A = X^{-1}(\{1\}) = \{\omega : X(\omega) = 1\}$.

Eksempel 1.17: Ligefordelt stokastisk variabel
Hvis $x_1, x_2, \ldots, x_n$ er n forskellige reelle tal og X en stokastisk variabel som antager enhver af disse værdier med samme sandsynlighed, dvs.

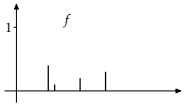
$$\mathrm{P}(X = x_i) = \frac{1}{n}, \quad i = 1, 2, \ldots, n,$$

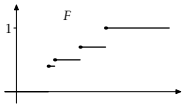
så siger man at X er ligefordelt på mængden $\{x_1, x_2, \ldots, x_n\}$.

Som måske allerede ovenstående eksempler antyder, er (grafen for) fordelingsfunktionen ikke overvældende velegnet til at give et informativt visuelt indtryk af fordelingen af en stokastisk variabel X. Det er langt bedre at beskæftige sig med sandsynlighedsfunktionen for X, dvs. funktionen $f : x \mapsto \mathrm{P}(X = x)$, betragtet som funktion defineret på X's værdimængde eller en ikke alt for voldsom udvidelse heraf. – I situationer der modelleres med endelige sandsynlighedsrum, er de interessante stokastiske variable meget ofte sådanne der tager værdier i de hele ikke-negative tal; i så fald kan man betragte sandsynlighedsfunktionen som defineret enten på X's faktiske værdimængde eller på mængden $\mathbb{N}_0$ af hele ikke-negative tal.

Definition 1.8: Sandsynlighedsfunktion
Sandsynlighedsfunktionen for en stokastisk variabel X er funktionen

$$f : x \longmapsto \mathrm{P}(X = x).$$





Sætning 1.7
Sammenhængen mellem fordelingsfunktion F og sandsynlighedsfunktion f er

$$f(x) = F(x) - F(x-),$$
$$F(x) = \sum_{z : z \le x} f(z).$$

Bevis
Udtrykket for f er en omformulering af Punkt 4 i sætning 1.6. Udtrykket for F følger af sætning 1.2 side 17. □

Uafhængige stokastiske variable

Man er ofte interesseret i at studere mere end én stokastisk variabel ad gangen.

Hvis $X_1, X_2, \ldots, X_n$ er stokastiske variable på det samme sandsynlighedsrum, så kalder man funktionen

$$f(x_1, x_2, \ldots, x_n) = \mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

for den *simultane* sandsynlighedsfunktion for X-erne, hvorimod eksempelvis funktionen

$$f_j(x_j) = \mathrm{P}(X_j = x_j)$$

kaldes den *marginale* sandsynlighedsfunktion for X_j. Hvis $\{i_1, i_2, \ldots, i_k\}$ er en ikke-triviel delmængde af $\{1, 2, \ldots, n\}$, så kaldes funktionen

$$f_{i_1, i_2, \ldots, i_k}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) = \mathrm{P}(X_{i_1} = x_{i_1}, X_{i_2} = x_{i_2}, \ldots, X_{i_k} = x_{i_k})$$

den *marginale* sandsynlighedsfunktion for $X_{i_1}, X_{i_2}, \ldots, X_{i_k}$.

Vi har tidligere (side 20f) defineret uafhængighed af hændelser. Det kan man bygge videre på i form af

Uafhængighed
Termen *uafhængig* bruges i forskellige betydninger i forskellige delområder af matematikken, så undertiden kan det være nødvendigt med en præcisere sprogbrug. Den her præsenterede form for uafhængighed er *stokastisk* uafhængighed.

Definition 1.9: Uafhængige stokastiske variable
Lad $(\Omega, \mathscr{F}, \mathrm{P})$ være et sandsynlighedsrum over en endelig mængde. De stokastiske variable $X_1, X_2, \ldots, X_n$ på $(\Omega, \mathscr{F}, \mathrm{P})$ siges at være uafhængige hvis det er sådan at hændelserne $\{X_1 \in B_1\}, \{X_2 \in B_2\}, \ldots, \{X_n \in B_n\}$ er uafhængige, ligegyldigt hvordan man vælger delmængderne $B_1, B_2, \ldots, B_n$ af $\mathbb{R}$.

Et nemt og mere overskueligt kriterium for uafhængighed er

Sætning 1.8
De stokastiske variable $X_1, X_2, \ldots, X_n$ er uafhængige hvis og kun hvis

$$\mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} \mathrm{P}(X_i = x_i) \tag{1.3}$$

for alle valg af tal $x_1, x_2, \ldots, x_n$ således at x_i tilhører X_i's værdimængde, $i = 1, 2, \ldots, n$.

Korollar 1.9
De stokastiske variable $X_1, X_2, \ldots, X_n$ er uafhængige hvis og kun hvis deres simultane sandsynlighedsfunktion er lig produktet af de marginale sandsynlighedsfunktioner:

$$f_{12\ldots n}(x_1, x_2, \ldots, x_n) = f_1(x_1) f_2(x_2) \ldots f_n(x_n).$$

Bevis for sætning 1.8
Lad os først vise »kun hvis«. Vi antager altså at $X_1, X_2, \ldots, X_n$ er uafhængige ifølge definition 1.9 og skal vise at (1.3) gælder. Sæt $B_i = \{x_i\}$; så er $\{X_i \in B_i\} = \{X_i = x_i\}$ $(i = 1, 2, \ldots, n)$, og vi har

$$\mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \mathrm{P}\Big(\bigcap_{i=1}^{n} \{X_i = x_i\}\Big) = \prod_{i=1}^{n} \mathrm{P}(X_i = x_i),$$

dvs. (1.3) gælder.

Derefter skal vi vise »hvis«, så vi antager at (1.3) gælder, og skal vise at for vilkårlige $B_1, B_2, \ldots, B_n$ er $\{X_1 \in B_1\}, \{X_2 \in B_2\}, \ldots, \{X_n \in B_n\}$ uafhængige. For en vilkårlig delmængde B af $\mathbb{R}^n$ er

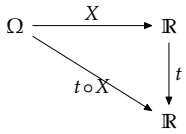
$$\mathrm{P}\Big((X_1, X_2, \ldots, X_n) \in B\Big) = \sum_{(x_1, x_2, \ldots, x_n) \in B} \mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

Anvendt på $B = B_1 \times B_2 \times \cdots \times B_n$ giver dette ved brug af (1.3) at

$$\begin{aligned}
\mathrm{P}\Big(\bigcap_{i=1}^{n}\{X_i \in B_i\}\Big) &= \mathrm{P}\big((X_1, X_2, \ldots, X_n) \in B\big) \\
&= \sum_{x_1 \in B_1} \sum_{x_2 \in B_2} \cdots \sum_{x_n \in B_n} \mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\
&= \sum_{x_1 \in B_1} \sum_{x_2 \in B_2} \cdots \sum_{x_n \in B_n} \prod_{i=1}^{n} \mathrm{P}(X_i = x_i) = \prod_{i=1}^{n} \sum_{x_i \in B_i} \mathrm{P}(X_i = x_i) \\
&= \prod_{i=1}^{n} \mathrm{P}(X_i \in B_i),
\end{aligned}$$

dvs. $\{X_1 \in B_1\}, \{X_2 \in B_2\}, \ldots, \{X_n \in B_n\}$ er uafhængige. □

Hvis de enkelte X-er har samme sandsynlighedsfunktion og dermed samme fordeling, taler man om at de er *identisk fordelte*, og i det følgende vil vi ofte møde vendingen *uafhængige identisk fordelte* stokastiske variable.



Funktioner af stokastiske variable

Hvis $(\Omega, \mathscr{F}, \mathrm{P})$ er et endeligt sandsynlighedsrum, X en stokastisk variabel på $(\Omega, \mathscr{F}, \mathrm{P})$ og t en funktion der afbilder X's værdimængde ind i de reelle tal, så er den sammensatte funktion $t \circ X$ igen en stokastisk variabel. Normalt skriver man ikke $t \circ X$, men $t(X)$. Fordelingen af $t(X)$ kan i princippet let findes, idet

$$\mathrm{P}(t(X) = y) = \mathrm{P}\big(X \in \{x : t(x) = y\}\big) = \sum_{x : t(x) = y} f(x) \tag{1.4}$$

hvor f er sandsynlighedsfunktionen hørende til X.

På tilsvarende måde taler man om $t(X_1, X_2, \ldots, X_n)$ hvor t en funktion af n variable og $X_1, X_2, \ldots, X_n$ er n stokastiske variable. Funktionen t behøver ikke være voldsomt avanceret; vi skal om lidt se på hvordan det ser ud når X-erne er uafhængige, og t er funktionen +, men først en ikke overraskende sætning.

Sætning 1.10
Lad $X_1, X_2, \ldots, X_m, X_{m+1}, X_{m+2}, \ldots, X_{m+n}$ være uafhængige stokastiske variable, og lad t_1 og t_2 være funktioner af henholdsvis m og n variable. Så er de stokastiske variable $Y_1 = t_1(X_1, X_2, \ldots, X_m)$ og $Y_2 = t_2(X_{m+1}, X_{m+2}, \ldots, X_{m+n})$ uafhængige.

Vedr. bevis for sætningen: opgave 1.22.

Fordelingen af en sum

Lad X_1 og X_2 være to stokastiske variable på $(\Omega, \mathscr{F}, \mathrm{P})$, og lad $f_{12}(x_1, x_2)$ være deres simultane sandsynlighedsfunktion. Da er $t(X_1, X_2) = X_1 + X_2$ ligeledes en stokastisk variabel på $(\Omega, \mathscr{F}, \mathrm{P})$, og den generelle formel (1.4) giver at

$$\begin{aligned}
\mathrm{P}(X_1 + X_2 = y) &= \sum_{x_2 \in X_2(\Omega)} \mathrm{P}(X_1 + X_2 = y \text{ og } X_2 = x_2) \\
&= \sum_{x_2 \in X_2(\Omega)} \mathrm{P}(X_1 = y - x_2 \text{ og } X_2 = x_2) \\
&= \sum_{x_2 \in X_2(\Omega)} f_{12}(y - x_2, x_2),
\end{aligned}$$

dvs. sandsynlighedsfunktionen for $X_1 + X_2$ fås ved at summere $f_{12}(x_1, x_2)$ over de talpar (x_1, x_2) for hvilke $x_1 + x_2 = y$. – Dette generaliseres uden videre til summer af mere end to stokastiske variable.

Hvis X_1 og X_2 er uafhængige, er deres simultane sandsynlighedsfunktion produktet af de marginale sandsynlighedsfunktioner, så vi får

Sætning 1.11
Hvis X_1 og X_2 er uafhængige stokastiske variable med sandsynlighedsfunktioner f_1 og f_2, så har $Y = X_1 + X_2$ sandsynlighedsfunktion

$$f(y) = \sum_{x_2 \in X_2(\Omega)} f_1(y - x_2) f_2(x_2).$$

Eksempel 1.18
Lad X_1 og X_2 være uafhængige identisk fordelte 01-variable med parameter p. Hvad er fordelingen af $X_1 + X_2$?

Den simultane sandsynlighedsfunktion f_{12} for (X_1, X_2) er (jf. bl.a. (1.2) side 27)

$$\begin{aligned}
f_{12}(x_1, x_2) &= f_1(x_1) f_2(x_2) \\
&= p^{x_1}(1-p)^{1-x_1} \cdot p^{x_2}(1-p)^{1-x_2} \\
&= p^{x_1+x_2}(1-p)^{2-(x_1+x_2)}
\end{aligned}$$

når $(x_1, x_2) \in \{0, 1\}^2$, og 0 ellers. Sandsynlighedsfunktionen for $X_1 + X_2$ er derfor

$$\begin{aligned}
f(0) &= f_{12}(0,0) &&= (1-p)^2, \\
f(1) &= f_{12}(1,0) + f_{12}(0,1) &&= 2p(1-p), \\
f(2) &= f_{12}(1,1) &&= p^2.
\end{aligned}$$

Som kontrol kan vi se efter om de fundne sandsynligheder summerer til 1:

$$f(0) + f(1) + f(2) = (1-p)^2 + 2p(1-p) + p^2 = \big((1-p) + p\big)^2 = 1.$$

Eksemplet rummer oplagte generalisationsmuligheder.

1.3 Eksempler

Sandsynlighedsfunktionen for en 01-variabel med parameter p er (fra formel (1.2) side 27)

$$f(x) = p^x(1-p)^{1-x}, \quad x = 0, 1.$$

Hvis $X_1, X_2, \ldots, X_n$ er uafhængige 01-variable med parameter p, er deres simultane sandsynlighedsfunktion derfor (når $(x_1, x_2, \ldots, x_n) \in \{0,1\}^n$)

$$f_{12\ldots n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^s(1-p)^{n-s} \tag{1.5}$$

hvor $s = x_1 + x_2 + \ldots + x_n$ (jf. korollar 1.9).

Binomialkoefficient
Antallet af forskellige k-delmængder, ɔ: delmængder med netop k elementer, som kan udtages fra en mængde G med n elementer, betegnes $\binom{n}{k}$. Denne størrelse kaldes en *binomialkoefficient*.

• Der gælder at $\binom{n}{0} = 1$ og at $\binom{n}{k} = \binom{n}{n-k}$ for $0 \le k \le n$.

Definition 1.10: Binomialfordeling
Fordelingen af summen af n uafhængige identisk fordelte 01*-variable med parameter p kaldes binomialfordelingen med antalsparameter* $n \in \mathbb{N}$ *og sandsynlighedsparameter* $p \in [0\,;1]$.

Vi vil finde sandsynlighedsfunktionen for binomialfordelingen med parametre n og p. Lad derfor $X_1, X_2, \ldots, X_n$ være uafhængige 01-variable med parameter p, dvs. deres simultane sandsynlighedsfunktion er givet ved (1.5). Pr. definition er $Y = X_1 + X_2 + \ldots + X_n$ binomialfordelt med parametre n og p. Hvis y er et heltal mellem 0 og n, er

$$\begin{aligned}
\mathrm{P}(Y = y) &= \sum_{x_1+x_2+\ldots+x_n=y} f_{12\ldots n}(x_1, x_2, \ldots, x_n) \\
&= \sum_{x_1+x_2+\ldots+x_n=y} p^y(1-p)^{n-y} \\
&= \left(\sum_{x_1+x_2+\ldots+x_n=y} 1 \right) p^y(1-p)^{n-y} \\
&= \binom{n}{y} p^y(1-p)^{n-y}
\end{aligned}$$

• Man kan udlede en rekursionsformel for binomialkoefficienterne: Lad g_0 være et element i G; der er nu to slags k-delmængder af G: 1) de k-delmængder der ikke indeholder g_0 og som derfor kan opfattes som k-delmængder af $G \setminus \{g_0\}$, og 2) de k-delmængder der indeholder g_0 og som derfor er af formen en $(k-1)$-delmængde af $G \setminus \{g_0\}$ forenet med $\{g_0\}$. Det samlede antal k-delmængder er lig summen af antallene af de to slags, altså:

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$$

(gælder når $0 < k < n$).

fordi der er $\binom{n}{y}$ forskellige talsæt $x_1, x_2, \ldots, x_n$ bestående af y 1-er og $n-y$ 0-er. Sandsynlighedsfunktionen for binomialfordelingen med parametre n og p er altså

$$f(y) = \binom{n}{y} p^y(1-p)^{n-y}, \quad y = 0, 1, 2, \ldots, n. \tag{1.6}$$

Sætning 1.12
Hvis Y_1 *og* Y_2 *er uafhængige binomialfordelte stokastiske variable med antalsparametre* n_1 *og* n_2 *og med samme sandsynlighedsparameter p, så er* $Y_1 + Y_2$ *binomialfordelt med parametre* $n_1 + n_2$ *og p.*

Bevis
Sætningen kan vises på to måder (mindst), en besværlig måde som går ud på at benytte sætning 1.11, og som vi ikke vil gennemregne her, og en smart måde:

Lad $X_1, X_2, \ldots, X_{n_1}, X_{n_1+1}, \ldots, X_{n_1+n_2}$ være $n_1 + n_2$ uafhængige identisk fordelte 01-variable. Pr. definition har $X_1 + X_2 + \ldots + X_{n_1}$ samme fordeling som Y_1, nemlig en binomialfordeling med parametre n_1 og p, og tilsvarende har $X_{n_1+1} + X_{n_1+2} + \ldots + X_{n_1+n_2}$ samme fordeling som Y_2. Ifølge sætning 1.10 er Y_1 og Y_2 uafhængige. Forudsætningen i sætning 1.12 er altså opfyldt for disse Y_1 og Y_2, og da $Y_1 + Y_2$ er en sum af $n_1 + n_2$ uafhængige identisk fordelte 01-variable med parameter p, er $Y_1 + Y_2$ binomialfordelt med parametre $n_1 + n_2$ og p. Hermed er sætningen vist!

Eller er den? Læseren vil måske synes at beviset er lidt underligt, for ikke at sige forkert. Er det der sker, ikke bare at der bliver tilvejebragt en situation med nogle helt specielle Y_i-er der er konstrueret sådan at konklusionen nærmest automatisk er sand? Skulle man ikke i stedet have begyndt med »vilkårlige« Y_1 og Y_2 med de nævnte fordelinger og så have ræsonneret ud fra dem?

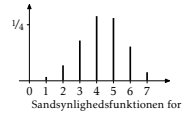
Nej, det er ikke nødvendigt. Pointen er at sætningen ikke handler om stokastiske variable *qua* reelle funktioner på et sandsynlighedsrum, men om hvad der sker når man transformerer bestemte sandsynlighedsfordelinger med afbildningen $(y_1, y_2) \mapsto y_1 + y_2$. Hvordan man tilvejebringer disse fordelinger, er i den forbindelse uden betydning, og de stokastiske variables rolle er alene at være symboler der gør at man kan skrive tingene op på en gennemskuelig måde. (Se evt. også side 24.) □

• Der gælder også at

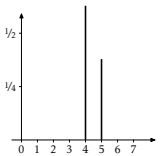
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

hvilket man kan indse på følgende måde: funktionen på højre side opfylder samme rekursionsformel som funktionen på venstre side (nemlig $f(n,k) = f(n-1,k) + f(n-1,k-1)$); desuden stemmer de overens når $k = n$ og $k = 0$; altså stemmer de overens for alle $k \le n$.

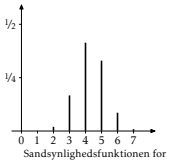
Binomialfordelingen kan benyttes til at modellere antal gunstige udfald i n uafhængige gentagelser af et forsøg der kan give enten gunstigt eller ikke-gunstigt udfald. Hvis gentagelserne strækker sig over to dage, f.eks. tirsdag og onsdag, kan man enten modellere totalantallet af gunstige udfald, eller man kan modellere antal gunstige om tirsdagen og antal gunstige om onsdagen og så lægge de to sammen; sætning 1.12 fortæller at det giver samme resultat. – Man kunne derefter overveje et problem af typen: Hvis man har foretaget n_1 gentagelser om tirsdagen og n_2 om onsdagen, og det samlede antal gunstige udfald er s, hvad kan man så sige om antal gunstige udfald om tirsdagen? Det handler den næste sætning om.



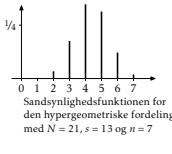
Sandsynlighedsfunktionen for binomialfordelingen med $n = 7$ og $p = 0.618$



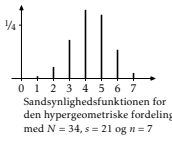
Sandsynlighedsfunktionen for den hypergeometriske fordeling med $N = 8$, $s = 5$ og $n = 7$



Sandsynlighedsfunktionen for den hypergeometriske fordeling med $N = 13$, $s = 8$ og $n = 7$



Sandsynlighedsfunktionen for den hypergeometriske fordeling med $N = 21$, $s = 13$ og $n = 7$



Sandsynlighedsfunktionen for den hypergeometriske fordeling med $N = 34$, $s = 21$ og $n = 7$

Sætning 1.13
Hvis Y_1 og Y_2 er uafhængige binomialfordelte stokastiske variable med antalsparametre n_1 og n_2 og samme sandsynlighedsparameter p, så har den betingede fordeling af Y_1 givet at $Y_1 + Y_2 = s$, sandsynlighedsfunktionen

$$\mathrm{P}\big(Y_1 = y \mid Y_1 + Y_2 = s\big) = \frac{\binom{n_1}{y}\binom{n_2}{s-y}}{\binom{n_1+n_2}{s}}, \tag{1.7}$$

som er forskellig fra 0 når $\max\{s-n_2, 0\} \le y \le \min\{n_1, s\}$.

Bemærk at den betingede fordeling ikke afhænger af p. – Vedr. bevis for sætningen: se opgave 1.17.

Sandsynlighedsfordelingen med sandsynlighedsfunktion (1.7) er en *hypergeometrisk fordeling*. Vi har mødt den allerede i eksempel 1.5 på side 17 i forbindelse med såkaldt stikprøveudtagning uden tilbagelægning. I forbindelse med stikprøveudtagning med tilbagelægning bliver der derimod tale om binomialfordelingen, således som det fremgår af indeværende afsnit.

Nu kan man jo sige, at hvis man tager en lille stikprøve fra en meget stor population, så må det være stort set lige meget om det sker med eller uden tilbagelægning. Dette udsagn præciseres i sætning 1.14; for at forstå begrundelsen for sætning 1.14 kan man tænke på følgende situation (jf. eksempel 1.5): Man har en kasse med N kugler, hvoraf s er sorte og resten hvide. Herfra udtages (uden tilbagelægning) en stikprøve på n; man ser på sandsynligheden for at denne stikprøve indeholder netop y sorte kugler, og interesserer sig for hvordan denne sandsynlighed opfører sig hvis N og s er meget store.

Sætning 1.14
Ved grænseovergangen hvor $N \to \infty$ og $s \to \infty$ på en sådan måde at $\frac{s}{N} \to p \in [0\,;1]$, vil

$$\frac{\binom{s}{y}\binom{N-s}{n-y}}{\binom{N}{n}} \longrightarrow \binom{n}{y} p^y (1-p)^{n-y}$$

for ethvert $y \in \{0, 1, 2, \ldots, n\}$.

Bevis
Ved almindelige omskrivninger fås

$$\begin{aligned}
\frac{\binom{s}{y}\binom{N-s}{n-y}}{\binom{N}{n}} &= \binom{n}{y} \frac{(N-n)!}{(s-y)!\,(N-n-(s-y))!} \frac{s!\,(N-s)!}{N!} \\
&= \binom{n}{y} \frac{\dfrac{s!}{(s-y)!} \cdot \dfrac{(N-s)!}{(N-s-(n-y))!}}{\dfrac{N!}{(N-n)!}} \\
&= \binom{n}{y} \frac{\overbrace{s(s-1)(s-2)\ldots(s-y+1)}^{y \text{ faktorer}} \cdot \overbrace{(N-s)(N-s-1)(N-s-2)\ldots(N-s-(n-y)+1)}^{n-y \text{ faktorer}}}{\underbrace{N(N-1)(N-2)\ldots(N-n+1)}_{n \text{ faktorer}}}.
\end{aligned}$$

Da der er n faktorer i både tæller og nævner, kan vi parre hver faktor i tælleren med en i nævneren og derved skrive den lange brøk som et produkt af n korte brøker. Under grænseovergangen vil hver af disse korte brøker have en grænseværdi: Hver af de y brøker af formen $\frac{s-\text{et eller andet}}{N-\text{et eller andet}}$ vil konvergere mod p, og hver af de $n-y$ brøker af formen $\frac{N-s-\text{et eller andet}}{N-\text{et eller andet}}$ vil konvergere mod $1-p$. Altså vil det samlede udtryk have den påståede grænseværdi. □

1.4 Middelværdi

Middelværdien af en reel funktion – for eksempel en stokastisk variabel – er det vægtede gennemsnit af de mulige funktionsværdier.

Definition 1.11: Middelværdi
Lad X være en reel stokastisk variabel på det endelige sandsynlighedsrum $(\Omega, \mathscr{F}, \mathrm{P})$. Middelværdien af X er tallet $\mathrm{E}(X) = \sum_{\omega\in\Omega} X(\omega)\,\mathrm{P}(\{\omega\})$.

Undertiden skriver man $\mathrm{E}\,X$ i stedet for $\mathrm{E}(X)$.

Hvis t er en funktion defineret på X's værdimængde, så er $t(X)$ igen en stokastisk variabel, og dens middelværdi er ifølge definition 1.11

$$\mathrm{E}\big(t(X)\big) = \sum_{\omega\in\Omega} t\big(X(\omega)\big)\,\mathrm{P}(\{\omega\}). \tag{1.8}$$

Sætning 1.15
Middelværdien af $t(X)$ kan også udregnes som

$$\mathrm{E}\big(t(X)\big) = \sum_x t(x)\,\mathrm{P}(X = x) = \sum_x t(x) f(x),$$

hvor der summeres over alle x i værdimængden $X(\Omega)$ for X, og hvor f er sandsynlighedsfunktionen for X.

Hvis specielt t er den identiske afbildning, får vi følgende alternative formel for middelværdien af X:

$$\mathrm{E}(X) = \sum_x x \,\mathrm{P}(X = x).$$

Bevis
Der gælder følgende omskrivninger der uddybes nedenfor:

$$\begin{aligned}
\sum_x t(x)\,\mathrm{P}(X=x) &\overset{1}{=} \sum_x t(x)\,\mathrm{P}\big(\{\omega : X(\omega) = x\}\big) \\
&\overset{2}{=} \sum_x t(x) \sum_{\omega\in\{X=x\}} \mathrm{P}(\{\omega\}) \\
&\overset{3}{=} \sum_x \sum_{\omega\in\{X=x\}} t(x)\,\mathrm{P}(\{\omega\}) \\
&\overset{4}{=} \sum_x \sum_{\omega\in\{X=x\}} t(X(\omega))\,\mathrm{P}(\{\omega\}) \\
&\overset{5}{=} \sum_{\omega\in\bigcup_x\{X=x\}} t(X(\omega))\,\mathrm{P}(\{\omega\}) \\
&\overset{6}{=} \sum_{\omega\in\Omega} t(X(\omega))\,\mathrm{P}(\{\omega\}) \\
&\overset{7}{=} \mathrm{E}(t(X)).
\end{aligned}$$

Uddybning:
Lighedstegn 1: en præcisering af hvad $\mathrm{P}(X = x)$ betyder.
Lighedstegn 2: følger af sætning 1.2 (side 17).
Lighedstegn 3: man ganger $t(x)$ ind i den inderste sum.
Lighedstegn 4: når $\omega \in \{X = x\}$, er $t(X(\omega)) = t(x)$.
Lighedstegn 5: mængderne $\{X = x\}, \omega \in \Omega$, er disjunkte.
Lighedstegn 6: $\bigcup_x \{X = x\}$ er lig Ω.
Lighedstegn 7: formel (1.8). □

Bemærkninger:

1. Sætning 1.15 er interessant og vigtig fordi den fortæller at middelværdien af $Y = t(X)$ kan udregnes på tre måder:

$$\mathrm{E}(t(X)) = \sum_{\omega\in\Omega} t(X(\omega))\,\mathrm{P}(\{\omega\}), \tag{1.9}$$

$$\mathrm{E}(t(X)) = \sum_x t(x)\,\mathrm{P}(X = x), \tag{1.10}$$

$$\mathrm{E}(t(X)) = \sum_y y\,\mathrm{P}(Y = y). \tag{1.11}$$

   Pointen er at udregningen efter behag kan foregå på Ω og med brug af P, formel (1.9), eller på (den udgave af de reelle tal som indeholder) værdimængden for X og med brug af X's fordeling, formel (1.10), eller på (den udgave af de reelle tal som indeholder) værdimængden for $Y = t(X)$ og med brug af Y's fordeling, formel (1.11).
2. I formuleringen af og beviset for sætningen var det underforstået at symbolet X stod for en almindelig endimensional stokastisk variabel, og at t var en funktion fra $\mathbb{R}$ til $\mathbb{R}$. Men der er intet som helst i vejen for at opfatte X som en n-dimensional stokastisk variabel, $X = (X_1, X_2, \ldots, X_n)$, og t som en afbildning fra $\mathbb{R}^n$ til $\mathbb{R}$. Sætningen forbliver rigtig, og beviset er uændret.
3. Matematikere vil kalde $\mathrm{E}(X)$ for *integralet* af funktionen X med hensyn til P og skrive $\mathrm{E}(X) = \int_\Omega X(\omega)\mathrm{P}(d\omega)$ eller kortere $\mathrm{E}(X) = \int_\Omega X d\mathrm{P}$, og de vil omtale sætning 1.15 som *integraltransformationssætningen*.

Man kan opfatte middelværdioperationen som en afbildning $X \mapsto \mathrm{E}(X)$ fra mængden af stokastiske variable (defineret på det givne endelige sandsynlighedsrum) ind i de reelle tal. Om denne afbildning gælder

Sætning 1.16
Afbildningen $X \mapsto \mathrm{E}(X)$ er en lineær afbildning fra vektorrummet af stokastiske variable på $(\Omega, \mathscr{F}, \mathrm{P})$ ind i de reelle tal.

Vedr. bevis: se opgave 1.18.

Der gælder således generelt at middelværdien af en sum er lig summen af middelværdierne. Derimod gælder kun i visse tilfælde en regel om middelværdien af et produkt.

Sætning 1.17
Hvis X og Y er uafhængige stokastiske variable på rummet $(\Omega, \mathscr{F}, \mathrm{P})$, så gælder at $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$.

Bevis
Uafhængigheden giver at $\mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x)\,\mathrm{P}(Y = y)$ for alle x og y, og vi kan så foretage disse omskrivninger:

$$\mathrm{E}(XY) = \sum_{x,y} xy\,\mathrm{P}(X = x, Y = y)$$

$$= \sum_{x,y} xy\,\mathrm{P}(X=x)\mathrm{P}(Y=y)$$
$$= \sum_{x}\sum_{y} xy\,\mathrm{P}(X=x)\mathrm{P}(Y=y)$$
$$= \sum_{x} x\,\mathrm{P}(X=x)\sum_{y} y\,\mathrm{P}(Y=y) = \mathrm{E}(X)\mathrm{E}(Y).$$

□

Sætning 1.18
Hvis $X(\omega) = a$ for alle ω, så er $\mathrm{E}(X) = a$, kort $\mathrm{E}(a) = a$. Specielt er $\mathrm{E}(1) = 1$.

Bevis
Indsæt i definitionen for middelværdi og regn ud. □

Sætning 1.19
Hvis A er en hændelse og $\mathbf{1}_A$ dens indikatorfunktion, så gælder at $\mathrm{E}(\mathbf{1}_A) = \mathrm{P}(A)$.

Bevis
Indsæt i definitionen for middelværdi og regn ud. (Indikatorfunktioner blev præsenteret i eksempel 1.16 på side 27.) □

Vi skal nu udlede forskellige egenskaber ved middelværdiafbildningen. Der er tale om egenskaber af formen: hvis der gælder ditten om X, så gælder der datten om $\mathrm{E}(X)$. Det der gælder om X, er typisk af formen »$u(X)$ med sandsynlighed 1«, dvs. det kræves ikke at $u(X(\omega))$ skal være opfyldt for *alle* ω, men kun at hændelsen $\{\omega : u(X(\omega))\}$ har sandsynlighed 1.

Lemma 1.20
Hvis A er en hændelse med $\mathrm{P}(A) = 0$, og X er en stokastisk variabel, så gælder at $\mathrm{E}(X\mathbf{1}_A) = 0$.

Bevis
Lad $\omega \in A$. Så er $\{\omega\} \subseteq A$ og dermed $\mathrm{P}(\{\omega\}) \le \mathrm{P}(A) = 0$ da P er voksende (lemma 1.1). Da også $\mathrm{P}(\{\omega\}) \ge 0$, kan vi slutte at $\mathrm{P}(\{\omega\}) = 0$. Derfor bliver den første af nedenstående to summer 0:

$$\mathrm{E}(X\mathbf{1}_A) = \sum_{\omega\in A} X(\omega)\,\mathbf{1}_A(\omega)\,\mathrm{P}(\{\omega\}) + \sum_{\omega\in\Omega\setminus A} X(\omega)\,\mathbf{1}_A(\omega)\,\mathrm{P}(\{\omega\}).$$

Den anden sum er 0 fordi når $\omega \in \Omega \setminus A$, er $\mathbf{1}_A(\omega) = 0$. □

Sætning 1.21
Middelværdiafbildningen er positiv, dvs. hvis $X \ge 0$ med sandsynlighed 1, så er $\mathrm{E}(X) \ge 0$.

Bevis
Sæt $B = \{\omega : X(\omega) \ge 0\}$ og $A = \{\omega : X(\omega) < 0\}$. Så er $1 = \mathbf{1}_B(\omega) + \mathbf{1}_A(\omega)$ for alle ω, og derfor er $X = X\mathbf{1}_B + X\mathbf{1}_A$ og dermed $\mathrm{E}(X) = \mathrm{E}(X\mathbf{1}_B) + \mathrm{E}(X\mathbf{1}_A)$. Middelværdien $\mathrm{E}(X\mathbf{1}_B)$ er ikke-negativ da den er en sum af ikke-negative led, og $\mathrm{E}(X\mathbf{1}_A)$ er 0 ifølge lemma 1.20. □

Korollar 1.22
Middelværdiafbildningen er voksende i den forstand at hvis $X \le Y$ med sandsynlighed 1, så er $\mathrm{E}(X) \le \mathrm{E}(Y)$. Specielt gælder at $|\mathrm{E}(X)| \le \mathrm{E}(|X|)$.

Bevis
Hvis $X \le Y$ med sandsynlighed 1, er $\mathrm{E}(Y) - \mathrm{E}(X) = \mathrm{E}(Y - X) \ge 0$ ifølge sætning 1.21.

Ved som Y at vælge $|X|$ får vi dernæst at $\mathrm{E}(X) \le \mathrm{E}(|X|)$ og $-\mathrm{E}(X) \le \mathrm{E}(|X|)$, dvs. $|\mathrm{E}(X)| \le \mathrm{E}(|X|)$. □

Korollar 1.23
Hvis $X = a$ med sandsynlighed 1, så er $\mathrm{E}(X) = a$.

Bevis
Med sandsynlighed 1 er $a \le X \le a$, så ifølge korollar 1.22 og sætning 1.18 er $a \le \mathrm{E}(X) \le a$, dvs. $\mathrm{E}(X) = a$. □

Lemma 1.24: Markovs ulighed
Hvis X er en ikke-negativ stokastisk variabel og c et positivt tal, så gælder at
$$\mathrm{P}(X \ge c) \le \frac{1}{c}\,\mathrm{E}(X).$$

Andrej Markov
russisk matematiker
(1856-1922).

Bevis
Da $c\,\mathbf{1}_{\{X\ge c\}}(\omega) \le X(\omega)$ for alle ω, er $c\,\mathrm{E}(\mathbf{1}_{\{X\ge c\}}) \le \mathrm{E}(X)$, dvs. $\mathrm{E}(\mathbf{1}_{\{X\ge c\}}) \le \frac{1}{c}\mathrm{E}(X)$. Ifølge sætning 1.19 er $\mathrm{E}(\mathbf{1}_{\{X\ge c\}}) = \mathrm{P}(X \ge c)$, så hermed er det ønskede vist. □

Korollar 1.25: Tjebysjovs ulighed
Hvis a er et positivt tal og X en stokastisk variabel, så gælder at
$$\mathrm{P}(|X| \ge a) \le \frac{1}{a^2}\,\mathrm{E}(X^2).$$

Panuftij Tjebysjov
russisk matematiker
(1821-94).

Bevis
$\mathrm{P}(|X| \ge a) = \mathrm{P}(X^2 \ge a^2) \le \frac{1}{a^2}\,\mathrm{E}(X^2)$ ifølge Markovs ulighed. □

Korollar 1.26
Hvis $\mathrm{E}|X| = 0$, så er $X = 0$ med sandsynlighed 1.

Bevis
Lad os vise at $\mathrm{P}(|X| > 0) = 0$. Ifølge Markovs ulighed gælder $\mathrm{P}(|X| \ge \varepsilon) \le$

$\frac{1}{\varepsilon}\,\mathrm{E}(|X|) = 0$ for ethvert $\varepsilon > 0$. Hvis vi vælger ε mindre end det mindste positive tal i værdimængden for $|X|$ (et sådant findes da mængden er endelig), er hændelsen $\{|X| \geq \varepsilon\}$ den samme som hændelsen $\{|X| > 0\}$. Altså er $\mathrm{P}(|X| > 0) = 0$. □

Sætning 1.27: Cauchy-Schwarz' ulighed
Hvis X og Y er stokastiske variable på et sandsynlighedsrum over en endelig mængde, så er

$$\Big(\mathrm{E}(XY)\Big)^2 \leq \mathrm{E}(X^2)\,\mathrm{E}(Y^2) \tag{1.12}$$

Lighedstegnet gælder hvis og kun hvis der findes et talpar $(a,b) \neq (0,0)$ således at $aX + bY = 0$ med sandsynlighed 1.

Augustin Louis Cauchy
fransk matematiker (1789-1857).

Hermann Amandus Schwarz
tysk matematiker (1843-1921).

Bevis
Hvis X og Y begge to er lig 0 med sandsynlighed 1, så er uligheden opfyldt (med lighed).

Antag at $Y \neq 0$ med positiv sandsynlighed. For alle $t \in \mathbb{R}$ er

$$0 \leq \mathrm{E}\Big((X + tY)^2\Big) = \mathrm{E}(X^2) + 2t\,\mathrm{E}(XY) + t^2\,\mathrm{E}(Y^2). \tag{1.13}$$

Højresiden er et andengradspolynomium i t (da Y ikke er konstant lig 0, er $\mathrm{E}(Y^2) > 0$). Da det altid er ikke-negativt, har det ikke to forskellige reelle rødder, og dets diskriminant er ikke-positiv, dvs. $(2\,\mathrm{E}(XY))^2 - 4\,\mathrm{E}(X^2)\,\mathrm{E}(Y^2) \leq 0$, hvilket er ensbetydende med (1.12). Endvidere gælder der lighedstegn hvis og kun hvis andengradspolynomiet har præcis én reel rod, dvs. hvis og kun hvis der findes et t så $\mathrm{E}(X + tY)^2 = 0$, og det er ifølge korollar 1.26 ensbetydende med at $X + tY$ er 0 med sandsynlighed 1.

Tilfældet $X \neq 0$ med positiv sandsynlighed behandles på samme måde. □

Varians og kovarians

Hvis man skal beskrive fordelingen af X med ét tal, kan man bruge $\mathrm{E}(X)$. Hvis man får lov til at bruge to tal, vil det være oplagt at lade det andet tal være et der fortæller noget om hvor store de tilfældige variationer omkring middelværdien er. Til det brug kan man bruge den såkaldte varians.

Definition 1.12: Varians og standardafvigelse
Variansen af den stokastiske variabel X er tallet

$$\mathrm{Var}(X) = \mathrm{E}\Big((X - \mathrm{E}\,X)^2\Big) = \mathrm{E}(X^2) - (\mathrm{E}\,X)^2.$$

Standardafvigelsen på X er tallet $\sqrt{\mathrm{Var}(X)}$.

Det følger umiddelbart af det første udtryk at $\mathrm{Var}(X)$ altid er et ikke-negativt tal. (Opgave 1.20 handler om at vise at de to udtryk for $\mathrm{Var}(X)$ er ens.)

Sætning 1.28
Der gælder at $\mathrm{Var}(X) = 0$ hvis og kun hvis X er konstant med sandsynlighed 1. I givet fald er konstanten lig $\mathrm{E}\,X$.

Bevis
Hvis $X = c$ med sandsynlighed 1, er $\mathrm{E}\,X = c$; så er $(X - \mathrm{E}\,X)^2 = 0$ med sandsynlighed 1, og dermed altså $\mathrm{Var}(X) = \mathrm{E}((X - \mathrm{E}\,X)^2) = 0$.

Hvis omvendt $\mathrm{Var}(X) = 0$, så fortæller korollar 1.26 at $(X - \mathrm{E}\,X)^2 = 0$ med sandsynlighed 1, dvs. X er med sandsynlighed 1 lig med $\mathrm{E}\,X$. □

Middelværdioperatoren er en lineær operator, dvs. der gælder altid $\mathrm{E}(aX) = a\,\mathrm{E}\,X$ og $\mathrm{E}(X + Y) = \mathrm{E}\,X + \mathrm{E}\,Y$. Noget tilsvarende er ikke tilfældet for variansoperatoren.

Sætning 1.29
Hvis X er en stokastisk variabel og a et reelt tal, så er $\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X)$.

Bevis
Indsæt i definitionen og benyt regnereglerne for middelværdi. □

Definition 1.13: Kovarians
Kovariansen mellem to stokastiske variable X og Y på samme sandsynlighedsrum er tallet $\mathrm{Cov}(X,Y) = \mathrm{E}\Big((X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y)\Big)$.

Man efterviser let følgende regneregler for kovarianser:

Sætning 1.30
For vilkårlige stokastiske variable X, Y, U, V og reelle konstanter a, b, c, d gælder

$$\begin{aligned}
\mathrm{Cov}(X,X) &= \mathrm{Var}(X),\\
\mathrm{Cov}(X,Y) &= \mathrm{Cov}(Y,X),\\
\mathrm{Cov}(X,a) &= 0,\\
\mathrm{Cov}(aX + bY, cU + dV) &= ac\,\mathrm{Cov}(X,U) + ad\,\mathrm{Cov}(X,V)\\
&\quad + bc\,\mathrm{Cov}(Y,U) + bd\,\mathrm{Cov}(Y,V).
\end{aligned}$$

Endvidere gælder

Sætning 1.31
Hvis to stokastiske variable er uafhængige, så er deres kovarians 0.

Bevis
Hvis X og Y er uafhængige, er også $X - \mathrm{E}\,X$ og $Y - \mathrm{E}\,Y$ uafhængige (sætning 1.10 side 30), og ifølge sætning 1.17 side 37 får vi dermed $\mathrm{E}\Big((X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y)\Big) = \mathrm{E}(X - \mathrm{E}\,X)\,\mathrm{E}(Y - \mathrm{E}\,Y) = 0 \cdot 0 = 0$. □

Sætning 1.32
For vilkårlige stokastiske variable X og Y er $|\mathrm{Cov}(X,Y)| \le \sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}$, og lighedstegnet gælder hvis og kun hvis der findes et talpar $(a,b) \neq (0,0)$ således at $aX+bY$ er konstant med sandsynlighed 1.

Bevis
Anvend sætning 1.27 på de to stokastiske variable $X - \mathrm{E}\,X$ og $Y - \mathrm{E}\,Y$. □

Definition 1.14: Korrelation
Korrelationen mellem to ikke-konstante stokastiske variable X og Y er tallet

$$\mathrm{corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}.$$

Korollar 1.33
Hvis X og Y er ikke-konstante stokastiske variable, så gælder

1. $-1 \le \mathrm{corr}(X,Y) \le +1$,
2. $\mathrm{corr}(X,Y) = +1$ *hvis og kun hvis der findes et talpar (a,b) med $ab<0$ således at $aX+bY$ er konstant med sandsynlighed 1,*
3. $\mathrm{corr}(X,Y) = -1$ *hvis og kun hvis der findes et talpar (a,b) med $ab>0$ således at $aX+bY$ er konstant med sandsynlighed 1.*

Bevis
Sætning 1.32 giver alle resultaterne på nær sammenhængen mellem korrelationens fortegn og ab's fortegn, men den følger af at $\mathrm{Var}(aX+bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X,Y)$ kun kan blive 0 hvis sidste led er negativt. □

Sætning 1.34
Hvis X og Y er stokastiske variable på samme sandsynlighedsrum, så er

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + 2\,\mathrm{Cov}(X,Y) + \mathrm{Var}(Y).$$

Hvis X og Y er uafhængige, så er $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Bevis
Den generelle formel vises ved almindelig udregning: Vi har

$$\big(X+Y-\mathrm{E}(X+Y)\big)^2 = (X-\mathrm{E}\,X)^2 + (Y-\mathrm{E}\,Y)^2 + 2(X-\mathrm{E}\,X)(Y-\mathrm{E}\,Y).$$

Vi tager middelværdi på begge sider og får

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{E}\big((X-\mathrm{E}\,X)(Y-\mathrm{E}\,Y)\big)\\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y).\end{aligned}$$

Ifølge sætning 1.31 er $\mathrm{Cov}(X,Y) = 0$ hvis X og Y er uafhængige, og vi har hermed vist det ønskede. □

Korollar 1.35
Lad $X_1, X_2, \ldots, X_n$ være indbyrdes uafhængige identisk fordelte stokastiske variable med middelværdi μ og varians σ^2, og lad $\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$ betegne deres gennemsnit. Der gælder at $\mathrm{E}\,\overline{X}_n = \mu$ og $\mathrm{Var}\,\overline{X}_n = \sigma^2/n$.

Korollaret fortæller hvordan den tilfældige variation af et gennemsnit af n observationer aftager med n – hvis man altså bruger variansen som et mål for tilfældig variation.

Eksempler

Eksempel 1.19: 01*-variable*
Lad X være en 01-variabel med $\mathrm{P}(X=1) = p$ og $\mathrm{P}(X=0) = 1-p$. Så har X middelværdi $\mathrm{E}\,X = 0\cdot\mathrm{P}(X=0) + 1\cdot\mathrm{P}(X=1) = p$. Variansen af X er i henhold til definition 1.12 $\mathrm{Var}(X) = \mathrm{E}(X^2) - p^2 = 0^2\cdot\mathrm{P}(X=0) + 1^2\cdot\mathrm{P}(X=1) - p^2 = p(1-p)$.

Eksempel 1.20: Binomialfordelingen
Binomialfordelingen med antalsparameter n og sandsynlighedsparameter p er fordelingen af en sum af n uafhængige 01-variable med parameter p (definition 1.10 side 32). Ifølge eksempel 1.19 og regnereglerne for middelværdi og varians er middelværdien i denne binomialfordeling derfor lig np og variansen er lig $np(1-p)$.

Man kan naturligvis også finde den søgte middelværdi som $\sum x f(x)$ hvor f er sandsynlighedsfunktionen for binomialfordelingen (formel (1.6) side 32), og tilsvarende kan man finde variansen som $\sum x^2 f(x) - \big(\sum x f(x)\big)^2$.

I eksempel 4.3 på side 80 bestemmes binomialfordelingens middelværdi og varians på en helt anden måde.

Store Tals Lov

Et af hovedområderne inden for sandsynlighedsregningen er at udlede resultater om den asymptotiske opførsel af følger af stokastiske variable. Her er et enkelt – det enkleste – eksempel.

Sætning 1.36: Store Tals Svage Lov
Lad $X_1, X_2, \ldots, X_n, \ldots$ være en følge af indbyrdes uafhængige identisk fordelte stokastiske variable med middelværdi μ og varians σ^2. Da gælder at gennemsnittet $\overline{X}_n = (X_1 + X_2 + \ldots + X_n)/n$ konvergerer mod μ i den forstand at

$$\lim_{n\to\infty} \mathrm{P}\big(\big|\overline{X}_n - \mu\big| < \varepsilon\big) = 1$$

for ethvert $\varepsilon > 0$.

Store Tals Stærke Lov
Der findes også en sætning der hedder *Store Tals Stærke Lov*. Den siger at under samme betingelser som i Store Tals Svage Lov gælder

$$\mathrm{P}\big(\lim_{n\to\infty} \overline{X}_n = \mu\big) = 1,$$

dvs. $\overline{X}_n$ konvergerer mod μ med sandsynlighed 1.

Bevis
Beviset er enkelt, takket være de mange forberedende øvelser: Ifølge korollar 1.35 er $\mathrm{Var}(\overline{X}_n - \mu) = \sigma^2/n$. Ifølge Tjebysjovs ulighed (korollar 1.25 side 39) er da

$$\mathrm{P}\left(\left|\overline{X}_n - \mu\right| < \varepsilon\right) = 1 - \mathrm{P}\left(\left|\overline{X}_n - \mu\right| \geq \varepsilon\right) \geq 1 - \frac{1}{\varepsilon^2}\frac{\sigma^2}{n},$$

hvilket går mod 1 for n gående mod ∞. □

Kommentar til beviset: Der er måske dem der synes at beviset (og sætningen) opererer med uendelig mange uafhængige identisk fordelte stokastiske variable, og kan man det (på dette sted i fremstillingen), og eksisterer der i det hele taget uendelige følger af stokastiske variable? Men der er ingen problemer. De eneste grænseovergange der sker, foregår i de reelle tal; vi opererer kun med n stokastiske variable ad gangen, og det er ganske uproblematisk – man kan f.eks. benytte et udfaldsrum som er et produktrum, jf. side 21f.

Store Tals Lov fortæller at gennemsnittet af et stort antal uafhængige identisk fordelte variable med stor sandsynlighed ligger tæt på middelværdien i fordelingen.

Lad os prøve at anvende Store Tals Lov på en følge af uafhængige identisk fordelte 01-variable $X_1, X_2, \ldots$ der er 1 med sandsynlighed p; middelværdien i deres fordeling er p (jf. eksempel 1.19). Gennemsnittet $\overline{X}_n$ bliver den relative hyppighed af 1-er blandt $X_1, X_2, \ldots, X_n$. Store Tals Lov fortæller da at den relative hyppighed af 1-er med stor sandsynlighed er tæt på p, altså at den relative hyppighed af 1-er er tæt på sandsynligheden for at få et 1. Det vil sige at vi nu inden for det matematiske univers har deduceret os frem til en overensstemmelse mellem sandsynlighed (i matematisk forstand) og relativ hyppighed. Dette må betragtes som et godt tegn på den matematiske sandsynlighedsregnings velegnethed.

Store Tals Lov åbner som vi har set, mulighed for at fortolke sandsynlighed som relativ hyppighed i det lange løb, og tilsvarende åbner den mulighed for at fortolke middelværdi som gennemsnit af et stort antal observationer.

1.5 Opgaver

Opgave 1.1
Angiv et sandsynlighedsrum der kan bruges som model for eksperimentet »man slår Plat eller Krone tre gange med en mønt«. Præcisér hændelserne »mindst en Krone«, »højst en Krone« og »ikke det samme resultat i to på hinanden følgende kast«. Hvor store er sandsynlighederne for disse hændelser?

Opgave 1.2
Giv et eksempel (et matematikeksempel, ikke (nødvendigvis) et »virkeligt« eksempel) på et sandsynlighedsrum hvor udfaldsrummet har tre elementer.

Opgave 1.3
I lemma 1.1 side 16 er en formel for sandsynligheden for at mindst en af to hændelser A og B indtræffer. Udled en formel for sandsynligheden for at netop en af de to hændelser indtræffer.

Opgave 1.4
Lad os sige at sandsynlighed kan fortolkes som relativ hyppighed i det lange løb. Argumentér for at definition 1.3 side 19 er en fornuftig måde at definere betinget sandsynlighed på.

Opgave 1.5
Vis at funktionen $\mathrm{P}(\ \ |B)$ i definition 1.4 side 19 opfylder betingelserne for at være et sandsynlighedsmål i henhold til definition 1.1 side 15. Hvordan finder man punktsandsynlighederne for $\mathrm{P}(\ \ |B)$ når man kender dem for P?

Opgave 1.6
Man kaster to almindelige terninger og ser hvor mange øjne de viser. Hvad er sandsynligheden for at antal øjne som terning nr. 1 viser, er et primtal, givet at summen af øjnene som de to terninger viser, er 8?

Opgave 1.7
En kasse indeholder fem røde og tre blå kugler. Hen under aften hvor det er halvmørkt og man ikke kan se forskel på rød og blå, tager Pedro fire tilfældige kugler op af kassen. Derefter tager Antonia én tilfældigt valgt kugle op.

1. Hvad er sandsynligheden for at Antonia tager en blå kugle?
2. Givet at Antonia tager en blå kugle, hvad er så sandsynligheden for at Pedro har taget fire røde?

Opgave 1.8
Lad os sige at i en bestemt familie er der sandsynligheden 1/4 for at børnene får blå øjne, og at de forskellige børn får blå øjne eller ej uafhængigt af hinanden. Familien får fem børn.

1. Hvis det vides at det yngste barn har blå øjne, hvad er da sandsynligheden for at mindst tre af børnene har blå øjne?
2. Hvis det vides at mindst et af de fem børn har blå øjne, hvad er da sandsynligheden for at mindst tre af børnene har blå øjne?

Opgave 1.9
Man bliver ofte præsenteret for spørgsmål af formen: »I en klinisk undersøgelse fandt man at ud af en stor gruppe lungekræftpatienter var 80% rygere, og ud af en tilsvarende kontrolgruppe var 40% rygere. Hvad kan man på den baggrund sige om rygnings betydning for lungekræft?«
[En kontrolgruppe er (i dette tilfælde) en gruppe personer der er søgt udvalgt sådan at den ligner patientgruppen på »alle« punkter, bortset fra at kontrolgruppens personer

ikke har fået diagnosticeret lungekræft. Man vil formodentlig tilstræbe at patientgruppe og kontrolgruppe blandt andet har nogenlunde samme aldersfordeling, samme kønsfordeling, samme socialgruppefordeling, osv.]

Idet vi helt ser bort fra de spørgsmål der kan stilles til hvad det vil sige at kontrolgruppen er »tilsvarende«, og fra de problemer der hænger sammen med den statistiske usikkerhed og biologiske variation, hvad kan man så fremsætte af interessante kvantitative udsagn der belyser rygnings betydning for lungekræft?

Vink: Ved *odds* for hændelsen A forstås tallet $\mathrm{P}(A)/\mathrm{P}(A^{\mathrm{c}})$. Udregn/opstil et udtryk for odds for lungekræft givet man er ryger, og det samme givet man er ikke-ryger.

Opgave 1.10
Undertiden taler man (sundhedsmyndigheder og politikere) om at der skal foretages *screeninger* af befolkningen for at opdage bestemte sygdomme meget tidligt i forløbet.

Antag at man vil teste hver enkelt person i en given befolkningsgruppe for en bestemt, temmelig sjælden sygdom. Testmetoden er ikke 100% pålidelig (det er testmetoder sjældent), så der er en vis lille sandsynlighed for en »falsk positiv«, ɔ: for at testen fejlagtigt siger at personen har sygdommen, og der er ligeledes en vis lille sandsynlighed for en »falsk negativ«, ɔ: for at testen fejlagtigt siger at personen ikke har sygdommen.

Opstil en egentlig matematisk model for det ovenfor skitserede. Hvilke parametre er det fornuftigt at lade indgå i modellen?

For den enkelte person er der to spørgsmål af umådelig interesse, nemlig: hvis testen er positiv, hvad er så sandsynligheden for at have sygdommen, og: hvis testen er negativ, hvad er så sandsynligheden for at have sygdommen?

Benyt den opstillede model til at besvare disse spørgsmål. Prøv også at indsætte talværdier.

Opgave 1.11
Vis at hvis hændelserne A og B er uafhængige, så er A og B^{c} også uafhængige.

Opgave 1.12
Giv et sæt nødvendige og tilstrækkelige betingelser for at en funktion f er en sandsynlighedsfunktion.

Opgave 1.13
Skitsér binomialfordelingens sandsynlighedsfunktion for forskellige værdier af n og p. Hvad sker der når $p \to 0$ eller $p \to 1$? Hvad sker der når man erstatter p med $1-p$?

Opgave 1.14
Antag at X er ligefordelt på $\{1, 2, 3, \ldots, 10\}$ (jf. definition 1.17 side 27). Opskriv og skitsér sandsynlighedsfunktionen og fordelingsfunktionen for X. Find $\mathrm{E}\,X$ og $\mathrm{Var}\,X$.

Opgave 1.15
Lad X og Y være stokastiske variable på samme sandsynlighedsrum $(\Omega, \mathscr{F}, \mathrm{P})$. Hvilken sammenhæng er der mellem udsagnene »$X = Y$ med sandsynlighed 1« og »X og Y har samme fordeling«? – Vink: benyt evt. eksempel 1.13 til inspiration.

Opgave 1.16
Lad X og Y være stokastiske variable på samme sandsynlighedsrum $(\Omega, \mathscr{F}, \mathrm{P})$. Vis at hvis X er konstant (ɔ: der findes et tal c således at $X(\omega) = c$ for alle $\omega \in \Omega$), så er X og Y uafhængige.

Vis også at hvis X er konstant med sandsynlighed 1 (ɔ: der findes et tal c således at $\mathrm{P}(X = c) = 1$), så er X og Y uafhængige.

Opgave 1.17
Bevis sætning 1.13 side 34.

Benyt først definitionen på betinget sandsynlighed og udnyt at $Y_1 = y$ og $Y_1 + Y_2 = s$ er ensbetydende med $Y_1 = y$ og $Y_2 = s - y$; benyt derefter uafhængigheden af Y_1 og Y_2, og benyt endelig at vi kender sandsynlighedsfunktionen for hver af de variable Y_1, Y_2 og $Y_1 + Y_2$.

Opgave 1.18
Vis at mængden af stokastiske variable defineret på et givet endeligt sandsynlighedsrum er et vektorrum over de reelle tal.

Vis at afbildningen $X \mapsto \mathrm{E}(X)$ er en lineær afbildning fra dette vektorrum ind i vektorrummet $\mathbb{R}$.

Opgave 1.19
Lad X være en stokastisk variabel hvis fordeling er givet ved $\mathrm{P}(X = 1) = \mathrm{P}(X = -1) = \frac{1}{2}$. Find $\mathrm{E}\,X$ og $\mathrm{Var}\,X$.

Lad desuden Y være en stokastisk variabel hvis fordeling er givet ved $\mathrm{P}(Y = 100) = \mathrm{P}(Y = -100) = \frac{1}{2}$, og find $\mathrm{E}\,Y$ og $\mathrm{Var}\,Y$.

Opgave 1.20
I definitionen af varians (definition 1.12 side 40) optræder helt ukommenteret to forskellige udtryk for variansen på en stokastisk variabel. – Vis at det faktisk er rigtigt at $\mathrm{E}\big((X - \mathrm{E}\,X)^2\big) = \mathrm{E}(X^2) - (\mathrm{E}\,X)^2$.

Opgave 1.21
Lad X være en stokastisk variabel således at

$$\mathrm{P}(X = 2) = 0.1 \qquad \mathrm{P}(X = -2) = 0.1$$
$$\mathrm{P}(X = 1) = 0.4 \qquad \mathrm{P}(X = -1) = 0.4,$$

og sæt $Y = t(X)$ hvor funktionen t (fra $\mathbb{R}$ til $\mathbb{R}$) er givet ved

$$t(x) = \begin{cases} -x & \text{hvis } |x| \leq 1 \\ x & \text{ellers.} \end{cases}$$

Vis at X og Y har samme fordeling. Find middelværdien af X og middelværdien af Y. Find variansen af X og variansen af Y. Find kovariansen mellem X og Y. Er X og Y uafhængige?

Opgave 1.22
Lav et bevis for sætning 1.10 (side 30): Med notationen fra sætningen skal det vises at

$$\mathrm{P}(Y_1 = y_1 \text{ og } Y_2 = y_2) = \mathrm{P}(Y_1 = y_1)\,\mathrm{P}(Y_2 = y_2)$$

for alle y_1 og y_2. Skriv den første sandsynlighed op som en sum hvor der summeres over $(x_1, x_2, \ldots, x_{m+n})$ tilhørende en vis mængde; denne mængde er en produktmængde af en mængde af $(x_1, x_2, \ldots, x_m)$-er og en mængde af $(x_{m+1}, x_{m+2}, \ldots, x_{m+n})$-er.

Opgave 1.23: Jensens ulighed
Lad X være en stokastisk variabel og ψ en konveks funktion.Vis Jensens ulighed

$$\mathrm{E}\,\psi(X) \geq \psi(\mathrm{E}\,X).$$

Opgave 1.24
Lad $x_1, x_2, \ldots, x_n$ være n positive tal. Tallet $G = (x_1\, x_2 \ldots x_n)^{1/n}$ kaldes det geometriske gennemsnit af x-erne, og $A = (x_1 + x_2 + \cdots + x_n)/n$ er det aritmetiske gennemsnit af x-erne. Vis at $G \leq A$.

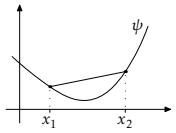
Vink: Lad X være en stokastisk variabel med $\mathrm{P}(X = x_i) = \frac{1}{n}$, $i = 1, 2, \ldots, n$, og anvend Jensens ulighed på X og den konvekse funktion $\psi = -\ln$.

Johan Ludwig William Valdemar Jensen (1859-1925), dansk matematiker og ingeniør ved Kjøbenhavns Telefon Aktieselskab. Blandt andet kendt for Jensens ulighed (se opgave 1.23).

Konvekse funktioner
En reel funktion ψ på et interval $I \subseteq \mathbb{R}$ er *konveks* hvis der for vilkårlige to punkter $(x_1, \psi(x_1))$ og $(x_2, \psi(x_2))$ på grafen for ψ gælder at hele det linjestykke der forbinder punkterne, ligger over eller på grafen, altså hvis der for vilkårlige $x_1, x_2 \in I$ gælder at

$$\lambda\psi(x_1) + (1-\lambda)\psi(x_2) \geq \psi\big(\lambda x_1 + (1-\lambda)x_2\big)$$

for alle $\lambda \in [0\,;1]$.



Hvis ψ er to gange kontinuert differentiabel, så gælder at ψ er konveks hvis og kun hvis $\psi'' \geq 0$.

2 Tællelige udfaldsrum

Dette kapitel præsenterer dele af teorien for sandsynligheder på tællelige udfaldsrum; på en række punkter er der tale om en uproblematisk generalisering af teorien for det endelige tilfælde, men der kommer også enkelte større modifikationer.

Som man vil erindre, opererer vi i det endelige tilfælde med mængden $\mathscr{F}$ af alle delmængder af det endelige udfaldsrum Ω, og ethvert element i $\mathscr{F}$, altså enhver delmængde af Ω (enhver hændelse), tillægges en sandsynlighed. På tilsvarende måde vil vi i det tællelige tilfælde tillægge enhver delmængde af udfaldsrummet en sandsynlighed. Dette er i fuld overensstemmelse med hvad man har brug for i konkrete modelbygningssituationer, men set fra det strengt formelle synspunkt er det lidt en forsimpling. – Den interesserede læser kan konsultere definition 5.2 side 91 for at se den generelle definition af sandsynlighedsrum.

2.1 Grundlæggende definitioner

Definition 2.1: Sandsynlighedsrum over tællelig mængde
Et sandsynlighedsrum over en tællelig mængde er et tripel $(\Omega, \mathscr{F}, \mathrm{P})$ *bestående af*

1. *et udfaldsrum* Ω, *som er en ikke-tom, tællelig mængde,*
2. *mængden* $\mathscr{F}$ *af alle delmængder af* Ω,
3. *et sandsynlighedsmål på* $(\Omega, \mathscr{F})$, *dvs. en afbildning* $\mathrm{P} : \mathscr{F} \to \mathbb{R}$ *som er*
   - *positiv:* $\mathrm{P}(A) \geq 0$ *for alle* $A \in \mathscr{F}$,
   - *normeret:* $\mathrm{P}(\Omega) = 1$, *og*
   - *σ-additiv: hvis* $A_1, A_2, \ldots$ *er en følge i* $\mathscr{F}$ *af parvis disjunkte hændelser, så er*

$$\mathrm{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathrm{P}(A_i). \tag{2.1}$$

En tælleligt uendelig mængde har overtælleligt mange delmængder, men som man ser, involverer betingelsen for at være et sandsynlighedsmål kun tælleligt mange hændelser ad gangen.

Lemma 2.1
Lad $(\Omega, \mathscr{F}, \mathrm{P})$ *være et sandsynlighedsrum over det tællelige udfaldsrum* Ω. *Der*

gælder

i. $\mathrm{P}(\varnothing) = 0$.

ii. *Hvis* $A_1, A_2, \ldots, A_n$ *er parvis disjunkte hændelser fra* $\mathscr{F}$*, så er*

$$\mathrm{P}\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} \mathrm{P}(A_i),$$

dvs. P *er også endeligt additiv.*

iii. $\mathrm{P}(A) + \mathrm{P}(A^{\mathrm{c}}) = 1$ *for enhver hændelse* A.

iv. *Hvis* $A \subseteq B$*, så er* $\mathrm{P}(B \setminus A) = \mathrm{P}(B) - \mathrm{P}(A)$*, og* $\mathrm{P}(A) \leq \mathrm{P}(B)$*, dvs.* P *er voksende.*

v. $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$.

Bevis

Hvis vi i (2.1) sætter $A_1 = \Omega$ og alle øvrige A-er til $\varnothing$, får man at $\mathrm{P}(\varnothing)$ må være lig 0. Derefter er det klart at σ-additivitet medfører endelig additivitet.

Den øvrige del af lemmaet vises på samme måde som i det endelige tilfælde (jf. beviset for lemma 1.1 side 16). □

Det næste lemma fortæller, at selv om vi har at gøre med et tælleligt uendeligt udfaldsrum, så kan stort set al sandsynlighedsmassen lokaliseres til en endelig mængde.

Lemma 2.2

Lad $(\Omega, \mathscr{F}, \mathrm{P})$ *være et sandsynlighedsrum over det tællelige udfaldsrum* Ω*. Da gælder at til ethvert* $\varepsilon > 0$ *findes en endelig delmængde* Ω_0 *af* Ω *således at* $\mathrm{P}(\Omega_0) \geq 1 - \varepsilon$.

Bevis

Skriv elementerne i Ω op i rækkefølge: $\omega_1, \omega_2, \omega_3, \ldots$ Så er $\bigcup_{i=1}^{\infty} \{\omega_i\} = \Omega$, og da P er σ-additiv, er $\sum_{i=1}^{\infty} \mathrm{P}(\{\omega_i\}) = \mathrm{P}(\Omega) = 1$. Da rækkens sum er 1, bliver afsnitssummen før eller senere større end $1 - \varepsilon$, dvs. der findes et n således at $\sum_{i=1}^{n} \mathrm{P}(\{\omega_i\}) \geq 1 - \varepsilon$. Mængden $\Omega_0 = \{\omega_1, \omega_2, \ldots, \omega_n\}$ har da den ønskede egenskab. □

Punktsandsynligheder

Punktsandsynligheder defineres på ganske tilsvarende måde som i det endelige tilfælde, jf. definition 1.2 side 17:

Definition 2.2: Punktsandsynligheder

Lad $(\Omega, \mathscr{F}, \mathrm{P})$ *være et sandsynlighedsrum over det tællelige udfaldsrum* Ω*. Funktionen*

$$\begin{aligned} p : \Omega &\longrightarrow [0\,;1] \\ \omega &\longmapsto \mathrm{P}(\{\omega\}) \end{aligned}$$

kaldes punktsandsynlighederne for P.

Nedenstående sætning kan vises på ganske samme måde som i det endelig tilfælde (side 17), dog kan summen nu have uendeligt mange led.

Sætning 2.3

Hvis p *er punktsandsynlighederne for sandsynlighedsmålet* P*, så gælder for en vilkårlig hændelse* $A \in \mathscr{F}$ *at* $\mathrm{P}(A) = \sum_{\omega \in A} p(\omega)$.

Bemærkninger: En konsekvens af sætningen er at to forskellige sandsynlighedsmål ikke kan have samme punktsandsynlighedsfunktion. En anden konsekvens er at p summerer til 1, dvs. $\sum_{\omega \in \Omega} p(\omega) = 1$; det ser man ved at sætte $A = \Omega$.

Den næste sætning kan vises efter samme opskrift som i det endelige tilfælde (sætning 1.3 side 18), idet man udnytter at når man summerer uendelige rækker med ikke-negative led, er det tilladt at bytte vilkårligt meget om på summationsrækkefølgen.

Sætning 2.4

Hvis $p : \Omega \to [0\,;1]$ *summerer til* 1*, dvs.* $\sum_{\omega \in \Omega} p(\omega) = 1$*, så findes præcis et sandsynlighedsmål på* Ω *der har* p *som sine punktsandsynligheder.*

Betingning, uafhængighed

Begreber som betinget sandsynlighed, betinget fordeling og uafhængighed defineres fuldstændig som i det endelige tilfælde, se side 18 og 20.

Det skal måske specielt bemærkes at Bayes' formel (jf. side 19) kan generaliseres på følgende måde: Hvis $B_1, B_2, \ldots$ er en klassedeling af Ω, dvs. B-erne er parvis disjunkte med $\bigcup_{i=1}^{\infty} B_i = \Omega$, og hvis A er en eller anden hændelse, så er

$$\mathrm{P}(B_j \mid A) = \frac{\mathrm{P}(A \mid B_j)\,\mathrm{P}(B_j)}{\sum_{i=1}^{\infty} \mathrm{P}(A \mid B_i)\,\mathrm{P}(B_i)}\,.$$

Stokastiske variable

Stokastiske variable defineres på samme måde som i det endelige tilfælde. Der er dog den forskel at de nu kan antage uendeligt mange værdier.

Her er nogle eksempler på fordelinger på en tællelig delmængde af de reelle tal. Først nogle fordelinger som anvendes i konkrete modelleringssituationer hvor man modellerer antalsobservationer, og som vi vil se mere til i afsnit 2.3:

- *Poissonfordelingen* med parameter $\mu > 0$ har sandsynlighedsfunktion

$$f(x) = \frac{\mu^x}{x!}\exp(-\mu), \quad x = 0, 1, 2, 3, \ldots$$

- *Den geometriske fordeling* med sandsynlighedsparameter $p \in]0\,;1[$ har sandsynlighedsfunktion

$$f(x) = p\,(1-p)^x, \quad x = 0, 1, 2, 3, \ldots$$

- *Den negative binomialfordeling* med sandsynlighedsparameter $p \in]0\,;1[$ og formparameter $k > 0$ har sandsynlighedsfunktion

$$f(x) = \binom{x+k-1}{x} p^k (1-p)^x, \quad x = 0, 1, 2, 3, \ldots$$

- *Den logaritmiske fordeling* med parameter $p \in]0\,;1[$ har sandsynlighedsfunktion

$$f(x) = \frac{1}{-\ln p}\,\frac{(1-p)^x}{x}, \quad x = 1, 2, 3, \ldots$$

Her er en anden type eksempel, der skal vise lidt om hvad den matematiske formalisme også handler om.

- En stokastisk variabels værdier kan godt være placeret lidt besynderligt på talaksen, man kan for eksempel definere X til at antage værdierne $\pm 1, \pm\frac{1}{2}, \pm\frac{1}{3}, \pm\frac{1}{4}, \ldots, 0$ med sandsynlighederne

$$\begin{aligned} \mathrm{P}\Big(X = \tfrac{1}{n}\Big) &= \tfrac{1}{3\cdot 2^n}, \qquad n = 1, 2, 3, \ldots \\ \mathrm{P}\Big(X = 0\Big) &= \tfrac{1}{3}, \\ \mathrm{P}\Big(X = -\tfrac{1}{n}\Big) &= \tfrac{1}{3\cdot 2^n}, \qquad n = 1, 2, 3, \ldots \end{aligned}$$

  Eksemplet viser blandt andet at selv om X antager uendeligt mange værdier, behøver disse ikke ligge uendeligt langt væk.

Begreberne fordelingsfunktion og sandsynlighedsfunktion defineres på samme måde som i det endelige tilfælde (definition 1.7 side 25 og definition 1.8 side 28), og lemma 1.5 og sætning 1.6 side 26 gælder uændret; beviset for sætning 1.6 skal ændres: enten kan man udnytte at en tællelig mængde »stort set« er endelig (lemma 2.2), eller også skal man benytte beviset for det generelle tilfælde (sætning 5.3).

Kriteriet for uafhængighed af stokastiske variable (sætning 1.8/korollar 1.9 side 29) gælder uforandret. Formlen for sandsynlighedsfunktionen for en sum af stokastiske variable (sætning 1.11 side 31) gælder ligeledes uændret.

2.2 Middelværdi

Lad $(\Omega, \mathscr{F}, \mathrm{P})$ være et sandsynlighedsrum over en tællelig mængde, og lad X være en reel stokastisk variabel på dette rum. Man kunne overveje at definere middelværdien af X på samme måde som i det endelige tilfælde (side 35), dvs. som tallet $\mathrm{E}\,X = \sum_{\omega\in\Omega} X(\omega)\,\mathrm{P}(\{\omega\})$ eller $\mathrm{E}\,X = \sum_x x\,f(x)$, hvor f er X's sandsynlighedsfunktion. Dette er for så vidt også en udmærket idé, bortset fra at de indgående summer nu er uendelige summer med hvad dertil kan høre af konvergensproblemer.

Eksempel 2.1

Hvis $\alpha > 1$, er $c(\alpha) = \sum_{x=1}^{\infty} x^{-\alpha}$ som bekendt et endeligt tal; derfor kan man have en stokastisk variabel X med sandsynlighedsfunktion $f_X(x) = \frac{1}{c(\alpha)}\,x^{-\alpha}$, $x = 1, 2, 3, \ldots$ Imidlertid er summen $\mathrm{E}\,X = \sum_{x=1}^{\infty} x f_X(x) = \frac{1}{c(\alpha)} \sum_{x=1}^{\infty} x^{-\alpha+1}$ kun et veldefineret endeligt tal når $\alpha > 2$. Når $\alpha \in]1\,;2]$, går sandsynlighedsfunktionen for langsomt mod 0 til at X kan have en middelværdi. Det er altså ikke alle stokastiske variable der kan tillægges en middelværdi.

Nogen ville måske foreslå at vi indførte $+\infty$ og $-\infty$ som tilladte værdier for $\mathrm{E}\,X$, men det løser ikke alle problemer. Se for eksempel på en stokastisk variabel Y med værdimængde $\mathbb{N} \cup -\mathbb{N}$ og sandsynlighedsfunktion $f_Y(y) = \frac{1}{2c(\alpha)}\,|y|^{-\alpha}$, $y = \pm 1, \pm 2, \pm 3, \ldots$; her er det ikke nemt at give et fornuftigt bud på hvad $\mathrm{E}\,Y$ skal betyde.

Definition 2.3: Middelværdi
Lad X være en stokastisk variabel på $(\Omega, \mathscr{F}, \mathrm{P})$.
Hvis summen $\sum_{\omega\in\Omega} |X(\omega)|\,\mathrm{P}(\{\omega\})$ er endelig, så siges X at have middelværdi, og middelværdien af X er i så fald tallet $\mathrm{E}\,X = \sum_{\omega\in\Omega} X(\omega)\,\mathrm{P}(\{\omega\})$.

Sætning 2.5
Lad X være en reel stokastisk variabel på $(\Omega, \mathscr{F}, \mathrm{P})$, og lad f være sandsynlighedsfunktionen for X. Da gælder at X har middelværdi hvis og kun hvis summen $\sum_x |x| f(x)$

er endelig, og i givet fald er $\mathrm{E}\,X = \sum_x x f(x)$.

Vedr. bevis: Sætningen bevises efter samme principper som sætning 1.15 (side 35), dog skal man nu jonglere med uendelige summer.

Eksempel 2.1 viser at det ikke er alle stokastiske variable der har middelværdi. For at have styr på tingene indfører vi en særlig betingelse for mængden af stokastiske variable som har middelværdi.

Definition 2.4
Mængden af stokastiske variable på $(\Omega, \mathscr{F}, \mathrm{P})$ *som har middelværdi, betegnes ofte* $\mathscr{L}^1(\Omega, \mathscr{F}, \mathrm{P})$ *eller* $\mathscr{L}^1(\mathrm{P})$ *eller* $\mathscr{L}^1$.

Der gælder (jf. sætning 1.16 side 37)

Sætning 2.6
Mængden $\mathscr{L}^1(\Omega, \mathscr{F}, \mathrm{P})$ *er et vektorrum (over* $\mathbb{R}$*), og afbildningen* $X \mapsto \mathrm{E}\,X$ *er en lineær afbildning af dette vektorrum ind i* $\mathbb{R}$.

Vedr. bevis: benyt regneregler for uendelige summer.

Resultaterne om middelværdier fra det endelige tilfælde generaliseres umiddelbart, dog skal det måske nævnes at sætning 1.17 side 37 kommer til at se sådan ud:

Sætning 2.7
Hvis X *og* Y *er uafhængige stokastiske variable som begge har middelværdi, så gælder at* XY *også har middelværdi, og* $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$.

Lemma 1.20 side 38 kommer til at se sådan ud:

Lemma 2.8
Hvis A *er en hændelse med* $\mathrm{P}(A) = 0$*, og* X *er en stokastisk variabel, så har den stokastiske variabel* $X\mathbf{1}_A$ *middelværdi, og* $\mathrm{E}(X\mathbf{1}_A) = 0$.

Den næste sætning fortæller at man kan ændre en stokastisk variabel på en mængde med sandsynlighed 0 uden at det spiller nogen rolle for dens middelværdi.

Sætning 2.9
Lad X *og* Y *være stokastiske variable. Hvis* $X = Y$ *med sandsynlighed* 1*, og hvis* $X \in \mathscr{L}^1$*, så er* $Y \in \mathscr{L}^1$ *og* $\mathrm{E}\,Y = \mathrm{E}\,X$.

Bevis
Da $Y = X + (Y - X)$, er det (i henhold til sætning 2.6) nok at vise at $Y - X \in \mathscr{L}^1$ og at $\mathrm{E}(Y - X) = 0$. Sæt $A = \{\omega : X(\omega) \neq Y(\omega)\}$. Eftersom der for alle ω gælder at $Y(\omega) - X(\omega) = (Y(\omega) - X(\omega))\mathbf{1}_A(\omega)$, følger det ønskede af lemma 2.8. □

Lemma 2.10 er en direkte konsekvens af majorantkriteriet for uendelige rækker:

Lemma 2.10
Lad Y *være en vilkårlig stokastisk variabel. Hvis der findes en ikke-negativ stokastisk variabel* $X \in \mathscr{L}^1$ *således at* $|Y| \leq X$*, så er* $Y \in \mathscr{L}^1$*, og* $\mathrm{E}\,Y \leq \mathrm{E}\,X$.

Varians og kovarians

Kort fortalt er variansen af en stokastisk variabel X defineret som $\mathrm{Var}(X) = \mathrm{E}(X - \mathrm{E}\,X)^2$, forudsat at dette er et endeligt tal.

Definition 2.5
Mængden af stokastiske variable X *på* $(\Omega, \mathscr{F}, \mathrm{P})$ *for hvilke* $\mathrm{E}(X^2) < +\infty$*, betegnes* $\mathscr{L}^2(\Omega, \mathscr{F}, \mathrm{P})$ *eller* $\mathscr{L}^2(\mathrm{P})$ *eller* $\mathscr{L}^2$.

Bemærk at $X \in \mathscr{L}^2(\mathrm{P})$ hvis og kun hvis $X^2 \in \mathscr{L}^1(\mathrm{P})$.

Lemma 2.11
Hvis $X \in \mathscr{L}^2$ *og* $Y \in \mathscr{L}^2$*, så er* $XY \in \mathscr{L}^1$.

Bevis
For vilkårlige reelle tal x og y er $|xy| \leq x^2 + y^2$. Dette benyttes sammen med lemma 2.10. □

Sætning 2.12
Mængden $\mathscr{L}^2(\Omega, \mathscr{F}, \mathrm{P})$ *er et underrum af* $\mathscr{L}^1(\Omega, \mathscr{F}, \mathrm{P})$*, og dette underrum indeholder alle konstante stokastiske variable.*

Bevis
Hvis vi i lemma 2.11 sætter $Y = X$, får vi at $\mathscr{L}^2 \subseteq \mathscr{L}^1$.

Det er klart at hvis $X \in \mathscr{L}^2$ og a er en konstant, så er $aX \in \mathscr{L}^2$, dvs. $\mathscr{L}^2$ er afsluttet over for multiplikation med skalarer.

Hvis $X, Y \in \mathscr{L}^2$, så ligger X^2, Y^2 og XY alle i $\mathscr{L}^1$ ifølge lemma 2.11. Dermed er $(X + Y)^2 = X^2 + 2XY + Y^2 \in \mathscr{L}^1$, dvs. $X + Y \in \mathscr{L}^2$. Da $\mathscr{L}^2$ således også er afsluttet over for addition, er det et underrum. □

Definition 2.6: Varians
Hvis $\mathrm{E}(X^2) < +\infty$*, så siges* X *at have en varians, og variansen af* X *er tallet*

$$\mathrm{Var}(X) = \mathrm{E}\big((X - \mathrm{E}\,X)^2\big) = \mathrm{E}(X^2) - (\mathrm{E}\,X)^2.$$

Definition 2.7: Kovarians
Hvis X *og* Y *har varians, så er deres kovarians tallet*

$$\mathrm{Cov}(X, Y) = \mathrm{E}\big((X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y)\big).$$

Der gælder samme regneregler for varianser og kovarianser som i det endelige tilfælde (side 40ff).

Sætning 2.13: Cauchy-Schwarz' ulighed
Hvis X og Y tilhører $\mathscr{L}^2(\mathrm{P})$, så er

$$\big(\mathrm{E}|XY|\big)^2 \le \mathrm{E}(X^2)\,\mathrm{E}(Y^2). \tag{2.2}$$

Bevis
Fra lemma 2.11 ved vi at $XY \in \mathscr{L}^1(\mathrm{P})$. Herefter kan man gå frem på samme måde som i det endelige tilfælde (sætning 1.27 side 40). □

Hvis man i (2.2) erstatter X med $X - \mathrm{E}\,X$ og Y med $Y - \mathrm{E}\,Y$, får man

$$\mathrm{Cov}(X,Y)^2 \le \mathrm{Var}(X)\,\mathrm{Var}(Y).$$

2.3 Eksempler

Den geometriske fordeling

Man har et tilfældighedseksperiment med to mulige udfald 0 og 1 (eller Ugunstig og Gunstig); lad p betegne sandsynligheden for 1. Man gentager eksperimentet indtil udfaldet 1 indtræffer for første gang. Hvad kan man sige om antallet af gentagelser der skal til, før der første gang optræder et 1? Det er klart at der principielt ikke er nogen øvre grænse for antallet, så dette eksempel kan næppe modelleres med et endeligt sandsynlighedrum. I øvrigt kan man heller ikke udelukke den mulighed at udfaldet 1 aldrig indtræffer!

Vi vil tænke på (og tale om) tingene på den måde at grundeksperimentet udføres til tidspunkterne $t = 1, 2, 3, \ldots$, og det der efterspørges, er ventetiden til der første gang kommer et 1. Vi sætter

$$T_1 = \text{ første tidspunkt hvor der kommer et 1,}$$
$$V_1 = T_1 - 1 = \text{ antallet af 0-er inden det første 1.}$$

Kvotientrækker
Hvis $|t| < 1$, er

$$\frac{1}{1-t} = \sum_{n=0}^{\infty} t^n.$$

Strengt taget kan vi ikke vide om der overhovedet kommer et 1; hvis der aldrig kommer et 1, sætter vi $T_1 = \infty$ og $V_1 = \infty$. De mulige værdier for T_1 er således $1, 2, 3, \ldots, \infty$, og de mulige værdier for V_1 er $0, 1, 2, 3, \ldots, \infty$.

Lad $t \in \mathbb{N}_0$. Hændelsen $\{V_1 = t\}$ (eller $\{T_1 = t+1\}$) indtræffer netop når de t første gange giver 0 og nr. $t+1$ giver 1, så $\mathrm{P}(V_1 = t) = p\,(1-p)^t$, $t = 0, 1, 2, \ldots$ Nu er (jf. formlen for summen af en kvotientrække)

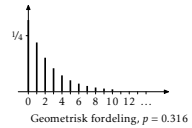
$$\mathrm{P}(V_1 \in \mathbb{N}_0) = \sum_{t=0}^{\infty} \mathrm{P}(V_1 = t) = \sum_{t=0}^{\infty} p\,(1-p)^t = 1,$$

dvs. V_1 (og T_1) er endelig med sandsynlighed 1, eller sagt på en anden måde: med sandsynlighed 1 vil der før eller senere komme et 1.

Fordelingen af V_1 er en geometrisk fordeling:

Definition 2.8: Geometrisk fordeling
Den geometriske fordeling med parameter p er den fordeling på $\mathbb{N}_0$ som har sandsynlighedsfunktion

$$f(t) = p\,(1-p)^t, \quad t \in \mathbb{N}_0.$$



Geometrisk fordeling, $p = 0.316$

Vi kan udregne middelværdien i den geometriske fordeling:

$$\begin{aligned}\mathrm{E}\,V_1 &= \sum_{t=0}^{\infty} tp(1-p)^t = p(1-p)\sum_{t=1}^{\infty} t\,(1-p)^{t-1} \\ &= p(1-p)\sum_{t=0}^{\infty}(t+1)(1-p)^t = \frac{1-p}{p}\end{aligned}$$

(jf. formlen for summen af en binomialrække (side 58). I middel vil der således komme $(1-p)/p$ gange 0 før det første 1, og middelventetiden til første 1 er $\mathrm{E}\,T_1 = \mathrm{E}\,V_1 + 1 = 1/p$. – Variansen i den geometriske fordeling udregnes på lignende måde til $(1-p)/p^2$.

Sandsynligheden for at vi skal vente længere end t på at det første 1 indtræffer, er

$$\mathrm{P}(T_1 > t) = \sum_{k=t+1}^{\infty} \mathrm{P}(T_1 = k) = (1-p)^t.$$

Sandsynligheden for at vi skal vente længere end $s+t$, givet at vi allerede har ventet s, er

$$\mathrm{P}(T_1 > s+t \mid T_1 > s) = \frac{\mathrm{P}(T_1 > s+t)}{\mathrm{P}(T_1 > s)} = (1-p)^t = \mathrm{P}(T_1 > t),$$

dvs. fordelingen af den resterende ventetid afhænger ikke af hvor længe vi allerede har ventet – man taler om at den proces der genererer 0-erne og 1-erne, er *uden hukommelse*. (Det kan i øvrigt nævnes at eksponentialfordelingen er den fordeling på den kontinuerte tidsskala som har den tilsvarende egenskab, se side 71.)

Den »forskudte« geometriske fordeling er den eneste fordeling på $\mathbb{N}$ med denne egenskab: Hvis T er en stokastisk variabel med værdier i $\mathbb{N}$ og sådan at $\mathrm{P}(T > s+t) = \mathrm{P}(T > s)\,\mathrm{P}(T > t)$ for alle $s, t \in \mathbb{N}$, så er

$$\mathrm{P}(T > t) = \mathrm{P}(T > 1 + 1 + \ldots + 1) = \big(\mathrm{P}(T > 1)\big)^t = (1-p)^t$$

hvor $p = 1 - \mathrm{P}(T > 1) = \mathrm{P}(T = 1)$.

Den negative binomialfordeling

Vi betragter samme situation som i afsnittet om den geometriske fordeling, blot ser vi nu på de stokastiske variable

$$T_k = \text{ tidspunktet for det } k\text{-te 1}$$
$$V_k = T_k - k = \text{ antallet af 0-er inden det } k\text{-te 1.}$$

Hændelsen $\{V_k = t\}$ (eller $\{T_k = t+k\}$) svarer til at der til tid $t+k$ kommer et 1, og at der inden da er kommet netop $k-1$ gange et 1 og t gange et 0. Sandsynligheden for at der til tid $t+k$ kommer et 1, er p; sandsynligheden for at der de $t+k-1$ første gange kommer netop $k-1$ 1-er og t 0-er er binomialsandsynligheden $\binom{t+k-1}{t}p^{k-1}(1-p)^t$. Alt i alt er derfor

$$\mathrm{P}(V_k = t) = \binom{t+k-1}{t}p^k(1-p)^t, \quad t \in \mathbb{N}_0.$$

Fordelingen af V_k er en negativ binomialfordeling.

Definition 2.9: Negativ binomialfordeling
Den negative binomialfordeling med sandsynlighedsparameter $p \in]0;1[$ og formparameter $k > 0$ er den fordeling på $\mathbb{N}_0$ som har sandsynlighedsfunktion

$$f(t) = \binom{t+k-1}{t}p^k(1-p)^t, \quad t \in \mathbb{N}_0. \tag{2.3}$$

Bemærkninger: 1) For $k = 1$ fås den geometriske fordeling. 2) I udledningerne ovenfor er k af gode grunde et heltal, men faktisk er udtrykket (2.3) veldefineret og en sandsynlighedsfunktion for vilkårlige positive reelle k.

Sætning 2.14
Hvis X og Y er uafhængige negativt binomialfordelte stokastiske variable med samme sandsynlighedsparameter p og med formparametre j og k, så er $X+Y$ negativt binomialfordelt med sandsynlighedsparameter p og formparameter $j+k$.

Bevis
Vi har

$$\begin{aligned}\mathrm{P}(X+Y=s) &= \sum_{x=0}^{s}\mathrm{P}(X=x)\,\mathrm{P}(Y=s-x)\\ &= \sum_{x=0}^{s}\binom{x+j-1}{x}p^j(1-p)^x\cdot\binom{s-x+k-1}{s-x}p^k(1-p)^{s-x}\\ &= p^{j+k}A(s)(1-p)^s,\end{aligned}$$

Binomialrækker
1. Der gælder

$$(1+t)^n = \sum_{k=0}^{n}\binom{n}{k}t^k$$

for ethvert $n \in \mathbb{N}$ og $t \in \mathbb{R}$.

2. Der gælder

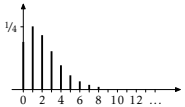
$$(1+t)^\alpha = \sum_{k=0}^{\infty}\binom{\alpha}{k}t^k$$

for ethvert $\alpha \in \mathbb{R}$ og $|t| < 1$.

3. Der gælder

$$\begin{aligned}&(1-t)^{-\alpha}\\ &= \sum_{k=0}^{\infty}\binom{-\alpha}{k}(-t)^k\\ &= \sum_{k=0}^{\infty}\binom{\alpha+k-1}{k}t^k\end{aligned}$$

for ethvert $\alpha \in \mathbb{R}$ og $|t| < 1$.



Negativ binomialfordeling med $p = 0.618$ og $k = 3.5$

hvor $A(s) = \sum_{x=0}^{s}\binom{x+j-1}{x}\binom{s-x+k-1}{s-x}$. Man kan nu give sig til at omskrive udtrykket for $A(s)$ i det håb at man kan nå frem til det rigtige, men man kan heldigvis også slippe nemmere om ved det: Summen af sandsynlighederne er jo 1, dvs. $\sum_{s=0}^{\infty}p^{j+k}A(s)(1-p)^s = 1$, hvoraf $p^{-(j+k)} = \sum_{s=0}^{\infty}A(s)(1-p)^s$. Men vi ved også at $p^{-(j+k)} = \sum_{s=0}^{\infty}\binom{s+j+k-1}{s}(1-p)^s$, enten fra formlen for summen af en binomialrække eller fra det forhold at sandsynlighedsfunktionen for den negative binomialfordeling med parametre p og $j+k$ summerer til 1. Da en funktions potensrækkeudvikling er entydig, kan vi heraf slutte at $A(s) = \binom{s+j+k-1}{s}$, og at $X+Y$ derfor har den påståede fordeling. □

Generaliserede binomialkoefficienter
For vilkårligt $r \in \mathbb{R}$ og $k \in \mathbb{N}$ defineres tallet $\binom{r}{k}$ som

$$\frac{r(r-1)\ldots(r-k+1)}{1\cdot 2\cdot 3\cdot\ldots\cdot k}.$$

(Hvis $r \in \{0,1,2,\ldots,n\}$, stemmer denne definition overens med den kombinatoriske definition af binomialkoefficient (side 32).)

Sætningen bliver i øvrigt gjort plausibel ved følgende ræsonnement (som inden for rammerne af en generel teori for stokastiske processer kan udbygges til et rigtigt bevis): Antal 0-er inden det $(j+k)$-te 1 er lig antal 0-er inden det j-te 1 plus antal 0-er i perioden fra det j-te 1 til det $(j+k)$-te 1; da værdierne til de enkelte tidspunkter genereres uafhængigt af hinanden, og da den regel der bestemmer ophøret af den første periode ikke afhænger af antallet af 0-er, er antal 0-er i den anden periode uafhængigt af antallet af 0-er i den første periode, og de er hver især negativt binomialfordelt.

Middelværdi og varians i den negative binomialfordeling er hhv. $k(1-p)/p$ og $k(1-p)/p^2$, jf. eksempel 4.5 side 81. Det forventede antal 0-er inden det k-te 1 er derfor $\mathrm{E}\,V_k = k(1-p)/p$, og den forventede ventetid til det k-te 1 er $\mathrm{E}\,T_k = k + \mathrm{E}\,V_k = k/p$.

Poissonfordelingen

Antag at en bestemt slags begivenheder indtræffer »helt tilfældigt« på noget vi vil kalde tidsaksen. Begivenhederne kan være jordskælv, dødsfald som følge af en bestemt ikke-epidemisk sygdom, trafikulykker i et bestemt vejkryds, partikler fra den kosmiske stråling, α-partikler der udsendes fra et radioaktivt stof, osv. osv. Man kan beskæftige sig med forskellige spørgsmål i den forbindelse, f.eks. hvad kan man sige om antallet af begivenheder i et bestemt tidsinterval, og hvad kan man sige om ventetiden fra en begivenhed til den næste. Det følgende handler om antallet af begivenheder i et bestemt interval.

Lad os se på et interval $]a;b]$ på tidsaksen. Vi vil gå ud fra at begivenhederne ikke alene indtræffer »helt tilfældigt«, men også at den rate hvormed de optræder, er konstant i de betragtede tidsrum (hvad dette nærmere skal betyde, vil

fremgå senere). Vi kan lave en approksimation til den »helt tilfældige« placering på den måde at vi deler intervallet op i et meget stort antal meget små intervaller, lad os sige n intervaller af længde $\Delta t = (b-a)/n$, og så for hvert enkelt interval lader en tilfældighedsmekanisme bestemme om det skal indeholde en begivenhed eller ej; sandsynligheden for at et givet interval får tildelt en begivenhed, skal være $p(\Delta t)$, og forskellige intervaller behandles uafhængigt af hinanden. Sandsynligheden $p(\Delta t)$ knyttet til det enkelte lille interval afhænger ikke af intervallets placering i det store interval, kun af det lille intervals længde Δt. På den måde bliver det samlede antal begivenheder i det store interval binomialfordelt med sandsynlighedsparameter $p(\Delta t)$ og antalsparameter n – og det er kun det samlede antal begivenheder vi er interesserede i.

Siméon-Denis Poisson
Fransk matematiker og fysiker (1781-1840).

Hvis n er meget stor, må man formode at vi får en god tilnærmelse til »det rigtige«. Vi vil derfor lade n gå mod uendelig; samtidig skal $p(\Delta t)$ gå mod 0 på en eller anden måde; en nærliggende måde at lade den gå mod 0 på er at lade det foregå sådan at middelværdien i binomialfordelingen er (eller går mod) et bestemt endeligt tal $\lambda(b-a)$. Middelværdien i vores binomialfordeling er $np(\Delta t) = \big((b-a)/\Delta t\big)p(\Delta t) = \frac{p(\Delta t)}{\Delta t}(b-a)$, så forslaget er at $p(\Delta t)$ skal gå mod 0 på en sådan måde at $\dfrac{p(\Delta t)}{\Delta t} \to \lambda$ hvor λ er en positiv konstant.

Eksponentialfunktionens rækkeudvikling
For alle (reelle og komplekse) t gælder at
$\exp(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!}$.

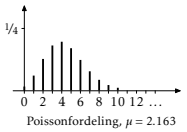
Parameteren λ skal fortolkes som den rate eller intensitet hvormed begivenhederne indtræffer, og den har dimension antal pr. tid. I demografiske og forsikringsmatematiske sammenhænge har den undertiden det maleriske navn *dødelighedsstyrke* (eng.: force of mortality).

Ved den beskrevne grænseovergang vil binomialsandsynlighederne ifølge nedenstående sætning (sætning 2.15) konvergere mod poissonsandsynligheder.

Definition 2.10: Poissonfordeling
Poissonfordelingen med parameter $\mu > 0$ er den fordeling på $\mathbb{N}_0$ som har sandsynlighedsfunktion

$$f(x) = \frac{\mu^x}{x!}\exp(-\mu), \quad x \in \mathbb{N}_0.$$

(At f summerer til 1, følger af eksponentialfunktionens rækkeudvikling.)



Poissonfordeling, $\mu = 2.163$

Middelværdien i poissonfordelingen med parameter μ er μ, og variansen er også lig μ; hvis X er poissonfordelt, er altså $\operatorname{Var} X = \operatorname{E} X$. – Vi har tidligere set at hvis X er binomialfordelt med parametre n og p, så er $\operatorname{Var} X = (1-p)\operatorname{E} X$, dvs. $\operatorname{Var} X < \operatorname{E} X$, og at hvis X er negativt binomialfordelt med parametre k og p, så er $p \operatorname{Var} X = \operatorname{E} X$, dvs. $\operatorname{Var} X > \operatorname{E} X$.

Sætning 2.15
Under den grænseovergang hvor $n \to \infty$ og $p \to 0$ på en sådan måde at np konvergerer mod $\mu > 0$, vil binomialfordelingen med parametre n og p konvergere mod poissonfordelingen med parameter μ i den forstand at

$$\binom{n}{x} p^x (1-p)^{n-x} \longrightarrow \frac{\mu^x}{x!}\exp(-\mu)$$

for ethvert $x \in \mathbb{N}_0$.

Bevis
Lad $x \in \mathbb{N}_0$ være givet. Vi har at

$$\begin{aligned}\binom{n}{x} p^x (1-p)^{n-x} &= \frac{n(n-1)\dots(n-x+1)}{x!}\, p^x (1-p)^{-x}(1-p)^n \\ &= \frac{1(1-\frac{1}{n})\dots(1-\frac{x-1}{n})}{x!}\,(np)^x (1-p)^{-x}(1-p)^n.\end{aligned}$$

Da x er fast, gælder at den store brøk konvergerer mod $1/x!$, $(np)^x$ konvergerer mod μ^x, og $(1-p)^{-x}$ konvergerer mod 1. Grænseværdien af $(1-p)^n$ findes let ved at se på logaritmen til den:

$$\ln\big((1-p)^n\big) = n\ln(1-p) = -np \cdot \frac{\ln(1-p)-\ln(1)}{-p} \longrightarrow -\mu$$

fordi brøken konvergerer mod differentialkvotienten af $\ln(t)$ i $t = 1$, som er 1. Altså konvergerer $(1-p)^n$ mod $\exp(-\mu)$, og sætningen er vist. □

Sætning 2.16
Hvis X_1 og X_2 er uafhængige poissonfordelte stokastiske variable med parametre μ_1 og μ_2, så er $X_1 + X_2$ poissonfordelt med parameter $\mu_1 + \mu_2$.

Bevis
Vi sætter $Y = X_1 + X_2$. Så er (jf. sætning 1.11 side 31)

$$\begin{aligned}\operatorname{P}(Y = y) &= \sum_{x=0}^{y} \operatorname{P}(X_1 = y-x)\operatorname{P}(X_2 = x) \\ &= \sum_{x=0}^{y} \frac{\mu_1^{y-x}}{(y-x)!}\exp(-\mu_1)\,\frac{\mu_2^x}{x!}\exp(-\mu_2) \\ &= \frac{1}{y!}\exp\big(-(\mu_1+\mu_2)\big)\sum_{x=0}^{y}\binom{y}{x}\mu_1^{y-x}\mu_2^x \\ &= \frac{1}{y!}\exp\big(-(\mu_1+\mu_2)\big)(\mu_1+\mu_2)^y.\end{aligned}$$

□

2.4 Opgaver

Opgave 2.1
Bevis sætning 2.4 på side 51.

Opgave 2.2
I definitionen på varians i det tællelige tilfælde (definition 2.6 side 55) går man tilsyneladende ud fra at $\mathrm{E}\big((X-\mathrm{E}\,X)^2\big) = \mathrm{E}(X^2) - (\mathrm{E}\,X)^2$. Gør rede for at denne formel er rigtig.

Opgave 2.3
Eftervis formlen $\operatorname{Var} X = \mathrm{E}(X(X-1)) - \xi(\xi-1)$, hvor $\xi = \mathrm{E}\,X$.

Opgave 2.4
Skitsér poissonfordelingens sandsynlighedsfunktion for forskellige værdier af parameteren μ.

Opgave 2.5
Udregn middelværdi og varians i poissonfordelingen med parameter μ. (Benyt evt. opgave 2.3.)

Opgave 2.6
Vis at hvis X_1 og X_2 er uafhængige poissonfordelte med parametre μ_1 og μ_2, så er den betingede fordeling af X_1 givet $X_1 + X_2$ en binomialfordeling.

Opgave 2.7
Skitsér den negative binomialfordelings sandsynlighedsfunktion for forskellige værdier af k og p.

Opgave 2.8
Betragt den negative binomialfordeling med parametre k og p, hvor $k > 0$ og $p \in]0\,;1[$. Som nævnt i teksten er middelværdi og varians i denne fordeling hhv. $k(1-p)/p$ og $k(1-p)/p^2$.

Gør rede for at man kan lave en grænseovergang med parametrene k og p sådan at middelværdien er konstant og variansen konvergerer mod middelværdien.

Vis at ved denne grænseovergang vil den negative binomialfordelings sandsynlighedsfunktion konvergere mod sandsynlighedsfunktionen for en poissonfordeling.

Opgave 2.9
Gør rede for at definitionen på kovarians (definition 2.7 side 55) er meningsfuld, dvs. at antagelsen om at X og Y har varians, sikrer at $\mathrm{E}\big((X-\mathrm{E}\,X)(Y-\mathrm{E}\,Y)\big)$ eksisterer.

Opgave 2.10
Lad os sige at på en given dag er antallet af personer der kører med S-toget uden billet, poissonfordelt med parameter μ, og lad os sige at der er sandsynligheden p for at en gratist bliver taget af billetkontrollen, og at de enkelte gratister bliver taget/ikke taget uafhængigt af hinanden. Hvad kan man heraf udlede om antallet af gratister der bliver taget af kontrollen?

Hvilke informationer skal man bruge for ud fra det konstaterede antal personer uden billet at kunne udtale sig om det faktiske antal uden billet?

Opgave 2.11: Sankt Petersborg-paradokset
En spiller vil deltage i et spil hvor de enkelt omgange foregår på følgende måde: hvis man betaler en indsats på k €, så vil man med sandsynlighed $1/2$ vinde $2k$ € og med sandsynlighed $1/2$ tabe sin indsats. Spilleren vil nu benytte en snedig strategi: han begynder med at satse 1 €; hvis han taber i en given omgang, satser han det dobbelte i næste omgang; hvis han vinder, får han udbetalt gevinsten og går hjem.

Hvor stor bliver hans nettogevinst?

Hvor mange spil skal der til før han kan gå hjem?

Selv om spilleren er sikker på at gå hjem med overskud, så skal han jo bruge noget kapital undervejs. Hvor mange penge skal han have med hjemmefra for at kunne »overleve« indtil han vinder?

Opgave 2.12
Bevis sætning 2.7 side 54.

Opgave 2.13
Som bekendt (sætning 2.7) kan middelværdien af et produkt af to uafhængige stokastiske variable udtrykkes på simpel vis ved de enkelte variables middelværdi.

Udled et lignende resultat om variansen af to uafhængige variable. – Er uafhængighedsantagelsen væsentlig?

3 Kontinuerte fordelinger

I DE FORRIGE KAPITLER har vi mødt stokastiske variable/sandsynlighedsfordelinger som rent faktisk er koncentreret på en endelig eller tællelig delmængde af de reelle tal; sådanne variable/fordelinger kaldes ofte diskrete variable/fordelinger. Der findes imidlertid også fordelinger med den egenskab, at der er et interval (eller flere intervaller) på $\mathbb{R}$ hvor enhver ikke-tom åben delmængde har positiv sandsynlighed, men ethvert punkt har sandsynlighed 0.

Eksempel: Når man leger flaskeleg, roterer man flasken for at få valgt en tilfældig retning; den person der uheldigvis sidder i den retning som flasken peger, skal gøre et eller andet nærmere aftalt som legen nu går ud på. Vi vil straks ødelægge det hele ved at modellere den roterende flaske ved »et punkt på enhedscirklen, valgt tilfældigt efter en ligefordeling«; hvad man helt præcist skal forstå ved det, er indtil videre ikke ganske klart, men det må blandt andet betyde at en cirkelbue af længde b får tildelt samme mængde sandsynlighed, lige meget hvor på cirklen den (altså buen) er placeret, og gad vidst om ikke den mængde sandsynlighed den skal tildeles, er $b/2\pi$. Da et givet punkt på cirkelperiferien er indeholdt i cirkelbuer af vilkårlig lille længde, må punktet selv have sandsynlighed 0.

Man kan bevise at der er en entydig sammenhæng mellem på den ene side sandsynlighedsmål på $\mathbb{R}$ og på den anden side fordelingsfunktioner, dvs. voksende og højrekontinuerte funktioner der går mod 0 hhv. 1 i $-\infty$ hhv. $+\infty$, jf. sætning 1.6 og/eller sætning 5.3. Sammenhængen er kort fortalt at for ethvert halvåbent interval $]a\,;b]$ er $\mathrm{P}(]a\,;b]) = F(b) - F(a)$. – Nu kan det matematiske raritetskabinet fremvise temmelig besynderlige funktioner der opfylder betingelserne for at være en fordelingsfunktion, men frygt ikke! I dette kapitel vil vi se på nogle yderst pæne fordelinger, nemlig de såkaldte kontinuerte fordelinger eller fordelinger med en tæthedsfunktion.

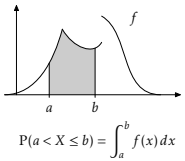
3.1 Grundlæggende definitioner

DEFINITION 3.1: SANDSYNLIGHEDSTÆTHEDSFUNKTION
En sandsynlighedstæthedsfunktion på $\mathbb{R}$ er en integrabel funktion $f : \mathbb{R} \to [0\,;+\infty[$ der integrerer til 1, dvs. $\int_{\mathbb{R}} f(x)\,dx = 1$.

Mere generelt er en sandsynlighedstæthedsfunktion på $\mathbb{R}^d$ en integrabel funktion $f : \mathbb{R}^d \to [0\,;+\infty[$ der integrerer til 1, dvs. $\int_{\mathbb{R}^d} f(x)\,dx = 1$.

Definition 3.2: Kontinuert fordeling

En kontinuert sandsynlighedsfordeling er en sandsynlighedsfordeling som har en sandsynlighedstæthedsfunktion: Hvis f er en sandsynlighedstæthedsfunktion på $\mathbb{R}$, så er den kontinuerte funktion $F(x) = \int_{-\infty}^{x} f(u)\,du$ fordelingsfunktion for en kontinuert fordeling på $\mathbb{R}$.



Man siger at en stokastisk variabel X har tæthedsfunktion f, hvis det er sådan at fordelingsfunktionen $F(x) = P(X \le x)$ for X er af formen $F(x) = \int_{-\infty}^{x} f(u)\,du$. Vi minder om at i så fald er $F'(x) = f(x)$ for alle kontinuitetspunkter x for f.

Sætning 3.1

Antag at X er en stokastisk variabel med tæthedsfunktion f. Så gælder

1. $P(a < X \le b) = \int_a^b f(x)\,dx$.
2. $P(X = x) = 0$ *for alle* $x \in \mathbb{R}$.
3. $P(a < X < b) = P(a < X \le b) = P(a \le X < b) = P(a \le X \le b)$.
4. *Hvis f er kontinuert i x, kan $f(x)\,dx$ opfattes som sandsynligheden for at X ligger i et infinitesimalt interval af længde dx omkring x.*

Bevis

Det første punkt følger af definitionerne på fordelingsfunktion og tæthedsfunktion. Det andet punkt følger af at

$$0 \le P(X = x) \le P(x - \tfrac{1}{n} < X \le x) = \int_{x-1/n}^{x} f(u)\,du,$$

hvor integralet går mod 0 når n går mod uendelig. Det tredje punkt følger af de to første. En præcis formulering af punkt 4 er at

$$\frac{1}{2\varepsilon} P(x - \varepsilon < X < x + \varepsilon) = \frac{1}{2\varepsilon} \int_{x-\varepsilon}^{x+\varepsilon} f(u)\,du \longrightarrow f(x)$$

når $\varepsilon \to 0$ og x er et kontinuitetspunkt for f; dette er et velkendt resultat fra analysen. □

Eksempel 3.1: Ligefordeling

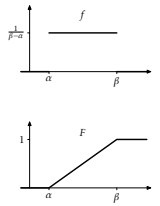
Ligefordelingen på intervallet fra α til β (hvor $-\infty < \alpha < \beta < +\infty$) er den fordeling som har tæthedsfunktionen

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

Dens fordelingsfunktion er den stykkevis lineære funktion

$$F(x) = \begin{cases} 0 & \text{for } x \le \alpha \\ \dfrac{x - \alpha}{\beta - \alpha} & \text{for } x_0 < x \le \beta \\ 1 & \text{for } x \ge \beta. \end{cases}$$

Ud fra en ligefordeling på et interval kan vi i øvrigt få den førnævnte ligefordeling på enhedscirklen, nemlig ved at tage ligefordelingen på $]0\,;2\pi[$ og flytte den over på enhedscirklen.



Tæthedsfunktion og fordelingsfunktion for ligefordelingen på intervallet fra α til β.

Definition 3.3: Flerdimensional kontinuert fordeling

Den d-dimensionale stokastiske variabel $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)$ siges at have tæthedsfunktion f, hvis det er sådan at f er en d-dimensional tæthedsfunktion, og det for alle intervaller $K_i =]a_i\,;b_i]$, $i = 1, 2, \ldots, d$, gælder at

$$P\left(\bigcap_{i=1}^{d} \{X_i \in K_i\}\right) = \int_{K_1 \times K_2 \times \ldots \times K_d} f(x_1, x_2, \ldots, x_d)\,dx_1\,dx_2 \ldots dx_d.$$

Funktionen f kaldes den simultane tæthedsfunktion for $X_1, X_2, \ldots, X_n$.

Mange begreber og definitioner (og sætninger) fra det diskrete tilfælde kan overføres til det kontinuerte tilfælde ved at man kort fortalt erstatter sandsynlighedsfunktioner med tæthedsfunktioner og summer med integraler. Eksempler:

Som pendant til sætning 1.8/korollar 1.9 har vi

Sætning 3.2

De stokastiske variable $X_1, X_2, \ldots, X_n$ er uafhængige hvis og kun hvis deres simultane tæthedsfunktion kan skrives som et produkt af tæthedsfunktionerne for de enkelte X_i-er.

Som pendant til sætning 1.11 har vi

Sætning 3.3

Hvis de stokastiske variable X_1 og X_2 er uafhængige og med tæthedsfunktioner f_1 og f_2, så har $Y = X_1 + X_2$ tæthedsfunktion

$$f(y) = \int_{\mathbb{R}} f_1(x)\,f_2(y - x)\,dx, \quad y \in \mathbb{R}.$$

(Se opgave 3.6.)

Transformation af fordelinger

Spørgsmålet om transformation af fordelinger handler nu som før om hvad man kan sige om fordelingen af $Y = t(X)$ når man kender fordelingen af X, og når t er en afbildning defineret på X's værdimængde. Man kan altid benytte den grundlæggende opskrift $\mathrm{P}(t(X) \le y) = \mathrm{P}\Big(X \in t^{-1}(]-\infty\,;y])\Big)$.

Eksempel 3.2
Antag at X er ligefordelt på $]-1\,;1[$. Vi vil finde fordelingen af $Y = X^2$.

For $y \in [0\,;1]$ er $\mathrm{P}(X^2 \le y) = \mathrm{P}\Big(X \in [-\sqrt{y}\,;\sqrt{y}]\Big) = \displaystyle\int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{2}\,dx = \sqrt{y}$, så Y har fordelingsfunktion

$$F_Y(y) = \begin{cases} 0 & \text{for } y \le 0, \\ \sqrt{y} & \text{for } 0 < y \le 1, \\ 1 & \text{for } y > 1. \end{cases}$$

Vi kan finde en tæthedsfunktion f for Y ved at differentiere F (i de punkter hvor den er differentiabel):

$$f(y) = \begin{cases} \frac{1}{2}y^{-1/2} & \text{for } 0 < y < 1, \\ 0 & \text{ellers.} \end{cases}$$

Hvis t er differentiabel, kan man i visse situationer opskrive en sammenhæng mellem tæthedsfunktionen for X og tæthedsfunktionen for $Y = t(X)$. Der er tale om en direkte overførsel af resultater om substitution i integraler, hentet fra den matematiske analyse:

Sætning 3.4
Antag at X er en reel stokastisk variabel, I en åben delmængde af $\mathbb{R}$ således at $\mathrm{P}(X \in I) = 1$, og $t : I \to \mathbb{R}$ en C^1-funktion defineret på I med $t' \neq 0$ overalt på I. Hvis X har tæthedsfunktion f_X, så har $Y = t(X)$ tæthedsfunktion

$$f_Y(y) = f_X(x)\Big|t'(x)\Big|^{-1} = f_X(x)\Big|(t^{-1})'(y)\Big|$$

hvor $x = t^{-1}(y)$ og $y \in t(I)$. Formlen skrives undertiden på den korte form

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|.$$

Der findes også en flerdimensional udgave:

Sætning 3.5
Antag at $\boldsymbol{X}$ er en d-dimensional stokastisk variabel, I en åben delmængde af $\mathbb{R}^d$ således at $\mathrm{P}(\boldsymbol{X} \in I) = 1$, og $\boldsymbol{t} : I \to \mathbb{R}^d$ en C^1-funktion defineret på I og med en

funktionalmatrix $D\boldsymbol{t} = \dfrac{d\boldsymbol{y}}{d\boldsymbol{x}}$ som er regulær på hele I. Hvis $\boldsymbol{X}$ har tæthedsfunktion $f_{\boldsymbol{X}}$, så har $\boldsymbol{Y} = \boldsymbol{t}(\boldsymbol{X})$ tæthedsfunktion

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(\boldsymbol{x})\Big|\det D\boldsymbol{t}(\boldsymbol{x})\Big|^{-1} = f_{\boldsymbol{X}}(\boldsymbol{x})\Big|\det D\boldsymbol{t}^{-1}(\boldsymbol{y})\Big|$$

hvor $\boldsymbol{x} = \boldsymbol{t}^{-1}(\boldsymbol{y})$ og $\boldsymbol{y} \in \boldsymbol{t}(I)$. Formlen skrives undertiden på den korte form

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(\boldsymbol{x})\left|\det\frac{d\boldsymbol{x}}{d\boldsymbol{y}}\right|.$$

Eksempel 3.3
Hvis X er ligefordelt på $]0\,;1[$, så har $Y = -\ln X$ tæthedsfunktion

$$f_Y(y) = 1 \cdot \left|\frac{dx}{dy}\right| = \exp(-y)$$

når $0 < x < 1$, dvs. når $0 < y < +\infty$, og $f_Y(y) = 0$ ellers. – Fordelingen af Y er således en eksponentialfordeling, se afsnit 3.3.

Bevis for sætning 3.4
Sætningen er en omformulering af et resultat fra den matematiske analyse om substitution i integraler. – Da den afledede af funktionen t er kontinuert og altid forskellig fra 0, er t enten strengt voksende eller strengt aftagende; vi antager at den er strengt voksende. Da er $Y \le b \Leftrightarrow X \le t^{-1}(b)$, så vi får følgende udtryk for fordelingsfunktionen F_Y for Y:

$$\begin{aligned} F_Y(b) &= \mathrm{P}(Y \le b) \\ &= \mathrm{P}\Big(X \le t^{-1}(b)\Big) \\ &= \int_{-\infty}^{t^{-1}(b)} f_X(x)\,dx \qquad [\text{ substitution } x = t^{-1}(y) \text{ }] \\ &= \int_{-\infty}^{b} f_X\Big(t^{-1}(y)\Big)(t^{-1})'(y)\,dy. \end{aligned}$$

Dette viser at funktionen $f_Y(y) = f_X\Big(t^{-1}(y)\Big)(t^{-1})'(y)$ har den egenskab at $F_Y(y) = \displaystyle\int_{-\infty}^{y} f_Y(u)\,du$, dvs. Y har tæthedsfunktion f_Y. □

Betingning

Hvis man har to kontinuerte stokastiske variable X og Y og søger den betingede fordeling af X givet Y, kan man ikke bare gå ud fra at udtrykket $\mathrm{P}(X = x \mid Y = y)$ uden videre er veldefineret og lig med $\mathrm{P}(X = x, Y = y)/\mathrm{P}(Y = y)$; nævneren $\mathrm{P}(Y = y)$ er altid nul, og det er tælleren ofte også.

I stedet kan man forsøge sig med at betinge med en velvalgt hændelse med positiv sandsynlighed, f.eks $y-\varepsilon < Y < y+\varepsilon$, og undersøge om den (veldefinerede) betingede sandsynlighed $\mathrm{P}(X \le x \mid y-\varepsilon < Y < y+\varepsilon)$ har en grænseværdi for $\varepsilon \to 0$; hvis det er tilfældet, kunne man formode at grænseværdien ville være $\mathrm{P}(X \le x \mid Y = y)$.

En anden, mere heuristisk tilgang til problemet er som følger: lad f_{XY} være den simultane tæthedsfunktion for X og Y og lad f_Y være tæthedsfunktionen for Y. Sandsynligheden for at (X,Y) ligger i et infinitesimalt rektangel med kantlængder dx og dy omkring (x,y) er da $f_{XY}(x,y)\,dx\,dy$, og sandsynligheden for at Y ligger i et infinitesimalt interval af længde dy omkring y er $f_Y(y)\,dy$; den betingede sandsynlighed for at X ligger i et infinitesimalt interval af længde dx omkring x, givet at Y ligger i et infinitesimalt interval af længde dy omkring y, er dermed $\dfrac{f_{XY}(x,y)\,dx\,dy}{f_Y(y)\,dy} = \dfrac{f_{XY}(x,y)}{f_Y(y)}\,dx$, så den betingede tæthed kunne formodes at være

$$f(x \mid y) = \frac{f_{XY}(x,y)}{f_Y(y)}.$$

Dette er faktisk også rigtigt i mange situationer.

Ovenstående omtale af betingede fordelinger i forbindelse med kontinuerte fordelinger er naturligvis overordentlig lemfældig; der findes matematiske rammer inden for hvilke man kan fremsætte præcise og rigtige udsagn om disse spørgsmål, men dem kommer vi ikke ind på i nærværende fremstilling.

3.2 Middelværdi

Middelværdien af en stokastisk variabel med tæthedsfunktion defineres på lignende måde som for stokastiske variable på et tælleligt udfaldsrum, blot erstattes sumtegnet med et integraltegn.

Definition 3.4: Middelværdi
Lad X være en stokastisk variabel med tæthedsfunktion f.
Hvis $\displaystyle\int_{-\infty}^{+\infty} |x| f(x)\,dx < +\infty$, *så siges X at have en middelværdi, og middelværdien af X defineres da som tallet* $\mathrm{E}\,X = \displaystyle\int_{-\infty}^{+\infty} x f(x)\,dx$.

Sætninger/regneregler for middelværdi og varians fra det tællelige tilfælde kan uden videre overføres.

3.3 Eksempler

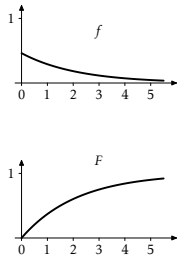
Eksponentialfordelingen

Definition 3.5: Eksponentialfordeling
Eksponentialfordelingen med skalaparameter $\beta > 0$ er fordelingen med tæthedsfunktion

$$f(x) = \begin{cases} \dfrac{1}{\beta}\exp(-x/\beta) & \text{for } x > 0 \\ 0 & \text{for } x \le 0. \end{cases}$$

Fordelingsfunktionen er

$$F(x) = \begin{cases} 1 - \exp(-x/\beta) & \text{for } x > 0 \\ 0 & \text{for } x \le 0. \end{cases}$$



Tæthedsfunktion og fordelingsfunktion for eksponentialfordelingen med $\beta = 2.163$.

At β er en *skala*parameter, fremgår af følgende sætning:

Sætning 3.6
Hvis X er eksponentialfordelt med skalaparameter β, og hvis $a > 0$, så er aX eksponentialfordelt med skalaparameter $a\beta$.

Bevis
Ifølge sætning 3.4 er tætheden for $Y = aX$ funktionen f_Y givet ved

$$f_Y(y) = \begin{cases} \dfrac{1}{\beta}\exp\Big(-(y/a)/\beta\Big)\cdot\dfrac{1}{a} = \dfrac{1}{a\beta}\exp\Big(-y/(a\beta)\Big) & \text{for } y > 0 \\ 0 & \text{for } y \le 0. \end{cases}$$

□

Sætning 3.7
Hvis X er eksponentialfordelt med skalaparameter β, er $\mathrm{E}\,X = \beta$ og $\mathrm{Var}\,X = \beta^2$.

Bevis
Da β er en skalaparameter, er det (jf. regnereglerne for middelværdi og varians) nok at vise påstanden for $\beta = 1$. Dette kan gøres ved almindelig udregning, brug evt. formlen sidst i sidebemærkningen om gammafunktionen. □

Eksponentialfordelingen benyttes ofte til at modellere ventetider mellem på hinanden følgende begivenheder, når disse begivenheder indtræffer »helt tilfældigt«. Eksponentialfordelingen er »uden hukommelse« i den forstand at sandsynligheden for at man skal vente længere end $s+t$, givet at man allerede har ventet s, er den samme som sandsynligheden for at man skal vente længere end t, dvs. hvis X er eksponentialfordelt, så gælder

$$\mathrm{P}(X > s+t \mid X > s) = \mathrm{P}(X > t), \qquad s,t > 0. \tag{3.1}$$

Gamma-funktionen
Gammafunktionen er funktionen

$$\Gamma(t) = \int_0^{+\infty} x^{t-1}\exp(-x)\,dx$$

hvor $t > 0$. (Faktisk er det muligt at definere $\Gamma(t)$ for alle $t \in \mathbb{C} \setminus \{0,-1,-2,-3,\ldots\}$.)
Der gælder

- $\Gamma(1) = 1$
- $\Gamma(t+1) = t\Gamma(t)$ for alle t
- $\Gamma(n) = (n-1)!$ for $n = 1,2,3,\ldots$.
- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. (Se også side 76.)

Ved almindelig integralsubstitution får man følgende formel:

$$\Gamma(k)\beta^k = \int_0^{+\infty} x^{k-1}\exp(-x/\beta)\,dx$$

der gælder for $k > 0$ og $\beta > 0$.

Bevis for formel (3.1)
Da $X > s+t \Rightarrow X > s$, er $\{X > s+t\} \cap \{X > s\} = \{X > s+t\}$ og dermed

$$\begin{aligned} \mathrm{P}(X > s+t \mid X > s) &= \frac{\mathrm{P}(X > s+t)}{\mathrm{P}(X > s)} = \frac{1-F(s+t)}{1-F(s)} \\ &= \frac{\exp(-(s+t)/\beta)}{\exp(-s/\beta)} = \exp(-t/\beta) = \mathrm{P}(X > t) \end{aligned}$$

for vilkårlige $s, t > 0$. □

Eksponentialfordelingen er på denne måde det kontinuerte modstykke til den geometriske fordeling, se side 57.

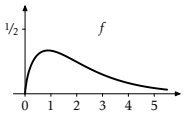
Endvidere er eksponentialfordelingen også beslægtet med poissonfordelingen. Antag at en type begivenheder indtræffer tilfældigt på den måde som er beskrevet i afsnittet om poissonfordelingen (side 59). Det at man skal vente mere end t på den første begivenhed, er det samme som at der i intervallet fra tid 0 til tid t indtræffer 0 begivenheder. Sandsynligheden for det sidstnævnte kan udregnes inden for poissonmodellens rammer til $\frac{(\lambda t)^0}{0!}\exp(-\lambda t) = \exp(-\lambda t)$, dvs. ventetiden til første begivenhed er eksponentialfordelt med parameter $1/\lambda$.

Gammafordelingen

Definition 3.6: Gammafordeling
Gammafordelingen med formparameter $k > 0$ og skalaparameter $\beta > 0$ er fordelingen med tæthedsfunktion

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(k)\beta^k}\,x^{k-1}\exp(-x/\beta) & \text{for } x > 0 \\ 0 & \text{for } x \le 0. \end{cases}$$

Tilfældet $k = 1$ giver eksponentialfordelingen.



Gammafordeling med $k = 1.7$ og $\beta = 1.27$

Sætning 3.8
Hvis X er gammafordelt med formparameter k og skalaparameter β, og hvis $a > 0$, så er $Y = aX$ gammafordelt med formparameter k og skalaparameter $a\beta$.

Bevis
På samme måde som sætning 3.6, dvs. ved brug af sætning 3.4. □

Sætning 3.9
Hvis X_1 og X_2 er uafhængige gammafordelte med formparametre hhv. k_1 og k_2 og med samme skalaparameter β, så er $X_1 + X_2$ gammafordelt med formparameter $k_1 + k_2$ og skalaparameter β.

Bevis
Ifølge sætning 3.3 er tætheden for $Y = X_1 + X_2$ (når $y > 0$)

$$f(y) = \int_0^y \frac{1}{\Gamma(k_1)\beta^{k_1}}x^{k_1-1}\exp(-x/\beta)\cdot\frac{1}{\Gamma(k_2)\beta^{k_2}}(y-x)^{k_2-1}\exp(-(y-x)/\beta)\,dx.$$

Ved at foretage substitutionen $u = x/y$ i integralet får vi

$$f(y) = \left(\frac{1}{\Gamma(k_1)\Gamma(k_2)\beta^{k_1+k_2}}\int_0^1 u^{k_1-1}(1-u)^{k_2-1}\,du\right)y^{k_1+k_2-1}\exp(-y/\beta).$$

Tætheden er altså af formen $f(y) = \text{konst}\,y^{k_1+k_2-1}\exp(-y/\beta)$ hvor konstanten sørger for at f integrerer til 1; ifølge formlen sidst i sidebemærkningen side 72 skal konstanten være $\dfrac{1}{\Gamma(k_1+k_2)\beta^{k_1+k_2}}$. Hermed er sætningen vist. □

Af sætning 3.9 følger at middelværdien og variansen i gammafordelingen må være lineære funktioner af formparameteren, og desuden må middelværdien være lineær i skalaparameteren og variansen kvadratisk i skalaparameteren. Det der gælder, er

Sætning 3.10
Hvis X er gammafordelt med parametre k og β, så er $\mathrm{E}\,X = k\beta$ og $\mathrm{Var}\,X = k\beta^2$.

Inden for den matematiske såvel som den praktiske statistik optræder nogle specielle gammafordelinger, nemlig χ^2-fordelinger.

Definition 3.7: χ^2-fordeling
En χ^2-fordeling med n frihedsgrader er det samme som en gammafordeling med formparameter $n/2$ og skalaparameter 2, dvs. med tæthedsfunktion

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

Beta-funktionen
Beta-funktionen er funktionen

$$B(k_1,k_2) = \int_0^1 u^{k_1-1}(1-u)^{k_2-1}\,du$$

hvor $k_1, k_2 > 0$. Af beviset for sætning 3.9 får vi som en sidegevinst at

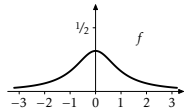
$$B(k_1,k_2) = \frac{\Gamma(k_1)\Gamma(k_2)}{\Gamma(k_1+k_2)}$$

for $k_1, k_2 > 0$.

Cauchyfordelingen

Definition 3.8: Cauchyfordelingen
Cauchyfordelingen med skalaparameter $\beta > 0$ er fordelingen med tæthedsfunktion

$$f(x) = \frac{1}{\pi\beta}\,\frac{1}{1+(x/\beta)^2}, \quad x \in \mathbb{R}.$$



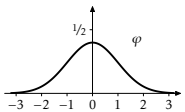
Cauchyfordeling med $\beta = 1$

Denne fordeling er meget anvendt som modeksempel (!), blandt andet er den et eksempel på en fordeling som ikke har nogen middelværdi (fordi funktionen $|x|/(1+x^2)$ er ikke integrabel). Fordelingen kan dog også forekomme i »det virkelige liv«, se f.eks. opgave 3.4.

Normalfordelingen

Definition 3.9: Standardnormalfordeling
Standardnormalfordelingen er fordelingen med tæthedsfunktion

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\tfrac{1}{2}x^2\Big), \quad x \in \mathbb{R}.$$



Standardnormalfordeling

Fordelingsfunktionen for standardnormalfordelingen er

$$\Phi(x) = \int_{-\infty}^{x} \varphi(u)\,du, \quad x \in \mathbb{R}.$$

Carl Friedrich Gauss
tysk matematiker (1777-1855).
Kendt blandt andet for sit arbejde inden for geometri og matematisk statistik (herunder mindste kvadraters metode). Han beskæftigede sig også med de praktiske sider af matematikken, f.eks. landmåling.
På de tyske 10 DM-sedler der var i omløb de sidste decennier før overgangen til euroen, kunne man se et portræt af Gauß, en tegning af normalfordelingstætheden og hans triangulering af et landområde i Nordtyskland.

Definition 3.10: Normalfordeling
Normalfordelingen (eller Gaußfordelingen) med positionsparameter μ og kvadratisk skalaparameter $\sigma^2 > 0$ er fordelingen med tæthedsfunktion

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\tfrac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\Big) = \frac{1}{\sigma}\,\varphi\Big(\frac{x-\mu}{\sigma}\Big), \qquad x \in \mathbb{R}.$$

Betegnelserne positionsparameter og kvadratisk skalaparameter retfærdiggøres af følgende sætning, der vises ved anvendelse af sætning 3.4:

Sætning 3.11
Hvis X er normalfordelt med positionsparameter μ og kvadratisk skalaparameter σ^2, og hvis a og b er reelle tal med $a \neq 0$, så er $Y = aX + b$ normalfordelt med positionsparameter $a\mu + b$ og kvadratisk skalaparameter $a^2\sigma^2$.

Sætning 3.12
Hvis X_1 og X_2 er uafhængige og normalfordelte med parametre μ_1 og σ_1^2 hhv. μ_2 og σ_2^2, så er $Y = X_1 + X_2$ normalfordelt med positionsparameter $\mu_1 + \mu_2$ og kvadratisk skalaparameter $\sigma_1^2 + \sigma_2^2$.

Bevis
På grund af sætning 3.11 er det nok at vise at hvis X_1 er standardnormalfordelt, og hvis X_2 er normalfordelt med positionsparameter 0 og kvadratisk skalaparameter σ^2, så er $X_1 + X_2$ normalfordelt med positionsparameter 0 og kvadratisk skalaparameter $1 + \sigma^2$. Det vises ved at benytte sætning 3.3 og omskrive. □

Bevis for at φ er en tæthedsfunktion
Vi mangler at gøre rede for at φ faktisk er en sandsynlighedstæthedsfunktion, altså at $\int_{\mathbb{R}} \varphi(x)\,dx = 1$, eller sagt på en anden måde: Hvis vi lader c betegne den konstant der gør $f(x) = c\exp(-\frac{1}{2}x^2)$ til en sandsynlighedstæthedsfunktion, skal det vises at $c = 1/\sqrt{2\pi}$.

Det snedige trick er at se på to uafhængige stokastiske variable X_1 og X_2 der hver især har tæthed f. Deres simultane tæthedsfunktion er

$$f(x_1)f(x_2) = c^2 \exp\Big(-\tfrac{1}{2}(x_1^2 + x_2^2)\Big).$$

Nu ser vi på en funktion $(y_1, y_2) = t(x_1, x_2)$ hvor sammenhængen mellem x-er og y-er er givet ved $x_1 = \sqrt{y_2}\cos y_1$ og $x_2 = \sqrt{y_2}\sin y_1$ for $(x_1, x_2) \in \mathbb{R}^2$ og $(y_1, y_2) \in \,]0\,;2\pi[\,\times\,]0\,;+\infty[$. Sætning 3.5 giver en anvisning på hvordan man finder tæthedsfunktionen for den todimensionale stokastiske variabel $(Y_1, Y_2) = t(X_1, X_2)$; almindelig udregning giver at den er $\frac{1}{2}c^2\exp(-\frac{1}{2}y_2)$ når $0 < y_1 < 2\pi$ og $y_2 > 0$, og 0 ellers. Da det er en sandsynlighedstæthedsfunktion, integrerer den til 1:

$$\begin{aligned} 1 &= \int_0^{+\infty}\int_0^{2\pi} \tfrac{1}{2}c^2 \exp(-\tfrac{1}{2}y_2)\,dy_1\,dy_2 \\ &= 2\pi c^2 \int_0^{+\infty} \tfrac{1}{2}\exp(-\tfrac{1}{2}y_2)\,dy_2 \\ &= 2\pi c^2, \end{aligned}$$

hvoraf $c = 1/\sqrt{2\pi}$. □

Sætning 3.13
Hvis $X_1, X_2, \ldots, X_n$ er uafhængige standardnormalfordelte stokastiske variable, så er fordelingen af $Y = X_1^2 + X_2^2 + \ldots + X_n^2$ en χ^2-fordeling med n frihedsgrader.

Bevis
Da χ^2-fordelingen er en gammafordeling, kan vi nøjes med at vise påstanden for $n = 1$ og derefter henvise til sætning 3.9. Tilfældet $n = 1$ behandles sålunde: For $x > 0$ er

$$\begin{aligned} \mathrm{P}(X_1^2 \le x) &= \mathrm{P}(-\sqrt{x} < X_1 \le \sqrt{x}) \\ &= \int_{-\sqrt{x}}^{\sqrt{x}} \varphi(u)\,du \qquad [\ \varphi \text{ er lige }] \\ &= 2\int_0^{\sqrt{x}} \varphi(u)\,du \qquad [\ \text{substitution } t = u^2\ ] \\ &= \int_0^x \frac{1}{\sqrt{2\pi}}\,t^{-1/2}\exp(-\tfrac{1}{2}t)\,dt, \end{aligned}$$

som ved differentiation giver at tæthedsfunktionen for X_1^2 er

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

Tæthedsfunktionen for χ^2-fordelingen med 1 frihedsgrad er ifølge definition 3.7

$$\frac{1}{\Gamma(1/2)\,2^{1/2}}\,x^{-1/2}\exp(-\tfrac{1}{2}x), \quad x>0.$$

Da de to tæthedsfunktioner er proportionale, er de ens, og vi har dermed fuldført beviset for sætningen. Som en sidegevinst får vi at $\Gamma(1/2)=\sqrt{\pi}$. □

Sætning 3.14
Hvis X er normalfordelt med positionsparameter μ og kvadratisk skalaparameter σ^2, så er $\mathrm{E}\,X=\mu$ *og* $\mathrm{Var}\,X=\sigma^2$.

Bevis
På grund af sætning 3.11 og regnereglerne for middelværdi og varians er det nok at vise sætningen i tilfældet $\mu=0$, $\sigma^2=1$. Antag derfor at $\mu=0$, $\sigma^2=1$. Af symmetrigrunde er $\mathrm{E}\,X=0$, og dermed $\mathrm{Var}\,X=\mathrm{E}((X-\mathrm{E}\,X)^2)=\mathrm{E}(X^2)$; da X^2 er χ^2-fordelt med 1 frihedsgrad, dvs. gammafordelt med parametre $1/2$ og 2, er $\mathrm{E}(X^2)=1$ (sætning 3.10). □

Normalfordelingen bliver anvendt meget i statistiske modeller. En af grundene – og en af grundene til at normalfordelingen i det hele taget er genstand for så stor opmærksomhed – er Den Centrale Grænseværdisætning. (En ganske anderledes begrundelse for og udledning af normalfordelingen kan ses på side 199ff.)

Sætning 3.15: Den Centrale Grænseværdisætning
Antag at $X_1, X_2, X_3, \ldots$ er en følge af indbyrdes uafhængige identisk fordelte stokastiske variable med middelværdi μ og varians $\sigma^2>0$, og lad S_n betegne summen af de n første: $S_n=X_1+X_2+\ldots+X_n$.

Da gælder at for $n\to\infty$ er $\dfrac{S_n-n\mu}{\sqrt{n\sigma^2}}$ asymptotisk standardnormalfordelt i den forstand at

$$\lim_{n\to\infty}\mathrm{P}\left(\frac{S_n-n\mu}{\sqrt{n\sigma^2}}\le x\right)=\Phi(x)$$

for ethvert $x\in\mathbb{R}$.

Bemærkninger:

- Størrelsen $\dfrac{S_n-n\mu}{\sqrt{n\sigma^2}}=\dfrac{\frac{1}{n}S_n-\mu}{\sqrt{\sigma^2/n}}$ er summen hhv. gennemsnittet af de n første variable minus middelværdien deraf, divideret med standardafvigelsen, dvs. udtrykket har middelværdi 0 og varians 1.
- Ifølge Store Tals Lov (side 43) vil $\frac{1}{n}S_n-\mu$ være meget lille (med meget stor sandsynlighed) når n vokser. Den Centrale Grænseværdisætning fortæller hvordan $\frac{1}{n}S_n-\mu$ bliver meget lille, nemlig på en sådan måde at $\frac{1}{n}S_n-\mu$ divideret med $\sqrt{\sigma^2/n}$ er standardnormalfordelt.

- Sætningen er yderst generel, idet der ikke gøres andre antagelser om X-ernes fordeling end at den har en middelværdi og en varians.

Beviset for sætningen forbigås.

3.4 Opgaver

Opgave 3.1
Skitsér gammafordelingens tæthed for forskellige værdier af formparameter og skalaparameter. (Find blandt andet ud af hvordan udseendet nær $x=0$ afhænger af parametrene; se f.eks. på $\lim_{x\to 0} f(x)$ og $\lim_{x\to 0} f'(x)$.)

Opgave 3.2
Skitsér normalfordelingens tæthed for forskellige værdier af μ og σ^2.

Opgave 3.3
Lad F være en kontinuert og strengt voksende funktion defineret på et åbent interval $I\subseteq\mathbb{R}$, og sådan at $F(x)$ konvergerer mod 0 hhv. 1 når x konvergerer mod venstre hhv. højre endepunkt af I (dvs. F kan – eventuelt efter en lille udvidelse – være en fordelingsfunktion).

1. Antag at den stokastiske variabel U er ligefordelt på $]0\,;1[$. Vis at fordelingsfunktionen for $X=F^{-1}(U)$ er F.
2. Antag at Y har fordelingsfunktion F. Vis at $V=F(Y)$ er ligefordelt på $]0\,;1[$.

Kommentar: Computerprogrammer har ofte en funktion der leverer tilfældige (eller i det mindste pseudo-tilfældige) tal fra ligefordelingen på $]0\,;1[$. Denne opgave anviser én mulig måde til ud fra tilfældige ligefordelte tal at fremstille tilfældige tal fra en vilkårlig kontinuert fordeling.

Opgave 3.4
En todimensional cowboy affyrer sin pistol i en tilfældig retning. Et stykke fra cowboyen er der et meget langt (f.eks. uendelig langt) plankeværk.

1. Hvad er sandsynligheden for at kuglen rammer plankeværket?
2. Givet at han rammer, hvad kan man så sige om fordelingen af det sted kuglen rammer?

Opgave 3.5
Antag at (X_1,X_2) er en todimensional stokastisk variabel med tæthedsfunktion $f(x_1,x_2)$, jf. definition 3.3. Vis at funktionen $f_1(x_1)=\displaystyle\int_{\mathbb{R}} f(x_1,x_2)\,dx_2$ er en tæthedsfunktion for X_1.

Opgave 3.6
Antag at X_1 og X_2 er uafhængige stokastiske variable med tæthedsfunktioner f_1 og f_2. Sæt $Y_1=X_1+X_2$ og $Y_2=X_1$.

1. Find tæthedsfunktionen for (Y_1,Y_2). (Benyt sætning 3.5.)
2. Find fordelingen af Y_1. (Benyt evt. opgave 3.5.)
3. Gøre rede for at det foregående er et bevis for sætning 3.3.

Opgave 3.7
På side 73 står der at det følger af sætning 3.9 at middelværdien og variansen i gammafordelingen er lineære funktioner af formparameteren, og at middelværdien er lineær i skalaparameteren og variansen kvadratisk i skalaparameteren. Gør rede for hvordan det følger af sætningen.

Opgave 3.8
Kviksølvindholdet i sværdfisk er i visse egne af USA normalfordelt med middelværdi 1.1 ppm og varians $0.25\,\text{ppm}^2$ (se f.eks. Lee and Krutchkoff (1980) og/eller opgave 5.2 og 5.3 i Larsen (2006)), men efter sundhedsmyndighedernes vurdering bør gennemsnitsindholdet af kviksølv i konsumfisk ikke overstige 1 ppm. De fisk der sælges via de autoriserede salgskanaler, bliver kontrolleret af sundhedsmyndighederne, og hvis kviksølvindholdet i en fisk er for højt, bliver den kasseret. Imidlertid bliver ca. 25% af de fangede fisk solgt på det sorte marked og altså uden om kontrollen; derfor er det ikke tilstrækkeligt at myndighederne anvender en kassationsregel der hedder »hvis fisken indeholder over 1 ppm kviksølv, så kasseres den«.

Hvor lille skal myndighedernes kassationsgrænse være for at det forventede kviksølvindhold i en solgt fisk (solgt legalt eller illegalt) bliver under 1 ppm, og hvordan skal grænsen vælges hvis det forventede indhold skal være så lille som muligt?

Opgave 3.9
Lad X være en stokastisk variabel. Vi definerer *støtten* for X (eller mere præcist: støtten for fordelingen af X) som den mindste afsluttede delmængde $B \subseteq \mathbb{R}$ med den egenskab at $\mathrm{P}(X \in B) = 1$. Støtten for X betegnes $\mathrm{supp}(X)$. – Man kan også karakterisere $\mathrm{supp}(X)$ på følgende måde: $x \notin \mathrm{supp}(X)$ hvis og kun hvis der findes en åben omegn U af x således at $\mathrm{P}(X \in U) = 0$.

1. Antag at X er diskret med sandsynlighedsfunktion f.
   Vis at $\mathrm{supp}(X) = \{x : f(x) > 0\}$ (dvs. støtten for f).
2. Antag at X er tællelig med sandsynlighedsfunktion f. Hvad er $\mathrm{supp}(X)$ i dette tilfælde?
   (Testeksempel: En stokastisk variabel der antager værdien $\frac{1}{n}$ med sandsynlighed 2^{-n}, $n = 1, 2, 3, \ldots$)
3. Antag at X er kontinuert med tæthedsfunktion f. Hvad er $\mathrm{supp}(X)$ da?

4 Frembringende funktioner

Inden for matematikkens verden finder man mange eksempler på at man konstruerer og benytter en-entydige oversættelser, repræsentationer, af én matematisk struktur til en anden. En af pointerne hermed er at visse ræsonnementer eller beviser eller udregninger er nemmere i nogle af strukturerne end i andre. Et klassisk eksempel er logaritmefunktionen, der som bekendt er en isomorfi mellem $(\mathbb{R}_+, \cdot)$ og $(\mathbb{R}, +)$, og som altså laver gange-stykker om til plus-stykker.

I dette kapitel skal vi møde en repræsentation af mængden af stokastiske variable med værdier i $\mathbb{N}_0$ (eller mere rigtigt: af mængden af sandsynlighedsfordelinger på $\mathbb{N}_0$) ved hjælp af såkaldte frembringende funktioner.

Bemærk: Alle stokastiske variable der optræder i dette kapitel, er stokastiske variable med værdier i $\mathbb{N}_0$.

4.1 Grundlæggende egenskaber

Definition 4.1: Frembringende funktion
Lad X være en stokastisk variabel med værdier i $\mathbb{N}_0$. Den frembringende funktion for X (eller mere præcist: for fordelingen af X) er funktionen

$$G(s) = \mathrm{E}\, s^X = \sum_{x=0}^{\infty} s^x \,\mathrm{P}(X = x), \quad s \in [0\,;1].$$

Teorien for frembringende funktionen trækker i høj grad på resultater fra teorien for potensrækker. For den frembringende funktion G for X gælder blandt andet:

1. G er kontinuert og har værdier i $[0\,;1]$, specielt er $G(0) = \mathrm{P}(X = 0)$ og $G(1) = 1$.
2. G er vilkårligt mange gange differentiabel på $]0\,;1[$, og den k-te afledede er

$$\begin{aligned} G^{(k)}(s) &= \sum_{x=k}^{\infty} x(x-1)(x-2)\ldots(x-k+1)\, s^{x-k}\,\mathrm{P}(X = x) \\ &= \mathrm{E}\Big(X(X-1)(X-2)\ldots(X-k+1)\, s^{X-k}\Big). \end{aligned}$$

   Specielt er $G^{(k)}(s) \geq 0$ for alle $s \in [0\,;1]$.

3. Sættes $s = 0$ i udtrykket for $G^{(k)}(s)$, fås $G^{(k)}(0) = k!\,\mathrm{P}(X = k)$, dvs. det er muligt at regne fra den frembringende funktion tilbage til sandsynlighedsfordelingen.
   Heraf følger at to forskellige sandsynlighedsfordelinger ikke kan have samme frembringende funktion, eller sagt på en anden måde: den afbildning der til en sandsynlighedsfordeling knytter dens frembringende funktion, er injektiv.
4. Ved at lade $s \to 1$ i udtrykket for $G^{(k)}(s)$ får vi at

$$G^{(k)}(1) = \mathrm{E}\Big(X(X-1)(X-2)\ldots(X-k+1)\Big), \tag{4.1}$$

   mere præcist gælder der at $\lim\limits_{s\to 1} G^{(k)}(s)$ er et endeligt tal hvis og kun hvis X^k har middelværdi, og i givet fald er (4.1) opfyldt.
   Vi noterer specielt at

$$\mathrm{E}\,X = G'(1) \tag{4.2}$$

   og

$$\begin{aligned}\mathrm{Var}\,X &= G''(1) - G'(1)\Big(G'(1)-1\Big)\\ &= G''(1) - \Big(G'(1)\Big)^2 + G'(1).\end{aligned} \tag{4.3}$$

   (formlen for variansen følger ved brug af opgave 2.3).

Et nyttigt og vigtigt resultat er

Sætning 4.1
Hvis X og Y er uafhængige stokastiske variable med frembringende funktioner G_X og G_Y, så er den frembringende funktion for summen af X og Y produktet af de frembringende funktioner: $G_{X+Y} = G_X\,G_Y$.

Bevis
For $|s| < 1$ er $G_{X+Y}(s) = \mathrm{E}\,s^{X+Y} = \mathrm{E}\big(s^X s^Y\big) = \mathrm{E}\,s^X\,\mathrm{E}\,s^Y = G_X(s)G_Y(s)$ hvor vi først har brugt eksponentialfunktionernes funktionalligning og dernæst sætning 1.17/2.7 (side 37/54). □

Eksempel 4.1: Etpunktsfordeling
Hvis X er udartet i a, så er dens frembringende funktion $G(s) = s^a$.

Eksempel 4.2: 01-variabel
Hvis X er en 01-variabel med $\mathrm{P}(X = 1) = p$, så er dens frembringende funktion $G(s) = 1 - p + sp$.

Eksempel 4.3: Binomialfordelingen
Binomialfordelingen med parametre n og p er fordelingen af en sum af n uafhængige identisk fordelte 01-variable med parameter p (definition 1.10 side 32). Ifølge sætning 4.1 og eksempel 4.2 er den frembringende funktion for en binomialfordelt stokastisk variabel Y med parametre n og p derfor $G(s) = (1 - p + sp)^n$.

I eksempel 1.20 side 43 fandt vi binomialfordelingens middelværdi og varians til np og $np(1-p)$. Nu prøver vi at finde disse størrelser ud fra den frembringende funktion. Vi har

$$G'(s) = np(1-p+sp)^{n-1} \qquad \text{og} \qquad G''(s) = n(n-1)p^2(1-p+sp)^{n-2},$$

så

$$G'(1) = \mathrm{E}\,Y = np \qquad \text{og} \qquad G''(1) = \mathrm{E}(Y(Y-1)) = n(n-1)p^2.$$

Ifølge (4.2) og (4.3) er da

$$\mathrm{E}\,Y = np \qquad \text{og} \qquad \mathrm{Var}\,Y = n(n-1)p^2 - np(np-1) = np(1-p).$$

Vi kan også få et nyt bevis for sætning 1.12 (side 33): Ifølge sætning 4.1 er den frembringende funktion for summen af to uafhængige binomialfordelte variable med parametre n_1 og p hhv. n_2 og p,

$$(1-p+sp)^{n_1}\cdot(1-p+sp)^{n_2} = (1-p+sp)^{n_1+n_2}$$

og højresiden ses at være den frembringende funktion for binomialfordelingen med parametre $n_1 + n_2$ og p.

Eksempel 4.4: Poissonfordelingen
Poissonfordelingen med parameter μ (definition 2.10 side 60) har frembringende funktion

$$G(s) = \sum_{x=0}^{\infty} s^x \frac{\mu^x}{x!}\exp(-\mu) = \exp(-\mu)\sum_{x=0}^{\infty}\frac{(\mu s)^x}{x!} = \exp\Big(\mu(s-1)\Big).$$

På samme måde som for binomialfordelingen kan vi ved hjælp af frembringende funktioner uhyre let vise resultatet i sætning 2.16 (side 61) om fordelingen af en sum af to uafhængige poissonfordelte variable.

Eksempel 4.5: Negativ binomialfordeling
Den frembringende funktion for en negativt binomialfordelt stokastisk variabel X (definition 2.9 side 58) med sandsynlighedsparameter $p \in]0\,;1[$ og formparameter $k > 0$ er

$$G(s) = \sum_{x=0}^{\infty}\binom{x+k-1}{x}p^k(1-p)^x s^x = p^k\sum_{x=0}^{\infty}\binom{x+k-1}{x}\Big((1-p)s\Big)^x = \left(\frac{1}{p} - \frac{1-p}{p}s\right)^{-k}.$$

(Til det sidste lighedstegn har vi benyttet nr. 3 af binomialrækkerne på side 58.) Så er

$$G'(s) = k\frac{1-p}{p}\left(\frac{1}{p} - \frac{1-p}{p}s\right)^{-k-1} \quad \text{og} \quad G''(s) = k(k+1)\left(\frac{1-p}{p}\right)^2\left(\frac{1}{p} - \frac{1-p}{p}s\right)^{-k-2}$$

Det sætter vi ind i (4.2) og (4.3) og får

$$\mathrm{E}\,X = k\frac{1-p}{p} \qquad \text{og} \qquad \mathrm{Var}\,X = k\frac{1-p}{p^2}$$

som påstået side 59.

Ligeledes kan man ved hjælp af frembringende funktioner helt uden videre vise sætning 2.14 side 58 om fordelingen af en sum af negativt binomialfordelte variable.

Som det fremgår af eksemplerne, er det ofte sådan at når man først har fundet den frembringende funktion for en fordeling (og det kan undertiden godt være besværligt), så er der mange ting der går uhyre nemt, eksempelvis kan man altså finde middelværdi og varians blot ved at differentiere et par gange og foretage en simpel udregning.

Vi har tidligere set nogle eksempler på konvergens af sandsynlighedsfordelinger: binomialfordelingen konvergerer under visse omstændigheder mod en poissonfordeling (sætning 2.15 side 60), og den negative binomialfordeling konvergerer ligeledes under visse omstændigheder mod en poissonfordeling (opgave 2.8 side 62). Konvergens af sandsynlighedsfordelinger på $\mathbb{N}_0$ hænger sammen med konvergens af frembringende funktioner:

Sætning 4.2: Kontinuitetssætningen
Antag at vi for hvert $n \in \{1,2,3,\ldots,\infty\}$ har en stokastisk variabel X_n med frembringende funktion G_n og sandsynlighedsfunktion f_n. Da gælder at $\lim_{n\to\infty} f_n(x) = f_\infty(x)$ for alle $x \in \mathbb{N}_0$ hvis og kun hvis $\lim_{n\to\infty} G_n(s) = G_\infty(s)$ for alle $s \in [0\,;1[$.

Beviset for sætningen udelades (det er mere en øvelse i matematisk analyse end i sandsynlighedsregning).

Nu vil vi gå over til at undersøge nogle nye problemer, dvs. problemer der ikke er behandlet tidligere i denne fremstilling.

4.2 Sum af et stokastisk antal stokastiske variable

Sætning 4.3
Antag at $N, X_1, X_2, \ldots$ er indbyrdes uafhængige stokastiske variable, og at alle X-erne har samme fordeling. Sæt $Y = X_1 + X_2 + \ldots + X_N$. Da gælder at den frembringende funktion G_Y for Y er

$$G_Y = G_N \circ G_X\,, \tag{4.4}$$

hvor G_N er den frembringende funktion for N og G_X den frembringende funktion for X-ernes fordeling.

Bevis
For de udfald $\omega \in \Omega$ for hvilke $N(\omega) = n$, er $Y(\omega) = X_1(\omega) + X_2(\omega) + \cdots + X_n(\omega)$,

dvs. $\{Y = y \text{ og } N = n\} = \{X_1 + X_2 + \cdots + X_n = y \text{ og } N = n\}$, så derfor er

$$\begin{aligned} G_Y(s) &= \sum_{y=0}^{\infty} s^y \,\mathrm{P}(Y = y) = \sum_{y=0}^{\infty} s^y \sum_{n=0}^{\infty} \mathrm{P}(Y = y \text{ og } N = n) \\ &= \sum_{y=0}^{\infty} s^y \sum_{n=0}^{\infty} \mathrm{P}\big(X_1 + X_2 + \cdots + X_n = y\big)\,\mathrm{P}(N = n) \\ &= \sum_{n=0}^{\infty} \Bigg(\sum_{y=0}^{\infty} s^y \,\mathrm{P}\big(X_1 + X_2 + \cdots + X_n = y\big) \Bigg) \mathrm{P}(N = n), \end{aligned}$$

hvor vi undervejs har benyttet at N er uafhængig af $X_1 + X_2 + \cdots + X_n$. Udtrykket i den store parentes er den frembringende funktion for $X_1 + X_2 + \cdots + X_n$, og den er ifølge sætning 4.1 lig $\big(G_X(s)\big)^n$. Det indsætter vi og får

$$G_Y(s) = \sum_{n=0}^{\infty} \big(G_X(s)\big)^n \,\mathrm{P}(N = n).$$

Her er højresiden nu faktisk den frembringende funktion for N, udregnet i punktet $G_X(s)$, så alt i alt er vi nået frem til at for vilkårligt s er

$$G_Y(s) = G_N\big(G_X(s)\big),$$

hvilket var det der skulle vises. □

Eksempel 4.6: Mider på æbleblade
Den 18. juli 1951 udtog man tilfældigt 25 blade på hvert af seks McIntosh-æbletræer i en æbleplantage i Connecticut og talte hvor mange røde mider (voksne hunner) der var på hvert blad. Derved fik man tallene i tabel 4.1 (Bliss and Fisher, 1953).

Hvis miderne var drysset tilfældigt ud over æbletræerne, ville det formentlig give sig udslag i at antal mider pr. blad var poissonfordelt. Statistikerne kan undersøge om talmaterialet med rimelighed kan beskrives med en poissonfordeling, og svaret er at poissonfordelingen *ikke* giver en god beskrivelse. Derfor må man finde på noget andet.

Man kan argumentere for at miderne er placeret i kolonier eller klumper, så man kunne overveje en model der afspejler dette. Eksempelvis kunne man tænke sig at miderne blev placeret på bladene ved hjælp af en totrins-tilfældighedsmekanisme: først bestemmes hvor mange mideklumper der skal være på det foreliggende blad, og dernæst bestemmes hvor mange individer der skal være i hver af klumperne. Lad N være en stokastisk variabel der angiver antal klumper på bladet, og lad $X_1, X_2, \ldots$ være stokastiske variable der angiver antal individer i klump nr. $1, 2, \ldots$ Det man har observeret, er så 150 værdier af $Y = X_1 + X_2 + \ldots + X_N$ (dvs. Y er en sum af et stokastisk antal stokastiske variable).

Da det altid er en god idé at forsøge sig med simple modeller før man kaster sig ud i de indviklede, vil vi antage at de stokastiske variable $N, X_1, X_2, \ldots$ er uafhængige,

Tabel 4.1
Mider på æbleblade

y	antal blade med y mider
0	70
1	38
2	17
3	10
4	9
5	3
6	2
7	1
8+	0
	150

og at alle X-erne har samme fordeling. – Herefter er de eneste uafklarede elementer i modellen fordelingen af N og fordelingen af X-erne.

Den først foreslåede model at antal mider på et blad er poissonfordelt, dur som nævnt ikke, men måske så i stedet antal klumper pr. blad er poissonfordelt? Lad os antage at N er poissonfordelt med parameter $\mu > 0$. Antal individer pr. klump modellerer vi med en *logaritmisk fordeling* med parameter $\alpha \in]0\,;1[$, dvs. fordelingen med punktsandsynligheder

$$f(x) = \frac{1}{-\ln(1-\alpha)}\,\frac{\alpha^x}{x}, \qquad x = 1, 2, 3, \ldots$$

(der er nødvendigvis et positivt antal mider i en klump). – Det er ikke særlig klart hvorfor man lige vælger denne fordeling, men der kommer et pænt slutresultat ud af det.

Sætning 4.3 fortæller hvordan man udtrykker den frembringende funktion for Y ved hjælp af de frembringende funktioner for N og X. Den frembringende funktion for N fandt vi i eksempel 4.4 til $G_N(s) = \exp\bigl(\mu(s-1)\bigr)$. Den frembringende funktion for X-ernes fordeling er

$$G_X(s) = \sum_{x=1}^{\infty} s^x f(x) = \frac{1}{-\ln(1-\alpha)} \sum_{x=1}^{\infty} \frac{(\alpha s)^x}{x} = \frac{-\ln(1-\alpha s)}{-\ln(1-\alpha)}.$$

Den frembringende funktion for Y er ifølge sætning 4.3 $G_Y(s) = G_N\bigl(G_X(s)\bigr)$, og vi får derfor

$$\begin{aligned} G_Y(s) &= \exp\left(\mu\left(\frac{-\ln(1-\alpha s)}{-\ln(1-\alpha)} - 1\right)\right) \\ &= \exp\left(\frac{\mu}{\ln(1-\alpha)} \ln\frac{1-\alpha s}{1-\alpha}\right) = \left(\frac{1-\alpha s}{1-\alpha}\right)^{\mu/\ln(1-\alpha)} \\ &= \left(\frac{1}{1-\alpha} - \frac{\alpha}{1-\alpha}s\right)^{\mu/\ln(1-\alpha)} = \left(\frac{1}{p} - \frac{1-p}{p}s\right)^{-k} \end{aligned}$$

hvor $p = 1-\alpha$ og $k = -\mu/\ln p$. Denne funktion genkender vi som den frembringende funktion for den negative binomialfordeling med sandsynlighedsparameter p og formparameter k (eksempel 4.5). Antallet Y af mider på et blad er således negativt binomialfordelt.

Korollar 4.4
I den i sætning 4.3 beskrevne situation er

$$\begin{aligned} \mathrm{E}\,Y &= \mathrm{E}\,N\,\mathrm{E}\,X \\ \mathrm{Var}\,Y &= \mathrm{E}\,N\,\mathrm{Var}\,X + \mathrm{Var}\,N\,(\mathrm{E}\,X)^2. \end{aligned}$$

Bevis
Formlerne (4.2) og (4.3) angiver sammenhængen mellem på den ene side den frembringende funktion og på den anden side fordelingens middelværdi og varians. Hvis man differentierer (4.4), får man at $G'_Y = (G'_N \circ G_X)\,G'_X$ der udregnet i punktet 1 giver $\mathrm{E}\,Y = \mathrm{E}\,N\,\mathrm{E}\,X$. Tilsvarende kan man udregne G''_Y i punktet 1 og efter nogle omskrivninger nå frem til det påstående resultat. □

4.3 Forgreningsprocesser

En forgreningsproces er en særlig slags stokastisk proces. Generelt er en stokastisk proces en familie $(X(t) : t \in T)$ af stokastiske variable, indiceret ved en størrelse t som man plejer at kalde for tiden, og som varierer i en mængde T der ofte er $[0\,;+\infty[$ (»kontinuert tid«) eller $\mathbb{N}_0$ (»diskret tid«). Værdimængden for de stokastiske variable er normalt en delmængde af enten $\mathbb{Z}$ (»diskret tilstandsrum« eller $\mathbb{R}$ (»kontinuert tilstandsrum«).

En forgreningsproces med diskret tid og diskret tilstandsrum kan nu præsenteres på følgende måde. Lad os sige at vi har at gøre med en særlig slags individer der alle har levetid 1 tidsenhed; når individets berammede levetid er udløbet, vil det enten dø eller spaltes i et stokastisk antal nye individer; de nye individer er af samme slags som deres stamfar, dvs. hvert af dem lever 1 tidsenhed hvorefter det enten dør eller spaltes i et antal nye, osv. Individerne dør/formerer sig stokastisk uafhængigt af hverandre, og den sandsynlighedsfordeling der bestemmer hvor mange efterkommere et individ får, er den samme for alle individer og til alle tider. Hvis vi lader $Y(t)$ betegne det samlede antal individer til tid t (opgjort efter at de dødsfald og spaltninger der skal ske til tid t, er sket), så er $(Y(t) : t \in \mathbb{N}_0)$ et eksempel på en forgreningsproces.

De spørgsmål man stiller til en forgreningsproces, er blandt andet: givet at $Y(0) = y_0$, hvad kan man så sige om fordelingen af $Y(t)$? hvor stor er middelværdien og variansen af $Y(t)$? hvor stor er sandsynligheden for at $Y(t) = 0$, dvs.

Om notationen
I de første kapitler gjorde vi meget ud af at stokastiske variable er funktioner defineret på et udfaldsrum Ω; efterhånden forsvandt både ω-erne og Ω ud af billedet. Når der nu står $Y(t)$, er det så fordi t har overtaget ω's plads? Nej, der er stadig et underforstået Ω, og det ville være mere korrekt at skrive $Y(\omega, t)$ og ikke bare $Y(t)$. Meningen er, at for hvert t har vi en almindelig stokastisk variabel $\omega \mapsto Y(\omega, t)$, og for hvert ω har vi en funktion $t \mapsto Y(\omega, t)$ af »tiden« t (som antager heltalsværdier).

populationen er uddød til tid t? hvor lang tid går der inden populationen uddør (hvis den uddør)?

Vi vil opstille en matematisk model for en forgreningsproces og besvare nogle af spørgsmålene.

Den såkaldte *afkomstfordeling* spiller en afgørende rolle i modellen. Afkomstfordelingen er den sandsynlighedsfordeling der modellerer antal efterkommere som det enkelte individ får når der går 1 tidsenhed. Afkomstfordelingen er en fordeling på $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$; værdien 0 svarer til at individet dør uden at få efterkommere, værdien 1 svarer til at individet dør men får 1 efterkommer (eller at individet lever videre), værdien 2 svarer til at individet dør men får 2 efterkommere, osv.

Lad G være afkomstfordelingens frembringende funktion. Processen startes til tid $t = 0$ med ét individ. Til tid $t = 1$ bliver dette individ til $Y(1)$ nye, og den frembringende funktion for $Y(1)$ er

$$s \longmapsto \mathrm{E}\left(s^{Y(1)}\right) = G(s).$$

Til tid $t = 2$ bliver hver af de $Y(1)$ individer til et stokastisk antal nye, så $Y(2)$ er en sum af $Y(1)$ uafhængige stokastiske variable, hver med frembringende funktion G, og ifølge sætning 4.3 er den frembringende funktion for $Y(2)$ da

$$s \longmapsto \mathrm{E}\left(s^{Y(2)}\right) = (G \circ G)(s).$$

Til tid $t = 3$ bliver hver af de $Y(2)$ individer til et stokastisk antal nye, så $Y(3)$ er en sum af $Y(2)$ uafhængige stokastiske variable, hver med frembringende funktion G, og den frembringende funktion for $Y(3)$ er da ifølge sætning 4.3

$$s \longmapsto \mathrm{E}\left(s^{Y(3)}\right) = (G \circ G \circ G)(s).$$

Fortsættes ræsonnementet, får man at den frembringende funktion for $Y(t)$ er

$$s \longmapsto \mathrm{E}\left(s^{Y(t)}\right) = (\underbrace{G \circ G \circ \ldots \circ G}_{t})(s). \tag{4.5}$$

(Bemærk i øvrigt at hvis vi var startet med y_0 individer til tid 0, så var den frembringende funktion blevet $\left((G \circ G \circ \ldots \circ G)(s)\right)^{y_0}$; det er en konsekvens af sætning 4.1.)

Lad os kalde middelværdien i afkomstfordelingen for μ; vi minder om at $\mu = G'(1)$ (formel (4.2)). Da formel (4.5) som nævnt bare er en anvendelse af sætning 4.3, kan vi anvende korollaret til denne sætning og få det ikke voldsomt overraskende resultat at $\mathrm{E}\,Y(t) = \mu^t$, dvs. i middel er der eksponentiel vækst (eller uddøen). Mere interessant er

Sætning 4.5
Antag at $Y(0) = 1$ og at afkomstfordelingen ikke er etpunktsfordelingen i 1.
Så gælder:

- *Hvis $\mu \le 1$, vil populationen uddø med sandsynlighed 1, nærmere bestemt er*

$$\lim_{t\to\infty} \mathrm{P}(Y(t) = y) = \begin{cases} 1 & \text{for } y = 0 \\ 0 & \text{for } y = 1, 2, 3, \ldots \end{cases}$$

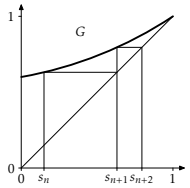
- *Hvis $\mu > 1$, vil populationen uddø med sandsynlighed $s^\star$ og vokse ud over alle grænser med sandsynlighed $1 - s^\star$, nærmere bestemt er*

$$\lim_{t\to\infty} \mathrm{P}(Y(t) = y) = \begin{cases} s^\star & \text{for } y = 0 \\ 0 & \text{for } y = 1, 2, 3, \ldots \end{cases}$$

Her betegner $s^\star$ den løsning til ligningen $G(s) = s$ som er mindre end 1.

Bevis
Vi vil vise sætningen ved at vise konvergens af den frembringende funktion for $Y(t)$: for hvert givet fast s_0 vil vi finde grænseværdien af $\mathrm{E}\,s_0^{Y(t)}$ når $t \to \infty$. Ifølge formel (4.5) er talfølgen $\left(\mathrm{E}\,s_0^{Y(t)}\right)_{t\in\mathbb{N}}$ den samme som den talfølge $(s_n)_{n\in\mathbb{N}}$ der er fastlagt ved

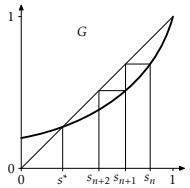
$$s_{n+1} = G(s_n), \quad n = 0, 1, 2, 3, \ldots$$



Tilfældet $\mu \le 1$

Denne talfølges opførsel afhænger af udseendet af G og af startværdien s_0. På intervallet $]0\,;1[$ er G og alle dens afledede som tidligere nævnt ikke-negative, og desuden er $G(1) = 1$ og $G'(1) = \mu$. Da vi pr. forudsætning har udelukket muligheden $G(s) = s$ for alle s (svarende til at afkomstfordelingen er etpunktsfordelingen i værdien 1), kan man konkludere at

- Hvis $\mu \le 1$, så er $G(s) > s$ for alle $s \in]0\,;1[$.
  I så fald er talfølgen (s_n) voksende og således konvergent. Kald grænseværdien $\bar{s}$; så er $\bar{s} = \lim_{n\to\infty} s_n = \lim_{n\to\infty} G(s_{n+1}) = G(\bar{s})$, hvor det midterste lighedstegn følger af definitionen på (s_n) og det sidste lighedstegn følger af at G er kontinuert i $\bar{s}$. Det eneste punkt $\bar{s} \in [0\,;1]$ for hvilket $G(\bar{s}) = \bar{s}$, er $\bar{s} = 1$, så vi er nået frem til at følgen (s_n) er voksende og har grænseværdi 1.
- Hvis $\mu > 1$, så findes præcis et tal $s^\star \in [0\,;1[$ således at $G(s^\star) = s^\star$.
  - Hvis $s^\star < s < 1$, så er $s^\star \le G(s) < s$.
    Deraf følger at hvis $s^\star < s_0 < 1$, så er talfølgen (s_n) aftagende og derfor konvergent med en grænseværdi $\bar{s} \ge s^\star$. På samme måde som i tilfældet $\mu \le 1$ indses at $G(\bar{s}) = \bar{s}$, og man kan derfor konkludere at grænseværdien faktisk er $s^\star$.
  - Hvis $0 < s < s^\star$, så er $s < G(s) \le s^\star$.



Tilfældet $\mu > 1$

Deraf følger at hvis $0 < s_0 < s^\star$, så er talfølgen (s_n) voksende, og dens grænseværdi må (med et lignende argument som i forrige punkt) være $s^\star$.

– Hvis s_0 er enten $s^\star$ eller 1, så er talfølgen (s_n) konstant lig s_0.

I tilfældet $\mu \le 1$ viser ovenstående betragtninger at den frembringende funktion for $Y(t)$ konvergerer mod den frembringende funktion for etpunktsfordelingen i punktet 0 (nemlig den konstante funktion 1), dvs. populationen uddør med sandsynlighed 1.

I tilfældet $\mu > 1$ viser ovenstående betragtninger at den frembringende funktion for $Y(t)$ konvergerer mod den funktion som har værdien $s^\star$ på intervallet $[0\,;1[$ og værdien 1 i punktet 1. Denne funktion er *ikke* frembringende funktion for en almindelig sandsynlighedsfordeling (så skulle funktionen have været kontinuert i 1), men med lidt god vilje kan man sige at den er frembringende funktion for en sandsynlighedsfordeling der placerer sandsynlighedsmassen $s^\star$ i punktet 0 og sandsynlighedsmassen $1 - s^\star$ i et uendelig fjernt punkt.

Man kan faktisk godt udvide teorien for frembringende funktioner til at omfatte fordelinger der placerer noget af sandsynlighedsmassen i et uendelig fjernt punkt, men dette kræver en nærmere redegørelse (som ikke vil blive bragt her); i stedet nøjes vi med at vise følgende to påstande:

$$\lim_{t\to\infty} \mathrm{P}(Y(t) = 0) = s^\star \tag{4.6}$$

$$\lim_{t\to\infty} \mathrm{P}(Y(t) \le c) = s^\star \quad \text{for alle } c > 0. \tag{4.7}$$

Da $\mathrm{P}(Y(t) = 0)$ er lig værdien af den frembringende funktion for $Y(t)$ udregnet i $s = 0$, følger (4.6) uden videre af det foregående. For at vise anden del foretager vi nogle omskrivninger hvor vi blandt andet benytter Markovs ulighed (lemma 1.24 side 39):

$$\mathrm{P}(Y(t) \le c) = \mathrm{P}\left(s^{Y(t)} \ge s^c\right) \le s^{-c}\,\mathrm{E}\,s^{Y(t)}$$

hvor $0 < s < 1$. For $t \to \infty$ går $s^{-c}\,\mathrm{E}\,s^{Y(t)}$ ifølge det tidligere viste mod $s^{-c}s^\star$; ved at vælge s tilstrækkelig tæt på 1, kan $s^{-c}s^\star$ komme lige så tæt på $s^\star$ som vi ønsker, så $\limsup_{t\to\infty} \mathrm{P}(Y(t) \le c) \le s^\star$. På den anden side er $\mathrm{P}(Y(t) = 0) \le \mathrm{P}(Y(t) \le c)$ og $\lim_{t\to\infty} \mathrm{P}(Y(t) = 0) = s^\star$, hvoraf $s^\star \le \liminf_{t\to\infty} \mathrm{P}(Y(t) \le c)$, så alt i alt eksisterer grænseværdien $\lim_{t\to\infty} \mathrm{P}(Y(t) \le c)$ og er lig $s^\star$.

Formlerne (4.6) og (4.7) viser at uanset hvor stort et interval $[0\,;c]$ vi vælger, så vil det i grænsen når $t \to \infty$ ikke indeholde andet sandsynlighedsmasse end den der befinder sig i punktet $y = 0$. □

4.4 Opgaver

Opgave 4.1
Den kosmiske stråling antages – i hvert fald i denne opgave – at optræde i form af partikler der ankommer »helt tilfældigt« til det måleapparat der skal registrere dem. At de kommer »helt tilfældigt«, betyder at antal partikler (med en given energi) der ankommer i et tidsrum af længde t, er poissonfordelt med parameter λt, hvor λ er en strålingsintensitet.

Lad os nu sige at det apparat der skal registrere partiklerne, er lettere defekt således at de enkelte partikler ikke med sikkerhed registreres, men at der er en vis konstant sandsynlighed for at partiklen ikke registreres. Hvad kan man sige om fordelingen af antal registrerede partikler?

Opgave 4.2
Benyt sætning 4.2 til at vise at binomialfordelingen med parametre n og p konvergerer mod poissonfordelingen med parameter μ når $n \to \infty$ og $p \to 0$ på en sådan måde at $np \to \mu$. (Dette er vist på en anden måde i sætning 2.15 side 60.)

Opgave 4.3
Benyt sætning 4.2 til at vise at den negative binomialfordeling med parametre k og p konvergerer mod poissonfordelingen med parameter μ når $k \to \infty$ og $p \to 1$ på en sådan måde at $k(1-p)/p \to \mu$. (Dette var også genstand for behandling i opgave 2.8 side 62.)

Tip:
Hvis (z_n) er en (reel eller kompleks) talfølge som konvergerer mod det endelige tal z, så er

$$\lim_{n\to\infty}\left(1 + \frac{z_n}{n}\right)^n = \exp(z).$$

Opgave 4.4
I sætning 4.5 antages det at processen starter med ét individ. Hvad vil der gælde hvis man i stedet starter med y_0 individer?

Opgave 4.5
Betragt en forgreningsproces hvor afkomstfordelingen placerer sandsynlighederne p_0, p_1 og p_2 på tallene 0, 1 og 2 ($p_0 + p_1 + p_2 = 1$), dvs. når der går et tidsskridt, bliver hvert individ til 0, 1 eller 2 med de nævnte sandsynligheder.

Hvordan afhænger sandsynligheden for at processen før eller senere uddør, af p_0, p_1 og p_2?

5 Generel teori

Dette kapitel åbner en smal sprække ind til teorien for sandsynlighedsmål på generelle udfaldsrum.

Andrej Kolmogorov
russisk matematiker (1903-87).
I 1933 udkom hans *Grundbegriffe der Wahrscheinlichkeitsrechnung* som er den første tilfredsstillende aksiomatiske fremstilling af sandsynlighedsregningen.

Definition 5.1: σ-algebra
Lad Ω være en ikke-tom mængde. En mængde $\mathscr{F}$ af delmængder af Ω kaldes en σ-algebra på Ω hvis der gælder:

1. $\Omega \in \mathscr{F}$.
2. *$\mathscr{F}$ er afsluttet over for komplementærmængdedannelse, dvs: hvis $A \in \mathscr{F}$, så er $A^c \in \mathscr{F}$.*
3. *$\mathscr{F}$ er afsluttet over for tællelige foreningsmængdedannelser, dvs: hvis $A_1, A_2, \ldots$ er en følge i $\mathscr{F}$, så er foreningsmængden $\bigcup_{n=1}^{\infty} A_n$ også i $\mathscr{F}$.*

Bemærkninger: Lad $\mathscr{F}$ være en σ-algebra. Da $\varnothing = \Omega^c$, følger det af 1 og 2 at $\varnothing \in \mathscr{F}$. Da $\bigcap_{n=1}^{\infty} A_n = \left(\bigcup_{n=1}^{\infty} A_n^c\right)^c$, følger det af 2 og 3 at $\mathscr{F}$ også er afsluttet over for tællelige fællesmængdedannelser.

Émile Borel
fransk matematiker (1871-1956).
Har blandt andet beskæftiget sig med målteori og med sandsynlighedsregning.

På de reelle tal $\mathbb{R}$ opererer man med en særlig σ-algebra $\mathscr{B}$ kaldet *Borel-σ-algebraen*, som er den mindste σ-algebra på $\mathbb{R}$ som indeholder alle åbne delmængder af $\mathbb{R}$; $\mathscr{B}$ er også den mindste σ-algebra på $\mathbb{R}$ som indeholder alle intervaller.

Definition 5.2: Sandsynlighedsrum
Et sandsynlighedrum er et tripel $(\Omega, \mathscr{F}, \mathrm{P})$ bestående af

1. *et udfaldsrum Ω som er en ikke-tom mængde,*
2. *en σ-algebra $\mathscr{F}$ af delmængder af Ω,*
3. *et sandsynlighedsmål på $(\Omega, \mathscr{F})$, dvs. en afbildning $\mathrm{P} : \mathscr{F} \to \mathbb{R}$ som er*
   - *positiv:* $\mathrm{P}(A) \geq 0$ *for alle* $A \in \mathscr{F}$,
   - *normeret:* $\mathrm{P}(\Omega) = 1$, *og*
   - *σ-additiv: hvis $A_1, A_2, \ldots$ er en følge af parvis disjunkte hændelser fra $\mathscr{F}$, så er* $\mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(A_i)$.

Sætning 5.1
Lad $(\Omega, \mathscr{F}, \mathrm{P})$ være et sandsynlighedsrum. Der gælder at sandsynlighedsmålet P *er*

monoton-kontinuert i den forstand at hvis $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$ er en voksende følge i $\mathscr{F}$, så er $\bigcup_{n=1}^{\infty} A_n \in \mathscr{F}$, og $\lim_{n\to\infty} \mathrm{P}(A_n) = \mathrm{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big)$; og tilsvarende gælder der at hvis $B_1 \supseteq B_2 \supseteq B_3 \supseteq \ldots$ er en aftagende følge i $\mathscr{F}$, så er $\bigcap_{n=1}^{\infty} B_n \in \mathscr{F}$, og $\lim_{n\to\infty} \mathrm{P}(B_n) = \mathrm{P}\Big(\bigcap_{n=1}^{\infty} B_n\Big)$.

Bevis
På grund af den endelige additivitet er

$$\begin{aligned}\mathrm{P}(A_n) &= \mathrm{P}\Big((A_n \setminus A_{n-1}) \cup (A_{n-1} \setminus A_{n-2}) \cup \ldots \cup (A_2 \setminus A_1) \cup (A_1 \setminus \varnothing)\Big) \\ &= \mathrm{P}(A_n \setminus A_{n-1}) + \mathrm{P}(A_{n-1} \setminus A_{n-2}) + \ldots + \mathrm{P}(A_2 \setminus A_1) + \mathrm{P}(A_1 \setminus \varnothing).\end{aligned}$$

For $n \to \infty$ fås derfor (idet vi sætter $A_0 = \varnothing$)

$$\lim_{n\to\infty} \mathrm{P}(A_n) = \sum_{n=1}^{\infty} \mathrm{P}(A_n \setminus A_{n-1}) = \mathrm{P}\Big(\bigcup_{n=1}^{\infty} (A_n \setminus A_{n-1})\Big) = \mathrm{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big),$$

hvor det midterste lighedstegn følger af at $\mathscr{F}$ er en σ-algebra og P er σ-additiv. Hermed er sætningens første påstand vist. Den anden påstand følger ved at se på den voksende følge $A_n = B_n^{\mathrm{c}}$. □

Definitionen af stokastisk variabel ser nu sådan ud (slgn. definition 1.6 side 24):

Definition 5.3: Stokastisk variabel
En stokastisk variabel på $(\Omega, \mathscr{F}, \mathrm{P})$ er en afbildning X af Ω ind i $\mathbb{R}$ med den egenskab at $\{X \in B\} \in \mathscr{F}$ for ethvert $B \in \mathscr{B}$.

Bemærkninger:

1. Betingelsen $\{X \in B\} \in \mathscr{F}$ sikrer at $\mathrm{P}(X \in B)$ er en meningsfuld størrelse.
2. Undertiden skriver man $X^{-1}(B)$ i stedet for $\{X \in B\}$ (der jo er en kort betegnelse for $\{\omega \in \Omega : X(\omega) \in B\}$), og man opfatter X^{-1} som en afbildning der afbilder delmængder af $\mathbb{R}$ over i delmængder af Ω. De to krav der stilles til X for at det er en stokastisk variabel, er da at X afbilder Ω over i $\mathbb{R}$ og X^{-1} afbilder $\mathscr{B}$ over i $\mathscr{F}$. Man taler om at $X : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$ er en *målelig afbildning*.

Fordelingsfunktion defineres ganske som tidligere (definition 1.7 side 25):

Definition 5.4: Fordelingsfunktion
Fordelingsfunktionen for en stokastisk variabel X er funktionen

$$\begin{aligned} F : \mathbb{R} &\longrightarrow [0\,;1] \\ x &\longmapsto \mathrm{P}(X \le x)\end{aligned}$$

Der gælder de samme resultater som vi tidligere har set:

Lemma 5.2
Hvis den stokastiske variabel X har fordelingsfunktion F, så er

$$\begin{aligned}\mathrm{P}(X \le x) &= F(x), \\ \mathrm{P}(X > x) &= 1 - F(x), \\ \mathrm{P}(a < X \le b) &= F(b) - F(a),\end{aligned}$$

for vilkårlige reelle tal x og $a < b$.

Sætning 5.3
Fordelingsfunktionen F for en stokastisk variabel X har følgende egenskaber:

1. *Den er ikke-aftagende, dvs. hvis $x \le y$, så er $F(x) \le F(y)$.*
2. *$\lim_{x\to-\infty} F(x) = 0$ og $\lim_{x\to+\infty} F(x) = 1$.*
3. *Den er højrekontinuert, dvs. $F(x+) = F(x)$ for alle x.*
4. *I ethvert punkt x gælder $\mathrm{P}(X = x) = F(x) - F(x-)$.*
5. *Et punkt x er et diskontinuitetspunkt for F hvis og kun hvis $\mathrm{P}(X = x) > 0$.*

Bevis
De enkelte punkter i sætningen vises således:
Ad 1: Det gamle bevis fra sætning 1.6 kan overføres.
Ad 2: Mængderne $A_n = \{X \in]-n\,;n]\}$ vokser op mod hele Ω, så derfor er ifølge sætning 5.1 $\lim_{n\to\infty}\big(F(n)-F(-n)\big) = \lim_{n\to\infty} \mathrm{P}(A_n) = \mathrm{P}(X \in \Omega) = 1$. Dette i forening med at F er ikke-aftagende og har værdier mellem 0 og 1, medfører påstand 2.
Ad 3: Mængderne $\{X \le x + \frac{1}{n}\}$ aftager mod $\{X \le x\}$. Derfor er $\lim_{n\to\infty} F(x + \frac{1}{n}) = \lim_{n\to\infty} \mathrm{P}(X \le x + \frac{1}{n}) = \mathrm{P}\Big(\bigcap_{n=1}^{\infty} \{X \le x + \frac{1}{n}\}\Big) = \mathrm{P}(X \le x) = F(x)$ ifølge sætning 5.1.
Ad 4: Der gælder at $\mathrm{P}(X = x) = \mathrm{P}(X \le x) - \mathrm{P}(X < x) = F(x) - \mathrm{P}\Big(\bigcup_{n=1}^{\infty} \{X \le x - \frac{1}{n}\}\Big) = F(x) - \lim_{n\to\infty} \mathrm{P}(X \le x - \frac{1}{n}) = F(x) - \lim_{n\to\infty} F(x - \frac{1}{n}) = F(x) - F(x-)$ ifølge sætning 5.1.
Ad 5: Følger af det foregående. □

Der gælder også den »omvendte« sætning, som vi dog ikke vil bevise her:

Sætning 5.4
Hvis der er givet en funktion $F : \mathbb{R} \to [0\,;1]$ som er ikke-aftagende og højrekontinuert, og $\lim_{x\to-\infty} F(x) = 0$ og $\lim_{x\to+\infty} F(x) = 1$, så er F fordelingsfunktion for en stokastisk variabel.

5.1 Hvorfor generalisere og aksiomatisere?

1. Én grund til at søge at lave en samlet aksiomatisk fremstilling af sandsynlighedsregningen er at det fra et matematisk-æstetisk synspunkt er noget sjusk hvis man er nødt til at behandle diskrete og kontinuerte sandsynligheder som to helt forskellige typer objekter (sådan som vi har gjort i nærværende fremstilling) – og hvad stiller man op med fordelinger der f.eks. er en blanding af en diskret og en kontinuert fordeling?

Eksempel 5.1
Antag at man vil modellere levetiden T af nogle dimser der er således beskafne at de enten går i stykker med det samme (ɔ: $T = 0$) hvilket sker med en sandsynlighed som vi kan kalde p, eller også holder de et eksponentialfordelt stykke tid; fordelingen af T kan da specificeres ved at sige at $\mathrm{P}(T = 0) = p$ og $\mathrm{P}(T > t \mid T > 0) = \exp(-t/\beta)$, $t > 0$ hvor β er eksponentialfordelingens parameter.

Fordelingen af T har en diskret komponent (sandsynlighedsmassen p placeret i 0) og en kontinuert komponent (sandsynlighedsmassen $1 - p$ smurt ud på den positive halvakse efter en eksponentialfordeling).

2. Man vil gerne have en matematisk ramme der gør det muligt at tale om *konvergens af fordelinger*. Eksempelvis er det »umiddelbart indlysende« at den diskrete ligefordeling på mængden $\{0, \frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \ldots, \frac{n-1}{n}, 1\}$ konvergerer mod den kontinuerte ligefordeling på $[0\,;1]$ når $n \to \infty$, og at normalfordelingen med middelværdi μ og varians $\sigma^2 > 0$ konvergerer mod etpunktsfordelingen i μ når $\sigma^2 \to 0$. Ligeledes fortæller den centrale grænseværdisætning (side 76) at med passende skaleringer konvergerer fordelingen af summer af uafhængige variable mod en normalfordeling. Teorien skal være i stand til at præcisere hvad det er for et konvergensbegreb der er i spil her.

3. Vi har i simple situationer set hvordan man modellerer sammensatte forsøg med uafhængige komponenter (se bl.a. side 21f og side 29), men hvordan gør man det generelt, og kan det lade sig gøre med uendelig mange komponenter?

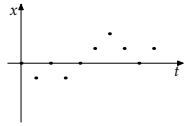
Eksempel 5.2
Store Tals Stærke Lov siger at hvis (X_n) er en følge af uafhængige identisk fordelte stokastiske variable med middelværdi μ, så er $\mathrm{P}\bigl(\lim_{n\to\infty} \overline{X}_n = \mu\bigr) = 1$ hvor $\overline{X}_n$ er gennemsnittet af $X_1, X_2, \ldots, X_n$ (se evt. også side 43). Hændelsen $\{\lim_{n\to\infty} \overline{X}_n = \mu\}$ vedrører uendelig mange

X-er, og man kan spørge om der overhovedet findes et sandsynlighedsrum hvor det er muligt at have uendelig mange stokastiske variable med en ønsket fordeling.

Eksempel 5.3
I afsnittene om den geometriske fordeling (side 56) og den negative binomialfordeling (side 58) ser man på hvor mange gange man skal gentage et 01-forsøg, inden der for første (eller for k-te) gang kommer et 1. Der optræder blandt andet en stokastisk variabel T_k som betegner nummeret på den gentagelse hvor man for k-te gang får resultatet 1. Det kan være af interesse at spørge om sandsynligheden for at der før eller senere vil være kommet i alt k 1-ere, altså $\mathrm{P}(T_k < \infty)$. Hvis $X_1, X_2, X_3, \ldots$ er stokastiske variable der repræsenterer 1., 2., 3., … gentagelse af 01-forsøget, er T_k en enkel funktion af X-erne: $T_k = \inf\{t \in \mathbb{N} : X_t = k\}$. – Hvordan kan man konstruere et sandsynlighedsrum der tillader hændelser der vedrører egenskaber ved uendelige følger af stokastiske variable?

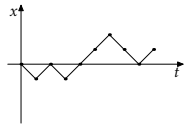
Eksempel 5.4
Tag en følge $(U(t) : t = 1, 2, 3\ldots)$ af uafhængige identisk fordelte stokastiske variable som er ligefordelte på topunktsmængden $\{-1, 1\}$, og definér derudfra en ny følge af stokastiske variable $(X(t) : t = 0, 1, 2, \ldots)$ ved

$$\begin{aligned} X(0) &= 0 \\ X(t) &= X(t-1) + U(t), \quad t = 1, 2, 3, \ldots \end{aligned}$$



dvs. $X(t) = U(1) + U(2) + \cdots + U(t)$. Den derved fremkommende stokastiske proces er en *simpel random walk* i én dimension. Traditionelt tænker man på stokastiske processer som afbildet i et koordinatsystem hvor den vandrette akse er t-aksen (tiden) og den lodrette akse x-aksen (stedet). I overensstemmelse hermed vil vi sige at den simple random walk $(X(t) : t \in \mathbb{N}_0)$ starter i $x = 0$ til tid $t = 0$, og derefter bevæger den sig et skridt op eller et skridt ned hver gang der er gået et tidsskridt.

Skridtlængderne behøver ikke være 1. Vi kan sagtens have en random walk der bevæger sig i skridt af længde Δx, og som bevæger sig til tidspunkterne $\Delta t, 2\Delta t, 3\Delta t, \ldots$ (her er $\Delta x \neq 0$ og $\Delta t > 0$). Vi tager denne gang en følge $(U(t) : t = \Delta t, 2\Delta t, 3\Delta t, \ldots)$ og sætter



$$\begin{aligned} X(0) &= 0 \\ X(t) &= X(t-\Delta t) + U(t)\Delta x, \quad t = \Delta t, 2\Delta t, 3\Delta t, \ldots \end{aligned}$$

dvs. $X(n\Delta t) = U(\Delta t)\Delta x + U(2\Delta t)\Delta x + \cdots + U(n\Delta t)\Delta x$. På denne måde har vi fået defineret $X(t)$ når t er et helt ikke-negativt multiplum af Δt (ɔ: når $t \in \mathbb{N}_0\Delta t$). Derefter kan vi i hvert delinterval $]k\Delta t\,;(k+1)\Delta t[$ definere $X(t)$ ved lineær interpolation; derved får vi alt i alt en kontinuert (stykkevis lineær) funktion $t \mapsto X(t)$, $t \geq 0$.

Man finder uden videre at $\mathrm{E}\bigl(X(n\Delta t)\bigr) = 0$ og $\mathrm{Var}\bigl(X(n\Delta t)\bigr) = n(\Delta x)^2$, dvs. hvis $t = n\Delta t$ (og dermed $n = t/\Delta t$), så er $\mathrm{Var}\, X(t) = \frac{(\Delta x)^2}{\Delta t} t$. Det antyder at hvis man har planer om at lade Δt og Δx gå mod 0, er det nok en god idé at gøre det på en måde så $(\Delta x)^2/\Delta t$ har en grænseværdi $\sigma^2 > 0$, hvorved altså $\mathrm{Var}\, X(t) \to \sigma^2 t$.

Spørgsmålet er nu hvordan man kan konstruere en matematisk ramme der gør den skitserede »grænseovergang« meningsfuld – blandt andet skal det præciseres hvad det er for nogle matematiske objekter der konvergerer, og hvad det er for en konvergens.

Løsningen bliver at man skal operere med sandsynlighedsfordelinger på rum af kontinuerte funktioner (de interpolerede random walks er alle kontinuerte funktioner af t, og grænseprocessen må også skulle være kontinuert). Hvis vi for nemheds skyld antager at $\sigma^2 = 1$, så bliver »grænseværdien« den såkaldte Wiener-proces $(W(t) : t \geq 0)$. Wiener-processen, der altid er en kontinuert funktion af t, har en del bemærkelsesværdige egenskaber:

- Hvis $0 \leq s_1 < t_1 \leq s_2 < t_2$, er tilvæksterne $W(t_1) - W(s_1)$ og $W(t_2) - W(s_2)$ uafhængige og normalfordelte med middelværdi 0 og varians hhv. $(t_1 - s_1)$ og $(t_2 - s_2)$.
- Med sandsynlighed 1 er funktionen $t \mapsto W(t)$ intetsteds differentiabel.
- Med sandsynlighed 1 vil W antage enhver reel værdi (dvs. med sandsynlighed 1 er billedmængden for $t \mapsto W(t)$ hele $\mathbb{R}$).
- For ethvert x gælder at W med sandsynlighed 1 vil passere x uendelig ofte (dvs. $\{t > 0 : W(t) = x\}$ indeholder med sandsynlighed 1 uendelig mange elementer).

Pionerværkerne i det matematiske arbejde med sådanne processer er Bachelier (1900) og Norbert Wieners arbejder fra begyndelsen af 1920-erne, se Wiener (1976, side 435-519).

Del II

Statistik

Indledning

Hvor sandsynlighedsregningen handler om at opstille og analysere sandsynlighedsmodeller for tilfældighedsfænomener (samt om at etablere det fornødne begrebsapparat), handler disciplinen *matematisk statistik* grundlæggende om at etablere og undersøge metoder til at uddrage informationer af talmaterialer der er behæftede med en usikkerhed der antages at kunne beskrives med passende sandsynlighedsfordelinger, og disciplinen *anvendt statistik* handler om hvordan disse metoder indgår i konkrete typer af modelleringsprocesser.

Den type af problemstillinger som det kommer til at handle om, kan kort skitseres således: Der foreligger et sæt tal $x_1, x_2, \ldots, x_n$ der vil blive omtalt som *observationerne* (det kunne f.eks. være resultaterne af 115 målinger af kviksølvindholdet i sværdfisk). Man opstiller en *statistisk model* gående ud på at observationerne er observerede værdier af stokastiske variable $X_1, X_2, \ldots, X_n$ der har en eller anden nærmere præciseret simultan fordeling der er kendt på nær nogle få ukendte *parametre* (f.eks. kunne X-erne være uafhængige identisk normalfordelte med de to ukendte parametre μ og σ^2). Der er herefter tre hoved-problemstillinger:

1. *Estimation*. På grundlag af observationer plus model skal der udregnes et *estimat* (dvs. et skøn eller overslag) over de ukendte parametres værdier; estimatet skal naturligvis være så godt som muligt (i en eller anden forstand der skal præciseres nærmere).
2. *Hypoteseprøvning*. I forbindelse med den faglige problemstilling kan der være forskellige interessante *statistiske hypoteser* man ønsker at teste. En statistisk hypotese er et udsagn om at parameterstrukturen er simplere end først påstået (f.eks. at visse parametre er kendte eller er ens).
3. *Modelkontrol*. Statistiske modeller er ofte ikke det fjerneste »naturtro« i den forstand at de søger at efterligne de »virkelige« mekanismer der har frembragt observationerne. Arbejdet med tilpasning og kontrol af modellen og vurdering af modellens brugbarhed får derfor et anderledes indhold og en anderledes betydning end det er tilfældet ved så mange andre typer af matematiske modeller.

-oOo-

Ronald Aylmer Fisher
engelsk statistiker og genetiker (1890-1962). Grundlæggeren af faget teoretisk statistik (i hvert fald faget i den udgave som præsenteres i denne bog).

I 1922 forklarede Fisher formålet med statistiske metoder således (Fisher (1922), her citeret efter Kotz and Johnson (1992)):

> *In order to arrive at a distinct formulation of statistical problems, it is necessary to define the task which the statistician sets himself: briefly, and in its most concrete form, the object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapabel of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.*

6 Den statistiske model

Vi vil først give en forholdsvis abstrakt præsentation af begrebet en statistisk model og nogle af de tilhørende begrebdannelser, sidenhen kommer en række illustrative eksempler.

Der foreligger en *observation* $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ som antages at være en observeret værdi af en stokastisk variabel $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ med værdier i *observationsrummet* $\mathfrak{X}$; mængden $\mathfrak{X}$ er ofte $\mathbb{R}^n$ eller $\mathbb{N}_0^n$.

For hver værdi af en *parameter* $\boldsymbol{\theta}$ tilhørende *parameterrummet* Θ har man et sandsynlighedsmål $\mathrm{P}_{\boldsymbol{\theta}}$ på $\mathfrak{X}$. Disse $\mathrm{P}_{\boldsymbol{\theta}}$-er er alle sammen kandidater til at være $\boldsymbol{X}$'s fordeling, og for (mindst) én værdi $\boldsymbol{\theta}$ er det rigtigt at $\boldsymbol{X}$'s fordeling er $\mathrm{P}_{\boldsymbol{\theta}}$.

Udsagnet »$\boldsymbol{x}$ er en observeret værdi af den stokastiske variabel $\boldsymbol{X}$, og for mindst én værdi af $\boldsymbol{\theta} \in \Theta$ er det rigtigt at fordelingen af $\boldsymbol{X}$ er $\mathrm{P}_{\boldsymbol{\theta}}$« er (en formulering af) den *statistiske model*.

Modelfunktionen er en funktion $f : \mathfrak{X} \times \Theta \to [0\,; +\infty[$ sådan at for hvert fast $\boldsymbol{\theta} \in \Theta$ er funktionen $\boldsymbol{x} \mapsto f(\boldsymbol{x}, \boldsymbol{\theta})$ den sandsynligheds(tætheds)funktion som $\boldsymbol{X}$ har hvis θ er den rigtige værdi af parameteren.

Likelihoodfunktionen svarende til $\boldsymbol{x}$ er funktionen $L : \Theta \to [0\,; +\infty[$ givet ved $L(\boldsymbol{\theta}) = f(\boldsymbol{x}, \boldsymbol{\theta})$. – Likelihoodfunktionen kommer til at spille en central rolle i forbindelse med teorien for estimation af parametre og test af statistiske hypoteser.

Hvis man kan skrive likelihoodfunktionen som $L(\boldsymbol{\theta}) = g(\boldsymbol{x})\, h(t(\boldsymbol{x}), \boldsymbol{\theta})$ for passende valgte funktioner g, h og t, så siges t at være *sufficient* (eller at give en *sufficient datareduktion*). – Sufficiensbegrebet er især interessant når t afbilder ind i et rum af meget mindre dimension end $\mathfrak{X}$, typisk $\mathbb{R}$ eller $\mathbb{R}^2$.

Bemærkninger:

1. Parametre betegnes ofte med græske bogstaver.
2. Parameterrummet Θ er normalt af meget mindre dimension end observationsrummet $\mathfrak{X}$.
3. Man vil sædvanligvis tilstræbe at parametriseringen er *injektiv*, dvs. at afbildningen $\boldsymbol{\theta} \mapsto \mathrm{P}_{\boldsymbol{\theta}}$ er injektiv.
4. Likelihoodfunktionen skal *ikke* summere eller integrere til 1.
5. Man opererer ofte med *log-likelihoodfunktionen*, dvs. logaritmen til likelihoodfunktionen; man benytter altid den naturlige logaritme.

Lidt mere notation:

1. Undertiden har vi brug for at præcisere at L er likelihoodfunktionen hørende til netop $\boldsymbol{x}$, og vi vil så skrive $L(\boldsymbol{\theta};\boldsymbol{x})$ i stedet for $L(\boldsymbol{\theta})$.
2. Symbolerne E_θ og Var_θ anvendes når der er behov for at præcisere at middelværdien hhv. variansen udregnes med hensyn til den sandsynlighedsfordeling der svarer til parameterværdien $\boldsymbol{\theta}$, altså med hensyn til P_θ.

6.1 Eksempler

Enstikprøveproblemet for 01-variable

Punkt-notationen
Hvis man har nogle indicerede værdier, f.eks. $a_1, a_2, \ldots, a_n$, bruger man som betegnelse for summen af dem det samme symbol, men med et punkt på indeksets plads:

$$a_\bullet = \sum_{i=1}^{n} a_i$$

Tilsvarende hvis der er mere end et indeks:

$$b_{i\bullet} = \sum_j b_{ij}$$

og

$$b_{\bullet j} = \sum_i b_{ij}$$

I den generelle formulering af enstikprøveproblemet for 01-variable har man en observation $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ der antages at være en observeret værdi af en stokastisk variabel $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ med værdier i $\mathfrak{X} = \{0,1\}^n$. De enkelte X-er antages at være uafhængige identisk fordelte 01-variable, og $\mathrm{P}(X_i = 1) = \theta$ hvor θ er den ukendte parameter. Parameterrummet er $\Theta = [0\,;1]$. Modelfunktionen er (jf. (1.5) side 32)

$$f(\boldsymbol{x},\theta) = \theta^{x_\bullet}(1-\theta)^{n-x_\bullet}, \quad (\boldsymbol{x},\theta) \in \mathfrak{X}\times\Theta,$$

og likelihoodfunktionen er

$$L(\theta) = \theta^{x_\bullet}(1-\theta)^{n-x_\bullet}, \quad \theta \in \Theta.$$

Som det ses, afhænger likelihoodfunktionen kun af $\boldsymbol{x}$ gennem $x_\bullet$, dvs. $x_\bullet$ er sufficient, eller mere præcist: funktionen der afbilder $\boldsymbol{x}$ over i $x_\bullet$, er sufficient.

▷ [Læs fortsættelsen side 117.]

Eksempel 6.1
Lad os sige at man har udført 7 gentagelser af et forsøg der har de to mulige udfald 0 og 1, og at man har opnået værdierne $1,1,0,1,1,0,0$. Vi vil opstille en statistisk model herfor.

Der foreligger observationen $\boldsymbol{x} = (1,1,0,1,1,0,0)$ som antages at være en værdi af en 7-dimensional stokastisk variabel $\boldsymbol{X} = (X_1, X_2, \ldots, X_7)$ med værdier i $\mathfrak{X} = \{0,1\}^7$. De enkelte X_i-er antages at være uafhængige identisk fordelte 01-variable, og $\mathrm{P}(X_i = 1) = \theta$ hvor θ er den ukendte parameter. Parameterrummet er $\Theta = [0\,;1]$. Modelfunktionen er

$$f(\boldsymbol{x},\theta) = \prod_{i=1}^{7} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{x_\bullet}(1-\theta)^{7-x_\bullet}.$$

Likelihoodfunktionen svarende til observationen $\boldsymbol{x} = (1,1,0,1,1,0,0)$ er

$$L(\theta) = \theta^4(1-\theta)^3, \quad \theta \in [0\,;1].$$

Den simple binomialfordelingsmodel

Binomialfordelingen fremkommer som fordelingen af en sum af uafhængige identisk fordelte 01-variable, så det vil næppe overraske at statistisk analyse af binomialfordelte observationer minder særdeles meget om statistisk analyse af 01-variable.

Hvis Y er binomialfordelt med (kendt) antalsparameter n og ukendt sandsynlighedsparameter $\theta \in [0\,;1]$, er modelfunktionen

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

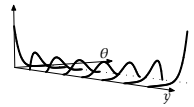
hvor $(y,\theta) \in \mathfrak{X}\times\Theta = \{0,1,2,\ldots,n\}\times[0\,;1]$, og likelihoodfunktionen er

$$L(\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \quad \theta \in [0\,;1].$$

▷ [Læs fortsættelsen side 118.]

Eksempel 6.2
Hvis man i eksempel 6.1 ikke interesserede sig for udfaldene af de syv enkeltforsøg, men kun for det samlede antal 1-er, så ville situationen være den at man havde en observation $y = 4$ af en binomialfordelt stokastisk variabel Y med antalsparameter $n = 7$ og ukendt sandsynlighedsparameter θ. Observationsrummet er $\mathfrak{X} = \{0,1,2,3,4,5,6,7\}$ og parameterrummet er $\Theta = [0\,;1]$. Modelfunktionen er

$$f(y,\theta) = \binom{7}{y}\theta^y(1-\theta)^{7-y}, \quad (y,\theta) \in \{0,1,2,3,4,5,6,7\}\times[0\,;1].$$



$f(y,\theta) = \binom{7}{y}\theta^y(1-\theta)^{7-y}$

Likelihoodfunktionen svarende til observationen $y = 4$ er

$$L(\theta) = \binom{7}{4}\theta^4(1-\theta)^3, \quad \theta \in [0\,;1].$$

Eksempel 6.3: Rismelsbiller I
I en del af et eksempel der omtales nærmere i afsnit 9.1, optræder 144 rismelsbiller (*Tribolium castaneum*) som udsættes for en bestemt dosis af insektgiften pyrethrum, hvorved 43 af dem dør i løbet af den fastsatte observationsperiode. Hvis vi går ud fra at billerne er »ens«, og at de dør eller ikke dør uafhængigt af hinanden, så kan vi tillade os at formode at antallet $y = 43$ er en observation af en binomialfordelt stokastisk variabel Y der har antalsparameter $n = 144$ og ukendt sandsynlighedsparameter θ. Den statistiske model er da givet ved modelfunktionen

$$f(y,\theta) = \binom{144}{y}\theta^y(1-\theta)^{144-y}, \quad (y,\theta) \in \{0,1,2,\ldots,144\}\times[0\,;1].$$

Likelihoodfunktionen svarende til observationen $y = 43$ er

$$L(\theta) = \binom{144}{43}\theta^{43}(1-\theta)^{144-43}, \quad \theta \in [0\,;1].$$

▷ [Eksemplet fortsætter i eksempel 7.1 side 118.]

Tabel 6.1
Skematisk opstilling ved sammenligning af binomialfordelinger

	gruppe nr.				
	1	2	3	…	s
antal gunstige	y_1	y_2	y_3	…	y_s
antal ikke-gunstige	$n_1 - y_1$	$n_2 - y_2$	$n_3 - y_3$	…	$n_s - y_s$
i alt	n_1	n_2	n_3	…	n_s

Sammenligning af binomialfordelinger

Man har observationer $y_1, y_2, \ldots, y_s$ af stokastiske variable $Y_1, Y_2, \ldots, Y_s$ der er indbyrdes uafhængige binomialfordelte således at Y_j har antalsparameter n_j (kendt) og sandsynlighedsparameter $\theta_j \in [0\,;1]$. Man kan med fordel tænke på observationerne som foreliggende i et skema som i tabel 6.1. Modelfunktionen er

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = \prod_{j=1}^{s} \binom{n_j}{y_j} \theta_j^{y_j} (1-\theta_j)^{n_j - y_j}$$

hvor parametervariablen $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_s)$ varierer i $\Theta = [0\,;1]^s$, og observationsvariablen $\boldsymbol{y} = (y_1, y_2, \ldots, y_s)$ varierer i $\mathfrak{X} = \prod_{j=1}^{s} \{0, 1, \ldots, n_j\}$. Likelihoodfunktionen og log-likelihoodfunktionen svarende til $\boldsymbol{y}$ er

$$L(\boldsymbol{\theta}) = \text{konst}_1 \cdot \prod_{j=1}^{s} \theta_j^{y_j} (1-\theta_j)^{n_j - y_j},$$

$$\ln L(\boldsymbol{\theta}) = \text{konst}_2 + \sum_{j=1}^{s} \Big(y_j \ln \theta_j + (n_j - y_j) \ln(1-\theta_j) \Big)$$

hvor konst_1 er produktet af de s binomialkoefficienter, og konst_2 er $\ln(\text{konst}_1)$.

▷ [Fortsættes side 118.]

Eksempel 6.4: Rismelsbiller II
Man har udsat nogle rismelsbiller for gift i forskellige koncentrationer, nemlig 0.20, 0.32, 0.50 og 0.80 mg/cm^2, og dernæst set hvor mange af billerne der var døde efter 13 dages forløb. Forsøgsresultaterne er vist i tabel 6.2. (Giften strøs ud på gulvet hvor billerne færdes, derfor måles koncentrationen i mængde pr. areal.) Man er interesseret i at undersøge om der er forskel på virkningen af de forskellige koncentrationer. Vi vil derfor opstille en statistisk model der gør en sådan undersøgelse mulig.

Som i eksempel 6.3 vil vi antage at antal døde biller ved hver af de fire giftkoncentrationer kan opfattes som observerede værdier af binomialfordelte stokastiske variable. For hvert j lader vi y_j betegne antal døde biller og n_j betegne antal biller i alt ved koncentration nr. j, $j = 1, 2, 3, 4$. Den statistiske model er da at $\boldsymbol{y} = (y_1, y_2, y_3, y_4) = (43, 50, 47, 48)$ er en observeret værdi af en firedimensional stokastisk variabel $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)$

Tabel 6.2
Rismelsbillers overlevelse ved forskellige giftdoser.

	koncentration			
	0.20	0.32	0.50	0.80
antal døde	43	50	47	48
antal ikke døde	101	19	7	2
i alt	144	69	54	50

hvor Y_1, Y_2, Y_3 og Y_4 er indbyrdes uafhængige binomialfordelte med antalsparametre $n_1 = 144$, $n_2 = 69$, $n_3 = 54$ og $n_4 = 50$ og sandsynlighedsparametre θ_1, θ_2, θ_3 og θ_4. Modelfunktionen er

$$f(y_1, y_2, y_3, y_4; \theta_1, \theta_2, \theta_3, \theta_4) = \binom{144}{y_1} \theta_1^{y_1} (1-\theta_1)^{144-y_1} \binom{69}{y_2} \theta_2^{y_2} (1-\theta_2)^{69-y_2} \times \binom{54}{y_3} \theta_3^{y_3} (1-\theta_3)^{54-y_3} \binom{50}{y_4} \theta_4^{y_4} (1-\theta_4)^{50-y_4}.$$

Log-likelihoodfunktionen svarende til observationen $\boldsymbol{y}$ er

$$\begin{aligned} \ln L(\theta_1, \theta_2, \theta_3, \theta_4) = \text{konst} \\ + 43 \ln \theta_1 + 101 \ln(1-\theta_1) + 50 \ln \theta_2 + 19 \ln(1-\theta_2) \\ + 47 \ln \theta_3 + 7 \ln(1-\theta_3) + 48 \ln \theta_4 + 2 \ln(1-\theta_4). \end{aligned}$$

▷ [Eksemplet fortsætter i eksempel 7.2 side 119.]

Multinomialfordelingen

Multinomialfordelingen er en generalisation af binomialfordelingen: I situationer hvor man har at gøre med n uafhængige gentagelser af et grundforsøg der kan resultere i et af *to* mulige udfald, vil antallet af gange man får den ene slags udfald, være *binomial*fordelt, jf. definition 1.10 side 32. I situationer hvor man har at gøre med n uafhængige gentagelser af et grundforsøg der kan resultere i et af r mulige udfald $\omega_1, \omega_2, \ldots, \omega_r$, kan man interessere sig for de stokastiske variable Y_i der er lig antal gange man får udfaldet ω_i, $i = 1, 2, \ldots, r$. Den r-dimensionale stokastiske variabel $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_r)$ vil blive *multinomial*fordelt.

Under de beskrevne omstændigheder er fordelingen af $\boldsymbol{Y}$ af formen

$$\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y}) = \binom{n}{y_1\ y_2\ \ldots\ y_r} \prod_{i=1}^{r} \theta_i^{y_i} \tag{6.1}$$

når $\boldsymbol{y} = (y_1, y_2, \ldots, y_r)$ er et sæt af ikke-negative heltal der summerer til n, og $\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y}) = 0$ ellers. Parameteren $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_r)$ er et sæt af r ikke-negative

Tabel 6.3
Genotypefordeling af torsk fra tre lokaliteter i Østersøen.

	Lolland	Bornholm	Ålandsøerne
AA	27	14	0
Aa	30	20	5
aa	12	52	75
i alt	69	86	80

reelle tal der summerer til 1, og hvor θ_i er sandsynligheden for at grundforsøget giver udfaldet ω_i. Størrelsen

$$\binom{n}{y_1\ y_2\ \dots\ y_r} = \frac{n!}{\prod_{i=1}^{r} y_i!}$$

er en såkaldt *multinomialkoefficient* og er lig med antallet af måder hvorpå man kan dele en mængde med n elementer op i r delmængder sådan at delmængde nr. i indeholder netop y_i elementer, $i = 1, 2, \dots, r$.

Sandsynlighedsfordelingen givet ved sandsynlighedsfunktionen (6.1) omtales som multinomialfordelingen med r klasser (eller kategorier) og med antalsparameter n (et kendt tal) og sandsynlighedsparameter $\boldsymbol{\theta}$.

▷ [Læs fortsættelsen side 119.]

Eksempel 6.5: Torsk i Østersøen
Den 6. marts 1961 fangede nogle havbiologer 69 torsk ved Lolland og undersøgte arten af blodets hæmoglobin i hver enkelt torsk. Senere på året fangede man desuden nogle torsk ved Bornholm og ved Ålandsøerne og undersøgte dem på samme måde (Sick, 1965).

Man mener at hæmoglobin-arten bestemmes af ét enkelt gen, og det som biologerne undersøgte, var torskenes genotype for så vidt angår dette gen. Genet kan optræde i to udgaver som traditionen tro kaldes for A og a, og de mulige genotyper er da AA, Aa og aa. Den fundne fordeling på genotyper for hver af de tre lokaliteter ses i tabel 6.3.

På hver geografisk lokalitet har man klassificeret et antal torsk i tre mulige klasser, så på hver lokalitet er der tale om en multinomialfordelingssituation. (Når der er tre klasser, taler man også om en *tri*nomialfordeling.) Som grundmodel benytter vi derfor den model der siger at de tre observerede tripler

$$\boldsymbol{y}_{\mathrm{L}} = \begin{pmatrix} y_{1\mathrm{L}} \\ y_{2\mathrm{L}} \\ y_{3\mathrm{L}} \end{pmatrix} = \begin{pmatrix} 27 \\ 30 \\ 12 \end{pmatrix}, \quad \boldsymbol{y}_{\mathrm{B}} = \begin{pmatrix} y_{1\mathrm{B}} \\ y_{2\mathrm{B}} \\ y_{3\mathrm{B}} \end{pmatrix} = \begin{pmatrix} 14 \\ 20 \\ 52 \end{pmatrix}, \quad \boldsymbol{y}_{\mathrm{Å}} = \begin{pmatrix} y_{1\mathrm{Å}} \\ y_{2\mathrm{Å}} \\ y_{3\mathrm{Å}} \end{pmatrix} = \begin{pmatrix} 0 \\ 5 \\ 75 \end{pmatrix}$$

stammer fra hver sin multinomialfordeling med antalsparametre henholdsvis $n_{\mathrm{L}} = 69$, $n_{\mathrm{B}} = 86$ og $n_{\mathrm{Å}} = 80$, og med sandsynlighedsparametre henholdsvis

$$\boldsymbol{\theta}_{\mathrm{L}} = \begin{pmatrix} \theta_{1\mathrm{L}} \\ \theta_{2\mathrm{L}} \\ \theta_{3\mathrm{L}} \end{pmatrix}, \quad \boldsymbol{\theta}_{\mathrm{B}} = \begin{pmatrix} \theta_{1\mathrm{B}} \\ \theta_{2\mathrm{B}} \\ \theta_{3\mathrm{B}} \end{pmatrix}, \quad \boldsymbol{\theta}_{\mathrm{Å}} = \begin{pmatrix} \theta_{1\mathrm{Å}} \\ \theta_{2\mathrm{Å}} \\ \theta_{3\mathrm{Å}} \end{pmatrix}.$$

▷ [Eksemplet fortsættes i eksempel 7.3 side 119.]

Enstikprøveproblemet i poissonfordelingen

Den simpleste situation er som følger. Man har observationer $y_1, y_2, \dots, y_n$ af uafhængige identisk poissonfordelte stokastiske variable $Y_1, Y_2, \dots, Y_n$ med parameter μ. Modelfunktionen er

$$f(\boldsymbol{y}, \mu) = \prod_{j=1}^{n} \frac{\mu^{y_j}}{y_j!} \exp(-\mu) = \frac{\mu^{y_\cdot}}{\prod_{j=1}^{n} y_j!} \exp(-n\mu),$$

hvor $\mu \geq 0$ og $\boldsymbol{y} \in \mathbb{N}_0^n$. Likelihoodfunktionen er $L(\mu) = \text{konst}\, \mu^{y_\cdot} \exp(-n\mu)$, og log-likelihoodfunktionen er $\ln L(\mu) = \text{konst} + y_\cdot \ln \mu - n\mu$.

▷ [Læs fortsættelsen side 120.]

Eksempel 6.6: Hestespark
For hvert af de 20 år fra 1875 til 1894 har man for hvert af den prøjsiske armés 10 regimenter registreret hvor mange soldater der døde fordi de blev sparket af en hest (Bortkiewicz, 1898). 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 angive 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. Det viste sig at det største antal dødsfald pr. regiment-år var 4, og der bliver derfor fem klasser svarende til 0, 1, 2, 3 og 4 døde pr. år. De faktiske tal ses i tabel 6.4.

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. Set fra en passende stor »flyvehøjde« kan man måske godt finde på at antage at dødsfaldene indtræffer uafhængigt af hinanden og med samme intensitet året igennem, således at betingelsene for en poissonfordelingsmodel er til stede.

Vi vil derfor forsøge os med den statistiske model der siger at de 200 observationer $y_1, y_2, \dots, y_{200}$ er observationer af indbyrdes uafhængige identisk poissonfordelte stokastiske variable $Y_1, Y_2, \dots, Y_{200}$ med parameter μ.

▷ [Eksemplet fortsætter i eksempel 7.4 side 120.]

Tabel 6.4
Antal dødsfald som følge af hestespark i den prøjsiske armé.

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

Ligefordeling på et interval

Dette eksempel har så vidt vides ikke den store praktiske anvendelse, men det kan være nyttigt for at afprøve teorien.

Antag at $x_1, x_2, \ldots, x_n$ er observationer af indbyrdes uafhængige identisk fordelte stokastiske variable $X_1, X_2, \ldots, X_n$ som er ligefordelte på intervallet $]0\,;\theta[$, hvor $\theta > 0$ er den ukendte parameter. Tæthedsfunktionen for X_i er

$$f(x,\theta) = \begin{cases} 1/\theta & \text{når } 0 < x < \theta \\ 0 & \text{ellers,} \end{cases}$$

så modelfunktionen er

$$f(x_1, x_2, \ldots, x_n, \theta) = \begin{cases} 1/\theta^n & \text{når } 0 < x_{\min} \text{ og } x_{\max} < \theta \\ 0 & \text{ellers.} \end{cases}$$

Her er $x_{\min} = \min\{x_1, x_2, \ldots, x_n\}$ og $x_{\max} = \max\{x_1, x_2, \ldots, x_n\}$, altså henholdsvis den mindste og den største observation.

▷ [Læs fortsættelsen side 120.]

Enstikprøveproblemet i normalfordelingen

Man har observationer $y_1, y_2, \ldots, y_n$ af uafhængige identisk normalfordelte stokastiske variable med middelværdi μ og varians σ^2, altså fra fordelingen med tæthedsfunktion $y \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right)$. Modelfunktionen er

$$\begin{aligned} f(\boldsymbol{y}, \mu, \sigma^2) &= \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2}\frac{(y_j-\mu)^2}{\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\mu)^2\right) \end{aligned}$$

hvor $\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$, $\mu \in \mathbb{R}$ og $\sigma^2 > 0$. Standardomskrivninger giver

$$\begin{aligned} \sum_{j=1}^{n}(y_j-\mu)^2 &= \sum_{j=1}^{n}\Big((y_j-\overline{y}) + (\overline{y}-\mu)\Big)^2 \\ &= \sum_{j=1}^{n}(y_j-\overline{y})^2 + 2(\overline{y}-\mu)\sum_{j=1}^{n}(y_j-\overline{y}) + n(\overline{y}-\mu)^2 \\ &= \sum_{j=1}^{n}(y_j-\overline{y})^2 + n(\overline{y}-\mu)^2, \end{aligned}$$

altså

$$\sum_{j=1}^{n}(y_j-\mu)^2 = \sum_{j=1}^{n}(y_j-\overline{y})^2 + n(\overline{y}-\mu)^2. \tag{6.2}$$

Ved hjælp heraf får vi log-likelihoodfunktionen til

$$\begin{aligned} \ln L(\mu,\sigma^2) &= \text{konst} - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\mu)^2 \\ &= \text{konst} - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\overline{y})^2 - \frac{n(\overline{y}-\mu)^2}{2\sigma^2}. \end{aligned} \tag{6.3}$$

Vi kan udnytte formel (6.2) til endnu et formål: hvis vi indsætter $\mu = 0$, får vi

$$\sum_{j=1}^{n}(y_j-\overline{y})^2 = \sum_{j=1}^{n}y_j^2 - n\overline{y}^2 = \sum_{j=1}^{n}y_j^2 - \frac{1}{n}y_\bullet^2$$

dvs. summen af de kvadratiske afvigelser af y-erne fra $\overline{y}$ kan udregnes ud fra summen af y-erne og summen af kvadraterne på y-erne. For at udregne likelihoodfunktionen behøver man altså ikke kende de enkelte observationer, det er nok at kende summen og summen af kvadraterne (dvs. stikprøvefunktionen $t(\boldsymbol{y}) = (\sum y, \sum y^2)$ er *sufficient*, jf. side 101).

▷ [Læs fortsættelsen side 121.]

Eksempel 6.7: Lysets hastighed
I årene 1880-82 foretog den amerikanske fysiker A.A. Michelson og den amerikanske matematiker og astronom S. Newcomb en række efter den tids forhold temmelig nøjagtige bestemmelser af lysets hastighed i luft (Newcomb, 1891). Deres metoder var baseret på Foucaults idé med at sende en lysstråle fra et hurtigt roterende spejl hen på et fjernt fast spejl som returnerer lysstrålen til det roterende spejl, hvor man måler dens vinkelforskydning i forhold til den oprindelige lysstråle. Hvis man kender rotationshastigheden samt afstanden mellem spejlene, kan man derved bestemme lyshastigheden.

28	26	33	24	34	−44
27	16	40	−2	29	22
24	21	25	30	23	29
31	19	24	20	36	32
36	28	25	21	28	29
37	25	28	26	30	32
36	26	30	22	36	23
27	27	28	27	31	27
26	33	26	32	32	24
39	28	24	25	32	25
29	27	28	29	16	23

Tabel 6.5
Newcombs bestemmelser af lysets passagetid af en strækning på 7442 m. Tabelværdierne × 10^{-3} + 24.8 er passagetiden i 10^{-6} sek.

I tabel 6.5 (fra Stigler (1977)) er vist resultaterne af de 66 målinger som Newcomb foretog i perioden 24. juli til 5. september 1882 i Washington, D.C. I Newcombs opstilling var der 3721 m mellem det roterende spejl der var placeret i Fort Myer på vestbredden af Potomac-floden, og det faste spejl der var anbragt på George Washington-monumentets fundament. Den størrelse som Newcomb rapporterer, er lysets passagetid, altså den tid som det er om at tilbagelægge den pågældende distance.

Af de 66 værdier i tabellen skiller to sig ud, nemlig −44 og −2, der synes at være *outliers*, altså tal der tilsyneladende ligger *for* langt væk fra flertallet af observationerne. I den efterfølgende analyse af tallene vil vi vælge at se bort fra de to nævnte observationer, og der indgår herefter kun 64 observationer.

▷ [Eksemplet fortsættes i eksempel 7.5 side 122.]

Tostikprøveproblemet i normalfordelingen

Man har to grupper af individer, og på hvert individ har man målt værdien af en bestemt variabel Y. Individerne i den ene gruppe hører ikke sammen med dem i den anden gruppe på nogen måde, de er *uparrede*. Der behøver heller ikke være lige mange observationer i de to grupper. Skematisk ser situationen sådan ud:

	observationer					
gruppe 1	y_{11}	y_{12}	⋯	y_{1j}	⋯	y_{1n_1}
gruppe 2	y_{21}	y_{22}	⋯	y_{2j}	⋯	y_{2n_2}

Her betegner y_{ij} observation nr. j i gruppe nr. i, $i = 1, 2$. Grupperne har henholdsvis n_1 og n_2 observationer. Vi vil gå ud fra at forskellen mellem observationer inden for en gruppe er tilfældig, hvorimod der er en *systematisk forskel* på to de grupper – det er derfor at observationerne er inddelt i grupper! Endelig antages at y_{ij}-erne er observerede værdier af uafhængige stokastiske variable Y_{ij} som er normalfordelte med samme varians σ^2 og med $\mathrm{E}\, Y_{ij} = \mu_i$, $j = 1, 2, \ldots, n_i$, $i = 1, 2$. På denne måde beskriver de to middelværdiparametre μ_1 og μ_2 den *systematiske*

variation, dvs. de to gruppers niveauer, medens variansparameteren σ^2 (samt normalfordelingen) beskriver den *tilfældige variation*, der altså er den samme i begge grupper (denne antagelse kan man eventuelt teste, se opgave 8.2 side 138). Modelfunktionen er

$$f(\boldsymbol{y}, \mu_1, \mu_2, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2\right)$$

hvor $\boldsymbol{y} = (y_{11}, y_{12}, y_{13}, \ldots, y_{1n_1}, y_{21}, y_{22}, \ldots, y_{2n_2}) \in \mathbb{R}^n$, $(\mu_1, \mu_2) \in \mathbb{R}^2$ og $\sigma^2 > 0$; vi har her sat $n = n_1 + n_2$. Den til (6.2) analoge spaltning af kvadratsummen er

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2 + \sum_{i=1}^{2} n_i(\overline{y}_i-\mu_i)^2$$

hvor $\overline{y}_i = \dfrac{1}{n}\displaystyle\sum_{j=1}^{n_i} y_{ij}$ er gennemsnittet i gruppe nr. i.

Log-likelihoodfunktionen er

$$\begin{aligned}&\ln L(\mu_1, \mu_2, \sigma^2) \qquad (6.4)\\ &= \text{konst} - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2 + \sum_{i=1}^{2} n_i(\overline{y}_i-\mu_i)^2\right).\end{aligned}$$

▷ [Fortsættes side 122.]

Eksempel 6.8: C-vitamin
C-vitamin (ascorbinsyre) er et veldefineret kemisk stof som man sagtens kan fremstille industrielt, og man skulle tro at det industrielt fremstillede virker på nøjagtig samme måde som »naturligt« C-vitamin. For at undersøge om det nu også forholder sig sådan, har man foretaget et eksperiment med nogle marsvin (små gnavere).

Man delte 20 nogenlunde ens marsvin op i to grupper, hvoraf den ene fik appelsinsaft, og den anden fik en tilsvarende mængde »kunstigt« C-vitamin. Efter seks ugers behandling målte man længden af fortændernes odontoblaster (det tandbensdannende væv). Man fik da disse resultater (i hver gruppe er observationerne ordnet efter størrelse):

appelsinsaft:	8.2	9.4	9.6	9.7	10.0	14.5	15.2	16.1	17.6	21.5
kunstigt C-vitamin:	4.2	5.2	5.8	6.4	7.0	7.3	10.1	11.2	11.3	11.5

Man kan fastslå at der må være tale om en art tostikprøveproblem. Karakteren af observationerne gør at det ikke er urimeligt at forsøge sig med en normalfordelingsmodel af en slags, og det er alt i alt nærliggende at sige at der er tale om et »tostikprøveproblem med normalfordelte observationer«. Vi vil analysere observationerne ved brug af denne model, mere nøjagtigt vil vi undersøge om odontoblasternes middelvækst er den samme i de to grupper.

▷ [Eksemplet fortsætter som eksempel 7.6 side 123.]

Tabel 6.6
Forbes' barometriske målinger. – Kogepunktet er angivet i °F, lufttrykket i 'inches Kviksølv'.

Kogepunkt	Lufttryk	Kogepunkt	Lufttryk
194.5	20.79	201.3	24.01
194.3	20.79	203.6	25.14
197.9	22.40	204.6	26.57
198.4	22.67	209.5	28.49
199.4	23.15	208.6	27.76
199.9	23.35	210.7	29.04
200.9	23.89	211.9	29.88
201.1	23.99	212.2	30.06
201.4	24.02		

Simpel lineær regression

Regressionsanalyse, der er en stor underafdeling inden for statistik, handler om at modellere middelværdistrukturen for (det som modellen opfatter som) de stokastiske variable, idet man inddrager et større eller mindre antal kvantitative variable. Her ser vi på det simpleste tilfælde.

Der foreligger et antal sammenhørende værdier (x_i, y_i), $i = 1, 2, \ldots, n$, hvor y-erne opfattes som observerede værdier af stokastiske variable $Y_1, Y_2, \ldots, Y_n$, og x-erne er såkaldte *baggrundsvariable* eller *forklarende variable*. Det er en væsentlig pointe at x-erne ifølge modellen er ikke-stokastiske.

Den simple lineære regressionsmodel går ud på at Y-erne er indbyrdes uafhængige normalfordelte stokastiske variable med samme varians σ^2 og med en middelværdistruktur af formen $\mathrm{E}\, Y_i = \alpha + \beta x_i$, eller sagt mere præcist: der findes konstanter α og β således at $\mathrm{E}\, Y_i = \alpha + \beta x_i$ for alle i. Modellen indeholder således tre ukendte parametre, α, β og σ^2. Modelfunktionen er

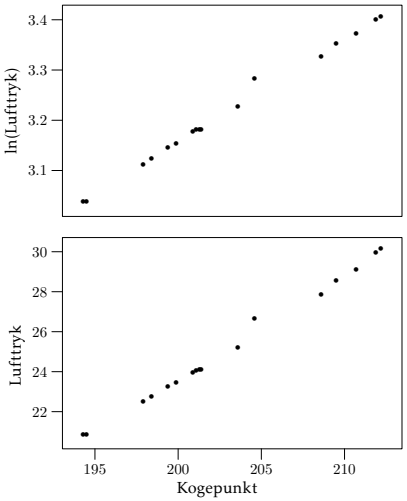
$$f(\boldsymbol{y}, \alpha, \beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{\big(y_i - (\alpha + \beta x_i)\big)^2}{\sigma^2} \right)$$
$$= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \big(y_i - (\alpha + \beta x_i)\big)^2 \right)$$

hvor $\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$, $\alpha, \beta \in \mathbb{R}$ og $\sigma^2 > 0$. Log-likelihoodfunktionen er

$$\ln L(\alpha, \beta, \sigma^2) = \text{konst} - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \big(y_i - (\alpha + \beta x_i)\big)^2. \qquad (6.5)$$

▷ [Fortsættes side 123.]

Eksempel 6.9: Forbes' barometriske målinger
Som bekendt aftager lufttrykket med højden over havets overflade, og derfor kan et



Figur 6.1
Forbes' målinger. Øverst: logaritmen til lufttryk afsat mod kogepunkt. Nederst: lufttryk afsat mod kogepunkt. Trykket er målt i 'inches Kviksølv', temperaturen i °F.

barometer benyttes som højdemåler. Da vands kogepunkt aftager med lufttrykket, kan man imidlertid også bestemme højden ved at koge vand. – I 1840erne og 1850erne foretog den skotske fysiker James D. Forbes på 17 forskellige lokaliteter i Alperne og i Skotland en række målinger hvor han bestemte dels vands kogepunkt, dels luftens tryk (omregnet til lufttrykket ved en standardlufttemperatur) (Forbes, 1857). Resultaterne er vist i tabel 6.6 (fra Weisberg (1980)).

Hvis man på en tegning afsætter lufttryk som funktion af kogepunkt, ser man at der er en tydelig sammenhæng (figur 6.1, nederst). Man kunne derfor overveje at opstille en lineær regressionsmodel med lufttryk som y og kogepunkt som x. Hvis man konsulterer en fysiker, kan man dog få at vide at man i højere grad skulle forvente en lineær sammenhæng mellem kogepunkt og logaritmen til lufttryk, hvilket også bekræftes af en figur (figur 6.1, øverst), så vi vil i stedet prøve at beskrive data ved hjælp af en regressionsmodel hvor man som y bruger logaritmen til lufttrykket og som x kogepunktet.

▷ [Eksemplet fortsættes i eksempel 7.7 side 125.]

6.2 Opgaver

Opgave 6.1
Gør rede for at binomialfordelingen faktisk er en instans af multinomialfordelingen.

Opgave 6.2
Binomialfordelingen blev defineret som fordelingen af en sum af uafhængige identisk fordelte 01-variable (definition 1.10 side 32).

Overvej hvordan man kan generalisere denne definition til en definition af multinomialfordelingen som fordelingen af en sum af uafhængige variable.

7 Estimation

En statistisk model er et udsagn om at det foreliggende datamateriale kan opfattes som en observation fra en bestemt sandsynlighedsfordeling der er fuldstændig specificeret på nær nogle få ukendte parametre. I dette kapitel skal vi beskæftige os med *estimationsproblemet*, dvs. spørgsmålet om hvordan man ud fra model plus observationer bærer sig ad med at udregne et *skøn* eller *estimat* over modellens ukendte parametre.

Man kan ikke inden for matematikkens rammer deducere sig frem til en løsning, det er nødvendigt undervejs at inddrage et eller flere udefra kommende principper. Afhængigt af hvilke principper man vælger at gå ud fra, kan man få forskellige estimationsmetoder. I det følgende præsenterer vi den fremgangsmåde som man »plejer« at bruge her i landet (og i mange andre lande). Først noget terminologi:

- En *stikprøvefunktion* er en funktion der er defineret på observationsrummet $\mathfrak{X}$ (og som afbilder ind i $\mathbb{R}$ eller $\mathbb{R}^n$).
  Hvis t er en stikprøvefunktion, er $t(\boldsymbol{X})$ en stokastisk variabel; ofte skelner man ikke så voldsomt meget mellem t og $t(\boldsymbol{X})$.
- En *estimator* for θ er en stikprøvefunktion (eller stokastisk variabel) med værdier i parameterrummet Θ. En estimator for $g(\theta)$ (hvor g er en funktion defineret på Θ) er en stikprøvefunktion (eller stokastisk variabel) med værdier i $g(\Theta)$. – Det er underforstået at estimatoren skal være et nogenlunde godt bud på den sande værdi af parameteren.
- Et *estimat* er en værdi som estimatoren antager, dvs. hvis t (eller $t(\boldsymbol{X})$) er en estimator, så er $t(\boldsymbol{x})$ et estimat.
- En *central* estimator for $g(\boldsymbol{\theta})$ er en estimator t med den egenskab at $\mathrm{E}_{\boldsymbol{\theta}}(t(\boldsymbol{X})) = g(\boldsymbol{\theta})$ for ethvert $\boldsymbol{\theta}$, dvs. som »i middel rammer rigtigt«.

7.1 Maksimaliseringsestimatoren

Antag at der foreligger en observation $\boldsymbol{x}$ der antages at kunne beskrives med en statistisk model der er specificeret ved modelfunktionen $f(\boldsymbol{x}, \boldsymbol{\theta})$. Hvis man skal vurdere de forskellige mulige $\boldsymbol{\theta}$-værdier for at finde en der kan udnævnes til at være et godt bud på »den sande værdi«, kan man basere vurderingen på værdierne af likelihoodfunktionen $L(\boldsymbol{\theta}) = f(\boldsymbol{x}, \boldsymbol{\theta})$: hvis $L(\boldsymbol{\theta}_1) > L(\boldsymbol{\theta}_2)$, så er $\boldsymbol{\theta}_1$ et bedre

bud på den sande værdi end $\boldsymbol{\theta}_2$ er; hvis man godtager dette ræsonnement, så må konsekvensen være at $\boldsymbol{\theta}$ skal estimeres som den værdi $\widehat{\boldsymbol{\theta}}$ der maksimaliserer L.

Definition 7.1
Maksimaliseringsestimatoren er den funktion der til en observation $\boldsymbol{x} \in \mathfrak{X}$ giver maksimumspunktet $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ for likelihoodfunktionen svarende til $\boldsymbol{x}$.

Maksimaliseringsestimatet er den værdi som maksimaliseringsestimatoren antager.

Ovenstående definition er naturligvis noget sjusket og ufuldstændig: der er ingen der siger at likelihoodfunktionen har netop ét maksimumspunkt, man kan godt komme ud for at der er flere maksimumspunkter, eller slet ingen. En lidt bedre definition kunne se sådan ud:

Definition 7.2
Et maksimaliseringsestimat hørende til observationen $\boldsymbol{x}$ er et maksimumspunkt $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ for likelihoodfunktionen hørende til $\boldsymbol{x}$.

En maksimaliseringsestimator er en (ikke nødvendigvis overalt defineret) funktion af $\mathfrak{X}$ ind i Θ, der til en observation $\boldsymbol{x}$ leverer et maksimaliseringsestimat.

Maksimaliseringsestimatoren er et bud på en generel metode til udregning af estimatorer. For at vurdere om det er et fornuftigt bud, kan man stille forskellige spørgsmål og se hvordan de besvares.

1. Hvor nemt er det at anvende metoden i konkrete modeller?
   Metoden går i praksis ud på at man skal bestemme maksimumspunkt/maksimumspunkter for funktionen L; bestemmelse af maksimumspunkter for en reel funktion er en almindelig og velforstået matematisk problemstilling som kan angribes (og løses) med standardmetoder.
   Det er i øvrigt oftest en fordel at bestemme $\widehat{\boldsymbol{\theta}}$ som maksimumspunkt for *log-likelihoodfunktionen* $\ln L$. Hvis $\ln L$ er en differentiabel funktion, skal maksimumspunkter i det indre af Θ som bekendt søges blandt løsningerne til ligningen $D \ln L(\boldsymbol{\theta}) = \mathbf{0}$. Her er D differentialoperatoren der betegner differentiation med hensyn til $\boldsymbol{\theta}$.
2. Findes der matematiske sætninger om maksimaliseringsestimatorens egenskaber, f.eks. om eksistens og entydighed, og om hvor tæt $\widehat{\boldsymbol{\theta}}$ ligger på $\boldsymbol{\theta}$?
   Ja, det gør der. Der findes en række generelle resultater om at når visse betingelser er opfyldt, og antallet af observationer går mod uendelig, så vil sandsynligheden for at der eksisterer et entydigt maksimaliseringsestimat, gå mod 1, og $\mathrm{P}_{\boldsymbol{\theta}}\bigl(|\widehat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}| < \varepsilon\bigr)$ går mod 1 (for ethvert $\varepsilon > 0$).
   Når nogle flere betingelser er opfyldt, blandt andet skal Θ være en åben mængde, og de tre første afledede af $\ln L$ skal eksistere og opfylde visse regularitetsbetingelser, så gælder at for $n \to \infty$ er $\widehat{\boldsymbol{\theta}}(\boldsymbol{X})$ asymptotisk normalfordelt med asymptotisk middelværdi $\boldsymbol{\theta}$ og en asymptotisk varians som er den inverse til $\mathrm{E}_{\boldsymbol{\theta}}\bigl(-D^2 \ln L(\boldsymbol{\theta};\boldsymbol{X})\bigr)$; desuden gælder at $\mathrm{E}_{\boldsymbol{\theta}}\bigl(-D^2 \ln L(\boldsymbol{\theta};\boldsymbol{X})\bigr) = \mathrm{Var}_{\boldsymbol{\theta}}\bigl(D \ln L(\boldsymbol{\theta};\boldsymbol{X})\bigr)$. (Ifølge den såkaldte Cramér-Rao-ulighed er dette den nedre grænse for variansen af en central estimator, så i den forstand er maksimaliseringsestimatoren asymptotisk optimal.)
   I situationer med en endimensional parameter θ betyder den asymptotiske normalitet af $\widehat{\theta}$ mere præcist at den stokastiske variabel

$$U = \frac{\widehat{\theta} - \theta}{\sqrt{\mathrm{E}_\theta\bigl(-D^2 \ln L(\theta;\boldsymbol{X})\bigr)}}$$

   er asymptotisk standardnormalfordelt når $n \to \infty$, og det er det samme som at sige at $\lim_{n\to\infty} \mathrm{P}(U \le u) = \Phi(u)$ for ethvert $u \in \mathbb{R}$. (Φ er fordelingsfunktionen for standardnormalfordelingen, jf. definition 3.9 side 74.)
3. Giver metoden estimater der ser fornuftige ud i de (få og simple) tilfælde hvor man er i stand til at overskue situationen?
   Det lader sig kun afgøre ved at se på eksempler.

7.2 Eksempler

Enstikprøveproblemet for 01-variable

◁ [Fortsat fra side 102.]

I denne model er log-likelihoodfunktionen og dens to første afledede

$$\begin{aligned} \ln L(\theta) &= x_\bullet \ln\theta + (n - x_\bullet)\ln(1-\theta), \\ D \ln L(\theta) &= \frac{x_\bullet - n\theta}{\theta(1-\theta)}, \\ D^2 \ln L(\theta) &= -\frac{x_\bullet}{\theta^2} - \frac{n - x_\bullet}{(1-\theta)^2} \end{aligned}$$

når $0 < \theta < 1$. Hvis $0 < x_\bullet < n$, har ligningen $D \ln L(\theta) = 0$ den entydige løsning $\widehat{\theta} = x_\bullet/n$, og da den anden afledede er negativ, er dette det entydige maksimumspunkt. Hvis $x_\bullet = n$, er L og $\ln L$ strengt voksende, og hvis $x_\bullet = 0$, er L og $\ln L$ strengt aftagende, så også i disse tilfælde er der et entydigt maksimumspunkt der er givet ved $\widehat{\theta} = x_\bullet/n$. Vi er således nået frem til at maksimaliseringsestimatet for θ er den relative hyppighed af 1-er – og det er jo meget fornuftigt.

Middelværdi og varians af estimatoren $\widehat{\theta} = \widehat{\theta}(\boldsymbol{X})$ er, jf. eksempel 1.19 side 43 og regnereglerne for middelværdi og varians, $\mathrm{E}_\theta\,\widehat{\theta} = \theta$ og $\mathrm{Var}_\theta\,\widehat{\theta} = \theta(1-\theta)/n$.

Den generelle teori (jf. ovenfor) oplyser at for store n er $\mathrm{E}_\theta\,\widehat{\theta} \approx \theta$ og $\mathrm{Var}_\theta\,\widehat{\theta} \approx \Bigl(\mathrm{E}_\theta\bigl(-D^2 \ln L(\theta,\boldsymbol{X})\bigr)\Bigr)^{-1} = \Bigl(\mathrm{E}_\theta\Bigl(\frac{X_\bullet}{\theta^2} + \frac{n - X_\bullet}{(1-\theta)^2}\Bigr)\Bigr)^{-1} = \theta(1-\theta)/n$.

▷ [Læs fortsættelsen side 129.]

Den simple binomialfordelingsmodel

◁ [Fortsat fra side 103.]

I den simple binomialfordelingsmodel er likelihoodfunktionen

$$L(\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \quad \theta \in [0\,;1]$$

og log-likelihoodfunktionen

$$\ln L(\theta) = \ln\binom{n}{y} + y\ln\theta + (n-y)\ln(1-\theta), \quad \theta \in [0\,;1].$$

På nær en konstant er denne funktion magen til den tilsvarende i enstikprøveproblemet med 01-variable. Vi kan derfor straks konstatere at maksimaliseringsestimatoren er $\widehat{\theta} = Y/n$. Da Y har samme fordeling som $X_\bullet$, er fordelingen af maksimaliseringsestimatoren den samme i de to modeller, specielt er også her $\mathrm{E}_\theta\,\widehat{\theta} = \theta$ og $\mathrm{Var}_\theta\,\widehat{\theta} = \theta(1-\theta)/n$.

▷ [Læs fortsættelsen side 129.]

Eksempel 7.1: Rismelsbiller I

◁ [Fortsat fra eksempel 6.3 side 103.]

I taleksemplet med rismelsbiller er $\widehat{\theta} = 43/144 = 0.30$. Den estimerede standardafvigelse er $\sqrt{\widehat{\theta}(1-\widehat{\theta})/144} = 0.04$.

▷ [Eksemplet fortsætter som eksempel 8.1 side 129.]

Sammenligning af binomialfordelinger

◁ [Fortsat fra side 104.]

Log-likelihoodfunktionen svarende til $\boldsymbol{y}$ er

$$\ln L(\boldsymbol{\theta}) = \text{konst} + \sum_{j=1}^{s}\Big(y_j\ln\theta_j + (n_j - y_j)\ln(1-\theta_j)\Big). \tag{7.1}$$

Det ses at $\ln L$ er en sum af led der hver især (på nær en konstant) er en log-likelihoodfunktion fra en simpel binomialfordelingsmodel, og parameteren θ_j optræder kun i det j-te led. Vi kan derfor uden videre opskrive maksimaliseringsestimatoren som

$$\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_s) = \left(\frac{Y_1}{n_1}, \frac{Y_2}{n_2}, \ldots, \frac{Y_s}{n_s}\right).$$

Da $Y_1, Y_2, \ldots, Y_s$ er uafhængige, bliver estimatorerne $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_s$ også uafhængige, og som i den simple binomialfordelingsmodel er $\mathrm{E}\,\widehat{\theta}_j = \theta_j$ og $\mathrm{Var}\,\widehat{\theta}_j = \theta_j(1-\theta_j)/n_j$, $j = 1, 2, \ldots, s$.

▷ [Læs fortsættelsen side 130.]

Eksempel 7.2: Rismelsbiller II

◁ [Fortsat fra eksempel 6.4 side 104.]

I rismelsbille-eksemplet hvor hver gruppe (koncentration) har sin egen binomialfordelingsparameter, estimeres denne som brøkdel døde i den pågældende gruppe, dvs. $(\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3, \widehat{\theta}_4) = (0.30, 0.72, 0.87, 0.96)$.

De estimerede standardafvigelser er $\sqrt{\widehat{\theta}_j(1-\widehat{\theta}_j)/n_j}$, dvs. 0.04, 0.05, 0.05 og 0.03.

▷ [Eksemplet fortsætter som eksempel 8.2 side 131.]

Multinomialfordelingen

◁ [Fortsat fra side 106.]

Hvis $\boldsymbol{y} = (y_1, y_2, \ldots, y_r)$ er en observation fra en multinomialfordeling med r klasser, antalsparameter n og sandsynlighedsparameter $\boldsymbol{\theta}$, så er log-likelihoodfunktionen

$$\ln L(\boldsymbol{\theta}) = \text{konst} + \sum_{i=1}^{r} y_i \ln\theta_i.$$

Parameteren $\boldsymbol{\theta}$ skal estimeres som maksimumspunktet $\widehat{\boldsymbol{\theta}}$ (i Θ) for $\ln L$; parameterrummet Θ er mængden af talsæt $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_r)$ for hvilke $\theta_i \geq 0$, $i = 1, 2, \ldots, r$, og $\theta_1 + \theta_2 + \cdots + \theta_r = 1$. Man ville vel umiddelbart formode at θ_i skal estimeres ved y_i/n, og det er da også det rigtige svar; men hvordan viser man det?

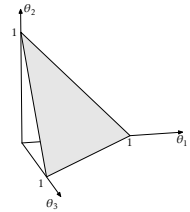
Én mulighed er at benytte en af de generelle metoder til bestemmelse af ekstremum under bibetingelser. En anden mulighed er at vise at vores formodning er rigtig. Vi vælger den sidste mulighed og skal altså vise at hvis vi sætter $\widehat{\theta}_i = y_i/n$, $i = 1, 2, \ldots, r$, og $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_r)$, så er $\ln L(\boldsymbol{\theta}) \leq \ln L(\widehat{\boldsymbol{\theta}})$ for alle $\boldsymbol{\theta} \in \Theta$.

Det snedige trick der skal bruges hertil, er at $\ln t \leq t - 1$ for alle t (og med lighedstegn hvis og kun hvis $t = 1$). Der gælder derfor

$$\begin{aligned}\ln L(\boldsymbol{\theta}) - \ln L(\widehat{\boldsymbol{\theta}}) &= \sum_{i=1}^{r} y_i \ln\frac{\theta_i}{\widehat{\theta}_i} \leq \sum_{i=1}^{r} y_i\left(\frac{\theta_i}{\widehat{\theta}_i} - 1\right)\\ &= \sum_{i=1}^{r}\left(y_i\frac{\theta_i}{y_i/n} - y_i\right) = \sum_{i=1}^{r}(n\theta_i - y_i) = n - n = 0.\end{aligned}$$

Ulighedstegnet er skarpt medmindre $\theta_i = \widehat{\theta}_i$ for alle $i = 1, 2, \ldots, r$.

▷ [Fortsættes side 132.]



Sandsynlighedssimplexet i det tredimensionale rum, dvs. mængden af ikke-negative tripler $(\theta_1, \theta_2, \theta_3)$ med $\theta_1 + \theta_2 + \theta_3 = 1$.

Eksempel 7.3: Torsk i Østersøen

◁ [Fortsat fra eksempel 6.5 side 106.]

Hvis vi indskrænker os til at studere torskene ved Lolland, er opgaven at bestemme det punkt $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_1)$ i det tredimensionale sandsynlighedssimplex som maksimaliserer

log-likelihoodfunktionen

$$\ln L(\boldsymbol{\theta}) = \text{konst} + 27\ln\theta_1 + 30\ln\theta_2 + 12\ln\theta_3.$$

Ifølge det foregående er $\widehat{\theta}_1 = 27/69 = 0.39$, $\widehat{\theta}_2 = 30/69 = 0.43$ og $\widehat{\theta}_3 = 12/69 = 0.17$.

▷ [Eksemplet fortsættes i eksempel 8.3 side 132.]

Enstikprøveproblemet i poissonfordelingen

◁ [Fortsat fra side 107.]

Log-likelihoodfunktionen og dens to første afledede er for $\mu > 0$

$$\begin{aligned} \ln L(\mu) &= \text{konst} + y_\bullet \ln\mu - n\mu, \\ D\ln L(\mu) &= \frac{y_\bullet}{\mu} - n, \\ D^2\ln L(\mu) &= -\frac{y_\bullet}{\mu^2}. \end{aligned}$$

Hvis $y_\bullet > 0$, har ligningen $D\ln L(\mu) = 0$ den entydige løsning $\widehat{\mu} = y_\bullet/n$, og da $D^2\ln L$ er negativ, er dette det entydige maksimumspunkt. Man ser desuden at formlen $\widehat{\mu} = y_\bullet/n$ også giver maksimumspunktet i den situation hvor $y_\bullet = 0$.

I poissonfordelingen er variansen lig med middelværdien, så ifølge de sædvanlige regneregler er middelværdi og varians af estimatoren $\widehat{\mu} = Y_\bullet/n$

$$\mathrm{E}_\mu\widehat{\mu} = \mu, \qquad \mathrm{Var}_\mu\widehat{\mu} = \mu/n. \tag{7.2}$$

Den generelle teori (jf. side 116) oplyser at for store n er $\mathrm{E}_\mu\widehat{\mu} \approx \mu$ og $\mathrm{Var}_\mu\widehat{\mu} \approx \left(\mathrm{E}_\mu\left(-D^2\ln L(\mu, \boldsymbol{Y})\right)\right)^{-1} = \left(\mathrm{E}_\mu\left(\frac{Y_\bullet}{\mu^2}\right)\right)^{-1} = \mu/n.$

Eksempel 7.4: Hestespark

◁ [Fortsat fra eksempel 6.6 side 107.]

I hestesparkeksemplet er $y_\bullet = 0\cdot109+1\cdot65+2\cdot22+3\cdot3+4\cdot1 = 122$, så $\widehat{\mu} = 122/200 = 0.61$.

Antallet af soldater i et givet regiment der i et givet år dør som følge af at være sparket af en hest, er altså (ifølge modellen) poissonfordelt med en parameter der estimeres til 0.61. Den estimerede standardafvigelse på estimatet er $\sqrt{\widehat{\mu}/n} = 0.06$, jf. formel (7.2).

Ligefordeling på et interval

◁ [Fortsat fra side 108.]

Likelihoodfunktionen er

$$L(\theta) = \begin{cases} 1/\theta^n & \text{når } x_{\max} < \theta \\ 0 & \text{ellers.} \end{cases}$$

Denne funktion antager ikke sit maksimum, men det er dog fristende at udnævne $\widehat{\theta} = x_{\max}$ til maksimaliseringsestimatet. – Tingene ville se pænere ud hvis vi gik over til at betragte ligefordelingen på det afsluttede interval fra 0 til θ.

Likelihoodfunktionen er ikke differentiabel i hele sit definitionsområde (som er $]0\,;+\infty[$); de regularitetsbetingelser der sikrer at maksimaliseringsestimatoren er asymptotisk normalfordelt (side 116), er derfor ikke opfyldt, og $\widehat{\theta} = X_{\max}$ er faktisk heller ikke asymptotisk normalfordelt (se opgave 7.4).

Enstikprøveproblemet i normalfordelingen

◁ [Fortsat fra side 109.]

Ved at løse ligningen $D\ln L = \mathbf{0}$ hvor $\ln L$ er log-likelihoodfunktionen (6.3) på side 109, finder man maksimaliseringsestimaterne for μ og σ^2 til

$$\widehat{\mu} = \overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j, \qquad \widehat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j - \overline{y})^2.$$

Man plejer dog at benytte et andet estimat for variansparameteren σ^2, nemlig

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \overline{y})^2,$$

hvilket kan begrundes med at s^2 er en *central* estimator; det fremgår af nedenstående sætning der er et specialtilfælde af sætning 11.1 side 171, jf. også afsnit 11.2.

Sætning 7.1

Antag at $X_1, X_2, \ldots, X_n$ er indbyrdes uafhængige identisk normalfordelte stokastiske variable med middelværdi μ og varians σ^2. Så gælder

1. *Den stokastiske variabel $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ er normalfordelt med middelværdi μ og varians σ^2/n.*
2. *Den stokastiske variabel $s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$ er gammafordelt med formparameter $f/2$ og skalaparameter $2\sigma^2/f$ hvor $f = n-1$, eller sagt på en anden måde: $(f/\sigma^2)s^2$ er χ^2-fordelt med f frihedsgrader.*
   Heraf følger blandt andet at $\mathrm{E}\,s^2 = \sigma^2$.
3. *De to stokastiske variable $\overline{X}$ og s^2 er stokastisk uafhængige.*

Bemærkninger: Antallet af frihedsgrader for variansskønnet i en normalfordelingsmodel er typisk antal observationer minus antal estimerede frie middelværdiparametre; antallet af frihedsgrader fortæller noget om præcisionen af variansskønnet, jf. opgave 7.3.

▷ [Fortsættes side 133.]

Eksempel 7.5: Lysets hastighed

◁ [Fortsat fra eksempel 6.7 side 109.]

Hvis vi går ud fra at de 64 positive værdier i tabel 6.5 side 110 kan betragtes som observationer fra en og samme normalfordeling, så skal denne normalfordelings middelværdi estimeres til $\overline{y} = 27.75$ og dens varians til $s^2 = 25.8$ med 63 frihedsgrader. Det betyder at passagetidens middelværdi estimeres til $(27.75 \times 10^{-3} + 24.8) \times 10^{-6}\text{sek} = 24.828 \times 10^{-6}\text{sek}$, og passagetidens varians estimeres til $25.8 \times (10^{-3} \times 10^{-6}\text{sek})^2 = 25.8 \times 10^{-6}(10^{-6}\text{sek})^2$ med 63 frihedsgrader, dvs. standardafvigelsen estimeres til $\sqrt{25.8 \times 10^{-6}}10^{-6}\text{sek} = 0.005 \times 10^{-6}\text{sek}$.

▷ [Eksemplet fortsætter som eksempel 8.4 side 134.]

Tostikprøveproblemet i normalfordelingen

◁ [Fortsat fra side 111.]

Denne models log-likelihoodfunktion (6.4) side 111 antager sit maksimum i punktet $(\overline{y}_1, \overline{y}_2, \widehat{\sigma}^2)$, hvor $\overline{y}_1$ og $\overline{y}_2$ er gennemsnittene i de to grupper, og $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$ (hvor $n = n_1 + n_2$). Ofte anvender man ikke $\widehat{\sigma}^2$ som estimat over σ^2, men derimod

$$s_0^2 = \frac{1}{n-2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2.$$

Nævneren $n-2$, *antallet af frihedsgrader*, bevirker at estimatoren bliver central; det fremgår af nedenstående sætning der er et specialtilfælde af sætning 11.1 side 171, jf. også afsnit 11.3.

Sætning 7.2
Antag at de stokastiske variable X_{ij} er indbyrdes uafhængige normalfordelte med samme varians σ^2 og med $\mathrm{E}\,X_{ij} = \mu_i$, $j = 1, 2, \ldots, n_i$, $i = 1, 2$. Så gælder

1. *De stokastiske variable $\overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$, $i = 1, 2$, er normalfordelte med middelværdi μ_i og varians σ^2/n_i.*

Tabel 7.1
C-vitamin-eksemplet, beregnede størrelser. n står for antal observationer y, S for Sum af y-er, $\overline{y}$ for gennemsnit af y-er, f for antal frihedsgrader, SS for Sum af kvadratiske afvigelser ('Sum of Squared deviations'), og s^2 for variansestimater ($s^2 = SS/f$).

	n	S	$\overline{y}$	f	SS	s^2
appelsinsaft	10	131.8	13.18	9	177.236	19.69
kunstigt C-vit.	10	80.0	8.00	9	68.960	7.66
sum	20	211.8		18	246.196	
gennemsnit			10.59			13.68

2. *Den stokastiske variabel $s_0^2 = \frac{1}{n-2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2$ er gammafordelt med formparameter $f/2$ og skalaparameter $2\sigma^2/f$ hvor $f = n-2$ og $n = n_1 + n_2$, eller sagt på en anden måde: $(f/\sigma^2)s_0^2$ er χ^2-fordelt med f frihedsgrader. Heraf følger blandt andet at $\mathrm{E}\,s_0^2 = \sigma^2$.*
3. *De tre stokastiske variable $\overline{X}_1$, $\overline{X}_2$ og s_0^2 er stokastisk uafhængige.*

Supplerende bemærkninger:

- Generelt er antallet af frihedsgrader for et variansskøn antal observationer minus antal estimerede middelværdiparametre.
- En størrelse som $y_{ij} - \overline{y}_i$ der er forskellen mellem den faktiske observation og det bedst mulige »fit« under den aktuelle model, kaldes undertiden for et *residual*. Som følge deraf kaldes en størrelse som $\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$ for en *residualkvadratsum*.

▷ [Fortsættes side 135.]

Eksempel 7.6: C-vitamin

◁ [Fortsat fra eksempel 6.8 side 111.]

Vi udregner forskellige hjælpestørrelser samt estimaterne over parametrene, se tabel 7.1. Middelværdien i appelsinsaft-gruppen estimeres til 13.18 og i den gruppe der har fået det kunstige C-vitamin, til 8.00. Den fælles varians estimeres til 13.68 med 18 frihedsgrader, og da hver af grupperne har 10 observationer, er den estimerede standardafvigelse på hver af de to middelværdiestimatorer $\sqrt{13.68/10} = 1.17$.

▷ [Eksemplet fortsætter som eksempel 8.5 side 137.]

Simpel lineær regression

◁ [Fortsat fra side 112.]

Vi skal bestemme estimater for parametrene α, β og σ^2 i den lineære regres-

sionsmodel. Log-likelihoodfunktionen er opskrevet i formel (6.5) på side 112. Vi kan spalte kvadratsummen på følgende måde:

$$\begin{aligned}\sum_{i=1}^{n}\bigl(y_i-(\alpha+\beta x_i)\bigr)^2 &= \sum_{i=1}^{n}\bigl((y_i-\overline{y})+(\overline{y}-\alpha-\beta\overline{x})-\beta(x_i-\overline{x})\bigr)^2\\ &= \sum_{i=1}^{n}(y_i-\overline{y})^2+\beta^2\sum_{i=1}^{n}(x_i-\overline{x})^2\\ &\quad -2\beta\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})+n(\overline{y}-\alpha-\beta\overline{x})^2,\end{aligned}$$

idet de øvrige to dobbelte produkter fra kvadreringen af den treleddede størrelse summerer til 0. Ved omskrivningen har vi opnået at α kun optræder i det sidste led, og dette antager sin mindsteværdi 0 netop når $\alpha=\overline{y}-\beta\overline{x}$. De resterende led udgør en andengradsfunktion af β, og denne funktion antager sit minimum når differentialkvotienten er 0, dvs. når $\beta=\dfrac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2}$. Konklusionen bliver således at maksimaliseringsestimaterne er

$$\widehat{\beta}=\frac{\displaystyle\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\displaystyle\sum_{i=1}^{n}(x_i-\overline{x})^2}\qquad\text{og}\qquad\widehat{\alpha}=\overline{y}-\widehat{\beta}\overline{x}.$$

(Det er her forudsat at $\sum(x_i-\overline{x})^2$ ikke er 0, dvs. at ikke alle x-erne er ens. – Hvis alle x-erne er ens, har det næppe nogen mening at prøve at estimere en funktion der skal vise hvordan y afhænger af x.) – Den *estimerede regressionslinje* er (den linje hvis ligning er) $y=\widehat{\alpha}+\widehat{\beta}x$.

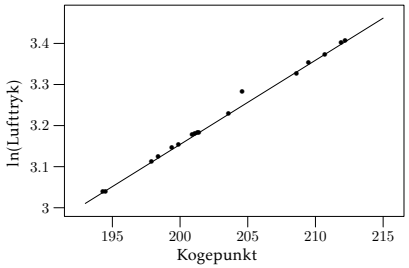
Den værdi af σ^2 der maksimaliserer log-likelihoodfunktionen, er

$$\widehat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\bigl(y_i-(\widehat{\alpha}+\widehat{\beta}x_i)\bigr)^2.$$

Som oftest angiver man dog i stedet det *centrale* variansestimat

$$s_{02}^2=\frac{1}{n-2}\sum_{i=1}^{n}\bigl(y_i-(\widehat{\alpha}+\widehat{\beta}x_i)\bigr)^2.$$

der har $n-2$ frihedsgrader. Med betegnelsen $SS_x=\displaystyle\sum_{i=1}^{n}(x_i-\overline{x})^2$ gælder følgende (jf. sætning 11.1 side 171 samt afsnit 11.6):



Figur 7.1
Forbes' målinger: Datapunkter plus estimeret regressionslinje.

Sætning 7.3
Om estimatorerne $\widehat{\alpha}$, $\widehat{\beta}$ og s_{02}^2 i den lineære regressionsmodel gælder:

1. *$\widehat{\beta}$ er normalfordelt med middelværdi β og varians σ^2/SS_x.*
2. *a) $\widehat{\alpha}$ er normalfordelt med middelværdi α og varians $\sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{SS_x}\right)$.*
   b) $\widehat{\alpha}$ og $\widehat{\beta}$ er korrelerede med korrelation $-1\Big/\sqrt{1+\frac{SS_x}{n\overline{x}^2}}$.
3. *a) $\widehat{\alpha}+\widehat{\beta}\overline{x}$ er normalfordelt med middelværdi $\alpha+\beta\overline{x}$ og varians σ^2/n.*
   b) $\widehat{\alpha}+\widehat{\beta}\overline{x}$ og $\widehat{\beta}$ er uafhængige.
4. *Variansestimatoren s_{02}^2 er stokastisk uafhængig af middelværdiestimatorerne, og den er gammafordelt med formparameter $f/2$ og skalaparameter $2\sigma^2/f$ hvor $f=n-2$, eller sagt på en anden måde: $(f/\sigma^2)s_{02}^2$ er χ^2-fordelt med f frihedsgrader.*
   Heraf følger blandt andet at $\mathrm{E}\,s_{02}^2=\sigma^2$.

Eksempel 7.7: Forbes' barometriske målinger
◁ [Fortsat fra eksempel 6.9 side 112.]
Som man ser af figur 6.1, er der et enkelt punkt der ser ud til at afvige temmelig meget fra det almindelige mønster, så vi vælger at se bort fra dette punkt og altså kun regne med de 16 resterende punkter.

Man finder den estimerede regressionslinje til

$$\ln(\text{lufttryk})=0.0205\cdot\text{kogepunkt}-0.95$$

og den estimerede varians er $s_{02}^2=6.84\times10^{-6}$ med 14 frihedsgrader. Figur 7.1 viser de observerede punkter og den estimerede linje. Umiddelbart ser det ud til at linjen beskriver punkterne udmærket.

Hvis man skal have nogen praktisk fornøjelse af sådanne kogepunktsbestemmelser, skal man også kende sammenhængen mellem højde og lufttryk. Sålænge vi holder os til bjerghøjder, aftager lufttrykket eksponentielt med højden, og der gælder at hvis lufttrykket ved havets overflade er p_0 (f.eks. 1013.25 hPa) og lufttrykket i højden h er p_h, så er $h\approx 8150\,\mathrm{m}\cdot(\ln p_0-\ln p_h)$.

7.3 Opgaver

Opgave 7.1
I eksempel 4.6 side 83 argumenteres der for at antal mider på et æbleblad er negativt binomialfordelt. Opskriv modelfunktion og likelihoodfunktion, og udregn på baggrund af de foreliggende observationer estimater over parametrene i fordelingen.

Opgave 7.2
Find middelværdien af maksimaliseringsestimatoren $\widehat{\sigma}^2$ for variansen, når vi har at gøre med et enstikprøveproblem i normalfordelingen.

Opgave 7.3
Find variansen på variansestimatoren s^2 i enstikprøveproblemet i normalfordelingen.

Opgave 7.4
På side 120 fandt vi at i den kontinuerte ligefordeling på $]0\,;\theta[$ er maksimaliseringsestimatoren $\widehat{\theta} = X_{\max}$ (hvor $X_{\max} = \max\{X_1, X_2, \ldots, X_n\}$).

1. Vis at tæthedsfunktionen for $X_{\max}$ er

$$f(x) = \begin{cases} \dfrac{nx^{n-1}}{\theta^n} & \text{når } 0 < x < \theta \\ 0 & \text{ellers.} \end{cases}$$

   Tip: find først $\mathrm{P}(X_{\max} \leq x)$.
2. Find middelværdi og varians af $\widehat{\theta} = X_{\max}$ og vis at for store n er $\mathrm{E}\widehat{\theta} \approx \theta$ og $\mathrm{Var}\,\widehat{\theta} \approx \theta^2/n^2$.
3. Vi kan nu finde den asymptotiske fordeling af $\dfrac{\widehat{\theta} - \theta}{\sqrt{\mathrm{Var}\,\widehat{\theta}}}$. For store n er ifølge forrige punkt $\dfrac{\widehat{\theta} - \theta}{\sqrt{\mathrm{Var}\,\widehat{\theta}}} \approx -Y$ hvor $Y = \dfrac{\theta - X_{\max}}{\theta/n} = n\left(1 - \dfrac{X_{\max}}{\theta}\right)$.
   Vis at $\mathrm{P}(Y > y) = \left(1 - \dfrac{y}{n}\right)^n$, og konkludér herudfra at Y er asymptotisk eksponentialfordelt med skalaparameter 1.

8 Hypoteseprøvning

Antag at man opererer med en statistisk model som har en modelfunktion $f(\boldsymbol{x}, \boldsymbol{\theta})$ hvor $\boldsymbol{x} \in \mathfrak{X}$ og $\boldsymbol{\theta} \in \Theta$. En *statistisk hypotese* er en påstand om at den sande parameterværdi $\boldsymbol{\theta}$ faktisk er beliggende i delmængden Θ_0 af Θ, formelt

$$H_0 : \boldsymbol{\theta} \in \Theta_0$$

hvor $\Theta_0 \subset \Theta$. Den statistiske hypotese postulerer altså at man kan klare sig med en simplere model.

Når man *tester* hypotesen, undersøger man hvordan hypotesen og de faktisk foreliggende observationer stemmer overens. Det foregår ved at man finder på eller vælger en endimensional stikprøvefunktion t kaldet en *teststørrelse*, som er indrettet på en måde så den »måler« afvigelsen mellem observationer og hypotese. Herefter udregner man den såkaldte *testsandsynlighed*, dvs. sandsynligheden (forudsat at hypotesen er rigtig) for at få en værdi af $\boldsymbol{X}$ der stemmer dårligere overens (målt ved hjælp af t) med hypotesen end den foreliggende observation $\boldsymbol{x}$ gør; hvis overensstemmelsen er meget dårlig, så *forkaster* man hypotesen. – Hele proceduren benævnes et *test*.

8.1 Kvotienttestet

Ligesom likelihoodfunktionen kunne danne udgangspunkt for konstruktion af en estimator for $\boldsymbol{\theta}$, kan den bruges i forbindelse med hypoteseprøvning. Man kan nemlig benytte følgende generelle metode til at konstruere et test:

1. Find maksimaliseringsestimatoren $\widehat{\boldsymbol{\theta}}$ i grundmodellen og maksimaliseringsestimatoren $\widehat{\widehat{\boldsymbol{\theta}}}$ under hypotesen, dvs. $\widehat{\boldsymbol{\theta}}$ er et punkt hvor $\ln L$ er maksimal i Θ, og $\widehat{\widehat{\boldsymbol{\theta}}}$ er et punkt hvor $\ln L$ er maksimal i Θ_0.
2. For at teste hypotesen sammenligner vi likelihoodfunktionens maksimale værdi under hypotesen med dens maksimale værdi i grundmodellen, dvs. vi sammenligner den bedste beskrivelse vi kan få af $\boldsymbol{x}$ under hypotesen, med den bedste beskrivelse vi kan få i grundmodellen. Det gøres med kvotienttesteststørrelsen

$$Q = \frac{L(\widehat{\widehat{\boldsymbol{\theta}}})}{L(\widehat{\boldsymbol{\theta}})} = \frac{L(\widehat{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x}), \boldsymbol{x})}{L(\widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{x})}.$$

Der gælder pr. definition at $0 \le Q \le 1$, og jo længere Q er fra 1, desto dårligere stemmer observationen $\boldsymbol{x}$ overens med hypotesen. Ofte bruger man ikke Q som teststørrelse, men derimod $-2\ln Q$. Denne størrelse er altid ikke-negativ, og jo større værdi, desto dårligere stemmer observationen $\boldsymbol{x}$ overens med hypotesen.

3. Testsandsynligheden ε er sandsynligheden (udregnet under forudsætning af at hypotesen er rigtig) for at få en værdi af $\boldsymbol{X}$ der passer dårligere sammen med hypotesen end den faktisk foreliggende værdi $\boldsymbol{x}$ gør, i matematik-sprog:

$$\varepsilon = \mathrm{P}_0\Big(Q(\boldsymbol{X}) \le Q(\boldsymbol{x})\Big) = \mathrm{P}_0\Big(-2\ln Q(\boldsymbol{X}) \ge -2\ln Q(\boldsymbol{x})\Big)$$

eller lidt kortere

$$\varepsilon = \mathrm{P}_0(Q \le Q_{\mathrm{obs}}) = \mathrm{P}_0\Big(-2\ln Q \ge -2\ln Q_{\mathrm{obs}}\Big).$$

(Fodtegnet 0 på P skal markere at der er tale om sandsynligheden udregnet under forudsætning af at hypotesen er rigtig.)

4. Hvis testsandsynligheden er lille, så forkastes hypotesen.
I praksis bruger man tit en strategi gående ud på at forkaste hypotesen hvis og kun hvis testsandsynligheden er mindre end et på forhånd fastlagt *signifikansniveau* α. Typiske værdier for signifikansniveauet er 5%, 2.5% og 1%. På den måde vil man med sandsynlighed α komme til at forkaste hypotesen i det tilfælde hvor den er rigtig. (Forhåbentlig vil der være en langt større sandsynlighed for at forkaste hypotesen hvis den ikke er rigtig.)

5. For at kunne udregne testsandsynligheden skal man kende fordelingen af teststørrelsen når hypotesen er rigtig.
I nogle situationer, f.eks. i normalfordelingsmodeller (se side 133ff og kapitel 11), kan testsandsynligheden udregnes eksakt ved hjælp af kendte og tabellerede standardfordelinger (t-, χ^2- og F-fordelinger).
I andre situationer kan man trække på generelle resultater der siger at under visse omstændigheder (nemlig dem der sikrer at $\widehat{\boldsymbol{\theta}}$ og $\widehat{\widehat{\boldsymbol{\theta}}}$ er asymptotisk normalfordelte, jf. side 116) er $-2\ln Q$ asymptotisk χ^2-fordelt med et antal frihedsgrader som er $\dim\Theta - \dim\Theta_0$ (i en passende betydning af dimension, som også omfatter at Θ kan være f.eks. en differentiabel flade). Til almindelig daglig brug kan man sige at antal frihedsgrader er lig »det faktiske antal parametre i grundmodellen minus det faktiske antal parametre under hypotesen«. – I de tilfælde hvor $-2\ln Q$ er asymptotisk χ^2-fordelt, kan man finde man den omtrentlige testsandsynlighed ved opslag i en tabel over χ^2-fordelingen (en lille χ^2-tabel ses på side 210).

Test
Ordet test kommer fra latin *testa* som er en lille lerkrukke af den slags som alkymisterne og sidenhen kemikerne brugte i deres undersøgelser (vel svarende til vore dages reagensglas).

På dansk er ordet test normalt fælleskøn, det hedder f.eks. *en* graviditetstest, men i statistiksammenhæng er det intetkøn, så rigtige statistikere siger *et* F-test, *et* t-test, osv.

8.2 Eksempler

Enstikprøveproblemet for 01-variable

◁ [Fortsat fra side 117.]

Antag at det af den faglige problemstilling fremgår at parameteren θ egentlig burde have værdien θ_0, og at det derfor er interessant at teste den statistiske hypotese $H_0 : \theta = \theta_0$. (Hvis man vil opskrive hypotesen lidt mere i overensstemmelse med den generelle formulering, må det blive som $H_0 : \theta \in \Theta_0$ hvor $\Theta_0 = \{\theta_0\}$.)

Idet $\widehat{\theta} = x_\bullet/n$, bliver kvotienttteststørrelsen

$$Q = \frac{L(\theta_0)}{L(\widehat{\theta})} = \frac{\theta_0^{x_\bullet}(1-\theta_0)^{n-x_\bullet}}{\widehat{\theta}^{x_\bullet}(1-\widehat{\theta})^{n-x_\bullet}} = \left(\frac{n\theta_0}{x_\bullet}\right)^{x_\bullet}\left(\frac{n-n\theta_0}{n-x_\bullet}\right)^{n-x_\bullet}$$

og

$$-2\ln Q = 2\left(x_\bullet \ln\frac{x_\bullet}{n\theta_0} + (n-x_\bullet)\ln\frac{n-x_\bullet}{n-n\theta_0}\right).$$

Testsandsynligheden kan derfor udregnes eksakt som

$$\varepsilon = \mathrm{P}_{\theta_0}\Big(-2\ln Q \ge -2\ln Q_{\mathrm{obs}}\Big),$$

og den kan bestemmes approksimativt som sandsynligheden for i χ^2-fordelingen med $1-0=1$ frihedsgrad at få værdier større end $-2\ln Q_{\mathrm{obs}}$; approksimationen kan benyttes når de »forventede« antal $n\theta_0$ og $n-n\theta_0$ er mindst 5.

Den simple binomialfordelingsmodel

◁ [Fortsat fra side 118.]

Antag at man i den simple binomialfordelingsmodel ønsker at teste en statistisk hypotese $H_0 : \theta = \theta_0$. Da likelihoodfunktionen i den simple binomialfordelingsmodel på nær en konstant faktor er lig likelihoodfunktionen i enstikprøveproblemet for 01-variable, se ovenfor, bliver Q og $-2\ln Q$ de samme i de to modeller, altså specielt

$$-2\ln Q = 2\left(y\ln\frac{y}{n\theta_0} + (n-y)\ln\frac{n-y}{n-n\theta_0}\right).$$

Testsandsynlighederne udregnes ligeledes på præcis samme måde i de to modeller.

Eksempel 8.1: Rismelsbiller I

◁ [Fortsat fra eksempel 7.1 side 118.]

Antag at det i rismelsbilleeksemplet er sådan [men det er ikke sådan; denne del af eksemplet er opdigtet til lejligheden] at man har en referencegift hvorom man véd at

når man doserer den på samme måde som den afprøvede gift, så dør 23% af billerne. Spørgsmålet er om den afprøvede gift virker på samme måde som referencegiften. I forhold til den statistiske model svarer dette spørgsmål til den statistiske hypotese $H_0 : \theta = 0.23$.

Når $n = 144$ og $\theta_0 = 0.23$, bliver $n\theta_0 = 33.12$ og $n - n\theta_0 = 110.88$, så $-2\ln Q$ som funktion af y er $-2\ln Q(y) = 2\Big(y \ln \frac{y}{33.12} + (144-y)\ln \frac{144-y}{110.88}\Big)$ og dermed $-2\ln Q_{\text{obs}} = -2\ln Q(43) = 3.60$. Den eksakte testsandsynlighed er

$$\varepsilon = \mathrm{P}_0\big(-2\ln Q(Y) \geq 3.60\big) = \sum_{y\,:\,-2\ln Q(y)\geq 3.60} \binom{144}{y} 0.23^y\, 0.77^{144-y}.$$

Ved almindelig udregning finder man at uligheden $-2\ln Q(y) \geq 3.60$ er opfyldt for $y = 0, 1, 2, \ldots, 23$ og for $y = 43, 44, 45, \ldots, 144$. Endvidere finder man at $\mathrm{P}_0(Y \leq 23) = 0.0249$, og at $\mathrm{P}_0(Y \geq 43) = 0.0344$, så $\varepsilon = 0.0249 + 0.0344 = 0.0593$.

Hvis man vil slippe for en del af regneriet, kan man benytte χ^2-approksimationen til fordelingen af $-2\ln Q$ for at finde testsandsynligheden; man skal bruge χ^2-fordelingen med $1 - 0 = 1$ frihedsgrad: grundmodellen har 1 ukendt parameter, og under hypotesen er der 0 ukendte parametre. I en tabel over fraktiler i χ^2-fordelingen med 1 frihedsgrad (se f.eks. side 210) finder man at 90%-fraktilen er 2.71 og 95%-fraktilen 3.84, så der er et sted mellem 5% og 10% sandsynlighed for at få værdier større end 3.60; computeren siger at i χ^2-fordelingen med 1 frihedsgrad er der en sandsynlighed på 5.78% for at få værdier større end 3.60, altså ganske tæt på den eksakte værdi 5.93%.

Da testsandsynligheden er over 5%, vil man – hvis man bruger den sædvanlige tommelfingerregel med et 5% signifikansniveau – ikke kunne afvise hypotesen; de foreliggende observationer er således ikke i modstrid med hypotesen om at giften virker på samme måde som referencegiften.

Sammenligning af binomialfordelinger

◁ [Fortsat fra side 118.]

Antag at man ønsker at undersøge om det kan antages at s forskellige binomialfordelinger har samme sandsynlighedsparameter. Dette formuleres som den statistiske hypotese $H_0 : \theta_1 = \theta_2 = \ldots = \theta_s$, eller lidt mere præcist som $H_0 : \boldsymbol{\theta} \in \Theta_0$ hvor $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_s)$ og hvor parameterrummet Θ_0 er givet som

$$\Theta_0 = \{\boldsymbol{\theta} \in [0\,;1]^s : \theta_1 = \theta_2 = \ldots = \theta_s\}.$$

For at teste H_0 skal vi først finde maksimaliseringsestimatoren under H_0, dvs. vi skal finde maksimumspunktet for restriktionen af $\ln L$ til Θ_0. Log-likelihood-funktionen er givet ved formel (7.1) på side 118, og når alle θ-erne er ens, er

$$\ln L(\theta, \theta, \ldots, \theta) = \text{konst} + \sum_{j=1}^{s} \Big(y_j \ln\theta + (n_j - y_j)\ln(1-\theta)\Big)$$

$$= \text{konst} + y_\bullet \ln\theta + (n_\bullet - y_\bullet)\ln(1-\theta)$$

der antager sit maksimum i punktet $\widehat{\theta} = y_\bullet/n_\bullet$; her er $y_\bullet = y_1 + y_2 + \ldots + y_s$ og $n_\bullet = n_1 + n_2 + \ldots + n_s$. Derfor bliver teststørrelsen

$$\begin{aligned} -2\ln Q &= -2\ln \frac{L(\widehat{\theta}, \widehat{\theta}, \ldots, \widehat{\theta})}{L(\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_s)} \\ &= 2\sum_{j=1}^{s}\Big(y_j \ln\frac{\widehat{\theta}_j}{\widehat{\theta}} + (n_j - y_j)\ln\frac{1-\widehat{\theta}_j}{1-\widehat{\theta}}\Big) \\ &= 2\sum_{j=1}^{s}\Big(y_j \ln\frac{y_j}{\widehat{y}_j} + (n_j - y_j)\ln\frac{n_j - y_j}{n_j - \widehat{y}_j}\Big) \end{aligned} \tag{8.1}$$

hvor $\widehat{y}_j = n_j\widehat{\theta}$ er det »forventede antal« gunstige i gruppe j under H_0, altså når grupperne har samme sandsynlighedsparameter. Det sidste udtryk for $-2\ln Q$ viser hvordan teststørrelsen sammenligner de observerede antal gunstige hhv. ugunstige udfald (antallene $y_1, y_2, \ldots, y_s$ og $n_1 - y_1, n_2 - y_2, \ldots, n_s - y_s$) med de »forventede antal« gunstige hhv. ugunstige udfald (tallene $\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_s$ og $n_1 - \widehat{y}_1, n_2 - \widehat{y}_2, \ldots, n_s - \widehat{y}_s$). Testsandsynligheden er

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

hvor fodtegnet 0 på P angiver at sandsynligheden skal udregnes under antagelse af at hypotesen er rigtig.* I givet fald er $-2\ln Q$ asymptotisk χ^2-fordelt med $s-1$ frihedsgrader. Som tommelfingerregel kan man benytte χ^2-approksimationen når alle de $2s$ »forventede« antal er 5 eller derover.

Eksempel 8.2: Rismelsbiller II

◁ [Fortsat fra eksempel 7.2 side 119.]

Formålet med undersøgelsen er at finde ud af om giften virker forskelligt i forskellige koncentrationer. Det gøres på den måde at man ser efter om tallene tyder på at man kan tillade sig at antage at der *ikke* er forskel på virkningerne af de forskellige koncentrationer. Det svarer i modellens termer til at teste den statistiske hypotese

$$H_0 : \theta_1 = \theta_2 = \theta_3 = \theta_4.$$

For at gøre det skal vi først udregne estimatet over den af H_0 postulerede fælles parameterværdi; det er $\widehat{\theta}_{\text{fælles}} = y_\bullet/n_\bullet = 188/317 = 0.59$. Derefter udregner vi de »forventede« antal $\widehat{y}_j = n_j\widehat{\theta}$ og $n_j - \widehat{y}_j$ og får tallene i tabel 8.1. Kvotienttesteststørrelsen $-2\ln Q$ (formel (8.1)) sammenligner nu disse tal med de observerede antal (tabel 6.2 side 105):

$$-2\ln Q_{\text{obs}} = 2\Big(\quad 43\ln\frac{43}{85.4} + 50\ln\frac{50}{40.9} + 47\ln\frac{47}{32.0} + 48\ln\frac{48}{29.7}$$

* Dette er ikke helt så uskyldigt som det måske kunne se ud til, for selv når hypotesen er rigtig, er der stadig en ukendt parameter inde i billedet, nemlig den fælles værdi af θ-erne.

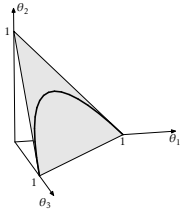
Tabel 8.1
Rismelsbillers overlevelse ved forskellige giftdoser: forventede antal hvis giften virker på samme måde for alle fire koncentrationer.

	koncentration			
	0.20	0.32	0.50	0.80
antal døde	85.4	40.9	32.0	29.7
antal ikke døde	58.6	28.1	22.0	20.3
i alt	144	69	54	50

$$+101\ln\frac{101}{58.6}+19\ln\frac{19}{28.1}+\ 7\ln\frac{7}{22.0}+\ 2\ln\frac{2}{20.3}\Big)$$
$$=113.1.$$

Grundmodellen har fire ukendte parametre, og hvis H_0 er rigtig, er der én ukendt parameter, så $-2\ln Q$-værdien skal sammenlignes med χ^2-fordelingen med $4-1=3$ frihedsgrader. I denne fordeling er 99.9%-fraktilen lig 16.27 (se f.eks. tabellen side 210), så der er langt under 0.01% sandsynlighed for at få en værre $-2\ln Q$-værdi hvis hypotesen er rigtig. På den baggrund må vi forkaste hypotesen og dermed konkludere at der er signifikant forskel på virkningen af de fire giftkoncentrationer.

Multinomialfordelingen



Det tonede område er sandsynlighedssimplexet, ɔ: mængden af tripler $\boldsymbol{\theta}=(\theta_1,\theta_2,\theta_3)$ af ikke-negative tal der summerer til 1. Den indtegnede kurve er Θ_0, dvs. de $\boldsymbol{\theta}$ der kan optræde hvis der er Hardy-Weinberg ligevægt.

◁ [Fortsat fra side 119.]

Det kan tænkes at den faglige problemstilling indebærer at sandsynlighedsparameteren $\boldsymbol{\theta}$ i virkeligheden kun kan variere i en vis delmængde Θ_0 af parameterrummet. Hvis Θ_0 er en differentiabel kurve eller flade, så er problemet »pænt« set fra den matematiske statistiks synspunkt.

Eksempel 8.3: Torsk i Østersøen

◁ [Fortsat fra eksempel 7.3 side 119.]

Ved simple ræsonnementer – som dog er af mindre interesse i nærværende sammenhæng – kan man vise at hvis torskene parrer sig med tilfældigt uden hensyntagen til genotype, og hvis populationen er lukket og i ligevægt, så vil de tre genotyper optræde med sandsynlighederne $\theta_1=\beta^2$, $\theta_2=2\beta(1-\beta)$ og $\theta_3=(1-\beta)^2$, hvor β er brøkdelen af A-gener i populationen (β er sandsynligheden for at et tilfældigt valgt gen er A). – Denne situation omtales som Hardy-Weinberg ligevægt.

Man kan nu undersøge om de foreliggende observationer tyder på at de tre torskepopulationer hver især er i Hardy-Weinberg ligevægt; her indskrænker vi os til Lollandspopulationen. På baggrund af observationen $\boldsymbol{y}_L=(27,30,12)$ fra en trinomialfordeling med sandsynlighedsparameter $\boldsymbol{\theta}_L$ skal vi derfor teste den statistiske hypotese $H_0:\boldsymbol{\theta}_L\in\Theta_0$ hvor Θ_0 er billedet ved afbildningen $\beta\mapsto\big(\beta^2,2\beta(1-\beta),(1-\beta)^2\big)$ af $[0\,;1]$ ind i sandsynlighedssimplexet, altså $\Theta_0=\Big\{\big(\beta^2,2\beta(1-\beta),(1-\beta)^2\big):\beta\in[0\,;1]\Big\}$.

Før vi kan teste hypotesen, skal vi estimere β. Under H_0 er log-likelihoodfunktionen

$$\ln L_0(\beta)=\ln L\big(\beta^2,2\beta(1-\beta),(1-\beta)^2\big)$$
$$=\text{konst}+2\cdot27\ln\beta+30\ln\beta+30\ln(1-\beta)+2\cdot12\ln(1-\beta)$$

$$=\text{konst}+(2\cdot27+30)\ln\beta+(30+2\cdot12)\ln(1-\beta),$$

som har maksimum i $\widehat{\beta}=\dfrac{2\cdot27+30}{2\cdot69}=\dfrac{84}{138}=0.609$ (dvs. β estimeres som det observerede antal A-gener divideret med det samlede antal gener). Teststørrelsen er

$$-2\ln Q=-2\ln\frac{L\big(\widehat{\beta}^2,2\widehat{\beta}(1-\widehat{\beta}),(1-\widehat{\beta})^2\big)}{L(\widehat{\theta}_1,\widehat{\theta}_2,\widehat{\theta}_3)}=2\sum_{i=1}^{3}y_i\ln\frac{y_i}{\widehat{y}_i}$$

hvor $\widehat{y}_1=69\cdot\widehat{\beta}^2=25.6$, $\widehat{y}_2=69\cdot2\widehat{\beta}(1-\widehat{\beta})=32.9$ og $\widehat{y}_3=69\cdot(1-\widehat{\beta})^2=10.6$ er de »forventede« antal under H_0.

I grundmodellen er der 2 frie parametre (der er tre θ-er, men de summerer til 1); under H_0 er der 1 fri parameter, nemlig β; $-2\ln Q$ får derfor $2-1=1$ frihedsgrad.

Man finder at $-2\ln Q=0.52$, der med 1 frihedgrad svarer til en testsandsynlighed på ca. 47%, så man kan sagtens antage at torskebestanden ved Lolland er i Hardy-Weinberg ligevægt.

Enstikprøveproblemet i normalfordelingen

◁ [Fortsat fra side 122.]

Antag at man ønsker at teste en hypotese om at middelværdien i normalfordelingen har en bestemt værdi; i det formelle sprog bliver det til at man ønsker at teste hypotesen $H_0:\mu=\mu_0$. Under H_0 er der én ukendt parameter, nemlig σ^2. Maksimaliseringsestimatet for σ^2 under H_0 findes som maksimumspunktet for log-likelihoodfunktionen (6.3) side 109, nu opfattet som en funktion af σ^2 alene (idet μ fikseres til μ_0), altså

$$\ln L(\mu_0,\sigma^2)=\text{konst}-\frac{n}{2}\ln(\sigma^2)-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\mu_0)^2$$

der antager sit maksimum når σ^2 har værdien $\widehat{\widehat{\sigma}}^2=\dfrac{1}{n}\displaystyle\sum_{j=1}^{n}(y_j-\mu_0)^2$. Kvotienttestørrelsen er $Q=\dfrac{L(\mu_0,\widehat{\widehat{\sigma}}^2)}{L(\widehat{\mu},\widehat{\sigma}^2)}$. Standardomskrivninger giver at

$$-2\ln Q=2\Big(\ln L(\overline{y},\widehat{\sigma}^2)-\ln L(\mu_0,\widehat{\widehat{\sigma}}^2)\Big)=n\ln\Big(1+\frac{t^2}{n-1}\Big)$$

hvor

$$t=\frac{\overline{y}-\mu_0}{\sqrt{s^2/n}}.\tag{8.2}$$

Hypotesen forkastes for store værdier af $-2\ln Q$, ɔ: for store værdier af $|t|$. Standardafvigelsen på $\overline{Y}$ er lig $\sqrt{\sigma^2/n}$ (sætning 7.1 side 121), så t-teststørrelsen kan

siges at måle forskellen mellem $\overline{y}$ og μ_0 i forhold til den estimerede standardafvigelse på $\overline{Y}$. Kvotienttestet er altså ækvivalent med et test baseret på den »umiddelbart forståelige« teststørrelse t.

Testsandsynligheden kan udregnes som $\varepsilon = \mathrm{P}_0(|t| > |t_{\mathrm{obs}}|)$. Om fordelingen af t under H_0 gælder

Sætning 8.1
Antag at $X_1, X_2, \ldots, X_n$ er indbyrdes uafhængige identisk normalfordelte stokastiske variable med middelværdi μ_0 og varians σ^2, og sæt $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ og $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$. Så følger den stokastiske variabel $t = \frac{\overline{X}-\mu_0}{\sqrt{s^2/n}}$ en t-fordeling med $n-1$ frihedsgrader.

t-fordelingen. t-fordelingen med f frihedsgrader er den kontinuerte fordeling som har tæthedsfunktion

$$\frac{\Gamma(\frac{f+1}{2})}{\sqrt{\pi f}\,\Gamma(\frac{f}{2})}\left(1+\frac{x^2}{f}\right)^{-\frac{f+1}{2}}$$

hvor $x \in \mathbb{R}$.

Supplerende bemærkninger:

- Det er en særlig – og væsentlig – pointe at fordelingen af t under H_0 ikke afhænger af den ukendte parameter σ^2 (og heller ikke af μ_0). Endvidere er t en dimensionsløs størrelse (– det bør teststørrelser altid være). Se også opgave 8.1.
- t-fordelingen er symmetrisk omkring 0, og testsandsynligheden kan derfor udregnes som $\mathrm{P}(|t| > |t_{\mathrm{obs}}|) = 2\,\mathrm{P}(t > |t_{\mathrm{obs}}|)$.
  I visse situationer kan man argumentere for det fornuftige i kun at forkaste hypotesen hvis t er stor og positiv (svarende til at $\overline{y}$ er større end μ_0); det kan for eksempel komme på tale hvis man på forhånd kan sige at hvis ikke $\mu = \mu_0$, så er $\mu > \mu_0$. I så fald skal testsandsynligheden udregnes som $\varepsilon = \mathrm{P}(t > t_{\mathrm{obs}})$.
  Tilsvarende, hvis man kun vil forkaste hypotesen når t er stor og negativ, skal testsandsynligheden udregnes som $\varepsilon = \mathrm{P}(t < t_{\mathrm{obs}})$.
  Et t-test der forkaster hypotesen når $|t|$ er stor, kaldes et *tosidet* t-test, og et test der forkaster hypotesen når t er stor og positiv [eller stor og negativ], kaldes et *ensidet* test.
- Fraktiler i t-fordelingen fås ud af computeren eller fra et statistisk tabelværk, se evt. side 216.
  Da fordelingen er symmetrisk om 0, vil tabelværkerne ofte kun indeholde fraktiler for sandsynligheder større end 0.5
- t-teststørrelsen kaldes undertiden for *Student's t* til ære for W.S. Gosset som i 1908 publicerede den første artikel om t-testet, og som skrev under pseudonymet *Student*.
- Se også afsnit 11.2 side 173.

William Sealy Gosset (1876-1937). Brygger ved Guinness-bryggeriet i Dublin, med interesser (og evner) for den nye statistiske videnskab. Han udviklede t-testet i forbindelse med kvalitetskontrolproblemer i bryggeprocessen, og fandt fordelingen af t-teststørrelsen (1908). Han publicerede sine statistiske arbejder under pseudonymet *Student*, vist nok fordi bryggeriet ikke ønskede at reklamere alt for meget med at dets medarbejdere beskæftigede sig med så suspekte emner som statistik.

Eksempel 8.4: Lysets hastighed

◁ [Fortsat fra eksempel 7.5 side 122.]

I vore dage er en meter pr. definition den strækning som lyset i vakuum gennemløber på 1/299 792 458 sekund, så lyset har den kendte hastighed 299 792 458 meter pr. sekund, og det vil derfor bruge $\tau_0 = 2.48238 \times 10^{-5}$ sekunder på at tilbagelægge en strækning på 7442 meter. Størrelsen τ_0 svarer til en tabelværdi på $((\tau_0 \times 10^6) - 24.8) \times 10^3 = 23.8$, så det ville være interessant at undersøge om de foreliggende data er forenelige med hypotesen om at den ukendte middelværdi μ har værdien $\mu_0 = 23.8$. Vi vil derfor teste den statistiske hypotese $H_0 : \mu = 23.8$.

Vi har tidligere fundet at $\overline{y} = 27.75$ og $s^2 = 25.8$, så t-teststørrelsen er

$$t = \frac{27.75 - 23.8}{\sqrt{25.8/64}} = 6.2\,.$$

Testsandsynligheden er sandsynligheden for at få t-værdier som enten er større end 6.2 eller mindre end −6.2. Ved tabelopslag kan man finde at i t-fordelingen med 63 frihedsgrader er 99.95%-fraktilen lidt over 3.4, dvs. der er mindre end 0.05% sandsynlighed for at få en værdi som er større end 6.2, og testsandsynligheden er dermed mindre $2 \times 0.05\% = 0.1\%$. En så lille testsandsynlighed betyder at man må forkaste hypotesen. Newcombs målinger af lysets passagetid stemmer altså ikke overens med hvad vi i dag véd om lysets hastighed.

Vi ser at Newcombs passagetider er en smule for store, og da den lyshastighed vi her har benyttet, er lysets hastighed i vakuum, kan noget af forklaringen eventuelt være at lyset bevæger sig en smule langsommere i luft end i vakuum.

Tostikprøveproblemet i normalfordelingen

◁ [Fortsat fra side 123.]

Antag at man ønsker at teste hypotesen $H_0 : \mu_1 = \mu_2$ om at de to stikprøver stammer fra samme normalfordeling (variansen er pr. antagelse den samme). Parameterestimaterne i grundmodellen blev fundet på side 122. Under H_0 er der kun to ukendte parametre, nemlig den fælles middelværdi μ og den fælles varians σ^2, og da H_0 svarer til den situation der er studeret under overskriften enstikprøveproblemet i normalfordelingen, kan vi uden videre opskrive maksimaliseringsestimaterne: for middelværdien μ er det totalgennemsnittet $\overline{y} = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i} y_{ij}$, og for variansen σ^2 er det $\widehat{\widehat{\sigma}}^2 = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2$. Kvotientteststørrelsen for H_0 er $Q = \frac{L(\overline{y}, \overline{y}, \widehat{\widehat{\sigma}}^2)}{L(\overline{y}_1, \overline{y}_2, \widehat{\sigma}^2)}$. Standardomskrivninger giver at

$$-2\ln Q = 2\left(\ln L(\overline{y}_1, \overline{y}_2, \widehat{\sigma}^2) - \ln L(\overline{y}, \overline{y}, \widehat{\widehat{\sigma}}^2)\right) = n\ln\left(1 + \frac{t^2}{n-2}\right)$$

hvor

$$t = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{s_0^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \tag{8.3}$$

Hypotesen forkastes for store værdier af $-2\ln Q$, dvs. for store værdier af $|t|$.

Ifølge regnereglerne for varianser er $\mathrm{Var}(\overline{Y}_1 - \overline{Y}_2) = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$, så t-størrelsen måler forskellen mellem de to middelværdiestimater i forhold til den estimerede standardafvigelse på denne forskel. Kvotienttestet er således ækvivalent med et test hvor teststørrelsen er »umiddelbart forståelig«.

Testsandsynligheden kan udregnes som $\varepsilon = \mathrm{P}_0(|t| > |t_{\mathrm{obs}}|)$. Om fordelingen af t under H_0 gælder

Sætning 8.2
Antag at de stokastiske variable X_{ij}, $j = 1, 2, \ldots, n_i$, $i = 1, 2$, er indbyrdes uafhængige og normalfordelte med samme middelværdi μ og med samme varians σ^2. Sæt

$$\overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}, \quad i = 1, 2, \qquad \textit{og}$$

$$s_0^2 = \frac{1}{n-2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2$$

Så er den stokastiske variabel

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_0^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

t-fordelt med $n-2$ frihedsgrader ($n = n_1 + n_2$).

Supplerende bemærkninger:

- Hvis hypotesen H_0 om ens middelværdier godkendes, er der ikke tale om to forskellige stikprøver med hver sin middelværdi, men om én stikprøve med $n_1 + n_2$ observationer. I konsekvens heraf skal middelværdi og varians estimeres som i et enstikprøveproblem, dvs. som $\overline{y}$ og s_{01}^2, hvor

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i} y_{ij} \qquad \text{og} \qquad s_{01}^2 = \frac{1}{n-1}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2.$$

- Fordelingen af t under H_0 afhænger ikke af de to ukendte parametre σ^2 og den fælles værdi af middelværdierne.

- t-fordelingen er symmetrisk omkring 0, og testsandsynligheden kan derfor udregnes som $\mathrm{P}(|t| > |t_{\mathrm{obs}}|) = 2\,\mathrm{P}(t > |t_{\mathrm{obs}}|)$.
  I visse situationer kan man argumentere for det fornuftige i kun at forkaste hypotesen hvis t er stor og positiv (svarende til at $\overline{y}_1$ er større end $\overline{y}_2$); det kan for eksempel komme på tale hvis man på forhånd kan sige at hvis ikke $\mu_1 = \mu_2$, så er $\mu_1 > \mu_2$. I så fald skal testsandsynligheden udregnes som $\varepsilon = \mathrm{P}(t > t_{\mathrm{obs}})$.
  Tilsvarende, hvis man kun vil forkaste hypotesen når t er stor og negativ, skal testsandsynligheden udregnes som $\varepsilon = \mathrm{P}(t < t_{\mathrm{obs}})$.
- Fraktiler i t-fordelingen fås ud af computeren eller fra et statistisk tabelværk, se evt. side 216.
  Da fordelingen er symmetrisk om 0, vil tabelværkerne ofte kun indeholde fraktiler for sandsynligheder større end 0.5
- Se også afsnit 11.3 side 174.

Eksempel 8.5: C-vitamin

◁ [Fortsat fra eksempel 7.6 side 123.]

Da metoden til sammenligning af middelværdierne i de to grupper forudsætter at de to grupper har samme varians, kan man eventuelt også teste hypotesen om varianshomogenitet (se opgave 8.2). Testet er baseret på varianskvotienten

$$R = \frac{s^2_{\text{appelsinsaft}}}{s^2_{\text{kunstigt}}} = \frac{19.69}{7.66} = 2.57\,.$$

Denne værdi skal sammenholdes med F-fordelingen med $(9, 9)$ frihedsgrader i et tosidet test. Tabelopslag (f.eks. side 212ff) viser at 95%-fraktilen er 3.18 og 90%-fraktilen 2.44; der er derfor mellem 10 og 20 procents chance for at få en værre R-værdi selvom hypotesen er rigtig, og på dette grundlag vil vi ikke afvise antagelsen om varianshomogenitet. Den fælles varians estimeres til $s_0^2 = 13.68$ med 18 frihedsgrader.

Vi kan nu gå over til det egentlige, nemlig at teste om der er signifikant forskel på to gruppers niveauer. Til det formål udregnes t-teststørrelsen

$$t = \frac{13.18 - 8.00}{\sqrt{13.68\,(\frac{1}{10} + \frac{1}{10})}} = \frac{5.18}{1.65} = 3.13\,.$$

Den fundne værdi skal sammenholdes med t-fordelingen med 18 frihedsgrader. I denne fordeling er 99.5%-fraktilen 2.878, så der er mindre end 1% chance for at få en værdi numerisk større end 3.13. Konklusionen bliver derfor at der er en klart signifikant forskel mellem de to grupper. – Som det ses af tallene, består forskellen i at den »kunstige« gruppe har mindre odontoblastvækst end appelsingruppen. Kunstigt C-vitamin synes altså ikke at virke så godt som det naturlige.

8.3 Opgaver

Opgave 8.1
Antag at $x_1, x_2, \ldots, x_n$ er en stikprøve fra normalfordelingen med middelværdi ξ og varians σ^2, og at man ønsker at teste hypotesen $H_{0x} : \xi = \xi_0$. Det gøres med et t-test. – Man kunne imidlertid også gøre følgende: Tag to tal a og $b \neq 0$ og sæt

$$\begin{aligned} \mu &= a + b\xi, \\ \mu_0 &= a + b\xi_0, \\ y_i &= a + bx_i, \quad i = 1, 2, \ldots, n. \end{aligned}$$

Hvis x-erne er en stikprøve fra normalfordelingen med parametre ξ og σ^2, så er y-erne en stikprøve fra normalfordelingen med parametre μ og $b^2\sigma^2$ (sætning 3.11 side 74), og hypotesen $H_{0x} : \xi = \xi_0$ er ensbetydende med hypotesen $H_{0y} : \mu = \mu_0$. Vis at t-teststørrelsen for at teste H_{0x} er lig med t-teststørrelsen for at teste H_{0y} (det vil sige at t-testet er invariant ved affine transformationer).

Opgave 8.2
I behandlingen af tostikprøveproblemet i normalfordelingen (side 110ff) antages uden videre at de to stikprøver har samme varians. Man kan imidlertid godt udføre et test for om de to grupper kan antages at have samme varians. For at gøre det udvider vi modellen til følgende: $y_{11}, y_{12}, \ldots, y_{1n_1}$ er observationer fra normalfordelingen med parametre μ_1 og σ_1^2, og $y_{21}, y_{22}, \ldots, y_{2n_2}$ er observationer fra normalfordelingen med parametre μ_2 og σ_2^2. I denne model tester vi så hypotesen $H : \sigma_1^2 = \sigma_2^2$.

Gør det! Dvs. opskriv likelihoodfunktionen, find maksimaliseringsestimaterne, og opskriv kvotienttesstørrelsen Q.

Vis at Q kan udtrykkes ved $R = \dfrac{s_1^2}{s_2^2}$, hvor $s_i^2 = \dfrac{1}{n_i - 1} \displaystyle\sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$ er det centrale variansskøn i gruppe i, $i = 1, 2$.

Man kan udlede fordelingen af R når antagelsen om om varianshomogenitet er rigtig; den er en F-fordeling (se evt. side 173) med $(n_1 - 1, n_2 - 1)$ frihedsgrader.

9 Nogle eksempler

9.1 Rismelsbiller

Dette eksempel viser blandt andet hvordan man kan udbygge en binomialfordelingsmodel sådan at sandsynlighedsparameteren tillades at afhænge af nogle baggrundsvariable; derved fås en såkaldt logistisk regressionsmodel.

I en undersøgelse (jf. Pack and Morgan (1990)) af insekters reaktion på insektgiften pyrethrum har man udsat nogle rismelsbiller, *Tribolium castaneum*, for forskellige mængder gift og derpå set hvor mange der var døde efter 13 dages forløb. Forsøget udførtes med fire forskellige giftkoncentrationer, dels på hanbiller, dels på hun-biller. Resultaterne (i reduceret form) ses i tabel 9.1.

Grundmodellen

Der indgår $144 + 69 + 54 + \ldots + 47 = 641$ biller i forsøget. Hver bille har et bestemt køn (to muligheder) og bliver udsat for en bestemt giftdosis (fire muligheder), og i løbet af forsøget er billen enten død eller har overlevet.

Første skridt i modelleringsprocessen består i at gøre sig klart hvilken status de forskellige størrelser skal have i modellen:

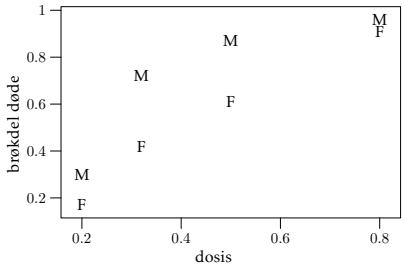
- Størrelserne dosis og køn er baggrundsvariable der er benyttet til at inddele billerne i $2 \cdot 4 = 8$ grupper, idet man forestiller sig at dosis og køn kan have betydning for billens overlevelse; det kan endda tænkes at selve talværdierne af dosis har betydning.
- Totalantallene $(144, 69, 54, 50, 152, 81, 44, 47)$ er kendte konstanter, nemlig antal biller i de enkelte grupper.
- Antal døde $(43, 50, 47, 48, 26, 34, 27, 43)$ er observerede værdier af stokastiske variable.

For at få en idé om talmaterialets beskaffenhed kan man lave nogle simple udregninger og tegninger (f.eks. tabel 9.1 og figur 9.1).

For hver af de otte grupper er det nærliggende at foreslå at beskrive »antal døde« som en observation fra en binomialfordeling med en antalsparameter der er det samlede antal biller i den pågældende gruppe, og med en (ukendt) sandsynlighedsparameter der skal fortolkes som sandsynligheden for at en bille af det pågældende køn dør af giften doseret i den pågældende koncentration.

Tabel 9.1
Rismelsbiller: For hvert køn og hver dosis er angivet antal døde/totalantal = obs. dødssandsynlighed. Dosis er anført i mg/cm².

dosis	M	F
0.20	43/144 = 0.30	26/152 = 0.17
0.32	50/ 69 = 0.72	34/ 81 = 0.42
0.50	47/ 54 = 0.87	27/ 44 = 0.61
0.80	48/ 50 = 0.96	43/ 47 = 0.91



Figur 9.1
Rismelsbiller: Observeret dødssandsynlighed (relativ hyppighed) som funktion af dosis, for hvert køn.

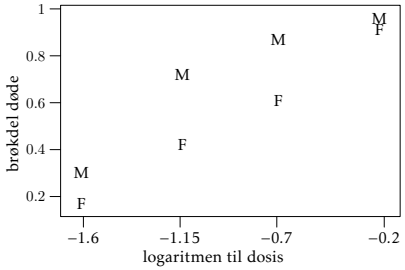
Her er det ikke så interessant blot at få at vide om der er en signifikant forskel på grupperne eller ej, det ville være langt mere spændende hvis man kunne give en nærmere beskrivelse af hvordan sandsynligheden for at dø afhænger af giftkoncentrationen, og hvis man kunne udtale sig om hvorvidt giften virker ens på hanner og hunner. Vi indfører noget notation og præciserer *grundmodellen*:

1. I den gruppe der svarer til dosis d og køn k, er der n_{dk} biller hvoraf y_{dk} døde; her er $k \in \{M, F\}$ og $d \in \{0.20, 0.32, 0.50, 0.80\}$.
2. Det antages at y_{dk} er en observation af en stokastisk variabel Y_{dk} som er binomialfordelt med kendt antalsparameter n_{dk} og med sandsynlighedsparameter p_{dk}.
3. Det antages desuden at de enkelte Y_{dk}-er er stokastisk uafhængige.

Likelihoodfunktionen i grundmodellen er

$$\begin{aligned} L &= \prod_k \prod_d \binom{n_{dk}}{y_{dk}} p_{dk}^{y_{dk}} (1-p_{dk})^{n_{dk}-y_{dk}} \\ &= \prod_k \prod_d \binom{n_{dk}}{y_{dk}} \cdot \prod_k \prod_d \Big(\frac{p_{dk}}{1-p_{dk}}\Big)^{y_{dk}} \cdot \prod_k \prod_d (1-p_{dk})^{n_{dk}} \\ &= \text{konst} \cdot \prod_k \prod_d \Big(\frac{p_{dk}}{1-p_{dk}}\Big)^{y_{dk}} \cdot \prod_k \prod_d (1-p_{dk})^{n_{dk}}, \end{aligned}$$

og log-likelihoodfunktionen er



Figur 9.2
Rismelsbiller: Observeret dødssandsynlighed (relativ hyppighed) som funktion af logaritmen til dosis, for hvert køn.

$$\begin{aligned} \ln L &= \text{konst} + \sum_k \sum_d y_{dk} \ln \frac{p_{dk}}{1-p_{dk}} + \sum_k \sum_d n_{dk} \ln(1-p_{dk}) \\ &= \text{konst} + \sum_k \sum_d y_{dk} \operatorname{logit}(p_{dk}) + \sum_k \sum_d n_{dk} \ln(1-p_{dk}). \end{aligned}$$

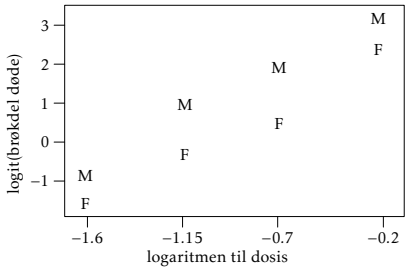
De otte parametre p_{dk} varierer uafhængigt af hinanden, og $\widehat{p}_{dk} = y_{dk}/n_{dk}$. Opgaven i det følgende er nu at modellere p's afhængighed af d og k.

En dosis-respons model

Hvordan er sammenhængen mellem giftkoncentrationen (dosis) d og sandsynligheden p_d for at en bille dør ved denne dosis? Hvis man vidste en hel masse om hvordan netop dette giftstof virker i billeorganismen, kunne man formentlig give et velbegrundet forslag til hvordan sandsynligheden afhænger af dosis. Men den statistiske modelbyggers tilgang til problemet er af langt mere jordbunden og pragmatisk karakter, som vi nu skal se.

I eksemplet har eksperimentator valgt nogle tilsyneladende mærkværdige dosis-værdier (0.20, 0.32, 0.50 og 0.80). Hvis man ser nærmere efter, opdager man dog at der (næsten) er tale om en kvotientrække, idet kvotienten mellem et tal og det næste er (næsten) den samme, nemlig 1.6. Det tager den statistiske modelbygger som et fingerpeg om at dosis antagelig skal måles på en logaritmisk skala, dvs. man skal interessere sig for hvordan sandsynligheden for at dø afhænger af logaritmen til dosis. Derfor tegner vi figur 9.1 en gang til, idet vi nu afsætter logaritmen til dosis ud ad abscisseaksen; resultatet ses i figur 9.2.

Vi skal modellere sandsynlighedernes afhængighed af baggrundsvariablen $\ln d$. En af de simpleste former for afhængighed er *lineær* afhængighed. Imidlertid ville det være en dårlig idé at foreslå at p_d skulle afhænge lineært af $\ln d$ (altså at $p_d = \alpha + \beta \ln d$ for passende valgte konstanter α og β) fordi dette ville



Figur 9.3
Rismelsbiller: Logit til estimeret dødssandsynlighed (relativ hyppighed) som funktion af logaritmen til dosis, for hvert køn.

være uforeneligt med kravet om at sandsynlighederne skal ligge mellem 0 og 1. Ofte gør man så det at man omregner p_d til en ny skala og postulerer at »p_d på den ny skala« afhænger lineært af $\ln d$. Omregningen kan foregå ved hjælp af logit-funktionen.

Logit-funktionen. Funktionen

$$\operatorname{logit}(p) = \ln \frac{p}{1-p}$$

afbilder intervallet $]0,1[$ bijektivt på den reelle akse $\mathbb{R}$. Den omvendte funktion er

$$p = \frac{\exp(z)}{1+\exp(z)}.$$

Når p er sandsynligheden for en bestemt hændelse (f.eks. at dø), så er $p/(1-p)$ forholdet mellem sandsynligheden for hændelsen og sandsynligheden for den modsatte hændelse; dette tal kaldes med et udtryk hentet fra spillebranchen for *odds* for den pågældende hændelse. Vi kan dermed sige at logit-funktionen udregner logaritmen til odds.

Vi vil foreslå/postulere følgende ofte anvendte model for sammenhængen mellem dosis og sandsynligheden for at dø: For hvert af de to køn afhænger $\operatorname{logit}(p_d)$ lineært af $x = \ln d$, eller mere udførligt: Der findes fire konstanter α_M, β_M og α_F, β_F således at

$$\operatorname{logit}(p_{dM}) = \alpha_M + \beta_M \ln d$$
$$\operatorname{logit}(p_{dF}) = \alpha_F + \beta_F \ln d.$$

En sådan model kaldes en *logistisk regressionsmodel*.

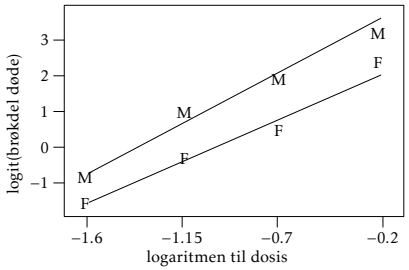
I figur 9.3 er logit til de relative hyppigheder afsat mod logaritmen til dosis; hvis modellen er rigtig, skal hvert af de to punktsæt fordele sig tilfældigt omkring en ret linje, og det ser jo ikke helt urimeligt ud; der behøves dog en nærmere undersøgelse for at afgøre om modellen giver en tilstrækkeligt god beskrivelse af datamaterialet.

I de følgende afsnit skal vi se hvordan man estimerer de ukendte parametre, hvordan man undersøger om modellen er god nok, og hvordan man sammenligner giftens virkning på han- og hunbiller.

Estimation

Likelihoodfunktionen L_0 i den logistiske model fås ved i grundmodellens likelihoodfunktion at opfatte p-erne som funktioner af α-erne og β-erne. Vi får

$$\ln L_0(\alpha_M, \alpha_F, \beta_M, \beta_F)$$



Figur 9.4
Rismelsbiller: Logit til estimeret dødssandsynlighed (relativ hyppighed) som funktion af logaritmen til dosis, for hvert køn, samt de estimerede regressionslinjer.

$$= \text{konst} + \sum_k \sum_d y_{dk}(\alpha_k + \beta_k \ln d) + \sum_k \sum_d n_{dk} \ln(1 - p_{dk})$$
$$= \text{konst} + \sum_k \Big(\alpha_k y_{\cdot k} + \beta_k \sum_d y_{dk} \ln d + \sum_d n_{dk} \ln(1 - p_{dk}) \Big),$$

så log-likelihoodfunktionen består altså af to separate bidrag, et for hanbillerne og et for hunbillerne. Parameterrummet er $\mathbb{R}^4$.

Hvis man differentierer $\ln L_0$ med hensyn til hver af de fire variable og sætter de partielle afledede lig 0, får man efter lidt omskrivning de fire ligninger

$$\sum_d y_{dM} = \sum_d n_{dM} p_{dM}, \qquad \sum_d y_{dF} = \sum_d n_{dF} p_{dF},$$
$$\sum_d y_{dM} \ln d = \sum_d n_{dM} p_{dM} \ln d, \qquad \sum_d y_{dF} \ln d = \sum_d n_{dF} p_{dF} \ln d,$$

hvor $p_{dk} = \dfrac{\exp(\alpha_k + \beta_k \ln d)}{1 + \exp(\alpha_k + \beta_k \ln d)}$. Man kan ikke løse disse ligninger eksplicit, så vi må klare os med en numerisk løsning.

Et almindeligt statistik-computerprogram kan finde følgende estimater: $\widehat{\alpha_M} = 4.27$ (med en standardafvigelse på 0.53) og $\widehat{\beta_M} = 3.14$ (standardafvigelse 0.39), og for hunbillerne er de $\widehat{\alpha_F} = 2.56$ (standardafvigelse 0.38) og $\widehat{\beta_F} = 2.58$ (standardafvigelse 0.30).

Hvis vi i figur 9.3 indtegner de estimerede regressionslinjer, får vi figur 9.4.

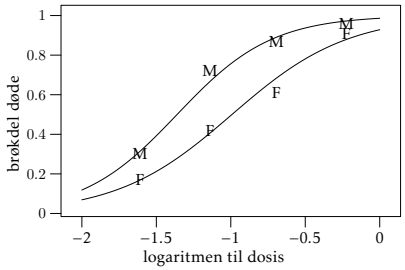
Modelkontrol

Vi har nu estimeret parametrene i den model der siger at

$$\operatorname{logit}(p_{dk}) = \alpha_k + \beta_k \ln d$$

eller

$$p_{dk} = \frac{\exp(\alpha_k + \beta_k \ln d)}{1 + \exp(\alpha_k + \beta_k \ln d)}.$$



Figur 9.5
Rismelsbiller: To forskellige kurver, samt de observerede relative hyppigheder.

En nærliggende form for modelkontrol er derfor at indtegne graferne for de to funktioner

$$x \longmapsto \frac{\exp(\alpha_M + \beta_M x)}{1+\exp(\alpha_M + \beta_M x)} \qquad \text{og} \qquad x \longmapsto \frac{\exp(\alpha_F + \beta_F x)}{1+\exp(\alpha_F + \beta_F x)}$$

i figur 9.2 hvorved man får figur 9.5. Den viser at modellen ikke er helt hen i vejret. Man kan desuden ved hjælp af likelihoodmetoden konstruere et numerisk test baseret på

$$Q = \frac{L(\widehat{\alpha}_M, \widehat{\alpha}_F, \widehat{\beta}_M, \widehat{\beta}_F)}{L_{\max}} \tag{9.1}$$

hvor $L_{\max}$ er likelihoodfunktionens maksimale værdi i grundmodellen (side 140). Med betegnelserne $\widehat{p}_{dk} = \text{logit}^{-1}(\widehat{\alpha}_k + \widehat{\beta}_k \ln d)$ og $\widehat{y}_{dk} = n_{dk}\widehat{p}_{dk}$ bliver

$$\begin{aligned} Q &= \frac{\prod_k \prod_d \binom{n_{dk}}{y_{dk}} \widehat{p}_{dk}^{y_{dk}} (1-\widehat{p}_{dk})^{n_{dk}-y_{dk}}}{\prod_k \prod_d \binom{n_{dk}}{y_{dk}} \left(\frac{y_{dk}}{n_{dk}}\right)^{y_{dk}} \left(1-\frac{y_{dk}}{n_{dk}}\right)^{n_{dk}-y_{dk}}} \\ &= \prod_k \prod_d \left(\frac{\widehat{y}_{dk}}{y_{dk}}\right)^{y_{dk}} \left(\frac{n_{dk}-\widehat{y}_{dk}}{n_{dk}-y_{dk}}\right)^{n_{dk}-y_{dk}} \end{aligned}$$

og

$$-2\ln Q = 2\sum_k \sum_d \left(y_{dk} \ln \frac{y_{dk}}{\widehat{y}_{dk}} + (n_{dk}-y_{dk}) \ln \frac{n_{dk}-y_{dk}}{n_{dk}-\widehat{y}_{dk}} \right).$$

Store værdier af $-2\ln Q$ (eller små værdier af Q) er tegn på at der er for stor uoverensstemmelse mellem de observerede antal (y_{kd} og $n_{kd}-y_{kd}$) og de forudsagte antal ($\widehat{y}_{kd}$ og $n_{kd}-\widehat{y}_{kd}$) til at modellen kan siges at være god nok. I det konkrete tilfælde er $-2\ln Q_{\text{obs}} = 3.36$, og testsandsynligheden, dvs. sandsynligheden (når den testede model er rigtig) for at få en $-2\ln Q$-værdi som er

større end $-2\ln Q_{\text{obs}}$, kan bestemmes ved at udnytte at når modellen er rigtig, er $-2\ln Q$ asymptotisk χ^2-fordelt med $8-4=4$ frihedsgrader; man finder at testsandsynligheden er ca. 0.50. (Antallet af frihedsgrader er antal frie parametre i grundmodellen minus antal frie parametre i den testede model.)

Da der således er henved 50% chance for at få et sæt observationer der harmonerer dårligere med den postulerede model, må vi konkludere at modellen ser ud til at være anvendelig.

Hypoteser om parametrene

Vi har opstillet en model som indeholder fire parametre, og som ser ud til at give en ganske god beskrivelse af observationerne. Næste punkt på dagsordenen er at undersøge om modellen kan forsimples.

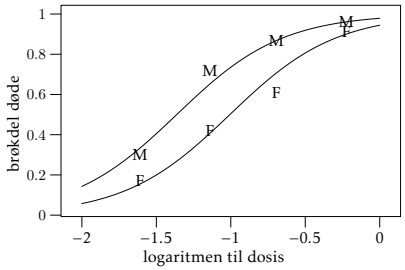
Eksempelvis kan man undersøge om de to kurver er parallelle, og hvis det kan accepteres, kan man derefter undersøge om kurverne er sammenfaldende. Vi formulerer derfor to statistiske hypoteser:

1. Hypotesen om parallelle kurver: $H_1 : \beta_M = \beta_F$, eller mere udførligt: Der findes konstanter α_M, α_F og β således at
$$\text{logit}(p_{dM}) = \alpha_M + \beta \ln d \qquad \text{og} \qquad \text{logit}(p_{dF}) = \alpha_F + \beta \ln d.$$
2. Hypotesen om sammenfaldende kurver: $H_2 : \alpha_M = \alpha_F$ og $\beta_M = \beta_F$, eller mere udførligt: Der findes konstanter α og β således at
$$\text{logit}(p_{dM}) = \alpha + \beta \ln d \qquad \text{og} \qquad \text{logit}(p_{dF}) = \alpha + \beta \ln d.$$

Vi undersøger først hypotesen H_1 om parallelle kurver. Maksimaliseringsestimaterne er $\widehat{\widehat{\alpha}}_M = 3.84$ (med en standardafvigelse på 0.34), $\widehat{\widehat{\alpha}}_F = 2.83$ (standardafvigelse 0.31) og $\widehat{\widehat{\beta}} = 2.81$ (standardafvigelse 0.24). Hypotesen testes med det sædvanlige kvotienttest hvor man sammenligner den maksimale likelihoodfunktion under antagelse af H_1 med den maksimale likelihoodfunktion i den senest accepterede model:

$$\begin{aligned} -2\ln Q &= -2\ln \frac{L(\widehat{\widehat{\alpha}}_M, \widehat{\widehat{\alpha}}_F, \widehat{\widehat{\beta}}, \widehat{\widehat{\beta}})}{L(\widehat{\alpha}_M, \widehat{\alpha}_F, \widehat{\beta}_M, \widehat{\beta}_F)} \\ &= 2\sum_k \sum_d \left(y_{dk} \ln \frac{\widehat{y}_{dk}}{\widehat{\widehat{y}}_{dk}} + (n_{dk}-y_{dk}) \ln \frac{n_{dk}-\widehat{y}_{dk}}{n_{dk}-\widehat{\widehat{y}}_{dk}} \right) \end{aligned}$$

hvor $\widehat{\widehat{y}}_{dk} = n_{dk} \dfrac{\exp(\widehat{\widehat{\alpha}}_k + \widehat{\widehat{\beta}} \ln d)}{1+\exp(\widehat{\widehat{\alpha}}_k + \widehat{\widehat{\beta}} \ln d)}$. Man får at $-2\ln Q_{\text{obs}} = 1.31$, der skal sammenlignes med χ^2-fordelingen med $4-3=1$ frihedsgrader (ændring i antal



Figur 9.6
Rismelsbiller:
Den endelige model.

parametre). Testsandsynligheden (ɔ: sandsynligheden for at få værdier større end 1.31) er ca. 25%, så værdien 1.31 er ikke usædvanligt stor. Modellen med parallelle kurver giver således ikke en signifikant dårligere beskrivelse af observationerne end den hidtidige model gør.

Efter at have accepteret hypotesen H_1 kan vi gå videre med hypotesen H_2 om sammenfaldende kurver. (Hvis H_1 var blevet forkastet, ville man ikke gå videre til H_2.) Når man tester H_2 i forhold til H_1, får man $-2\ln Q_{\text{obs}} = 27.50$ der skal sammenlignes med χ^2-fordelingen med et antal frihedsgrader på $3-2=1$; sandsynligheden for at få værdier større end 27.50 er lig nul med adskillige betydende cifre, hvilket viser at modellen med sammenfaldende kurver giver en væsentligt dårligere beskrivelse af observationerne end den forrige model gør. Vi må derfor forkaste hypotesen om sammenfaldende kurver.

Konklusionen på det hele er, at vi kan beskrive sammenhængen mellem dosis d og sandsynligheden p for at dø på den måde at for hvert køn afhænger logit p lineært af $\ln d$; de to kurver er parallelle, men ikke sammenfaldende. De estimerede kurver er

$$\text{logit}(p_{dM}) = 3.84 + 2.81\ln d \qquad \text{og} \qquad \text{logit}(p_{dF}) = 2.83 + 2.81\ln d\,,$$

svarende til at

$$p_{dM} = \frac{\exp(3.84+2.81\ln d)}{1+\exp(3.84+2.81\ln d)} \qquad \text{og} \qquad p_{dF} = \frac{\exp(2.83+2.81\ln d)}{1+\exp(2.83+2.81\ln d)}\,.$$

Figur 9.6 illustrerer situationen.

9.2 Lungekræft i Fredericia

Dette er et eksempel på en såkaldt multiplikativ poissonmodel. I øvrigt er eksemplet interessant på den måde at man tilsyneladende kan nå frem til modstridende konklusioner blot ved at ændre en smule på fremgangsmåden ved analysen af modellen.

Situationen

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 Fredericia havde en betydelig mængde forurenende industri midt inde i byen. For at belyse spørgsmålet indsamlede man data om lungekræfthyppigheden i perioden 1968-71, dels for Fredericia, dels for byerne Horsens, Kolding og Vejle. De tre sidste byer skulle tjene som sammenligningsgrundlag, idet det på nær den mistænkte industri var byer af nogenlunde samme art som Fredericia.

Lungekræft opstår tit som et resultat af daglige påvirkninger af skadelige stoffer gennem mange år. En eventuel større risiko i Fredericia kunne måske derfor vise sig ved at lungekræftpatienterne fra Fredericia var yngre end dem fra kontrolbyerne; desuden er det 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 er vist i tabel 9.2. Da antallene af lungekræfttilfælde i sig selv ikke siger noget så længe man ikke kender risikogruppernes størrelse, må man også rapportere antal indbyggere i de forskellige aldersklasser og byer, se tabel 9.3.

Det er nu statistikerens opgave at beskrive tallene i tabel 9.2 ved hjælp af en statistisk model hvori der indgår 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 »byvirkninger« (dvs. forskelle mellem byer) efter at man på en eller anden måde havde taget højde for forskellene mellem aldersgrupperne.

Modelopstilling

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 opfattes som observerede værdier af stokastiske variable, og det er fordelingen af disse stokastiske

Tabel 9.2
Lungekræfttilfælde i fire byer fordelt på aldersklasser (fra Andersen, 1977).

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

variable der skal specificeres af den statistiske model. Vi indfører noget notation:

$$y_{ij} = \text{antal tilfælde i aldersgruppe } i \text{ i by } j,$$
$$r_{ij} = \text{antal personer i aldersgruppe } i \text{ i by } j,$$

hvor $i = 1, 2, 3, 4, 5, 6$ nummererer aldersgrupperne, og $j = 1, 2, 3, 4$ nummererer byerne. Observationerne y_{ij} opfattes som observerede værdier af stokastiske variable Y_{ij}.

Hvilken fordeling skal Y-erne have? I første omgang ville man måske foreslå at Y_{ij} skulle være binomialfordelt med antalsparameter r_{ij}; nu er sandsynligheden for at få lungekræft temmelig lille, så med henvisning til sætningen om approksimation af binomialfordelinger med poissonfordelinger (sætning 2.15 side 60) vil vi i stedet foreslå at Y_{ij} skal være poissonfordelt med en parameter μ_{ij} der afhænger af aldersgruppe og by (modellen skal ikke indeholde observationsperiodens længde da denne er konstant lig 4 år). Hvis vi skriver μ_{ij} som $\mu_{ij} = \lambda_{ij} r_{ij}$, så kan intensiteten λ_{ij} fortolkes som »antal lungekræfttilfælde pr. person i aldersgruppe i i by j i den betragtede fireårsperiode«, dvs. λ er den *alders- og byspecifikke cancer-incidens*. Endvidere vil vi gå ud fra at de enkelte Y_{ij}-er er stokastisk uafhængige. Vi får dermed følgende grundmodel:

> De stokastiske variable Y_{ij} er stokastisk uafhængige og poissonfordelte således at Y_{ij} har parameter $\lambda_{ij} r_{ij}$ hvor λ_{ij}-erne er ukendte positive parametre.

Det er let 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). Den generelle opskrift er $\widehat{\lambda}_{ij} = y_{ij}/r_{ij}$.

Nu var det jo tanken at vi gerne ville blive i stand til at sammenligne byerne efter at have taget højde for deres forskellige aldersfordelinger, 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, nemlig en model hvor λ_{ij} er

Tabel 9.3
Antal indbyggere i de forskellige aldersklasser i de fire byer (fra Andersen, 1977).

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

spaltet op i et produkt $\alpha_i \beta_j$ af en *aldersvirkning* α_i og en *byvirkning* β_j. Hvis dette lader sig gøre, er vi heldigt stillede, for så kan vi sammenligne byerne ved at sammenligne byparametrene β_j. Vi vil derfor i første omgang teste den statistiske hypotese

$$H_0 : \lambda_{ij} = \alpha_i \beta_j$$

hvor $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_1, \beta_2, \beta_3, \beta_4$ er ukendte parametre. Mere udførligt er hypotesen at findes parameterværdier $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_1, \beta_2, \beta_3, \beta_4 \in \mathbb{R}_+$ således at der for by j og aldersgruppe i gælder at lungekræftrisikoen λ_{ij} fås som $\lambda_{ij} = \alpha_i \beta_j$. – Den model som H_0 specificerer, er en *multiplikativ* model, fordi aldersparametre og byparametre indgår multiplikativt.

En detalje vedrørende parametriseringen

Der er det særlige ved parametriseringen af modellen under H_0 at den ikke er injektiv: afbildningen

$$(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_1, \beta_2, \beta_3, \beta_4) \longmapsto (\alpha_i \beta_j)_{i=1,\ldots,6;\ j=1,\ldots,4}$$

fra $\mathbb{R}_+^{10}$ til $\mathbb{R}^{24}$ er ikke injektiv, idet billedmængden er faktisk kun en 9-dimensional flade. (Man overbeviser sig let om at $\alpha_i \beta_j = \alpha_i^* \beta_j^*$ for alle i og j, hvis og kun hvis der findes et $c > 0$ således at $\alpha_i = c\,\alpha_i^*$ for alle i og $\beta_j^* = c\,\beta_j$ for alle j.)

De 10 parametre skal derfor pålægges ét bånd for at få en injektiv parametrisering. Et sådant bånd kan være at $\alpha_1 = 1$, eller at $\alpha_1 + \alpha_2 + \cdots + \alpha_6 = 1$, eller at $\alpha_1 \alpha_2 \ldots \alpha_6 = 1$, eller det tilsvarende for β, osv. I det aktuelle eksempel vil vi benytte betingelsen $\beta_1 = 1$, dvs. vi fikserer Fredericia-parameteren til 1. Modellen indeholder herefter ni ukendte parametre.

Estimation i den multiplikative model

I den multiplikative model lader det sig ikke gøre at opskrive eksplicitte udtryk for estimaterne, man er henvist til i de konkrete tilfælde at benytte numeriske

metoder, f.eks. i form af hensigtsmæssige computerprogrammer. Likelihoodfunktionen i grundmodellen er

$$L = \prod_{i=1}^{6}\prod_{j=1}^{4} \frac{(\lambda_{ij} r_{ij})^{y_{ij}}}{y_{ij}!} \exp(-\lambda_{ij} r_{ij}) = \text{konst} \cdot \prod_{i=1}^{6}\prod_{j=1}^{4} \lambda_{ij}^{y_{ij}} \exp(-\lambda_{ij} r_{ij}).$$

Når man maksimaliserer med hensyn til λ_{ij}-erne, får man som allerede nævnt at maksimaliseringsestimaterne i grundmodellen er $\widehat{\lambda}_{ij} = y_{ij}/r_{ij}$.

Når vi i udtrykket for L erstatter λ_{ij} med $\alpha_i\beta_j$, får vi likelihoodfunktionen L_0 under H_0:

$$\begin{aligned} L_0 &= \text{konst} \cdot \prod_{i=1}^{6}\prod_{j=1}^{4} \alpha_i^{y_{ij}} \beta_j^{y_{ij}} \exp(-\alpha_i\beta_j r_{ij}) \\ &= \text{konst} \cdot \Big(\prod_{i=1}^{6} \alpha_i^{y_{i\bullet}}\Big)\Big(\prod_{j=1}^{4} \beta_j^{y_{\bullet j}}\Big) \exp\Big(-\sum_{i=1}^{6}\sum_{j=1}^{4} \alpha_i\beta_j r_{ij}\Big), \end{aligned}$$

og log-likelihoodfunktionen er

$$\ln L_0 = \text{konst} + \Big(\sum_{i=1}^{6} y_{i\bullet} \ln \alpha_i\Big) + \Big(\sum_{j=1}^{4} y_{\bullet j} \ln \beta_j\Big) - \sum_{i=1}^{6}\sum_{j=1}^{4} \alpha_i\beta_j r_{ij}.$$

Vi skal bestemme det parametersæt $(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3, \widehat{\alpha}_4, \widehat{\alpha}_5, \widehat{\alpha}_6, 1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4)$ der maksimaliserer $\ln L_0$. De partielle afledede er

$$\begin{aligned} \frac{\partial \ln L_0}{\partial \alpha_i} &= \frac{y_{i\bullet}}{\alpha_i} - \sum_j \beta_j r_{ij}, \quad i = 1,2,3,4,5,6 \\ \frac{\partial \ln L_0}{\partial \beta_j} &= \frac{y_{\bullet j}}{\beta_j} - \sum_i \alpha_i r_{ij}, \quad j = 1,2,3,4. \end{aligned}$$

Det ses at de partielle afledede er lig 0 hvis og kun hvis

$$y_{i\bullet} = \sum_j \alpha_i\beta_j r_{ij} \quad \text{for alle } i \qquad \text{og} \qquad y_{\bullet j} = \sum_i \alpha_i\beta_j r_{ij} \quad \text{for alle } j.$$

Vi noterer til senere brug at der heraf følger at

$$y_{\bullet\bullet} = \sum_i \sum_j \widehat{\alpha}_i\widehat{\beta}_j r_{ij}. \tag{9.2}$$

Man finder følgende maksimaliseringsestimater i den multiplikative model:

$$\begin{aligned} \widehat{\alpha}_1 &= 0.0036 & \widehat{\beta}_1 &= 1 \\ \widehat{\alpha}_2 &= 0.0108 & \widehat{\beta}_2 &= 0.719 \\ \widehat{\alpha}_3 &= 0.0164 & \widehat{\beta}_3 &= 0.690 \\ \widehat{\alpha}_4 &= 0.0210 & \widehat{\beta}_4 &= 0.762 \\ \widehat{\alpha}_5 &= 0.0229 & & \\ \widehat{\alpha}_6 &= 0.0148. & & \end{aligned}$$

Den multiplikative models beskrivelse af data

Herefter vil vi undersøge hvor god en beskrivelse den multiplikative model faktisk giver af datamaterialet. Det vil vi gøre ved at teste multiplikativitetshypotesen H_0 i forhold til grundmodellen, og det foregår med et almindeligt kvotienttest: Man udregner

$$-2\ln Q = -2\ln \frac{L_0(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3, \widehat{\alpha}_4, \widehat{\alpha}_5, \widehat{\alpha}_6, 1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4)}{L(\widehat{\lambda}_{11}, \widehat{\lambda}_{12}, \ldots, \widehat{\lambda}_{63}, \widehat{\lambda}_{64})}.$$

Store værdier af $-2\ln Q$ er signifikante, dvs. 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 $\varepsilon = \mathrm{P}_0\big(-2\ln Q \geq -2\ln Q_{\text{obs}}\big)$, altså sandsynligheden for at få en større $-2\ln Q$-værdi forudsat at H_0 er rigtig. Når H_0 er rigtig, er $-2\ln Q$ approksimativt χ^2-fordelt med $f = 24 - 9 = 15$ frihedsgrader (forudsat at de forventede antal alle er mindst 5). Det betyder at testsandsynligheden kan udregnes som $\varepsilon = \mathrm{P}\big(\chi^2_{15} \geq -2\ln Q_{\text{obs}}\big)$.

Ved almindelige omskrivninger, hvor man undervejs skal benytte (9.2), finder man at

$$Q = \prod_{i=1}^{6}\prod_{j=1}^{4} \Big(\frac{\widehat{y}_{ij}}{y_{ij}}\Big)^{y_{ij}},$$

og dermed

$$-2\ln Q = 2\sum_{i=1}^{6}\sum_{j=1}^{4} y_{ij} \ln \frac{y_{ij}}{\widehat{y}_{ij}}.$$

hvor $\widehat{y}_{ij} = \widehat{\alpha}_i\widehat{\beta}_j r_{ij}$ er det forventede antal lungekræfttilfælde i aldersklasse i i by j.

Som led i beregningerne af $\widehat{y}_{ij}$ udregnes de estimerede alders- og by-specifikke lungekræftintensiteter $\widehat{\alpha}_i\widehat{\beta}_j$. Værdierne af $1000\,\widehat{\alpha}_i\widehat{\beta}_j$, dvs. de forventede antal

Tabel 9.4
Estimerede alders- og byspecifikke lungekræftintensiteter i perioden 1968-71 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 9.5
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

tilfælde pr. 1000 indbyggere, ses i tabel 9.4. Selve de forventede antal $\widehat{y}_{ij}$ i de forskellige byer og aldersklasser ses i tabel 9.5, og den konkrete værdi af $-2\ln Q$ bliver $-2\ln Q_{\text{obs}} = 22.6$. 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, ɔ: *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 (β-erne) og nogle alders-parametre (α-erne), men uden parametre svarende til en vekselvirkning mellem by og alder. Det betyder at den forskel der er mellem byerne, er den samme for alle aldersklasser, og at den forskel der er 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 undersøge om der er signifikant forskel på byerne. Hvis der ikke er nogen forskel på byerne, er byparametrene ens, dvs. $\beta_1 = \beta_2 = \beta_3 = \beta_4$, og da $\beta_1 = 1$, må den fælles værdi være 1. Vi vil derfor 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

$$-2\ln Q = -2\ln \frac{L_1(\widehat{\widehat{\alpha}}_1, \widehat{\widehat{\alpha}}_2, \widehat{\widehat{\alpha}}_3, \widehat{\widehat{\alpha}}_4, \widehat{\widehat{\alpha}}_5, \widehat{\widehat{\alpha}}_6)}{L_0(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3, \widehat{\alpha}_4, \widehat{\alpha}_5, \widehat{\alpha}_6, 1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4)}$$

hvor $L_1(\alpha_1, \alpha_2, \ldots, \alpha_6) = L_0(\alpha_1, \alpha_2, \ldots, \alpha_6, 1, 1, 1, 1)$ er likelihoodfunktionen under H_1, og $\widehat{\widehat{\alpha}}_1, \widehat{\widehat{\alpha}}_2, \ldots, \widehat{\widehat{\alpha}}_6$ er estimaterne under H_1, dvs. $(\widehat{\widehat{\alpha}}_1, \widehat{\widehat{\alpha}}_2, \ldots, \widehat{\widehat{\alpha}}_6)$ er maksimumspunktet for L_1.

Funktionen L_1 er et produkt af seks funktioner med hver sit α:

$$\begin{aligned} L_1(\alpha_1, \alpha_2, \ldots, \alpha_6) &= \text{konst} \cdot \prod_{i=1}^{6} \prod_{j=1}^{4} \alpha_i^{y_{ij}} \exp(-\alpha_i r_{ij}) \\ &= \text{konst} \cdot \prod_{i=1}^{6} \alpha_i^{y_{i\cdot}} \exp(-\alpha_i r_{i\cdot}) \,. \end{aligned}$$

Maksimaliseringsestimaterne findes derfor til $\widehat{\widehat{\alpha}}_i = \dfrac{y_{i\cdot}}{r_{i\cdot}}$, $i = 1, 2, \ldots, 6$. Talværdierne bliver

$$\begin{aligned} \widehat{\widehat{\alpha}}_1 &= 33/11600 = 0.002845 & \widehat{\widehat{\alpha}}_4 &= 45/2748 = 0.0164 \\ \widehat{\widehat{\alpha}}_2 &= 32/3811 \;\; = 0.00840 & \widehat{\widehat{\alpha}}_5 &= 40/2217 = 0.0180 \\ \widehat{\widehat{\alpha}}_3 &= 43/3367 \;\; = 0.0128 & \widehat{\widehat{\alpha}}_6 &= 31/2665 = 0.0116. \end{aligned}$$

Almindelige omskrivninger giver at

$$Q = \frac{L_1(\widehat{\widehat{\alpha}}_1, \widehat{\widehat{\alpha}}_2, \widehat{\widehat{\alpha}}_3, \widehat{\widehat{\alpha}}_4, \widehat{\widehat{\alpha}}_5, \widehat{\widehat{\alpha}}_6)}{L_0(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3, \widehat{\alpha}_4, \widehat{\alpha}_5, \widehat{\alpha}_6, 1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4)} = \prod_{i=1}^{6} \prod_{j=1}^{4} \left(\frac{\widehat{\widehat{y}}_{ij}}{\widehat{y}_{ij}} \right)^{y_{ij}}$$

og dermed

$$-2\ln Q = 2 \sum_{i=1}^{6} \sum_{j=1}^{4} y_{ij} \ln \frac{\widehat{y}_{ij}}{\widehat{\widehat{y}}_{ij}} \,;$$

her er $\widehat{\widehat{y}}_{ij} = \widehat{\widehat{\alpha}}_i r_{ij}$ og som før $\widehat{y}_{ij} = \widehat{\alpha}_i \widehat{\beta}_j r_{ij}$. De forventede antal $\widehat{\widehat{y}}_{ij}$ er vist i tabel 9.6.

Store værdier af $-2\ln Q$ er signifikante. Man skal sammenholde $-2\ln Q$ med χ^2-fordelingen med $f = 9-6 = 3$ frihedsgrader.

Man finder at $-2\ln Q_{\text{obs}} = 5.67$. I χ^2-fordelingen med $f = 9-6 = 3$ frihedsgrader er 80%-fraktilen 4.64 og 90%-fraktilen 6.25, således at testsandsynligheden er næsten 20%. De foreliggende observationer er altså udmærket 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*.

Tabel 9.6
De forventede antal $\widehat{y}_{ij}$ 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

En anden mulighed

Det er sjældent sådan at der kun er én bestemt måde at undersøge en praktisk problemstilling på ved hjælp af en statistisk model og nogle statistiske hypoteser. Det aktuelle spørgsmål om der er en øget risiko for lungekræft ved at bo i Fredericia, blev i forrige afsnit belyst ved at vi testede hypotesen H_1 om ens byparametre. Det viste sig at H_1 kunne accepteres, og man kan således sige at der ikke er nogen signifikant forskel på de fire byer.

Nu kunne man imidlertid angribe problemet på en anden måde. Man kunne sige at det hele drejer sig om at vurdere om det er farligere at bo i Fredericia end i de øvrige byer. Dermed er det indirekte forudsat at de tre øvrige byer stort set er ens, hvilket man kan og bør teste. Man kunne derfor anlægge følgende strategi for formulering og test af hypoteser:

1. Vi benytter stadig 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 tester hypotesen
$$H_2 : \beta_2 = \beta_3 = \beta_4.$$
3. Hvis H_2 bliver accepteret, er der et fælles niveau β for de tre »kontrolbyer«. Vi kan derefter sammenligne Fredericia med dette fælles niveau ved at teste om $\beta_1 = \beta$. Da β_1 pr. definition er lig 1, er den hypotese der skal testes,
$$H_3 : \beta = 1.$$

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 teststørrelsen

$$Q = \frac{L_2(\widetilde{\alpha}_1, \widetilde{\alpha}_2, \ldots, \widetilde{\alpha}_6, \widetilde{\beta})}{L_0(\widehat{\alpha}_1, \widehat{\alpha}_2, \ldots, \widehat{\alpha}_6, 1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4)}$$

Tabel 9.7
De forventede antal $\widetilde{y}_{ij}$ 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

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 $\widetilde{\alpha}_1, \widetilde{\alpha}_2, \ldots, \widetilde{\alpha}_6, \widetilde{\beta}$ er maksimaliseringsestimaterne under H_2.

Når H_2 er rigtig, er $-2\ln Q$ approksimativt 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 ingen principielt nye problemer forbundet med at estimere parametrene under H_2. Man finder

$$\begin{array}{lll} \widetilde{\alpha}_1 = 0.00358 & \widetilde{\alpha}_4 = 0.0210 & \widetilde{\beta}_1 = 1 \\ \widetilde{\alpha}_2 = 0.0108 & \widetilde{\alpha}_5 = 0.0230 & \widetilde{\beta} = 0.7220 \\ \widetilde{\alpha}_3 = 0.0164 & \widetilde{\alpha}_6 = 0.0148 & \end{array}$$

Endvidere bliver

$$-2\ln Q = 2\sum_{i=1}^{6}\sum_{j=1}^{4} y_{ij} \ln \frac{\widehat{y}_{ij}}{\widetilde{y}_{ij}}$$

hvor $\widehat{y}_{ij} = \widehat{\alpha}_i \widehat{\beta}_j r_{ij}$, se tabel 9.5, og

$$\begin{aligned} \widetilde{y}_{i1} &= \widetilde{\alpha}_i \, r_{i1} \\ \widetilde{y}_{ij} &= \widetilde{\alpha}_i \widetilde{\beta} \, r_{ij}, \quad j = 2, 3, 4. \end{aligned}$$

De forventede antal $\widetilde{y}_{ij}$ ses i tabel 9.7. Man finder at $-2\ln Q_{\text{obs}} = 0.40$; denne værdi skal sammenholdes med χ^2-fordelingen med $f = 9 - 7 = 2$ frihedsgrader. I denne fordeling 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 sagtens tillade os at gå ud fra at der ikke er nogen signifikant forskel mellem de tre byer.

Herefter kan vi gå over til at teste H_3, der siger at alle fire byer er ens, og at der er de seks forskellige aldersgrupper med hver sin parameter α_i. Under forudsætning af H_2 er H_3 identisk med hypotesen H_1 fra tidligere, så estimaterne over aldersparametrene er $\widehat{\widehat{\alpha}}_1, \widehat{\widehat{\alpha}}_2, \ldots, \widehat{\widehat{\alpha}}_6$ fra side 153.

I denne omgang skal vi teste H_3 (= H_1) i forhold til den nu gældende grundmodel H_2. Teststørrelsen er $-2\ln Q$ hvor

$$Q = \frac{L_1(\widehat{\widetilde{\alpha}}_1, \widehat{\widetilde{\alpha}}_2, \ldots, \widehat{\widetilde{\alpha}}_6)}{L_2(\widetilde{\alpha}_1, \widetilde{\alpha}_2, \ldots, \widetilde{\alpha}_6, \widetilde{\beta})} = \frac{L_0(\widehat{\widetilde{\alpha}}_1, \widehat{\widetilde{\alpha}}_2, \ldots, \widehat{\widetilde{\alpha}}_6, 1, 1, 1, 1)}{L_0(\widetilde{\alpha}_1, \widetilde{\alpha}_2, \ldots, \widetilde{\alpha}_6, 1, \widetilde{\beta}, \widetilde{\beta}, \widetilde{\beta})}$$

der let omformes til

$$Q = \prod_{i=1}^{6} \prod_{j=1}^{4} \left(\frac{\widehat{\widetilde{y}}_{ij}}{\widetilde{\widetilde{y}}_{ij}} \right)^{y_{ij}}$$

så at

$$-2\ln Q = 2 \sum_{i=1}^{6} \sum_{j=1}^{4} y_{ij} \ln \frac{\widetilde{y}_{ij}}{\widehat{\widetilde{y}}_{ij}}.$$

Store værdier af $-2\ln Q$ er signifikante. Når H_3 er rigtig, er $-2\ln Q$ approksimativt χ^2-fordelt med $f = 7 - 6 = 1$ frihedsgrad (forudsat at alle de indgående forventede antal er mindst fem).

Ved at indsætte værdierne fra tabel 9.2, tabel 9.6 og tabel 9.7 i det seneste udtryk for $-2\ln Q$ 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 $\widetilde{\beta} = 0.7$, så lungekræfthyppigheden i Fredericia er altså signifikant større end i kontrolbyerne.

Se det var jo en pæn og klar konklusion, der blot er stik modsat den vi nåede frem til på side 153!

Sammenligning af de to fremgangsmåder

Vi har benyttet to fremgangsmåder der kun var en smule forskellige, men gav helt modsatte resultater. 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, har vi derved fået en ny aktuel grundmodel (nemlig den gamle med de simplifikationer som den accepterede hypotese giver). Fortsæt med punkt 2.

Model/Hypotese		$-2\ln Q$	f	ε
M:	vilkårlige parametre	22.65	24 – 9 = 15	godt 5%
H:	multiplikativitet			
M:	multiplikativitet	5.67	9 – 6 = 3	ca. 20%
H:	fire ens byer			

Oversigt over den første fremgangsmåde.

Model/Hypotese		$-2\ln Q$	f	ε
M:	vilkårlige parametre	22.65	24 – 9 = 15	godt 5%
H:	multiplikativitet			
M:	multiplikativitet	0.40	9 – 7 = 2	godt 80%
H:	de tre byer ens			
M:	de tre byer ens	5.27	7 – 6 = 1	ca. 2%
H:	de fire byer ens			

Oversigt over den anden fremgangsmåde.

   b) Hvis hypotesen forkastes, så slut. Data beskrives da ved den senest anvendte grundmodel.

Begge de anvendte fremgangsmåder tager udgangspunkt i den samme poissonmodel, de adskiller sig udelukkende ved valgene af hypoteser i punkt 2. I den første fremgangsmåde tages skridtet fra den multiplikative model til »fire ens« på én gang, hvilket giver en teststørrelse på 5.67, som, da den kan fordeles på 3 frihedsgrader, ikke er signifikant. I den anden fremgangsmåde spalter vi op i de to skridt »multiplikativitet → tre ens« og »tre ens → fire ens«, og det viser sig da 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 afsnittene 9.1 og 9.2. De er alle af formen

$$-2\ln Q = 2 \sum \text{obs.antal} \cdot \ln \frac{\text{modellens forventede antal}}{\text{hypotesens forventede antal}}$$

og er (tilnærmelsesvis) χ^2-fordelt med et antal frihedsgrader som er »antal frie parametre under modellen« minus »antal frie parametre under hypotesen«. Dette gælder faktisk helt generelt når man tester hypoteser om poissonfordelte observationer (dog under forudsætning af at summen af de forventede antal er lig summen af de observerede antal).

9.3 Ulykker på en granatfabrik

I dette eksempel virker det oplagt at forsøge sig med en poissonfordelingsmodel. Det viser sig dog at den ikke passer særlig godt, så man må finde på noget andet.

Situationen

Man har undersøgt hvor mange ulykkestilfælde hver enkelt arbejder på en granatfabrik i England kom ud for i løbet af en fem ugers periode. Det hele foregik under første verdenskrig, så de pågældende arbejdere var kvinder (mens mændene var soldater). I tabel 9.10 ses fordelingen af $n = 647$ kvinder efter antallet y af ulykkestilfælde i en fem ugers periode. Man søger en statistisk model der kan beskrive dette talmateriale. (Eksemplet stammer fra Greenwood and Yule (1920) og er her i landet især kendt via sin forekomst i Hald (1948, 1968) der gennem mere end en menneskealder har været en toneangivende dansk lærebog i statistik.)

Lad y_i betegne antal ulykker som kvinde nr. i kommer ud for. Vi benytter betegnelsen f_y for antallet af kvinder der har været ude for netop y ulykker, dvs. i det foreliggende tilfælde er $f_0 = 447$, $f_1 = 132$, osv. Det samlede antal ulykker er $0f_0 + 1f + 2f_2 + \ldots = \sum_{y=0}^{\infty} y f_y = 301$. Vi går ud fra at y_i er observation af en stokastisk variabel Y_i, $i = 1, 2, \ldots, n$, og vi vil antage at de stokastiske variable $Y_1, Y_2, \ldots, Y_n$ er indbyrdes uafhængige, omend dette måske er en lidt diskutabel antagelse.

Den første model

I første omgang kan man forsøge sig med en model gående ud på at $Y_1, Y_2, \ldots, Y_n$ er uafhængige og identisk poissonfordelte med parameter μ, således at

$$\mathrm{P}(Y_i = y) = \frac{\mu^y}{y!}\exp(-\mu).$$

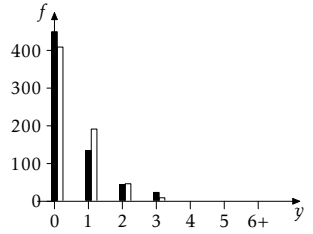
Poissonfordelingen kommer ind i billedet ud fra en forestilling om at ulykkerne sker »helt tilfældigt«, og man kan sige at parameteren μ beskriver kvindernes »ulykkestilbøjelighed«.

Poissonparameteren μ estimeres ved $\widehat{\mu} = \overline{y} = 301/647 = 0.465$ (der sker 0.465 ulykker pr. kvinde pr. fem uger). Det forventede antal kvinder med y ulykker er $\widehat{f_y} = n\frac{\widehat{\mu}^y}{y!}\exp(-\widehat{\mu})$; værdierne heraf ses i tabel 9.11. Der er tilsyneladende ikke nogen særlig god overensstemmelse mellem de observerede og de forventede antal. Man kan udregne det centrale variansestimat $s^2 = \frac{1}{n-1}\sum(y_i - \overline{y})^2$ til $s^2 =$

y	antal kvinder med y ulykker
0	447
1	132
2	42
3	21
4	3
5	2
6+	0
	647

Tabel 9.10
Fordelingen af 647 kvinder efter antallet y af ulykkestilfælde i en fem ugers periode (fra Greenwood and Yule (1920)).

y	f_y	$\widehat{f_y}$
0	447	406.3
1	132	189.0
2	42	44.0
3	21	6.8
4	3	0.8
5	2	0.1
6+	0	0.0
	647	647.0



Tabel 9.11
Model 1: Tabel og graf med de observerede antal f_y (sorte søjler) og de forventede antal $\widehat{f_y}$ (hvide søjler).

0.692, og det er næsten halvanden gange $\overline{y}$; da poissonfordelingen har den egenskab at middelværdien er lig variansen, har vi således endnu et tegn på at poissonmodellen er dårlig. Man må derfor overveje en anden model.

Den anden model

Man kan udvide model 1 på følgende måde:

- Det antages stadig at $Y_1, Y_2, \ldots, Y_n$ er uafhængige og poissonfordelte, men nu tillader vi at de har hver sin middelværdi, så nu er Y_i poissonfordelt med parameter μ_i, $i = 1, 2, \ldots, n$.
  Hvis modelopstillingen gjorde holdt her, ville der være en parameter for hver person; derved kunne man opnå et perfekt fit (med estimaterne $\widehat{\mu_i} = y_i$, $i = 1, 2, \ldots, n$), men man ville i høj grad være i strid med Fisher's maksime om at statistikkens formål er datareduktion (jf. side 100). Men der endnu et trin i modelopbygningen:
- Det antages endvidere at $\mu_1, \mu_2, \ldots, \mu_n$ er uafhængige observationer fra en og samme sandsynlighedsfordeling. Denne sandsynlighedsfordeling skal være en kontinuert fordeling på den positive halvakse, og det viser sig

bekvemt at benytte en gammafordeling med, lad os sige, formparameter κ og skalaparameter β, altså med tæthedsfunktion

$$g(\mu) = \frac{1}{\Gamma(\kappa)\beta^\kappa}\,\mu^{\kappa-1}\exp(-\mu/\beta), \qquad \mu > 0.$$

- Den betingede sandsynlighed for at en kvinde kommer ud for netop y ulykker, givet at hun har et bestemt μ, er $\dfrac{\mu^y}{y!}\exp(-\mu)$. Den ubetingede sandsynlighed fås ved at blande de betingede sandsynligheder med hensyn til μ's fordeling:

$$\begin{aligned}\mathrm{P}(Y=y) &= \int_0^{+\infty} \frac{\mu^y}{y!}\exp(-\mu)\cdot g(\mu)\,d\mu \\ &= \int_0^{+\infty} \frac{\mu^y}{y!}\exp(-\mu)\frac{1}{\Gamma(\kappa)\beta^\kappa}\,\mu^{\kappa-1}\exp(-\mu/\beta)\,d\mu \\ &= \frac{\Gamma(y+\kappa)}{y!\,\Gamma(\kappa)}\left(\frac{1}{\beta+1}\right)^\kappa\left(\frac{\beta}{\beta+1}\right)^y,\end{aligned}$$

hvor det sidste lighedstegn følger af definitionen på gammafunktionen (benyt f.eks. formlen sidst i sidebemærkningen side 72). Hvis κ er et naturligt tal, er $\Gamma(\kappa) = (\kappa-1)!$, og så er

$$\binom{y+\kappa-1}{y} = \frac{\Gamma(y+\kappa)}{y!\,\Gamma(\kappa)}.$$

Her er højresiden imidlertid defineret for alle $\kappa > 0$, og vi kan derfor bruge ligningen som en definitionsligning der definerer symbolet på venstre side af lighedstegnet for alle $\kappa > 0$ og alle $y \in \{0,1,2,\ldots\}$. Idet vi desuden indfører betegnelsen $p = 1/(\beta+1)$, kan vi alt i alt skrive den fundne sandsynlighed for y ulykker som

$$\mathrm{P}(Y=y) = \binom{y+\kappa-1}{y}p^\kappa(1-p)^y, \qquad y = 0,1,2,\ldots \tag{9.3}$$

Vi ser at Y er *negativt binomialfordelt* med formparameter κ og sandsynlighedsparameter $p = 1/(\beta+1)$ (jf. definition 2.9 side 58).

I den negative binomialfordeling er der to parametre man kan »skrue på«, og man kan håbe at det derved er muligt at få denne model til at passe bedre til observationerne end model 1 gjorde.

I den nye model er (jf. eksempel 4.5 side 81)

$$\mathrm{E}(Y) = \kappa(1-p)/p = \kappa\beta,$$

$$\mathrm{Var}(Y) = \kappa(1-p)/p^2 = \kappa\beta(\beta+1).$$

Heraf ses blandt andet at variansen er $(\beta+1)$ gange større end middelværdien. – I det foreliggende talmateriale fandt vi netop at variansen var større end middelværdien, så foreløbig kan det ikke udelukkes at den negative binomialfordelingsmodel er brugbar.

Estimation af parametrene i Model 2

Vi benytter som altid likelihoodmetoden til estimation af de ukendte parametre. Likelihoodfunktionen er et produkt af sandsynligheder af formen (9.3):

$$\begin{aligned}L(\kappa,p) &= \prod_{i=1}^n \binom{y_i+\kappa-1}{y_i}p^\kappa(1-p)^{y_i} \\ &= p^{n\kappa}(1-p)^{y_1+y_2+\cdots+y_n}\prod_{i=1}^n\binom{y_i+\kappa-1}{y_i} \\ &= \text{konst}\cdot p^{n\kappa}(1-p)^{y_\cdot}\cdot\prod_{k=1}^\infty(\kappa+k-1)^{f_k+f_{k+1}+f_{k+2}+\ldots}\end{aligned}$$

hvor f_k stadig betegner antal observationer som har værdien k. Logaritmen til likelihoodfunktionen bliver derfor (på nær en konstant)

$$\ln L(\kappa,p) = n\kappa\ln p + y_\cdot\ln(1-p) + \sum_{k=1}^\infty\Bigl(\sum_{j=k}^\infty f_j\Bigr)\ln(\kappa+k-1)$$

der i det konkrete eksempel antager det mere uskyldige udseende

$$\begin{aligned}\ln L(\kappa,p) = {}& 647\kappa\ln p + 301\ln(1-p) \\ &+ 200\ln\kappa + 68\ln(\kappa+1) + 26\ln(\kappa+2) \\ &+ 5\ln(\kappa+3) + 2\ln(\kappa+4).\end{aligned}$$

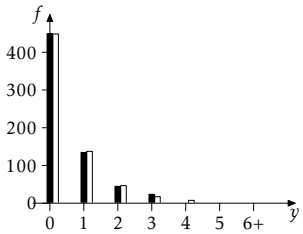
Denne funktion kan man let bestemme maksimum for med sædvanlige numeriske metoder til bestemmelse af ekstremumspunkter, for eksempel en generel simplexmetode. Disse numeriske metoder itererer sig frem til løsningen, og man kan finde et godt udgangspunkt for iterationen ved at løse de to ligninger

$$\begin{aligned}\text{teoretisk middelværdi} &= \text{empirisk middelværdi} \\ \text{teoretisk varians} &= \text{empirisk varians}\end{aligned}$$

der i det foreliggende tilfælde bliver $\kappa\beta = 0.465$ og $\kappa\beta(\beta+1) = 0.692$, hvor $\beta = (1-p)/p$. Ligningerne har løsningen $(\widetilde{\kappa},\widetilde{p}) = (0.953, 0.672)$ (og dermed $\widetilde{\beta} = 0.488$).

y	f_y	$\widehat{\widehat{f}}_y$
0	447	445.9
1	132	134.9
2	42	44.0
3	21	14.7
4	3	5.0
5	2	1.7
6+	0	0.9
	647	647.1

Tabel 9.12
Model 2: Tabel og graf med de observerede antal f_y (sorte søjler) og forventede antal $\widehat{\widehat{f}}_y$ (hvide søjler).



De fundne værdier benyttes som startværdier i en iteration der leder frem til likelihoodfunktionens maksimumspunkt; man finder estimaterne

$$\widehat{\kappa} = 0.8651$$
$$\widehat{p} = 0.6503 \quad \text{(og dermed } \widehat{\beta} = 0.5378\text{)}.$$

I tabel 9.12 ses de tilsvarende forventede antal

$$\widehat{\widehat{f}}_y = n\binom{y+\widehat{\kappa}-1}{y}\widehat{p}^{\widehat{\kappa}}(1-\widehat{p})^y$$

beregnet ud fra den estimerede negative binomialfordeling. På baggrund heraf tillader vi os at konkludere at den negative binomialfordelingsmodel beskriver observationerne godt nok.

10 Den flerdimensionale normalfordeling

Statistiske modeller for normalfordelte observationer kan, som vi skal se i kapitel 11, formuleres meget overskueligt og elegant ved brug af terminologi fra lineær algebra, og bestemmelse af estimatorer og teststørrelser og udledning af deres fordelinger kan med stor fordel foregå inden for disse rammer.

Inden vi for alvor kan gå i gang med normalfordelingsmodellerne, er der nogle forberedende ting der skal overstås, og undervejs vil læseren måske have glæde af at se i tillægget om lineær algebra (side 203ff). I afsnit 10.1 præciseres enkelte ting i forbindelse med flerdimensionale fordelinger, blandt andet om middelværdi og varians. Derefter (i afsnit 10.2) skal vi definere den flerdimensionale normalfordeling, hvilket viser sig at være besværligere end man måske umiddelbart skulle tro.

10.1 Flerdimensionale stokastiske variable

En n-dimensional stokastisk variabel $\boldsymbol{X}$ kan opfattes som et sæt bestående af n endimensionale stokastiske variable, eller som en stokastisk vektor (i vektorrummet $V = \mathbb{R}^n$) hvis koordinater i standardkoordinatsystemet er n endimensionale stokastiske variable. *Middelværdien* af den n-dimensionale stokastiske variabel $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$ er vektoren $\mathrm{E}\,\boldsymbol{X} = \begin{bmatrix} \mathrm{E}\,X_1 \\ \mathrm{E}\,X_2 \\ \vdots \\ \mathrm{E}\,X_n \end{bmatrix}$, altså talsættet bestående af middelværdierne af de enkelte koordinater – forudsat at alle de optrædende endimensionale stokastiske variable har middelværdi. *Variansen* af $\boldsymbol{X}$ er den symmetriske positivt semidefinitte $n \times n$-matrix $\mathrm{Var}\,\boldsymbol{X}$ hvis (i,j)-te element er $\mathrm{Cov}(X_i, X_j)$:

$$\mathrm{Var}\,\boldsymbol{X} = \begin{bmatrix} \mathrm{Var}\,X_1 & \mathrm{Cov}(X_1,X_2) & \cdots & \mathrm{Cov}(X_1,X_n) \\ \mathrm{Cov}(X_2,X_1) & \mathrm{Var}\,X_2 & \cdots & \mathrm{Cov}(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n,X_1) & \mathrm{Cov}(X_n,X_2) & \cdots & \mathrm{Var}\,X_n \end{bmatrix}$$

$$= \mathrm{E}\big((\boldsymbol{X} - \mathrm{E}\,\boldsymbol{X})(\boldsymbol{X} - \mathrm{E}\,\boldsymbol{X})'\big),$$

forudsat at alle de optrædende endimensionale stokastiske variable har en varians.
Ud fra definitionerne viser man let

Sætning 10.1
Lad $\boldsymbol{X}$ være en n-dimensional stokastisk variabel, og antag at $\boldsymbol{X}$ har middelværdi og varians. Hvis A er en lineær afbildning fra $\mathbb{R}^n$ til $\mathbb{R}^p$ og $\boldsymbol{b}$ en konstant vektor i $\mathbb{R}^p$ [eller A er en $p \times n$-matrix og $\boldsymbol{b}$ en $p \times 1$-matrix], så er

$$\mathrm{E}(A\boldsymbol{X} + \boldsymbol{b}) = A(\mathrm{E}\,\boldsymbol{X}) + \boldsymbol{b} \tag{10.1}$$

$$\mathrm{Var}(A\boldsymbol{X} + \boldsymbol{b}) = A(\mathrm{Var}\,\boldsymbol{X})A'. \tag{10.2}$$

Eksempel 10.1: Multinomialfordelingen
Multinomialfordelingen (side 105) er et eksempel på en flerdimensional fordeling. Hvis $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_r \end{bmatrix}$ er multinomialfordelt med parametre n og $\boldsymbol{p} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_r \end{bmatrix}$, så er $\mathrm{E}\,\boldsymbol{X} = \begin{bmatrix} np_1 \\ np_2 \\ \vdots \\ np_r \end{bmatrix}$ og

$$\mathrm{Var}\,\boldsymbol{X} = \begin{bmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_r \\ -np_2p_1 & np_2(1-p_2) & \cdots & -np_2p_r \\ \vdots & \vdots & \ddots & \vdots \\ -np_rp_1 & -np_rp_2 & \cdots & np_r(1-p_r) \end{bmatrix}.$$

Summen af X-erne er altid n, dvs. $A\boldsymbol{X} = n$, hvor A er den lineære afbildning

$$A : \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_r \end{bmatrix} \longmapsto \begin{bmatrix} 1 & 1 & \ldots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_r \end{bmatrix} = x_1 + x_2 + \cdots + x_r.$$

Derfor er $\mathrm{Var}(A\boldsymbol{X}) = 0$. Da $\mathrm{Var}(A\boldsymbol{X}) = A(\mathrm{Var}\,\boldsymbol{X})A'$ (formel (10.2)), har vi her et eksempel hvor $\mathrm{Var}\,\boldsymbol{X}$ ikke er positivt definit (men kun positivt semidefinit).

10.2 Definition og egenskaber

I dette afsnit definerer vi den flerdimensionale normalfordeling og viser den flerdimensionale udgave af sætning 3.11 side 74: hvis $\boldsymbol{X}$ er n-dimensionalt normalfordelt med parametre $\boldsymbol{\mu}$ og Σ, og hvis A er en $p \times n$-matrix og $\boldsymbol{b}$ en $p \times 1$-matrix, så er $A\boldsymbol{X} + \boldsymbol{b}$ p-dimensionalt normalfordelt med parametre $A\boldsymbol{\mu} + \boldsymbol{b}$ og $A\Sigma A'$. Det er imidlertid ikke helt trivielt at definere den n-dimensionale normalfordeling med parametre $\boldsymbol{\mu}$ og Σ i det tilfælde hvor Σ ikke er regulær. Derfor går vi frem i en række skridt.

Definition 10.1: n-dimensional standardnormalfordeling
Den n-dimensionale standardnormalfordeling er den n-dimensionale kontinuerte fordeling som har tæthedsfunktion

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}} \exp\Big(-\tfrac{1}{2}\|\boldsymbol{x}\|^2\Big), \quad \boldsymbol{x} \in \mathbb{R}^n. \tag{10.3}$$

Bemærkninger til definitionen:

1. For $n = 1$ stemmer definitionen overens med den tidligere definition af standardnormalfordeling (definition 3.9 side 74).
2. Den n-dimensionale stokastiske variabel $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$ er n-dimensionalt standardnormalfordelt hvis og kun hvis $X_1, X_2, \ldots, X_n$ er uafhængige og endimensionalt standardnormalfordelte. Det følger af at

$$\frac{1}{(2\pi)^{n/2}} \exp\Big(-\tfrac{1}{2}\|\boldsymbol{x}\|^2\Big) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}x_i^2)$$

   (jf. sætning 3.2 side 67).
3. Hvis $\boldsymbol{X}$ er standardnormalfordelt, så er $\mathrm{E}\,\boldsymbol{X} = \boldsymbol{0}$ og $\mathrm{Var}\,\boldsymbol{X} = \mathrm{I}$; det følger af punkt 2.

Sætning 10.2
Hvis A er en isometrisk lineær afbildning af $\mathbb{R}^n$ ind i sig selv, og $\boldsymbol{X}$ er n-dimensionalt standardnormalfordelt, så bliver $A\boldsymbol{X}$ igen n-dimensionalt standardnormalfordelt.

Bevis
Ifølge sætningen om transformation af tætheder (sætning 3.5 side 68) har $\boldsymbol{Y} = A\boldsymbol{X}$ tæthedsfunktionen $f(A^{-1}\boldsymbol{y})|\det A^{-1}|$ hvor f er givet ved (10.3). Da f kun afhænger af $\boldsymbol{x}$ gennem $\|\boldsymbol{x}\|$, og da $\|A^{-1}\boldsymbol{y}\| = \|\boldsymbol{y}\|$ fordi A er en isometri, er $f(A^{-1}\boldsymbol{y}) = f(\boldsymbol{y})$; desuden er $\det A^{-1} = 1$, igen fordi A er en isometri. Altså er $f(A^{-1}\boldsymbol{y})|\det A^{-1}| = f(\boldsymbol{y})$. □

Korollar 10.3
Hvis $\boldsymbol{X}$ er n-dimensionalt standardnormalfordelt og $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ er en ortonormalbasis for $\mathbb{R}^n$, så er $\boldsymbol{X}$'s koordinater $X_1, X_2, \ldots, X_n$ i denne basis uafhængige og endimensionalt standardnormalfordelte.

Koordinaterne $X_1, X_2, \ldots, X_n$ for $\boldsymbol{X}$ i basen $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ er som bekendt $X_i = \langle \boldsymbol{X}, \boldsymbol{e}_i \rangle$, $i = 1, 2, \ldots, n$.

Bevis
Da koordinattransformationsmatricen er en isometri, følger påstanden af sætning 10.2 og bemærkning 2 til definition 10.1. □

Definition 10.2: Regulær normalfordeling
Antag at $\boldsymbol{\mu} \in \mathbb{R}^n$ og at Σ er en positivt definit lineær afbildning af $\mathbb{R}^n$ ind i sig selv [*eller at $\boldsymbol{\mu}$ er en n-dimensional søjlevektor og Σ en positivt definit $n \times n$-matrix*].

Den n-dimensionale regulære normalfordeling med middelværdi $\boldsymbol{\mu}$ og varians Σ er den n-dimensionale kontinuerte fordeling hvis tæthedsfunktion er

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\,|\det \Sigma|^{1/2}} \exp\Big(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big), \quad \boldsymbol{x} \in \mathbb{R}^n.$$

Bemærkninger til definitionen:

1. For $n = 1$ fås den sædvanlige (endimensionale) normalfordeling med middelværdi μ og varians σ^2 (= det ene element i Σ).
2. Hvis $\Sigma = \sigma^2 \mathrm{I}$, reducerer tæthedsfunktionen til

$$f(\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\frac{1}{2}\frac{\|\boldsymbol{x}-\boldsymbol{\mu}\|^2}{\sigma^2}\Big), \quad \boldsymbol{x} \in \mathbb{R}^n.$$

   Med andre ord er $\boldsymbol{X}$ n-dimensionalt regulært normalfordelt med parametre $\boldsymbol{\mu}$ og $\sigma^2\mathrm{I}$, hvis og kun hvis $X_1, X_2, \ldots, X_n$ er uafhængige og endimensionalt normalfordelte sådan at X_i har middelværdi μ_i og varians σ^2.
3. Definitionen omtaler rask væk parametrene $\boldsymbol{\mu}$ og Σ som middelværdi og varians. Strengt taget burde man i første omgang have givet dem nogle neutrale navne og sidenhen vise at de faktisk er henholdsvis middelværdi og varians.
   Man kan vise at de virkelig er middelværdi og varians: Ifølge bemærkning 3 til definition 10.1 er påstanden rigtig når $\boldsymbol{\mu} = \mathbf{0}$ og $\Sigma = \mathrm{I}$. I det generelle tilfælde ser vi på $\boldsymbol{X} = \boldsymbol{\mu} + \Sigma^{1/2}\boldsymbol{U}$, hvor $\boldsymbol{U}$ er n-dimensionalt standardnormalfordelt og $\Sigma^{1/2}$ er som i sætning B.5 side 205 (med $A = \Sigma$). Ifølge nedenstående sætning 10.4 er $\boldsymbol{X}$ regulært normalfordelt med parametre $\boldsymbol{\mu}$ og Σ, og ifølge sætning 10.1 er $\mathrm{E}\,\boldsymbol{X} = \boldsymbol{\mu}$ og $\mathrm{Var}\,\boldsymbol{X} = \Sigma$.

Sætning 10.4
Hvis $\boldsymbol{X}$ er n-dimensionalt regulært normalfordelt med parametre $\boldsymbol{\mu}$ og Σ, hvis A er en bijektiv lineær afbildning af $\mathbb{R}^n$ ind i sig selv, og hvis $\boldsymbol{b} \in \mathbb{R}^n$ er en konstant vektor, så er $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b}$ n-dimensionalt regulært normalfordelt med parametre $A\boldsymbol{\mu} + \boldsymbol{b}$ og $A\Sigma A'$.

Bevis
Ifølge sætningen om transformation af tætheder (sætning 3.5 side 68) har

$\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b}$ tæthedsfunktionen $f_{\boldsymbol{Y}}(\boldsymbol{y}) = f(A^{-1}(\boldsymbol{y}-\boldsymbol{b}))|\det A^{-1}|$ hvor f er som i definition 10.2. Ved almindelig udregning fås dette til

$$\begin{aligned} &f_{\boldsymbol{Y}}(\boldsymbol{y}) \\ &= \frac{1}{(2\pi)^{n/2}\,|\det\Sigma|^{1/2}\,|\det A|} \exp\Big(-\tfrac{1}{2}\big(A^{-1}(\boldsymbol{y}-\boldsymbol{b})-\boldsymbol{\mu}\big)'\Sigma^{-1}\big(A^{-1}(\boldsymbol{y}-\boldsymbol{b})-\boldsymbol{\mu}\big)\Big) \\ &= \frac{1}{(2\pi)^{n/2}\,|\det(A\Sigma A')|^{1/2}} \exp\Big(-\tfrac{1}{2}\big(\boldsymbol{y}-(A\boldsymbol{\mu}+\boldsymbol{b})\big)'(A\Sigma A')^{-1}\big(\boldsymbol{y}-(A\boldsymbol{\mu}+\boldsymbol{b})\big)\Big) \end{aligned}$$

som ønsket. □

Eksempel 10.2

Lad os prøve med et simpelt eksempel med $n = 2$. Antag at $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ er todimensionalt regulært normalfordelt med parametre $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ og $\Sigma = \sigma^2\mathrm{I} = \sigma^2\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Vi vil finde fordelingen af $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 + X_2 \\ X_1 - X_2 \end{bmatrix}$, dvs. $\boldsymbol{Y} = A\boldsymbol{X}$ hvor $A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$.

Ifølge sætning 10.4 er $\boldsymbol{Y}$ todimensionalt regulært normalfordelt med $\mathrm{E}\,\boldsymbol{Y} = \begin{bmatrix} \mu_1 + \mu_2 \\ \mu_1 - \mu_2 \end{bmatrix}$ og varians $\sigma^2 AA' = 2\sigma^2\mathrm{I}$, dvs. $X_1 + X_2$ og $X_1 - X_2$ er uafhængige og normalfordelte med middelværdi hhv. $\mu_1 + \mu_2$ og $\mu_1 - \mu_2$, og med samme varians $2\sigma^2$.

Vi vil definere den generelle n-dimensionale normalfordeling på følgende måde:

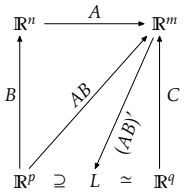
Definition 10.3: Den n-dimensionale normalfordeling
Antag at $\boldsymbol{\mu} \in \mathbb{R}^n$ og at Σ er en positivt semidefinit lineær afbildning af $\mathbb{R}^n$ ind i sig selv [*eller $\boldsymbol{\mu}$ er en n-dimensional søjlevektor og Σ en positivt semidefinit $n \times n$-matrix*]. *Lad p betegne rangen af Σ.*

Den n-dimensionale normalfordeling med middelværdi $\boldsymbol{\mu}$ og varians Σ er fordelingen af $\boldsymbol{\mu} + B\boldsymbol{U}$ hvor $\boldsymbol{U}$ er p-dimensionalt standardnormalfordelt og B er en injektiv lineær afbildning af $\mathbb{R}^p$ ind i $\mathbb{R}^n$ sådan at $BB' = \Sigma$.

Bemærkninger:

1. Vi ved fra sætning B.6 side 205 at der altid findes et B med de omtalte egenskaber, og at B er entydigt bestemt på nær isometri. Da standardnormalfordelingen er invariant ved isometrier af $\mathbb{R}^p$ (sætning 10.2), følger det nu at fordelingen af $\boldsymbol{\mu} + B\boldsymbol{U}$ ikke afhænger af hvordan man har valgt B, blot B er injektiv og $BB' = \Sigma$.
2. Definitionen generaliserer definition 10.2; det følger af sætning 10.4.

Sætning 10.5
Hvis $\boldsymbol{X}$ er n-dimensionalt normalfordelt med parametre $\boldsymbol{\mu}$ og Σ, og hvis A er en lineær afbildning af $\mathbb{R}^n$ ind i $\mathbb{R}^m$, så er $\boldsymbol{Y} = A\boldsymbol{X}$ m-dimensionalt normalfordelt med parametre $A\boldsymbol{\mu}$ og $A\Sigma A'$.



Bevis
Lad os sige at $\boldsymbol{X} = \boldsymbol{\mu} + B\boldsymbol{U}$ hvor $\boldsymbol{U}$ er p-dimensionalt standardnormalfordelt og B er en injektiv lineær afbildning af $\mathbb{R}^p$ ind i $\mathbb{R}^n$. Så er $\boldsymbol{Y} = A\boldsymbol{\mu} + AB\boldsymbol{U}$. Det vi skal vise, er at $AB\boldsymbol{U}$ har samme fordeling som $C\boldsymbol{V}$, hvor C er en velvalgt injektiv lineær afbildning af $\mathbb{R}^q$ ind i $\mathbb{R}^m$, og $\boldsymbol{V}$ er q-dimensionalt standardnormalfordelt; her er q rangen af AB.

Vi sætter $L = \mathcal{R}((AB)')$, altså billedrummet for $(AB)'$, og vi sætter $q = \dim L$. Ideen i beviset er at vi som C kan bruge restriktionen af AB til $L \simeq \mathbb{R}^q$. Ifølge sætning B.1 side 204 er L det ortogonale komplement til $\mathcal{N}(AB)$, nulrummet for AB (hvoraf følger at restriktionen af AB til L er injektiv). Vi kan derfor vælge en ortonormalbasis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$ for $\mathbb{R}^p$ sådan at $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_q$ er en basis for L, og resten af $\boldsymbol{e}$-erne er en basis for $\mathcal{N}(AB)$. Lad $U_1, U_2, \ldots, U_p$ betegne $\boldsymbol{U}$'s koordinater i forhold til denne basis, altså $U_i = \langle \boldsymbol{U}, \boldsymbol{e}_i \rangle$. Da $\boldsymbol{e}_{q+1}, \boldsymbol{e}_{q+2}, \ldots, \boldsymbol{e}_p \in \mathcal{N}(AB)$, er

$$AB\boldsymbol{U} = AB \sum_{i=1}^{p} U_i \boldsymbol{e}_i = \sum_{i=1}^{q} U_i AB \boldsymbol{e}_i .$$

Ifølge korollar 10.3 er $U_1, U_2, \ldots, U_p$ uafhængige endimensionalt standardnormalfordelte; derfor er også $U_1, U_2, \ldots, U_q$ uafhængige endimensionalt standardnormalfordelte, så hvis vi definerer $\boldsymbol{V}$ til at være den q-dimensionale stokastiske variabel som består af koordinaterne $U_1, U_2, \ldots, U_q$, så er $\boldsymbol{V}$ q-dimensionalt standardnormalfordelt.

Hvis vi definerer den lineære afbildning C af $\mathbb{R}^q$ ind i $\mathbb{R}^n$ som den afbildning der afbilder $\boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_q \end{bmatrix}$ over i $\sum_{i=1}^{q} v_i AB \boldsymbol{e}_i$, har vi nu angivet et $\boldsymbol{V}$ og et C med de ønskede egenskaber. □

Sætning 10.6: Spaltningssætningen
Antag at $\boldsymbol{X}$ er n-dimensionalt normalfordelt med middelværdi $\boldsymbol{0}$ og varians $\sigma^2 \mathrm{I}$. Hvis $\mathbb{R}^n = L_1 \oplus L_2 \oplus \cdots \oplus L_k$ er en ortogonal opspaltning og $\mathfrak{p}_1, \mathfrak{p}_2, \ldots, \mathfrak{p}_k$ de tilhørende projektioner, så er de stokastiske variable $\mathfrak{p}_1\boldsymbol{X}, \mathfrak{p}_2\boldsymbol{X}, \ldots, \mathfrak{p}_k\boldsymbol{X}$ uafhængige; $\mathfrak{p}_j\boldsymbol{X}$ er n-dimensionalt normalfordelt med middelværdi $\boldsymbol{0}$ og varians $\sigma^2\mathfrak{p}_j$, og $\|\mathfrak{p}_j\boldsymbol{X}\|^2$ er χ^2-fordelt med skalaparameter σ^2 og $f_j = \dim L_j$ frihedsgrader.

Bevis
Det er nok at se på tilfældet $\sigma^2 = 1$. – Man kan vælge en ortonormalbasis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ hvor hver basisvektor ligger i et af underrummene L_i. De stokastiske variable $X_i = \langle \boldsymbol{X}, \boldsymbol{e}_i \rangle$, $i = 1, 2, \ldots, n$, er uafhængige endimensionalt standardnormalfordelte ifølge korollar 10.3. Projektionen $\mathfrak{p}_j\boldsymbol{X}$ af $\boldsymbol{X}$ på L_j er summen af led af formen $X_i\boldsymbol{e}_i$ hvor der summeres over alle i for hvilke $\boldsymbol{e}_i \in L_j$, dvs. f_j led. Derved bliver $\mathfrak{p}_j\boldsymbol{X}$ iht. definition 10.3 n-dimensionalt normalfordelt med de påståede parametre. Da de enkelte $\mathfrak{p}_j\boldsymbol{X}$-er er funktioner af hver deres X_i-er, bliver de uafhængige af hinanden. Da $\|\mathfrak{p}_j\boldsymbol{X}\|^2$ er summen af de f_j X_i^2-er for hvilke $\boldsymbol{e}_i \in L_j$, er den χ^2-fordelt med f_j frihedsgrader (sætning 3.13 side 75). □

10.3 Opgaver

Opgave 10.1
På side 163 hævdes det at $\operatorname{Var}\boldsymbol{X}$ er positivt semidefinit. Vis at dette faktisk er tilfældet. – Tip: Benyt regnereglerne for kovarianser (sætning 1.30 side 41 – sætningen er rigtig for alle typer reelle stokastiske variable med varians) til at vise at $\operatorname{Var}(\boldsymbol{a}'\boldsymbol{X}\boldsymbol{a}) = \boldsymbol{a}'(\operatorname{Var}\boldsymbol{X})\boldsymbol{a}$ hvor $\boldsymbol{a}$ er en $n \times 1$-matrix (ɔ: en søjlevektor).

Opgave 10.2
Udfyld detaljerne i argumentationen for bemærkning 2 til definitionen af den n-dimensionale standardnormalfordeling.

Opgave 10.3
Antag at $\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ er todimensionalt normalfordelt. Vis at X_1 og X_2 er uafhængige hvis og kun hvis deres kovarians er 0. (Sammenlign med sætning 1.31 side 41 der gælder for alle slags stokastiske variable (med varians).)

Opgave 10.4
Lad t være funktionen givet ved $t(x) = -x$ når $|x| < a$ og $t(x) = x$ ellers; her er a en positiv konstant. Lad X_1 være en en-dimensional standardnormalfordelt stokastisk variabel, og sæt $X_2 = t(X_1)$.

Er X_1 og X_2 uafhængige? Vis at man kan vælge a på en sådan måde at $\operatorname{Cov}(X_1, X_2) = 0$. Hvordan harmonerer det med opgave 10.3? (En lignende problemstilling behandledes i opgave 1.21.)

11 Lineære normale modeller

Dette kapitel præsenterer de klassiske lineære normalfordelingsmodeller som variansanalyse og regressionsanalyse formuleret i lineær algebra-sprog.

11.1 Estimation og test, generelt

Vi vil studere den generelle lineære normale model gående ud på at $\boldsymbol{y}$ er en observation af en n-dimensionalt normalfordelt stokastisk variabel $\boldsymbol{Y}$ med middelværdivektor $\boldsymbol{\mu}$ og variansmatrix $\sigma^2\mathrm{I}$; det antages at parameteren $\boldsymbol{\mu}$ er et punkt i underrummet L af $V = \mathbb{R}^n$, og at $\sigma^2 > 0$, dvs. parameterrummet er $L \times]0\,;+\infty[$. Modellen kaldes en *lineær* normal model fordi middelværdien tilhører et lineært underrum L.

Likelihoodfunktionen svarende til observationen $\boldsymbol{y}$ er

$$L(\boldsymbol{\mu},\sigma^2) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left(-\tfrac{1}{2}\,\frac{\|\boldsymbol{y}-\boldsymbol{\mu}\|^2}{\sigma^2}\right), \qquad \boldsymbol{\mu} \in L,\ \ \sigma^2 > 0.$$

Estimation

Lad p betegne ortogonalprojektionen af $V = \mathbb{R}^n$ på L. For et vilkårligt $\boldsymbol{z} \in V$ er $\boldsymbol{z} = (\boldsymbol{z}-\mathrm{p}\boldsymbol{z})+\mathrm{p}\boldsymbol{z}$ hvor $\boldsymbol{z}-\mathrm{p}\boldsymbol{z} \perp \mathrm{p}\boldsymbol{z}$, og dermed er $\|\boldsymbol{z}\|^2 = \|\boldsymbol{z}-\mathrm{p}\boldsymbol{z}\|^2+\|\mathrm{p}\boldsymbol{z}\|^2$ (Pythagoras); anvendt på $\boldsymbol{z} = \boldsymbol{y}-\boldsymbol{\mu}$ giver det at

$$\|\boldsymbol{y}-\boldsymbol{\mu}\|^2 = \|\boldsymbol{y}-\mathrm{p}\boldsymbol{y}\|^2 + \|\mathrm{p}\boldsymbol{y}-\boldsymbol{\mu}\|^2,$$

hvoraf følger at $L(\boldsymbol{\mu},\sigma^2) \le L(\mathrm{p}\boldsymbol{y},\sigma^2)$ for ethvert σ^2, dvs. maksimaliseringsestimatet for $\boldsymbol{\mu}$ er $\mathrm{p}\boldsymbol{y}$. Ved sædvanlige metoder finder man at $L(\mathrm{p}\boldsymbol{y},\sigma^2)$ maksimaliseres med hensyn til σ^2 når σ^2 er lig $\frac{1}{n}\|\boldsymbol{y}-\mathrm{p}\boldsymbol{y}\|^2$. Maksimumsværdien $L(\mathrm{p}\boldsymbol{y},\frac{1}{n}\|\boldsymbol{y}-\mathrm{p}\boldsymbol{y}\|^2)$ findes i øvrigt ved almindelige omskrivninger at være $a_n(\|\boldsymbol{y}-\mathrm{p}\boldsymbol{y}\|^2)^{-n/2}$ hvor a_n er et tal der afhænger af n, men ikke af $\boldsymbol{y}$.

Ved nu at anvende sætning 10.6 på $\boldsymbol{X} = \boldsymbol{Y}-\boldsymbol{\mu}$ og den ortogonale opspaltning $\mathbb{R} = L \oplus L^\perp$ får vi

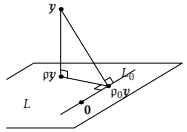
Sætning 11.1
I den generelle lineære model gælder

- *middelværdivektoren $\boldsymbol{\mu}$ estimeres ved $\widehat{\boldsymbol{\mu}} = \mathrm{p}\boldsymbol{y}$, altså projektionen af $\boldsymbol{y}$ vinkelret ned på L,*
- *estimatoren $\boldsymbol{\mu} = \mathrm{p}\boldsymbol{Y}$ er n-dimensionalt normalfordelt med middelværdi $\boldsymbol{\mu}$ og varians $\sigma^2\mathrm{p}$,*
- *variansen σ^2 estimeres centralt ved $s^2 = \frac{1}{n-\dim L}\|\boldsymbol{y}-\mathrm{p}\boldsymbol{y}\|^2$, og maksimaliseringsestimatoren for σ^2 er $\widehat{\sigma}^2 = \frac{1}{n}\|\boldsymbol{y}-\mathrm{p}\boldsymbol{y}\|^2$,*
- *estimatoren s^2 er χ^2-fordelt med skalaparameter $\sigma^2/(n-\dim L)$ og $n-\dim L$ frihedsgrader,*
- *de to estimatorer $\widehat{\boldsymbol{\mu}}$ og s^2 er stokastisk uafhængige (og de to estimatorer $\widehat{\boldsymbol{\mu}}$ og $\widehat{\sigma}^2$ er ligeledes stokastisk uafhængige).*

Vektoren $\boldsymbol{y}-\mathrm{p}\boldsymbol{y}$ kaldes *residualvektoren*, og størrelsen $\|\boldsymbol{y}-\mathrm{p}\boldsymbol{y}\|^2$ kaldes *residualkvadratsummen*. Tallet $(n-\dim L)$ er *antallet af frihedsgrader* for variansskønnet og/eller residualkvadratsummen. – Man kan bestemme $\widehat{\boldsymbol{\mu}}$ af relationen $\boldsymbol{y}-\widehat{\boldsymbol{\mu}} \perp L$ som i realiteten er $\dim L$ lineære ligninger med lige så mange ubekendte; disse ligninger kaldes *normalligningerne* (de udtrykker at $\boldsymbol{y}-\widehat{\boldsymbol{\mu}}$ er normal til L).

Test af hypotese om middelværdien

Antag at der foreligger en middelværdihypotese af formen $H_0 : \boldsymbol{\mu} \in L_0$ hvor L_0 er et underrum af L. Ifølge sætning 11.1 er maksimaliseringsestimaterne for $\boldsymbol{\mu}$ og σ^2 under H_0 hhv. $\mathrm{p}_0\boldsymbol{y}$ og $\frac{1}{n}\|\boldsymbol{y}-\mathrm{p}_0\boldsymbol{y}\|^2$. Kvotienttesstørrelsen bliver derfor



$$Q = \frac{L(\mathrm{p}_0\boldsymbol{y}, \frac{1}{n}\|\boldsymbol{y}-\mathrm{p}_0\boldsymbol{y}\|^2)}{L(\mathrm{p}\boldsymbol{y}, \frac{1}{n}\|\boldsymbol{y}-\mathrm{p}\boldsymbol{y}\|^2)} = \left(\frac{\|\boldsymbol{y}-\mathrm{p}\boldsymbol{y}\|^2}{\|\boldsymbol{y}-\mathrm{p}_0\boldsymbol{y}\|^2}\right)^{n/2}.$$

Da $L_0 \subseteq L$, er $\mathrm{p}\boldsymbol{y}-\mathrm{p}_0\boldsymbol{y} \in L$, så $\boldsymbol{y}-\mathrm{p}\boldsymbol{y} \perp \mathrm{p}\boldsymbol{y}-\mathrm{p}_0\boldsymbol{y}$. Ifølge Pythagoras er da $\|\boldsymbol{y}-\mathrm{p}_0\boldsymbol{y}\|^2 = \|\boldsymbol{y}-\mathrm{p}\boldsymbol{y}\|^2 + \|\mathrm{p}\boldsymbol{y}-\mathrm{p}_0\boldsymbol{y}\|^2$. Ved hjælp heraf omskriver vi Q videre:

$$\begin{aligned} Q &= \left(\frac{\|\boldsymbol{y}-\mathrm{p}\boldsymbol{y}\|^2}{\|\boldsymbol{y}-\mathrm{p}\boldsymbol{y}\|^2 + \|\mathrm{p}\boldsymbol{y}-\mathrm{p}_0\boldsymbol{y}\|^2}\right)^{n/2} \\ &= \left(1 + \frac{\|\mathrm{p}\boldsymbol{y}-\mathrm{p}_0\boldsymbol{y}\|^2}{\|\boldsymbol{y}-\mathrm{p}\boldsymbol{y}\|^2}\right)^{-n/2} = \left(1 + \frac{\dim L - \dim L_0}{n-\dim L}\,F\right)^{-n/2} \end{aligned}$$

hvor

$$F = \frac{\frac{1}{\dim L-\dim L_0}\|\mathrm{p}\boldsymbol{y}-\mathrm{p}_0\boldsymbol{y}\|^2}{\frac{1}{n-\dim L}\|\boldsymbol{y}-\mathrm{p}\boldsymbol{y}\|^2}$$

er den teststørrelse man i praksis benytter. – Da Q er en aftagende funktion af F, skal man forkaste for store værdier af F. Det følger af sætning 10.6 at under H_0 er tælleren og nævneren i F-teststørrelsen stokastisk uafhængige og χ^2-fordelte; tælleren er χ^2-fordelt med skalaparameter $\sigma^2/(\dim L-\dim L_0)$ og

$(\dim L-\dim L_0)$ frihedsgrader, og nævneren er χ^2-fordelt med skalaparameter $\sigma^2/(n-\dim L)$ og $(n-\dim L)$ frihedsgrader. Fordelingen af teststørrelsen er en F-fordeling med $(\dim L-\dim L_0)$ og $(n-\dim L)$ frihedsgrader.

F-fordelingen
F-fordelingen med $f_t = f_{\text{tæller}}$ og $f_n = f_{\text{nævner}}$ frihedsgrader er den kontinuerte fordeling på den positive halvakse som har tæthedsfunktion

$$C\,\frac{x^{(f_t-2)/2}}{(f_n+f_t x)^{(f_t+f_n)/2}}$$

hvor

$$C = \frac{\Gamma\left(\frac{f_t+f_n}{2}\right)}{\Gamma\left(\frac{f_t}{2}\right)\Gamma\left(\frac{f_n}{2}\right)} f_t^{f_t/2} f_n^{f_n/2}$$

Hermed er estimations- og testproblemerne i princippet løst (og sætningerne 7.1 og 7.2 side 121/122 vist). Vi kan så gå over til at se hvordan det tager sig ud i konkrete modeller.

11.2 Enstikprøveproblemet

Man har observationer $y_1, y_2, \ldots, y_n$ af stokastiske variable $Y_1, Y_2, \ldots, Y_n$ som er uafhængige og identisk (endimensionalt) normalfordelte med middelværdi μ og varians σ^2 hvor $\mu \in \mathbb{R}$ og $\sigma^2 > 0$. Man ønsker at estimere parametrene μ og σ^2 og sidenhen at teste hypotesen $H_0 : \mu = 0$.

Vi vil opfatte y_i-erne som arrangeret i en n-dimensional vektor $\boldsymbol{y}$ der opfattes som en observation af en n-dimensional stokastisk variabel $\boldsymbol{Y}$ som er normalfordelt med middelværdi $\boldsymbol{\mu}$ og varians $\sigma^2\mathrm{I}$, og hvor det ifølge modellen antages at $\boldsymbol{\mu}$ tilhører det endimensionale underrum $L = \{\mu\mathbf{1} : \mu \in \mathbb{R}\}$ af vektorer der har den samme værdi μ på alle pladser; her og i det følgende betegner $\mathbf{1}$ den vektor som har værdien 1 på alle pladser.

Ifølge sætning 11.1 kan maksimaliseringsestimatet $\widehat{\boldsymbol{\mu}}$ findes ved at projicere $\boldsymbol{y}$ vinkelret ned på L. Vi kan få en ligning til bestemmelse af $\widehat{\boldsymbol{\mu}}$ ved at bemærke at $\boldsymbol{y}-\widehat{\boldsymbol{\mu}}$ skal stå vinkelret på enhver vektor i L, specielt på vektoren $\mathbf{1}$, så

$$0 = \langle \boldsymbol{y}-\widehat{\boldsymbol{\mu}}, \mathbf{1}\rangle = \sum_{i=1}^{n}(y_i-\widehat{\mu}) = y_\bullet - n\widehat{\mu}$$

hvor $y_\bullet$ som sædvanlig er summen af y_i-erne. Heraf ses at $\widehat{\boldsymbol{\mu}} = \widehat{\mu}\mathbf{1}$ hvor $\widehat{\mu} = \overline{y}$, dvs. μ estimeres ved gennemsnittet af observationerne. Maksimaliseringsestimatet for σ^2 er derefter

$$\widehat{\sigma}^2 = \frac{1}{n}\|\boldsymbol{y}-\widehat{\boldsymbol{\mu}}\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2,$$

og det centrale variansskøn er

$$s^2 = \frac{1}{n-\dim L}\|\boldsymbol{y}-\widehat{\boldsymbol{\mu}}\|^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2.$$

Disse estimater blev fundet på anden vis på side 121.

Vi vil herefter teste hypotesen $H_0 : \mu = 0$ eller rettere $\boldsymbol{\mu} \in L_0$, hvor $L_0 = \{\mathbf{0}\}$. Under H_0 estimeres $\boldsymbol{\mu}$ ved projektionen af $\boldsymbol{y}$ på L_0, dvs. ved $\mathbf{0}$. Hypotesen kan derfor

testes med F-teststørrelsen

$$F = \frac{\frac{1}{1-0}\|\widehat{\boldsymbol{\mu}} - \mathbf{0}\|^2}{\frac{1}{n-1}\|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}\|^2} = \frac{n\widehat{\mu}^2}{s^2} = \left(\frac{\overline{y}}{\sqrt{s^2/n}}\right)^2,$$

dvs. $F = t^2$ hvor t er den sædvanlige t-teststørrelse, jf. side 133. Man kan derfor efter behag benytte F (med 1 og $n-1$ frihedsgrader) eller t (med $n-1$ frihedsgrader) som teststørrelse.

Hypotesen $\mu = 0$ er den eneste hypotese af formen $\boldsymbol{\mu} \in L_0$ hvor L_0 er et underrum af L; men hvad gør man så hvis den interessante hypotese er af formen $\mu = \mu_0$ hvor μ_0 ikke er 0? Svar: Træk μ_0 fra alle y-erne og benyt den netop beskrevne metode. Derved får man en F- eller t-størrelse hvor der i tælleren står $\overline{y} - \mu_0$ i stedet for $\overline{y}$, alt andet er uforandret.

11.3 Ensidet variansanalyse

Man har observationer der er delt ind i k grupper; observationerne benævnes y_{ij} hvor $i = 1, 2, \ldots, k$ er gruppenummer, og $j = 1, 2, \ldots, n_i$ nummererer observationerne inden for gruppen. Skematisk ser det sådan ud:

	observationer					
gruppe 1	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1n_1}
gruppe 2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2n_2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
gruppe i	y_{i1}	y_{i2}	$\cdots$	y_{ij}	$\cdots$	y_{in_i}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
gruppe k	y_{k1}	y_{k2}	$\cdots$	y_{kj}	$\cdots$	y_{kn_k}

Vi går ud fra at forskellen mellem observationerne inden for en gruppe er tilfældig, hvorimod der er en systematisk forskel mellem grupperne. Vi går endvidere ud fra at y_{ij}-erne er observerede værdier af uafhængige stokastiske variable Y_{ij}, og vi vil beskrive den tilfældige variation ved hjælp af en normalfordeling. Det skal derfor alt i alt være sådan at Y_{ij} er normalfordelt med middelværdi μ_i og varians σ^2. Formuleret mere omhyggeligt: vi antager at der findes reelle tal $\mu_1, \mu_2, \ldots, \mu_k$ og et positivt tal σ^2 således at Y_{ij} er normalfordelt med middelværdi μ_i og varians σ^2, $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, k$; desuden er alle Y_{ij}-erne uafhængige. På denne måde beskriver middelværdiparametrene $\mu_1, \mu_2, \ldots, \mu_k$ *den systematiske variation*, nemlig de enkelte gruppers niveauer, mens variansparameteren σ^2 (samt normalfordelingen) beskriver *den tilfældige variation* inden for grupperne. Den tilfældige variation antages at være den samme i alle grupperne, og denne antagelse kan man undertiden teste, se afsnit 11.4.

Man ønsker at estimere parametrene $\mu_1, \mu_2, \ldots, \mu_k$ og σ^2, og man ønsker at teste hypotesen $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ om at der ikke er forskel på de k grupper.

Vi vil opfatte y_{ij}-erne som en vektor $\boldsymbol{y} \in V = \mathbb{R}^n$ hvor $n = n_\bullet$ er antallet af observationer. Grundmodellen er da at $\boldsymbol{y}$ er en observeret værdi af en n-dimensionalt normalfordelt stokastisk variabel $\boldsymbol{Y}$ med middelværdi $\boldsymbol{\mu}$ og varians $\sigma^2\mathrm{I}$, hvor $\sigma^2 > 0$, og hvor $\boldsymbol{\mu}$ tilhører det underrum som man kort skriver som

$$L = \{\boldsymbol{\xi} \in V : \xi_{ij} = \mu_i\};$$

mere udførligt er L mængden af vektorer $\boldsymbol{\xi} \in V$ for hvilke der findes et talsæt $(\mu_1, \mu_2, \ldots, \mu_k) \in \mathbb{R}^k$ sådan at $\xi_{ij} = \mu_i$ for alle i og j. Dimensionen af L er k (idet vi går ud fra at alle n_i-erne er større end nul).

Kroneckers δ er symbolet $\delta_{ij} = \begin{cases} 1 & \text{hvis } i = j \\ 0 & \text{ellers} \end{cases}$

Ifølge sætning 11.1 estimeres $\boldsymbol{\mu}$ ved $\widehat{\boldsymbol{\mu}} = \mathrm{p}\boldsymbol{y}$ hvor p er projektionen af V på L. For at nå frem til et mere brugbart udtryk til bestemmelse af $\widehat{\boldsymbol{\mu}}$ udnytter vi at der skal gælde at $\widehat{\boldsymbol{\mu}} \in L$ og $\boldsymbol{y} - \widehat{\boldsymbol{\mu}} \perp L$, i særdeleshed skal $\boldsymbol{y} - \widehat{\boldsymbol{\mu}}$ stå vinkelret på de k vektorer $\boldsymbol{e}^1, \boldsymbol{e}^2, \ldots, \boldsymbol{e}^k \in L$ som er defineret på den måde at den (i, j)-te komponent i vektoren $\boldsymbol{e}^s$ er $(\boldsymbol{e}^s)_{ij} = \delta_{is}$, dvs. for $s = 1, 2, \ldots, k$ skal gælde

$$0 = \langle \boldsymbol{y} - \widehat{\boldsymbol{\mu}}, \boldsymbol{e}^s \rangle = \sum_{j=1}^{n_s} (y_{sj} - \widehat{\mu}_s) = y_{s\bullet} - n_s\widehat{\mu}_s,$$

hvoraf følger at $\widehat{\mu}_i = y_{i\bullet}/n_i = \overline{y}_i$, dvs. $\widehat{\mu}_i$ er gennemsnittet i gruppe i.

Variansparameteren σ^2 estimeres ved

$$s_0^2 = \frac{1}{n - \dim L}\|\boldsymbol{y} - \mathrm{p}\boldsymbol{y}\|^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$$

der har $n - k$ frihedsgrader og er stokastisk uafhængig af $\widehat{\boldsymbol{\mu}}$.

Hypotesen H_0 om ens grupper formuleres som $H_0 : \boldsymbol{\mu} \in L_0$, hvor

$$L_0 = \{\boldsymbol{\xi} \in V : \xi_{ij} = \mu\}$$

er det endimensionale underrum af vektorer hvor der står det samme på alle pladser. Fra afsnit 11.2 ved vi at projektionen af $\boldsymbol{y}$ på L_0 er den vektor $\mathrm{p}_0\boldsymbol{y}$ hvor der på alle pladser står totalgennemsnittet $\overline{y} = y_{\bullet\bullet}/n_\bullet$. Hypotesen H_0 testes med teststørrelsen

$$F = \frac{\frac{1}{\dim L - \dim L_0}\|\mathrm{p}\boldsymbol{y} - \mathrm{p}_0\boldsymbol{y}\|^2}{\frac{1}{n - \dim L}\|\boldsymbol{y} - \mathrm{p}\boldsymbol{y}\|^2} = \frac{s_1^2}{s_0^2}$$

hvor

$$s_1^2 = \frac{1}{\dim L - \dim L_0}\|p\boldsymbol{y} - p_0\boldsymbol{y}\|^2$$
$$= \frac{1}{k-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{y}_i - \overline{y})^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2,$$

der under H_0 er F-fordelt med $k-1$ og $n-k$ frihedsgrader. Man taler om at s_0^2 beskriver *variationen inden for grupperne*, og at s_1^2 beskriver *variationen mellem grupperne*. Teststørrelsen måler derfor variationen mellem grupper i forhold til variationen inden for grupper – heraf metodens navn: *ensidet variansanalyse*.

Hvis hypotesen accepteres, estimeres middelværdivektoren $\boldsymbol{\mu}$ ved den vektor $p_0\boldsymbol{y}$ hvor der står $\overline{y}$ på alle pladser, og variansen σ^2 ved

$$s_{01}^2 = \frac{1}{n - \dim L_0}\|\boldsymbol{y} - p_0\boldsymbol{y}\|^2 = \frac{\|\boldsymbol{y} - p\boldsymbol{y}\|^2 + \|p\boldsymbol{y} - p_0\boldsymbol{y}\|^2}{(n - \dim L) + (\dim L - \dim L_0)},$$

dvs.

$$s_{01}^2 = \frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 + \sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2\Big).$$

Eksempel 11.1: Kvælning af hunde
I en undersøgelse af sammenhængen mellem hypoxivarighed og hypoxantinkoncentration i cerebrospinalvæsken (se en nærmere beskrivelse i eksempel 11.2 side 193) har man målt hypoxantinkoncentrationen i nogle hunde efter fire forskellige hypoxivarigheder, se tabel 11.7 side 194. Vi vil her undersøge om der er signifikant forskel på de fire grupper svarende til de fire forskellige hypoxivarigheder.

Indledningsvis udregnes forskellige hjælpestørrelser samt estimater over de ukendte parametre, se tabel 11.1. De fire middelværdiparametre estimeres således til 1.46, 5.50, 7.48 og 12.81, og variansen estimeres til $s_0^2 = 4.82$ med 21 frihedsgrader. Variansen mellem grupper estimeres til $s_1^2 = 465.5/3 = 155.2$ med 3 frihedsgrader, så teststørrelsen for hypotesen om homogenitet mellem grupper er $F = 155.2/4.82 = 32.2$ der er overordentlig signifikant, dvs. der er i høj grad forskel på de fire gruppers middelværdier.

Traditionelt opsummerer man udregninger og testresultater i et variansanalyseskema, se tabel 11.2.

Variansanalysemodellen forudsætter at der er varianshomogenitet, og det kan man teste ved hjælp af Bartletts test (afsnit 11.4). Vi indsætter s^2-værdierne fra tabel 11.1 i Bartletts teststørrelse (11.1) og får $B = -\Big(6\ln\frac{1.27}{4.82} + 5\ln\frac{2.99}{4.82} + 4\ln\frac{7.63}{4.82} + 6\ln\frac{8.04}{4.82}\Big) = 5.5$ der skal sammenlignes med χ^2-fordelingen med $k-1=3$ frihedsgrader. Tabelopslag viser at der er over 10% chance for at få en større B-værdi end værdien 5.5 som derfor ikke er signifikant stor. Med andre ord kan vi opretholde antagelsen om varianshomogenitet.

▷ [Se også eksempel 11.2 side 193.]

i	n_i	$\sum_{j=1}^{n_i} y_{ij}$	$\overline{y}_i$	f_i	$\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2$	s_i^2
1	7	10.2	1.46	6	7.64	1.27
2	6	33.0	5.50	5	14.94	2.99
3	5	37.4	7.48	4	30.51	7.63
4	7	89.7	12.81	6	48.23	8.04
sum	25	170.3		21	101.32	
gennemsnit			6.81			4.82

Tabel 11.1
Kvælning af hunde: Nogle hjælpestørrelser til beregningerne.

	f	SS	s^2	test
variation inden for grupper	21	101.32	4.82	
variation mellem grupper	3	465.47	155.16	155.16/4.82=32.2
variation omkring fælles gn.snit	24	566.79	23.62	

Tabel 11.2
Kvælning af hunde: Variansanalyseskema. f står for antal frihedsgrader, SS er sum af kvadratiske afvigelser, og $s^2 = SS/f$.

Tostikprøveproblemet, uparrede observationer

Tilfældet $k = 2$ benævnes ikke overraskende tostikprøveproblemet. I mange lærebøger præsenterer man tostikprøveproblemet for sig selv, ofte endda før k-stikprøveproblemet. Det eneste interessante ved dette tostikprøveproblem er vel ellers at F-teststørrelsen kan skrives som t^2 hvor

$$t = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)s_0^2}},$$

jf. side 136. Under H_0 er t t-fordelt med $n-2$ frihedsgrader.

11.4 Bartletts test for varianshomogenitet

I normalfordelingsmodeller er det meget ofte en forudsætning at observationerne stammer fra normalfordelinger med samme varians (eller varians som er kendt på nær en ukendt faktor). I dette afsnit skal vi omtale et test der kan anvendes når man ønsker at vurdere om et antal grupper af normalfordelte observationer kan antages at have samme varians, det vil sige vurdere om der er *varianshomogenitet*. Testet kan ikke benyttes hvis en af grupperne kun indeholder en enkelt observation, og for at man skal kunne anvende den sædvanlige χ^2-approksimation til fordelingen af $-2\ln Q$, skal hver gruppe indeholde mindst seks observationer, eller rettere: i hver enkelt gruppe skal variansestimatet have mindst fem frihedsgrader.

Den generelle situation tænkes at være som i k-stikprøve-situationen (eller ensidet variansanalyse-situationen), jf. side 174, og vi ønsker altså nu at teste antagelsen om at grupperne har samme variansparameter σ^2.

Én måde at udlede teststørrelsen på er som følger. Antag at de k grupper har hver sin middelværdiparameter og hver sin variansparameter, dvs. normalfordelingen hørende til gruppe i har middelværdi μ_i og varians σ_i^2. Variansparametrene estimeres i henhold til det tidligere centralt ved $s_i^2 = \frac{1}{f_i}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2$, hvor $f_i = n_i - 1$ er antallet af frihedsgrader for variansskønnet, og n_i er antal observationer i gruppe i. Ifølge sætning 7.1 side 121 er s_i^2 gammafordelt med formparameter $f_i/2$ og skalaparameter $2\sigma_i^2/f_i$. Lad os derfor se på følgende statistiske problem: Antag at $s_1^2, s_2^2, \ldots, s_k^2$ er uafhængige observationer fra gammafordelinger således at s_i^2 stammer fra gammafordelingen med formparameter $f_i/2$ og skalaparameter $2\sigma_i^2/f_i$ hvor f_i er et kendt tal; i denne model ønsker vi at teste hypotesen

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2.$$

For at gøre det opskriver vi likelihoodfunktionen svarende til observationerne $s_1^2, s_2^2, \ldots, s_k^2$; den er

$$\begin{aligned} L(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2) &= \prod_{i=1}^{k} \frac{1}{\Gamma(\frac{f_i}{2})\left(\frac{2\sigma_i^2}{f_i}\right)^{f_i/2}} \left(s_i^2\right)^{f_i/2-1} \exp\left(-s_i^2 \Big/ \frac{2\sigma_i^2}{f_i}\right) \\ &= \text{konst} \cdot \prod_{i=1}^{k} \left(\sigma_i^2\right)^{-f_i/2} \exp\left(-\frac{f_i}{2}\frac{s_i^2}{\sigma_i^2}\right) \\ &= \text{konst} \cdot \left(\prod_{i=1}^{k} (\sigma_i^2)^{-f_i/2}\right) \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{k} f_i \frac{s_i^2}{\sigma_i^2}\right). \end{aligned}$$

Maksimaliseringsestimaterne for $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$ er $s_1^2, s_2^2, \ldots, s_k^2$. Maksimaliseringsestimatet for den fælles værdi σ^2 under H_0 er maksimumspunktet for funktionen $\sigma^2 \mapsto L(\sigma^2, \sigma^2, \ldots, \sigma^2)$, og det er $s_0^2 = \frac{1}{f_\bullet}\sum_{i=1}^{k} f_i s_i^2$, hvor $f_\bullet = \sum_{i=1}^{k} f_i$; s_0^2 er altså det vægtede gennemsnit af de enkelte variansskøn med frihedsgradstallene som vægte. Kvotienteststørrelsen er $Q = L(s_0^2, s_0^2, \ldots, s_0^2) \big/ L(s_1^2, s_2^2, \ldots, s_k^2)$. Man kan med fordel anvende $-2\ln Q$ som teststørrelse, og den betegnes i denne forbindelse ofte B og kaldes *Bartletts teststørrelse* for varianshomogenitet; den findes let at være

$$B = -\sum_{i=1}^{k} f_i \ln \frac{s_i^2}{s_0^2}. \tag{11.1}$$

Teststørrelsen B er altid et positivt tal, og store værdier af B er signifikante, dvs. tyder på at hypotesen om varianshomogenitet er forkert. Hvis hypotesen er rigtig, er B approksimativt χ^2-fordelt med $k-1$ frihedsgrader, dvs. man kan udregne den omtrentlige testsandsynlighed som $\varepsilon = \mathrm{P}(\chi^2_{k-1} \geq B_{\text{obs}})$. Denne χ^2-approksimation er god når alle f_i-erne er store; som tommelfingerregel siger man at de alle skal være mindst 5.

Hvis der kun er to grupper (dvs. $k = 2$), kan man alternativt teste hypotesen om varianshomogenitet med et test baseret på forholdet mellem de to variansestimater; dette er omtalt i forbindelse med tostikprøveproblemet i normalfordelingen, se opgave 8.2 side 138. (Dette tostikprøvetest er ikke baseret på nogen χ^2-approksimationer, så det har ingen restriktioner på antallene af frihedsgrader.)

11.5 Tosidet variansanalyse

Man har nogle observationer y der er arrangeret i et tosidet skema:

	1	2	…	j	…	s
1	y_{11k} $k=1,2,\ldots,n_{11}$	y_{12k} $k=1,2,\ldots,n_{12}$	⋯	y_{1jk} $k=1,2,\ldots,n_{1j}$	⋯	y_{1sk} $k=1,2,\ldots,n_{1s}$
2	y_{21k} $k=1,2,\ldots,n_{21}$	y_{22k} $k=1,2,\ldots,n_{22}$	⋯	y_{2jk} $k=1,2,\ldots,n_{2j}$	⋯	y_{2sk} $k=1,2,\ldots,n_{2s}$
⋮	⋮	⋮		⋮		⋮
i	y_{i1k} $k=1,2,\ldots,n_{i1}$	y_{i2k} $k=1,2,\ldots,n_{i2}$	⋯	y_{ijk} $k=1,2,\ldots,n_{ij}$	⋯	y_{isk} $k=1,2,\ldots,n_{is}$
⋮	⋮	⋮		⋮		⋮
r	y_{r1k} $k=1,2,\ldots,n_{r1}$	y_{r2k} $k=1,2,\ldots,n_{r2}$	⋯	y_{rjk} $k=1,2,\ldots,n_{rj}$	⋯	y_{rsk} $k=1,2,\ldots,n_{rs}$

Vi benytter betegnelsen y_{ijk} for observation nr. k i skemaets (i,j)-te *celle*, hvori der i alt er n_{ij} observationer. Den (i,j)-te celle befinder sig i skemaets i-te *række* og j-te *søjle*; der er i alt r rækker og s søjler (i engelske tekster vil der typisk være 'r rows and c columns').

Vi opstiller en statistisk model gående ud på at y_{ijk}-erne er observerede værdier af uafhængige normalfordelte stokastiske variable Y_{ijk} med samme varians σ^2 og med en middelværdistruktur der er bestemt ud fra den måde observationerne er inddelt på – eller måske er det inddelingen der er bestemt af den formodede middelværdistruktur. Her er en præsentation af grundmodel og

mulige interessante hypoteser (dog kun hypoteser om middelværdiparametrene; det er bestandigt underforstået at alle Y-er har samme ukendte varians σ^2):

- *Grundmodellen* G siger at Y-er der hører til samme celle, har samme middelværdi. Mere præcist siger den at der findes tal η_{ij}, $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, s$, således at $\mathrm{E}\, Y_{ijk} = \eta_{ij}$ for alle i, j og k. Vi formulerer grundmodellen kort som

$$G : \mathrm{E}\, Y_{ijk} = \eta_{ij}\,.$$

- *Additivitetshypotesen* eller hypotesen om *forsvindende vekselvirkning* siger at der ikke er nogen vekselvirkning mellem rækker og søjler, men at rækkevirkninger og søjlevirkninger indgår additivt. Mere præcist siger hypotesen at der findes tal $\alpha_1, \alpha_2, \ldots, \alpha_r$ og $\beta_1, \beta_2, \ldots, \beta_s$ således at $\mathrm{E}\, Y_{ijk} = \alpha_i + \beta_j$ for alle i, j og k. Den korte formulering af additivitetshypotesen er

$$H_0 : \mathrm{E}\, Y_{ijk} = \alpha_i + \beta_j\,.$$

- Hypotesen om *ens søjler* eller om *forsvindende søjlevirkning* siger at der ikke er nogen forskel på søjlerne, mere præcist siger den at der findes tal $\alpha_1, \alpha_2, \ldots, \alpha_r$ således at $\mathrm{E}\, Y_{ijk} = \alpha_i$ for alle i, j og k. Den korte formulering er

$$H_1 : \mathrm{E}\, Y_{ijk} = \alpha_i\,.$$

- Hypotesen om *ens rækker* eller om *forsvindende rækkevirkning* siger at der ikke er nogen forskel på rækkerne, mere præcist siger den at der findes tal $\beta_1, \beta_2, \ldots, \beta_s$ således at $\mathrm{E}\, Y_{ijk} = \beta_j$ for alle i, j og k. Den korte formulering er

$$H_2 : \mathrm{E}\, Y_{ijk} = \beta_j\,.$$

- Hypotesen om *total homogenitet* siger at der ikke er nogen forskel på cellerne overhovedet, mere præcist siger den at der findes et tal γ således at $\mathrm{E}\, Y_{ijk} = \gamma$ for alle i, j og k. Den korte formulering er

$$H_3 : \mathrm{E}\, Y_{ijk} = \gamma\,.$$

Vi vil skrive tingene op i lineær algebra-sprog. Vi opfatter da observationerne som udgørende en vektor $\boldsymbol{y} \in V = \mathbb{R}^n$, hvor $n = n_{\cdot\cdot}$ er antallet af observationer; vektorerne er struktureret i et todimensionalt skema som ovenfor. Grundmodellen og de fire hypoteser kan formuleres som udsagn om at middelværdivektoren

$\boldsymbol{\mu} = \mathrm{E}\, \boldsymbol{Y}$ tilhører bestemte underrum:

$$\begin{aligned}
G &: \boldsymbol{\mu} \in L && \text{hvor } L = \{\boldsymbol{\xi} : \xi_{ijk} = \eta_{ij}\},\\
H_0 &: \boldsymbol{\mu} \in L_0 && \text{hvor } L_0 = \{\boldsymbol{\xi} : \xi_{ijk} = \alpha_i + \beta_j\},\\
H_1 &: \boldsymbol{\mu} \in L_1 && \text{hvor } L_1 = \{\boldsymbol{\xi} : \xi_{ijk} = \alpha_i\},\\
H_2 &: \boldsymbol{\mu} \in L_2 && \text{hvor } L_2 = \{\boldsymbol{\xi} : \xi_{ijk} = \beta_j\},\\
H_3 &: \boldsymbol{\mu} \in L_3 && \text{hvor } L_3 = \{\boldsymbol{\xi} : \xi_{ijk} = \gamma\}.
\end{aligned}$$

Der gælder visse relationer mellem hypoteserne/underrummene:

$$G \Leftarrow H_0 \begin{array}{l} \swarrow H_1 \\ \nwarrow H_2 \end{array} \begin{array}{l} \nwarrow \\ \swarrow \end{array} H_3 \qquad\qquad L \supseteq L_0 \begin{array}{l} \supsetneq L_1 \\ \supsetneq L_2 \end{array} \begin{array}{l} \gtrsim \\ \gtrsim \end{array} L_3$$

Grundmodellen samt modellerne svarende til H_1 og H_2 er eksempler på k-stikprøveproblemer (k er hhv. rs, r og s), og modellen svarende til H_3 er et enstikprøveproblem. Derfor kan vi uden videre opskrive estimaterne over middelværdiparametrene i disse fire modeller:

$$\begin{aligned}
&\text{under } G \text{ er } \widehat{\eta_{ij}} = \overline{y}_{ij} && \text{altså gennemsnittet i celle (i, j),}\\
&\text{under } H_1 \text{ er } \widehat{\alpha_i} = \overline{y}_{i\cdot} && \text{altså gennemsnittet i den } i\text{-te række,}\\
&\text{under } H_2 \text{ er } \widehat{\beta_j} = \overline{y}_{\cdot j} && \text{altså gennemsnittet i den } j\text{-te søjle,}\\
&\text{under } H_3 \text{ er } \widehat{\gamma} = \overline{y}_{\cdot\cdot} && \text{altså totalgennemsnittet.}
\end{aligned}$$

Det er derimod ikke nær så enkelt at estimere parametrene under additivitetshypotesen H_0. Først vil vi indføre begrebet sammenhængende model.

Sammenhængende modeller

Hvilke krav/ønsker kan det være hensigtsmæssigt at stille til antallene af observationer i de forskellige celler, altså til nedenstående »antalstabel«?

$$\boldsymbol{n} = \begin{array}{|c|c|c|c|} \hline n_{11} & n_{12} & \cdots & n_{1s} \\ \hline n_{21} & n_{22} & \cdots & n_{2s} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline n_{r1} & n_{r2} & \cdots & n_{rs} \\ \hline \end{array}$$

Det siger sig selv at der ikke må være hele rækker eller søjler udelukkende med 0-er, men kan der tillades 0-er på enkelte pladser?

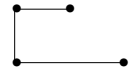
For klarlægge problemerne ser vi på et overskueligt eksempel. Lad os sige at $r = 2$ og $s = 2$, og at

$$\boldsymbol{n} = \begin{array}{|c|c|} \hline 0 & 9 \\ \hline 9 & 0 \\ \hline \end{array}$$

I dette tilfælde er additivitetsunderrummet L todimensionalt (der er kun de to parametre η_{12} og η_{21}), og faktisk er $L = L_0 = L_1 = L_2$. I særdeleshed er $L_1 \cap L_2 \neq L_3$, selv om mange måske ville have troet at der altid gjaldt at $L_1 \cap L_2 = L_3$. Sagen er, at denne model i realiteten består af to separate delmodeller (for hhv. celle $(1,2)$ og celle $(2,1)$), og derfor bliver $\dim(L_1 \cap L_2) > 1$ eller ensbetydende hermed $\dim(L_0) < r+s-1$ (det følger ved brug af den generelle formel $\dim(L_1 + L_2) = \dim L_1 + \dim L_2 - \dim(L_1 \cap L_2)$ og det faktum at $L_0 = L_1 + L_2$).

En model der ikke kan deles op i separate delmodeller, kaldes en *sammenhængende model*. Begrebet sammenhængende model kan præciseres på følgende måde: Ud fra antalstabellen $\boldsymbol{n}$ danner vi en graf hvor knuderne er de talpar (i,j) for hvilke $n_{ij} > 0$, og hvor kanterne forbinder par af »naboknuder« (dvs. knuder som enten har samme i eller samme j). Eksempel:

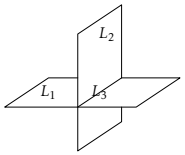
5	8	0
1	0	4



Vi siger at modellen er sammenhængende, hvis grafen er sammenhængende (som det er tilfældet i eksemplet).

Man overbeviser sig let om at modellen er sammenhængende hvis og kun hvis nulrummet for den lineære afbildning der afbilder den $(r+s)$-dimensionale vektor $(\alpha_1, \alpha_2, \ldots, \alpha_r, \beta_1, \beta_2, \ldots, \beta_s)$ over i den »tilsvarende« vektor i L_0, er endimensionalt, altså hvis og kun hvis $L_3 = L_1 \cap L_2$. Udtrykt på almindeligt dansk er en sammenhængende model derfor en model hvor følgende udsagn er korrekt: »hvis alle rækkeparametre er ens og alle søjleparametre er ens, så er der total homogenitet«.

I det følgende beskæftiger vi os kun med sammenhængende modeller.



$(L_1 \cap L_3^\perp) \perp (L_2 \cap L_3^\perp)$

Projektionen på L_0

Ifølge den generelle teori estimeres middelværdivektoren $\boldsymbol{\mu}$ under H_0 ved projektionen $\mathfrak{p}_0 \boldsymbol{y}$ af $\boldsymbol{y}$ på L_0. I visse tilfælde findes en nem formel til beregning af denne projektion. – Vi minder indledningsvis om at $L_0 = L_1 + L_2$ og $L_3 = L_1 \cap L_2$.

Lad os *antage* at

$$(L_1 \cap L_3^\perp) \perp (L_2 \cap L_3^\perp). \tag{11.2}$$

I så fald er

$$L_0 = (L_1 \cap L_3^\perp) \oplus (L_2 \cap L_3^\perp) \oplus L_3$$

og dermed

$$\mathfrak{p}_0 = (\mathfrak{p}_1 - \mathfrak{p}_3) + (\mathfrak{p}_2 - \mathfrak{p}_3) + \mathfrak{p}_3$$

dvs.

$$\begin{aligned}\widehat{\alpha_i + \beta_j} &= (\overline{y}_{i\cdot} - \overline{y}_{\cdots}) + (\overline{y}_{\cdot j} - \overline{y}_{\cdots}) + \overline{y}_{\cdots} \\ &= \overline{y}_{i\cdot} + \overline{y}_{\cdot j} - \overline{y}_{\cdots}\end{aligned}$$

Se, det er jo en meget fin formel til beregning af (koordinaterne for) projektionen på L_0. Spørgsmålet er blot hvornår forudsætningen er opfyldt. Nødvendigt og tilstrækkeligt for (11.2) er at

$$\langle \mathfrak{p}_1 \boldsymbol{e} - \mathfrak{p}_3 \boldsymbol{e}, \mathfrak{p}_2 \boldsymbol{f} - \mathfrak{p}_3 \boldsymbol{f} \rangle = 0 \tag{11.3}$$

for alle $\boldsymbol{e}, \boldsymbol{f} \in \mathfrak{B}$, hvor $\mathfrak{B}$ er en basis for L. Som $\mathfrak{B}$ kan man f.eks. bruge vektorerne $\boldsymbol{e}^{\rho\sigma}$ hvor $(\boldsymbol{e}^{\rho\sigma})_{ijk} = 1$ hvis $(i,j) = (\rho,\sigma)$, og 0 ellers. Hvis man indsætter sådanne vektorer i (11.3), finder man at den nødvendige og tilstrækkelige betingelse for (11.2) er at

$$n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}, \qquad i = 1,2,\ldots,r;\ j = 1,2,\ldots,s. \tag{11.4}$$

Når denne betingelse er opfyldt, siger man at der foreligger *det balancerede tilfælde*. Bemærk at hvis der er lige mange observationer i alle celler, så er (11.4) automatisk opfyldt.

Sammenfattende kan vi nu sige, at i det balancerede tilfælde, dvs. når (11.4) er opfyldt, kan estimatorerne under additivitetshypotesen udregnes efter opskriften

$$\widehat{\alpha_i + \beta_j} = \overline{y}_{i\cdot} + \overline{y}_{\cdot j} - \overline{y}_{\cdots}. \tag{11.5}$$

Test af hypoteser

De forskellige hypoteser om middelværdistrukturen kan nu testes; hver gang kan man benytte en F-teststørrelse hvor nævneren er variansskønnet i den aktuelle grundmodel, og tælleren er et skøn over »hypotesens variation omkring grundmodellen«. Sine to frihedsgradsantal arver F fra hhv. tæller- og nævnervariansskønnet.

1. *Additivitetshypotesen* H_0 testes med $F = s_1^2 / s_0^2$ hvor

$$s_0^2 = \frac{1}{n - \dim L} \|\boldsymbol{y} - \mathfrak{p}\boldsymbol{y}\|^2 = \frac{1}{n-g} \sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{k=1}^{n_{ij}} \left(y_{ijk} - \overline{y}_{ij}\right)^2 \tag{11.6}$$

beskriver *variationen inden for grupper* ($g = \dim L$ er antal ikke-tomme celler), og

$$s_1^2 = \frac{1}{\dim L - \dim L_0} \|\mathfrak{p}\boldsymbol{y} - \mathfrak{p}_0 \boldsymbol{y}\|^2$$

$$= \frac{1}{g-(r+s-1)} \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}} \left(\overline{y}_{ij} - (\widehat{\alpha_i+\beta_j})\right)^2$$

beskriver *vekselvirkningsvariationen*. – I det balancerede tilfælde er

$$s_1^2 = \frac{1}{g-(r+s-1)} \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n_{ij}} \left(\overline{y}_{ij} - \overline{y}_{i\bullet} - \overline{y}_{\bullet j} + \overline{y}_{\bullet\bullet\bullet}\right)^2.$$

Ovenstående forudsætter at $n > g$, dvs. der skal være celler med mere end én observation.

2. Hvis additivitetshypotesen accepteres, udnævnes den til aktuel grundmodel, og man udregner man et nyt variansskøn

$$\begin{aligned} s_{01}^2 &= \frac{1}{n-\dim L_0}\|\boldsymbol{y} - p_0\boldsymbol{y}\|^2 \\ &= \frac{1}{n-\dim L_0}\left(\|\boldsymbol{y} - p\boldsymbol{y}\|^2 + \|p\boldsymbol{y} - p_0\boldsymbol{y}\|^2\right). \end{aligned}$$

Alt efter de konkrete omstændigheder og problemstillinger vil man derefter teste H_1 og/eller H_2.

- For at teste hypotesen H_1 om *forsvindende søjlevirkning* benyttes teststørrelsen $F = s_1^2/s_{01}^2$ hvor

$$\begin{aligned} s_1^2 &= \frac{1}{\dim L_0 - \dim L_1}\|p_0\boldsymbol{y} - p_1\boldsymbol{y}\|^2 \\ &= \frac{1}{s-1}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}\left(\widehat{\alpha_i+\beta_j} - \overline{y}_{i\bullet}\right)^2. \end{aligned}$$

  I det balancerede tilfælde er

$$s_1^2 = \frac{1}{s-1}\sum_{j=1}^{s} n_{\bullet j}\left(\overline{y}_{\bullet j} - \overline{y}_{\bullet\bullet\bullet}\right)^2.$$

- For at teste hypotesen H_2 om *forsvindende rækkevirkning* benyttes teststørrelsen $F = s_2^2/s_{01}^2$ hvor

$$\begin{aligned} s_2^2 &= \frac{1}{\dim L_0 - \dim L_2}\|p_0\boldsymbol{y} - p_2\boldsymbol{y}\|^2 \\ &= \frac{1}{r-1}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}\left(\widehat{\alpha_i+\beta_j} - \overline{y}_{\bullet j}\right)^2. \end{aligned}$$

		kvælstof					
		0			1		
fosfor	0	591	450	584	619	618	524
		509	636	413	651	655	564
	1	722	689	625	801	688	682
		584	614	513	703	774	623
	2	702	677	684	814	757	810
		643	668	699	792	790	703

Tabel 11.3
Dyrkningsforsøg: Udbytte ved dyrkningsforsøg med kartofler på 36 parceller. Værdierne er $1000\times(\log(\textit{udbytte [lbs]})-3)$.

I det balancerede tilfælde er

$$s_2^2 = \frac{1}{r-1}\sum_{i=1}^{r} n_{i\bullet}\left(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet\bullet}\right)^2.$$

Bemærk at H_1 og H_2 testes sideordnet, altså begge i forhold til H_0 med variansskønnet s_{01}^2; i det balancerede tilfælde er de to F-tællere s_1^2 og s_2^2 stokastisk uafhængige (fordi $p_0\boldsymbol{y} - p_1\boldsymbol{y}$ og $p_0\boldsymbol{y} - p_2\boldsymbol{y}$ er ortogonale).

Et eksempel

For at opnå de optimale vækstbetingelser skal planter have de fornødne næringsstoffer i de rette forhold. Dette eksempel handler om at bestemme det rette forhold mellem mængden af tilført kvælstof- og fosforgødning til kartofler. Man har dyrket nogle kartoffelmarker på seks forskellige måder, svarende til seks forskellige kombinationer af mængde tilført fosfor (0, 1 eller 2 enheder) og mængde tilført kvælstof (0 eller 1 enhed), og derefter har man målt høstudbyttet. På den måde får man nogle observationer, høstudbytterne, som er inddelt i grupper efter dyrkningsmetode således at grupperne er fastlagt ved hjælp af to kriterier, nemlig tilført fosfor og tilført kvælstof.

Et sådant dyrkningsforsøg udført i 1932 ved Ely gav de resultater der er vist i tabel 11.3.* Opgaven er nu at undersøge hvordan de to faktorer kvælstof og fosfor virker hver for sig og sammen. Er det f.eks. sådan at virkningen af at gå fra en til to enheder fosfor afhænger af om der tilføres kvælstof eller ej? Det kan undersøges med en tosidet variansanalyse.

* Bemærk i øvrigt at høstudbytterne har undergået visse forandringer på deres vej til tabel 11.3. Den væsentligste er at man har taget logaritmen til tallene. Grunden hertil er, at erfaringsmæssigt er høstudbyttet af kartofler ikke særlig normalfordelt, hvorimod det det ser bedre ud med logaritmen til høstudbyttet. Da man havde taget logaritmen til tallene, viste det sig at alle resultaterne hed 3-komma-et-eller-andet, så for at få nogle pæne tal ud af det har man trukket 3 fra og ganget med 1000.

Tabel 11.4
Dyrkningsforsøg: Middeltal $\overline{y}$ (øverst) og variansestimat s^2 (nederst) i hver af de seks grupper, jf. tabel 11.3.

		kvælstof 0	kvælstof 1
fosfor	0	530.50	605.17
		7680.30	2644.57
	1	624.50	711.83
		5569.90	4248.57
	2	678.83	777.67
		474.97	1745.07

Man kan måske for den gode ordens skyld være interesseret i at teste grundmodellens antagelse om varianshomogenitet, og derfor beregnes for hver af de seks grupper ikke blot gennemsnit, men også variansestimater, se tabel 11.4. Man finder den samlede kvadratsum til 111817, således at den fælles varians inden for grupper estimeres ved $s_0^2 = 111817/(36-6) = 3727.2$ (jf. formel (11.6)). Bartletts teststørrelse bliver $B_{\text{obs}} = 9.5$ der skal sammenlignes med χ^2-fordelingen med $6-1=5$ frihedsgrader. Tabelopslag viser at der er knap 10% chance for at få en større værdi, og der er således ikke noget der taler alvorligt imod antagelsen om varianshomogenitet. Vi kan derfor basere de videre undersøgelser på den formodede grundmodel.

Vi vil derefter gå i gang med at undersøge om talmaterialet kan beskrives med en model hvor virkningerne af de to faktorer »tilført fosfor« og »tilført kvælstof« indgår additivt. Vi betegner den k-te observation i den i-te række og j-te søjle y_{ijk}. Grundmodellen er at y_{ijk}-erne opfattes som observerede værdier af uafhængige normalfordelte stokastiske variable Y_{ijk}, hvor Y_{ijk} er normalfordelt med middelværdi μ_{ij} og varians σ^2. Additivitetshypotesen kan formuleres som $H_0 : \mathrm{E}\, Y_{ijk} = \alpha_i + \beta_j$. Da vi har at gøre med et »balanceret tilfælde« (idet formel (11.4) er opfyldt), kan estimaterne $\widehat{\alpha_i + \beta_j}$ udregnes efter formel (11.5). Vi udregner hjælpestørrelserne

$$\begin{array}{lll} \overline{y}_{\cdot\cdot} = 654.75 & \overline{y}_{1\cdot} - \overline{y}_{\cdot\cdot} = -86.92 & \overline{y}_{\cdot 1} - \overline{y}_{\cdot\cdot} = -43.47 \\ & \overline{y}_{2\cdot} - \overline{y}_{\cdot\cdot} = 13.42 & \overline{y}_{\cdot 2} - \overline{y}_{\cdot\cdot} = 43.47 \\ & \overline{y}_{3\cdot} - \overline{y}_{\cdot\cdot} = 73.50 & \end{array} .$$

Ved hjælp heraf udregnes de estimerede gruppemiddelværdier $\widehat{\alpha_i + \beta_j}$ under additivitetshypotesen, se tabel 11.5. Det variansestimat der benyttes hvis additivitetshypotesen er rigtig, er $s_{01}^2 = \dfrac{112693.5}{36-(3+2-1)} = 3521.7$ med $36-(3+2-1) = 32$ frihedsgrader.

Vi kan nu teste om der er additivitet mellem fosfor og kvælstof i kartoffeldyrkningseksemplet. Vi har tidligere fundet at $s_0^2 = 3727.2$. Vekselvirkningsvariansen

Tabel 11.5
Dyrkningsforsøg: Gruppegennemsnit (øverst) og estimerede gruppemiddelværdier under additivitetshypotesen (nederst).

		kvælstof 0	kvælstof 1
fosfor	0	530.50	605.17
		524.36	611.30
	1	624.50	711.83
		624.70	711.64
	2	678.83	777.67
		684.78	771.72

Tabel 11.6
Dyrkningsforsøg med kartofler: Variansanalyseskema.
f står for antal frihedsgrader, SS for Sum af kvadratiske afvigelser, $s^2 = SS/f$.

variation	f	SS	s^2	test
inden for grupper	30	111817	3727	
vekselvirkning	2	877	439	439/3727=0.12
additivitetshypotesen	32	112694	3522	
mellem N-niveauer	1	68034	68034	68034/3522=19
mellem P-niveauer	2	157641	78820	78820/3522=22
omkring total-gns.	35	338369		

findes til

$$s_1^2 = \frac{877}{6-(3+2-1)} = \frac{877}{2} = 438.5 .$$

Teststørrelsen er dermed

$$F = \frac{s_1^2}{s_0^2} = \frac{438.5}{3727.2} = 0.12$$

der skal sammenlignes med F-fordelingen med 2 og 30 frihedsgrader. Tabelopslag viser at testsandsynligheden er lidt under 90%, så der er næppe nogen tvivl om at additivitetshypotesen kan godkendes.

Som forbedret estimat over den fælles varians benyttes herefter

$$s_{01}^2 = \frac{112694}{36-(3+2-1)} = 3521.7$$

med $36-(3+2-1) = 32$ frihedsgrader.

Nu da vi véd at den additive model giver en god beskrivelse af observationerne, og det således har mening at tale om en kvælstofvirkning og en fosforvirkning, kan det være af interesse at undersøge om der er en signifikant virkning af kvælstof hhv. fosfor.

For at undersøge om kvælstof har en virkning, testes hypotesen H_1 om forsvindende kvælstofvirkning (søjlevirkning). Variansen mellem kvælstofniveauer

udregnes til

$$s_2^2 = \frac{68034}{2-1} = 68034\,.$$

Variansestimatet i den additive model var $s_{01}^2 = 3522$ med 32 frihedsgrader, og F-teststørrelsen bliver derfor

$$F_{\text{obs}} = \frac{s_2^2}{s_{01}^2} = \frac{68034}{3522} = 19$$

der skal sammenlignes med $F_{1,32}$-fordelingen. Værdien $F_{\text{obs}} = 19$ er ganske utvetydigt signifikant stor, og hypotesen om forsvindende kvælstofvirkning må derfor forkastes, dvs. konklusionen bliver at det *har* en virkning at tilføre kvælstof.

Hvis man undersøger om der er en forsvindende fosforvirkning, så må også denne hypotese forkastes, dvs. det har også en signifikant virkning at tilføre fosforgødning.

Variansanalyseskemaet (tabel 11.6) giver en samlet oversigt over analysen.

11.6 Regressionsanalyse

Regressionsanalyse handler om at undersøge hvordan én målt størrelse afhænger af en eller flere andre. Antag at der foreligger et datamateriale som er fremkommet på den måde at man på hvert af nogle »individer« (f.eks. forsøgspersoner eller forsøgsdyr eller enkelt-laboratorieforsøg osv.) har målt værdien af et antal størrelser (variable). En af disse størrelser indtager en særstilling, idet man nemlig gerne vil »beskrive« eller »forklare« denne størrelse ved hjælp af de øvrige. Tit kalder man den variabel der skal beskrives, for y, og de variable ved hjælp af hvilke man vil beskrive, for $x_1, x_2, \ldots, x_p$. Andre betegnelser fremgår af følgende oversigt:

y	$x_1, x_2, \ldots, x_p$
den modellerede variabel	baggrundsvariable
den afhængige variabel	de uafhængige variable
den forklarede variabel	de forklarende variable
responsvariabel	

Her skitseres et par eksempler:

1. Lægen observerer den tid y som patienten overlever efter at være blevet behandlet for sygdommen, men lægen har også registreret en mængde baggrundsoplysninger om patienten, såsom køn, alder, vægt, detaljer om sygdommen osv. Nogle af baggrundsoplysningerne kan måske indeholde en eller anden form for information om hvor længe patienten kan forventes at overleve.
2. I en række nogenlunde ens i-lande har man bestemt mål for lungekræftforekomst, cigaretforbrug og forbrug af fossilt brændstof, altsammen pr. indbygger. Man kan da udnævne lungekræftforekomst til y-variabel og søge at »forklare« den ved hjælp af de andre to variable, der så får rollen som forklarende variable.
3. Man ønsker at undersøge et bestemt stofs giftighed. Derfor giver man det i forskellige koncentrationer til nogle grupper af forsøgsdyr og ser hvor mange af dyrene der dør. Her er koncentrationen x en uafhængig variabel hvis værdi eksperimentator bestemmer, og antallet y af døde er den afhængige variabel.

Regressionsanalyse går ud på at finde en statistisk model hvormed man kan beskrive en y-variabel ved hjælp af en kendt simpel funktion af nogle baggrundsvariable og nogle parametre. Parametrene er de samme for alle observationssæt, hvorimod baggrundsvariablene typisk ikke er det.

Man må naturligvis ikke forvente at den statistiske model leverer en perfekt beskrivelse, dels fordi den model man måtte finde frem til, næppe er fuldstændig rigtig, dels fordi en af pointerne med statistiske modeller jo netop er at de kun beskriver hovedtrækkene i datamaterialet og ser stort på de finere detaljer. Der vil derfor være en vis forskel mellem en observeret værdi y og den tilsvarende *fittede* værdi $\widehat{y}$, dvs. den værdi som man ifølge regressionsmodellen skulle få med de givne værdier af baggrundsvariablene. Denne forskel kaldes *residualet* og betegnes ofte e. Vi har så opspaltningen

$$\begin{array}{ccccc} y & = & \widehat{y} & + & e \\ \text{observeret værdi} & = & \text{fittet værdi} & + & \text{residual.} \end{array}$$

Residualerne er det som modellen *ikke* beskriver, og derfor er det naturligt at man (eller rettere modellen) anser dem for *tilfældige*, dvs. for at være tilfældige tal fra en vis sandsynlighedsfordeling.

-oOo-

Væsentlige forudsætninger for at kunne benytte regressionsanalyse er at

1. det er ikke x-erne, men kun y-erne og residualerne, der er behæftede med tilfældig variation (»usikkerhed«),
2. de enkelte målinger er *stokastisk uafhængige* af hinanden, det vil sige de tilfældigheder der indvirker på én bestemt y-værdi (efter at man har taget højde for baggrundsvariablene), har ikke nogen sammenhæng med de tilfældigheder der spiller ind på de øvrige y-værdier.

Det simpleste eksempel på regressionsanalyse er det hvor der kun er én enkelt baggrundsvariabel, som vi så betegner x. Opgaven bliver da at beskrive y-erne ved hjælp af en kendt simpel funktion af x. Det simpleste ikke-trivielle bud på en sådan funktion må vel være en funktion af typen $y = \alpha + x\beta$ hvor α og β er to parametre, dvs. man formoder at y er en affin funktion af x. Derved får man den såkaldte *simple lineære regressionsmodel*, jf. side 112.

En lidt mere avanceret model er den *multiple lineære regressionsmodel* hvor man har p forklarende variable $x_1, x_2, \ldots, x_p$ og søger at beskrive y-værdierne med en funktion af formen $y = \sum_{j=1}^{p} x_j\beta_j$.

Formulering af modellen

For at regressionsmodellen kan blive til en genuin statistisk model, skal man specificere den sandsynlighedsfordeling som skal beskrive y-ernes variation omkring deres middelværdi. I dette kapitel går vi ud fra at denne sandsynlighedsfordeling er en normalfordeling med varians σ^2 (den samme for alle observationer).

Vi vil formulere modellen mere præcist på følgende måde: Der foreligger n sammenhørende værdier af en afhængig variabel y med tilhørende p baggrundsvariable $x_1, x_2, \ldots, x_p$. Det i-te sæt værdier er $\bigl(y_i, (x_{i1}, x_{i2}, \ldots, x_{ip})\bigr)$. Det antages at $y_1, y_2, \ldots, y_n$ er observerede værdier af uafhængige normalfordelte stokastiske variable $Y_1, Y_2, \ldots, Y_n$ med samme varians σ^2 og med

$$\mathrm{E}\, Y_i = \sum_{j=1}^{p} x_{ij}\beta_j, \quad i = 1, 2, \ldots, n, \tag{11.7}$$

hvor $\beta_1, \beta_2, \ldots, \beta_p$ er ukendte parametre. Ofte vil en af de forklarende variable være konstanten 1, dvs. den har værdien 1 for alle i.

I matrixnotation kan (11.7) skrives som $\mathrm{E}\boldsymbol{Y} = X\boldsymbol{\beta}$, hvor $\boldsymbol{Y}$ er søjlevektoren bestående af de n Y-er, X er en kendt $n \times p$-matrix (den såkaldte *designmatrix*) indeholdende x_{ij}-værdierne, og $\boldsymbol{\beta}$ søjlevektoren bestående af de p ukendte β-er. Man kan naturligvis også formulere det ved hjælp af underrum: $\mathrm{E}\boldsymbol{Y} \in L_1$ hvor $L_1 = \{X\boldsymbol{\beta} \in \mathbb{R}^n : \boldsymbol{\beta} \in \mathbb{R}^p\}$. – Betegnelsen *lineær* regression skyldes at $\mathrm{E}\boldsymbol{Y}$ er en lineær funktion af $\boldsymbol{\beta}$.

Ovenstående kan generaliseres på flere måder. I stedet for observationer med samme varians kan man have observationer hvis varians er kendt på nær en konstant faktor, dvs. $\mathrm{Var}\,\boldsymbol{Y} = \sigma^2\,\Sigma$ hvor $\Sigma > 0$ er en kendt matrix og σ^2 en ukendt parameter; så bliver der tale om *vægtet* lineær regressionsanalyse. Man kan

udskifte normalfordelingen med f.eks. binomialfordelingen, poissonfordelingen eller gammafordelingen, og samtidig generalisere (11.7) til

$$g(\mathrm{E}\, Y_i) = \sum_{j=1}^{p} x_{ij}\beta_j, \quad i = 1, 2, \ldots, n$$

for en passende funktion g; så bliver der tale om *generaliseret* lineær regression. Logistisk regression (afsnit 9.1, specielt side 141ff) er et eksempel på generaliseret lineær regression.

I det følgende vil vi kun beskæftige os med ordinær lineær regression.

Estimation af parametrene

Ifølge den generelle teori estimerer man middelværdivektoren $X\boldsymbol{\beta}$ som projektionen af $\boldsymbol{y}$ vinkelret ned på $L_1 = \{X\boldsymbol{\beta} \in \mathbb{R}^n : \boldsymbol{\beta} \in \mathbb{R}^p\}$. Det betyder at $\boldsymbol{\beta}$ skal estimeres ved en vektor $\widehat{\boldsymbol{\beta}}$ med den egenskab at $\boldsymbol{y} - X\widehat{\boldsymbol{\beta}} \perp L_1$. Nu er $\boldsymbol{y} - X\widehat{\boldsymbol{\beta}} \perp L_1$ ensbetydende med at $\bigl\langle \boldsymbol{y} - X\widehat{\boldsymbol{\beta}}, X\boldsymbol{\beta} \bigr\rangle = 0$ for alle $\boldsymbol{\beta} \in \mathbb{R}^p$, som er ensbetydende med at $\bigl\langle X'\boldsymbol{y} - X'X\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta} \bigr\rangle = 0$ for alle $\boldsymbol{\beta} \in \mathbb{R}^p$, som igen er ensbetydende med at $X'X\widehat{\boldsymbol{\beta}} = X'\boldsymbol{y}$. Ligningssystemet $X'X\widehat{\boldsymbol{\beta}} = X'\boldsymbol{y}$ består af p lineære ligninger med p ubekendte, og det kaldes *normalligningerne* (se også side 172). Hvis $p \times p$-matricen $X'X$ er regulær, er der en entydig løsning, nemlig:

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}\,.$$

(Betingelsen at $X'X$ er regulær, kan formuleres på mange forskellige ensbetydende måder: dimensionen af L_1 er p; rangen af X er p; rangen af $X'X$ er p; søjlerne i X er lineært uafhængige; parametriseringen er injektiv.)

Variansparameteren estimeres ved

$$s^2 = \frac{\|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}\|^2}{n - \dim L_1}.$$

Ved at bruge regnereglerne for variansmatricer fås i øvrigt

$$\begin{aligned} \mathrm{Var}\,\widehat{\boldsymbol{\beta}} &= \mathrm{Var}\bigl((X'X)^{-1}X'\boldsymbol{Y}\bigr) \\ &= \bigl((X'X)^{-1}X'\bigr)\,\mathrm{Var}\,\boldsymbol{Y}\bigl((X'X)^{-1}X'\bigr)' \\ &= \bigl((X'X)^{-1}X'\bigr)\sigma^2\mathrm{I}\bigl((X'X)^{-1}X'\bigr)' \\ &= \sigma^2(X'X)^{-1} \end{aligned} \tag{11.8}$$

der estimeres ved $s^2(X'X)^{-1}$. Kvadratroden af diagonalelementerne heri er estimater over *middelfejlen* (standardafvigelsen) på de tilsvarende $\widehat{\beta}$-er.

Ethvert ordentligt computerprogram til statistik har en indbygget funktion til løsning af normalligningerne; funktionen vil returnere parameterestimaterne og deres middelfejl, og muligvis også hele den estimerede Var $\widehat{\boldsymbol{\beta}}$.

Simpel lineær regression

I simpel lineær regression (jf. bl.a. side 112) er

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix}, \quad X'X = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \quad \text{og} \quad X'\boldsymbol{y} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix},$$

og normalligningerner bliver

$$\begin{aligned} \alpha n + \beta \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i \\ \alpha \sum_{i=1}^n x_i + \beta \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i y_i \end{aligned}$$

som det ikke er uoverkommeligt at løse. Det er dog endnu lettere simpelthen at udregne projektionen $\mathrm{p}\boldsymbol{y}$ af $\boldsymbol{y}$ på L_1. Man ser umiddelbart at de to vektorer $\boldsymbol{u} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$ og $\boldsymbol{v} = \begin{bmatrix} x_1 - \overline{x} \\ x_2 - \overline{x} \\ \vdots \\ x_n - \overline{x} \end{bmatrix}$ er ortogonale og udspænder L_1. Dermed bliver

$$\mathrm{p}\boldsymbol{y} = \frac{\langle \boldsymbol{y}, \boldsymbol{u} \rangle}{\|\boldsymbol{u}\|^2}\boldsymbol{u} + \frac{\langle \boldsymbol{y}, \boldsymbol{v} \rangle}{\|\boldsymbol{v}\|^2}\boldsymbol{v} = \frac{\sum_{i=1}^n y_i}{n}\boldsymbol{u} + \frac{\sum_{i=1}^n (x_i - \overline{x}) y_i}{\sum_{i=1}^n (x_i - \overline{x})^2}\boldsymbol{v},$$

således at den j-te koordinat i $\mathrm{p}\boldsymbol{y}$ bliver $(\mathrm{p}\boldsymbol{y})_j = \overline{y} + \dfrac{\sum_{i=1}^n (x_i - \overline{x}) y_i}{\sum_{i=1}^n (x_i - \overline{x})^2} (x_j - \overline{x})$, jf. side 124.

Ved at udregne variansmatricen (11.8) finder man de varianser og korrelationer som postuleredes i sætning 7.3 side 125.

Hypoteseprøvning

Hypoteser af formen $H_0 : \mathrm{E}\,\boldsymbol{Y} \in L_0$ hvor L_0 er et underrum af L_1, testes på helt sædvanlig måde med et F-test.

Ofte vil man være interesseret i en hypotese af formen $H : \beta_j = 0$, svarende til at den tilsvarende forklarende variabel x_j er uden betydning. En sådan hypotese kan testes enten med et F-test eller med t-teststørrelsen

$$t = \frac{\widehat{\beta_j}}{\text{estimeret middelfejl på } \widehat{\beta_j}}.$$

Om faktorer

Der kan være to forskellige slags baggrundsvariable. I det foregående er omtalt eksempler på *kvantitative* baggrundsvariable, dvs. nogle der angiver en eller anden numerisk størrelse. Man kan imidlertid også operere med kvalitative baggrundsvariable, *faktorer*, der angiver tilhørsforhold til en klasse i forbindelse med en klassificering.

Eksempel: I ensidet variansanalyse optræder observationer y der er inddelt i et antal grupper, se evt. side 174; man kan opfatte data som bestående af sammenhørende værdier (y, f) af en observation y og en faktor f som simpelthen er navnet på den gruppe som y tilhører. Man kan formulere det som et regressionsproblem: Lad os sige at der er k forskellige niveauer af f (dvs. der er k grupper), og lad os kalde dem $1, 2, \ldots, k$. Så indfører vi nogle kunstige (kvantitative) forklarende variable $x_1, x_2, \ldots, x_k$ sådan at $x_i = 1$ hvis $f = i$ og $x_i = 0$ ellers. På den måde erstatter man (y, f) med $(y, (x_1, x_2, \ldots, x_p))$ hvor det er sådan at alle x-er på nær ét er lig 0, og det x som er lig 1, udpeger den gruppe som y tilhører. Ensidet variansanalyse-modellen kan nu skrives

$$\mathrm{E}\,Y = \sum_{j=1}^p x_j \beta_j$$

hvor β_j svarer til μ_j i den oprindelige formulering af modellen.

Ved at kombinere kvantitative baggrundsvariable og faktorer kan man formulere komplicerede modeller, eksempelvis med over- og underordnede grupper eller med forskellige lineære sammenhænge i forskellige delgrupper.

Et eksempel

Eksempel 11.2: Kvælning af hunde
Syv hunde er under bedøvelse blevet udsat for iltmangel ved sammenpresning af

Tabel 11.7
Kvælning af hunde: Målinger af hypoxantinkoncentration til de fire forskellige tidspunkter. I hver gruppe er observationerne ordnet efter størrelse.

varighed (min)	koncentration (μmol/l)						
0	0.0	0.0	1.2	1.8	2.1	2.1	3.0
6	3.0	4.9	5.1	5.1	7.0	7.9	
12	4.9	6.0	6.5	8.0	12.0		
18	9.5	10.1	12.0	12.0	13.0	16.0	17.1

luftrøret, og hypoxantinkoncentrationen måltes efter 0, 6, 12 og 18 minutters forløb. Det var af forskellige grunde ikke muligt at foretage målinger på alle syv hunde til alle fire tidspunkter, og det kan heller ikke afgøres hvordan målinger og hunde hører sammen. Resultaterne af forsøget er vist i tabel 11.7.

Man kan anskue situationen på den måde at der foreligger $n = 25$ par sammenhørende værdier af koncentration og varighed. Varighederne er kendte størrelser – de indgår i forsøgsplanen – hvorimod koncentrationerne kan betragtes som observerede værdier af stokastiske variable: tallene er ikke ens fordi der er en vis biologisk variation og en vis forsøgsusikkerhed, og det kan passende modelleres som tilfældig variation. Det er derfor nærliggende at søge at modellere tallene ved hjælp af en regressionsmodel med koncentration som y-variabel og varighed som x-variabel. Man kan naturligvis ikke på forhånd vide om varigheden i sig selv er en hensigtsmæssig forklarende variabel. Måske viser det sig at man bedre kan beskrive koncentrationen som en lineær funktion af kvadratroden af varigheden end som en lineær funktion af selve varigheden, men det betyder blot at der er tale om en lineær regressionsmodel med kvadratroden af varigheden som forklarende variabel.

Vi vil antage at hypoxantinkoncentrationen kan beskrives ved en lineær regressionsmodel med hypoxivarigheden som uafhængig variabel.

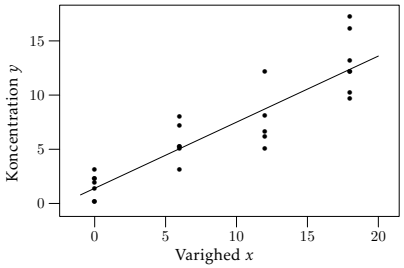
Vi lader x_1, x_2, x_3 og x_4 betegne de fire tidspunkter 0, 6, 12 og 18 min, og vi lader y_{ij} betegne den j-te koncentrationsværdi til tid x_i. Med de indførte betegnelser kan den foreslåede statistiske model for talmaterialet formuleres på den måde at y_{ij}-erne antages at være observerede værdier af uafhængige stokastiske variable Y_{ij}, hvor Y_{ij} er normalfordelt med $\mathrm{E}\, Y_{ij} = \alpha + \beta x_i$ og varians σ^2.

Estimaterne over α og β udregnes, f.eks. ved brug af formlerne på side 124, til $\widehat{\beta} = 0.61\,\mu\mathrm{mol\,l^{-1}\,min^{-1}}$ og $\widehat{\alpha} = 1.4\,\mu\mathrm{mol\,l^{-1}}$. Variansen estimeres til $s_{02}^2 = 4.85\,\mu\mathrm{mol^2\,l^{-2}}$ med $25 - 2 = 23$ frihedsgrader.

Middelfejlen på $\widehat{\beta}$ er $\sqrt{s_{02}^2/SS_x} = 0.06\,\mu\mathrm{mol\,l^{-1}\,min^{-1}}$, og på $\widehat{\alpha}$ er den $\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{SS_x}\right)s_{02}^2} = 0.7\,\mu\mathrm{mol\,l^{-1}}$ (sætning 7.3). Størrelsen af de to middelfejl viser at det er passende at angive $\widehat{\beta}$ med to og $\widehat{\alpha}$ med én decimal, så vi konkluderer at den estimerede regressionslinje er

$$y = 1.4\,\mu\mathrm{mol\,l^{-1}} + 0.61\,\mu\mathrm{mol\,l^{-1}\,min^{-1}}\,x\,.$$

Som en form for modelkontrol kan vi nu teste om hypoxantinkoncentrationen afhænger lineært (eller rettere affint) af hypoxiens varighed. Vi indlejrer derfor regressionsmodellen i en k-stikprøvemodel hvor grupperne bestemmes af x-erne på den måde at en gruppe består af observationer med samme x-værdi, jf. eksempel 11.1 side 176.



Figur 11.1
Kvælning af hunde: Sammenhørende værdier af hypoxantinkoncentration og hypoxivarighed, samt den estimerede regressionslinje.

Vi formulerer metoden i lineær algebra-sprog. Lad $L_1 = \{\boldsymbol{\xi} : \xi_{ij} = \alpha + \beta x_i\}$ være underrummet svarende til modellen med lineær sammenhæng mellem x og y, og lad $L = \{\boldsymbol{\xi} : \xi = \mu_i\}$ være underrummet svarende til k-stikprøvemodellen. Der gælder at $L_1 \subset L$. Fra afsnit 11.1 ved vi at teststørrelsen for at teste L_1 i forhold til L er

$$F = \frac{\frac{1}{\dim L - \dim L_1}\|\mathrm{p}\boldsymbol{y} - \mathrm{p}_1\boldsymbol{y}\|^2}{\frac{1}{n - \dim L}\|\boldsymbol{y} - \mathrm{p}\boldsymbol{y}\|^2}$$

hvor p og p_1 er projektionerne på L og L_1. Vi ved at $\dim L - \dim L_1 = 4 - 2 = 2$ og $n - \dim L = 25 - 4 = 21$. I den tidligere behandling af eksemplet (side 176) fandt vi $\|\boldsymbol{y} - \mathrm{p}\boldsymbol{y}\|^2$ til 101.32; $\|\mathrm{p}\boldsymbol{y} - \mathrm{p}_1\boldsymbol{y}\|^2$ kan f.eks. udregnes som $\|\boldsymbol{y} - \mathrm{p}_1\boldsymbol{y}\|^2 - \|\boldsymbol{y} - \mathrm{p}\boldsymbol{y}\|^2 = 111.50 - 101.32 = 10.18$. Teststørrelsen bliver dermed $F = 5.09/4.82 = 1.06$ der skal sammenholdes med F-fordelingen med 2 og 21 frihedsgrader. Tabelopslag viser at testsandsynligheden bliver over 30%, så hypotesen godtages, dvs. der synes at være en lineær sammenhæng mellem varigheden af hypoxien og hypoxantinkoncentrationen. Det fremgår også af figur 11.1.

11.7 Opgaver

Opgave 11.1: Indianere i Peru
Ændringer i menneskers livsbetingelser kan give sig udslag i fysiologiske ændringer, eksempelvis i ændret blodtryk.

En gruppe antropologer undersøgte hvordan blodtrykket ændrer sig hos peruvianske indianere der flyttes fra deres oprindelige primitive samfund i de høje Andesbjerge til den såkaldte civilisation, dvs. storbyen, der i øvrigt ligger i langt mindre højde over havets overflade end deres oprindelig bopæl (Davin (1975), her citeret efter Ryan et al. (1976)). Antropologerne udvalgte en stikprøve på 39 mænd over 21 år der havde undergået en sådan flytning. På hver af disse måltes blodtrykket (det systoliske og det diastoliske) samt en række baggrundsvariable, heriblandt alder, antal år siden flytningen, højde, vægt og puls. Desuden har man udregnet endnu en baggrundsvariabel, nemlig »brøkdel af livet levet i de nye omgivelser«, dvs. antal år siden flytning divideret med nuværende alder. Man forestillede sig at denne baggrundsvariabel kunne have stor »forklaringsevne«.

y	x_1	x_2	y	x_1	x_2
170	0.048	71.0	114	0.474	59.5
120	0.273	56.5	136	0.289	61.0
125	0.208	56.0	126	0.289	57.0
148	0.042	61.0	124	0.538	57.5
140	0.040	65.0	128	0.615	74.0
106	0.704	62.0	134	0.359	72.0
120	0.179	53.0	112	0.610	62.5
108	0.893	53.0	128	0.780	68.0
124	0.194	65.0	134	0.122	63.4
134	0.406	57.0	128	0.286	68.0
116	0.394	66.5	140	0.581	69.0
114	0.303	59.1	138	0.605	73.0
130	0.441	64.0	118	0.233	64.0
118	0.514	69.5	110	0.432	65.0
138	0.057	64.0	142	0.409	71.0
134	0.333	56.5	134	0.222	60.2
120	0.417	57.0	116	0.021	55.0
120	0.432	55.0	132	0.860	70.0
114	0.459	57.0	152	0.741	87.0
124	0.263	58.0			

Tabel 11.8
Indianere i Peru: Sammenhørende værdier af y: systolisk blodtryk (mm Hg), x_1: brøkdel af livet i de nye omgivelser, og x_2: vægt (kg).

Her vil vi ikke se på hele talmaterialet, men kun på *blodtrykket* (det systoliske) der skal optræde som y-variabel, og på de to x-variable *brøkdel af livet i de nye omgivelser* og *vægt*. Disse er angivet i tabel 11.8 (fra Ryan et al. (1976)).

1. Antropologerne mente at x_1, brøkdel levet i de nye omgivelser, var et godt mål for hvor længe personerne havde levet i de civiliserede omgivelser, og at det derfor måtte være interessant at se om x_1 kunne forklare variationen i blodtrykket y. Første skridt kunne derfor være at estimere en simpel lineær regressionsmodel med x_1 som forklarende variabel. Gør det!
2. Hvis man i et koordinatsystem afsætter y mod x_1, viser det sig imidlertid at det faktisk ikke virker særlig rimeligt at hævde at (middelværdien af) y afhænger lineært af x_1. Derfor må man give sig til at overveje om andre af de målte baggrundsvariable med fordel kan inddrages.
   Nu ved man at en persons *vægt* har betydning for den pågældendes blodtryk, så næste modelforslag kunne være en multipel regressionsmodel med både x_1 og x_2 som forklarende variable.
   a) Estimér parametrene i denne model. Hvad sker der med variansestimatet?
   b) Undersøg residualerne for at vurdere modellens kvalitet.
3. Giv en tolkning af slutmodellen i forhold til de peruvianske indianere.

Opgave 11.2
Dette er en berømt tosidet variansanalyse-opgave fra Københavns Universitet; som man vil se, indeholder opgaveteksten en vis portion underforstået lokalkendskab:

En student cykler hver dag fra sit hjem til H.C. Ørsted Institutet og tilbage igen. Han kan cykle to forskellige veje, én som han plejer at benytte, og én som han mistænker for at være en genvej. For at undersøge om det faktisk er en genvej, måler han nogle gange hvor lang tid han er om turen. Resultaterne fremgår af nedenstående skema hvor tiderne er opgivet med 10 sekunder som enhed og ud fra et beregningsnulpunkt på 9 minutter.

	genvej			sædvanlig vej				
udtur	4	−1	3	13	8	11	5	7
hjemtur	11	6		16	18	17	21	19

Da det som bekendt kan være vanskeligt at slippe væk fra H.C. Ørsted Institutet på cykel, tager hjemturen gennemsnitligt længere tid end udturen.

Havde studenten ret i sin mistanke?

Vejledning: Det er klart at disse resultater må kunne behandles ved tosidet variansanalyse; da cellerne imidlertid ikke indeholder lige mange observationer, kan den sædvanlige formel for projektionen på underrummet svarende til additivitetshypotesen ikke bruges.

Opgave 11.3
Betragt den simple lineære regressionsmodel $\mathrm{E}\,Y = \alpha + x\beta$, og antag at der foreligger et antal sammenhørende værdier (y_i, x_i), $i = 1, 2, \ldots, n$.

Hvordan ser designmatricen ud? Skriv normalligningerne op og løs dem.

Find formler for middelfejlene (ɔ: standardafvigelserne) på $\widehat{\alpha}$ og $\widehat{\beta}$, samt en formel for korrelationen mellem de to estimatorer. Tip: udnyt formel (11.8).

I visse typer forsøg kan eksperimentator (eller statistikeren) selv bestemme x-værdierne inden for visse grænser. Hvordan skal man vælge x-erne?

A En udledning af normalfordelingen

Normalfordelingen er meget benyttet i statistiske modeller. Nogle gange kan man begrunde brugen af den ved henvisninger til Den Centrale Grænseværdisætning, andre gange skyldes det nærmest matematiske bekvemmelighedsgrunde. Der findes imidlertid også argumentationer af formen »hvis man har tænkt sig at analysere data på den-og-den måde, så svarer det indirekte til at antage at observationerne stammer fra den-og-den fordeling«. I dette afsnit vil vi præsentere et eksempel på en sådan argumentation; grundideen hidrører fra Gauß (1809, bog II, afsnit 3).

Den overordnede opgave er at finde et forslag til en type sandsynlighedsfordelinger der kan bruges til at beskrive hvordan observationer fordeler sig tilfældigt omkring en bestemt ukendt værdi μ. Vi gør nogle antagelser:

Antagelse 1: Fordelingerne er kontinuerte fordelinger på $\mathbb{R}$, dvs. de har en kontinuert tæthedsfunktion.

Antagelse 2: Parameteren μ kan have en vilkårlig reel værdi og er en *positionsparameter*, dvs. modelfunktionen er af formen

$$f(x-\mu), \quad (x,\mu) \in \mathbb{R}^2,$$

for en passende funktion f som er defineret på hele $\mathbb{R}$ (og som er kontinuert ifølge antagelse 1).

Antagelse 3: Funktionen f er kontinuert differentiabel.

Antagelse 4: Maksimaliseringsestimatet for μ skal være gennemsnittet af observationerne, mere præcist: for enhver stikprøve $x_1, x_2, \ldots, x_n$ fra fordelingen med tæthedsfunktion $x \mapsto f(x-\mu)$, skal maksimaliseringsestimatet være gennemsnittet $\overline{x}$.

Vi skal nu se hvad man kan deducere herudfra.

Log-likelihoodfunktionen svarende til observationerne $x_1, x_2, \ldots, x_n$ bliver

$$\ln L(\mu) = \sum_{i=1}^{n} \ln f(x_i - \mu).$$

Den er differentiabel med

$$(\ln L)'(\mu) = \sum_{i=1}^{n} g(x_i - \mu)$$

hvor $g = -(\ln f)'$.

Da f er en tæthedsfunktion, vil der gælde at $\liminf_{x\to\pm\infty} f(x) = 0$, og derfor er $\liminf_{x\to\pm\infty} \ln L(\mu) = -\infty$; det betyder at $\ln L$ antager sit maksimum i (mindst) et punkt $\widehat{\mu}$, og at dette er et stationært punkt, dvs. $(\ln L)'(\widehat{\mu}) = 0$. Da det forlanges at $\widehat{\mu} = \overline{x}$, er vi hermed nået frem til at der skal gælde at

$$\sum_{i=1}^{n} g(x_i - \overline{x}) = 0 \tag{A.1}$$

for alle stikprøver $x_1, x_2, \ldots, x_n$.

Tag nu en stikprøve med to elementer x og $-x$; da gennemsnittet af disse to er 0, giver (A.1) at $g(x) + g(-x) = 0$, dvs. $g(-x) = -g(x)$ for ethvert x. Specielt er $g(0) = 0$.

Tag dernæst en stikprøve med tre elementer x, y og $-(x+y)$; da deres gennemsnit er 0, giver (A.1) at $g(x) + g(y) + g(-(x+y)) = 0$, og da vi netop har vist at g er en ulige funktion, kan vi konkludere at $g(x+y) = g(x) + g(y)$ for alle x og y. Det følger nu af sætning A.1 nedenfor at funktionen g må være af formen $g(x) = cx$ for en eller anden konstant c. Da g var defineret til at være $-(\ln f)'$, må $\ln f$ derfor være af formen $\ln f(x) = b - \frac{1}{2}cx^2$, hvor b er en integrationskonstant, og så bliver f af formen $f(x) = a\exp(-\frac{1}{2}cx^2)$, hvor $a = e^b$. Da f skal integrere til 1, må konstanten c nødvendigvis være positiv – og man kunne så passende omdøbe den til $1/\sigma^2$ – og konstanten a er entydigt bestemt (og der gælder at $a = 1/\sqrt{2\pi\sigma^2}$, jf. side 74). Funktionen f må altså nødvendigvis være tæthedsfunktionen for normalfordelingen med middelværdi 0 og varians σ^2, og den søgte type fordelinger er således normalfordelingerne med middelværdi μ og varians σ^2.

Opgave A.1
Ovenstående undersøgelse handler om en situation hvor der er tale om observationer fra en kontinuert fordeling på den reelle akse, og hvor man ønsker at maksimaliseringsestimatoren for *positions*parameteren skal være gennemsnittet af observationerne.

Antag at man i stedet har at gøre med observationer fra en kontinuert fordeling på den *positive* halvakse, at der indgår en *skala*parameter, og at maksimaliseringsestimatoren for *skala*parameteren skal være gennemsnittet af observationerne. – Hvad kan man i denne situation sige om observationernes fordeling?

Cauchys funktionalligning

I dette afsnit vises en sætning som anvendes i det forrige, og som derudover må siges at høre med til den »matematisk almendannelse«.

Sætning A.1
Antag at f er en reel funktion med den egenskab at

$$f(x+y) = f(x) + f(y) \tag{A.2}$$

for alle $x, y \in \mathbb{R}$. Hvis f er kontinuert i et punkt $x_0 \neq 0$, så findes der et tal $c \in \mathbb{R}$ sådan at $f(x) = cx$ for alle $x \in \mathbb{R}$.

Bevis
Vi ser først på hvad man kan deducere ud fra betingelsen $f(x+y) = f(x) + f(y)$.

1. Hvis man sætter $x = y = 0$, fås at $f(0) = f(0) + f(0)$, dvs. $f(0) = 0$.
2. Hvis man sætter $y = -x$, fås at for vilkårligt $x \in \mathbb{R}$ er $f(x) + f(-x) = f(0) = 0$, altså $f(-x) = -f(x)$.
3. Ved gentagne anvendelser af (A.2) fås at for vilkårlige $x_1, x_2, \ldots, x_n$ er

$$f(x_1 + x_2 + \ldots + x_n) = f(x_1) + f(x_2) + \ldots + f(x_n). \tag{A.3}$$

Tag nu et reelt tal $x \neq 0$ og naturlige tal p og q, og sæt $a = p/q$. Da

$$f(\underbrace{ax + ax + \ldots + ax}_{q \text{ led}}) = f(\underbrace{x + x + \ldots + x}_{p \text{ led}}),$$

giver (A.3) at $qf(ax) = pf(x)$ eller $f(ax) = af(x)$. Vi kan konkludere at for alle rationale tal a og alle reelle tal x er

$$f(ax) = af(x). \tag{A.4}$$

Indtil videre har vi ikke benyttet at f er kontinuert i x_0, men det kommer nu. Lad x være et reelt tal forskelligt fra 0. Der findes en følge (a_n) af rationale tal som konvergerer mod x_0/x, og som alle er forskellige fra 0. Da følgen $(a_n x)$ konvergerer mod x_0 og f er kontinuert i x_0, vil

$$\frac{f(a_n x)}{a_n} \longrightarrow \frac{f(x_0)}{x_0/x} = \frac{f(x_0)}{x_0}\, x.$$

Men ifølge (A.4) er $\dfrac{f(a_n x)}{a_n} = f(x)$ for ethvert a_n, så der må åbenbart gælde at $f(x) = \dfrac{f(x_0)}{x_0}\, x$, altså $f(x) = c\,x$ hvor $c = \dfrac{f(x_0)}{x_0}$. Da x var vilkårlig (og da c ikke afhænger af x), er sætningen hermed vist. □

B Nogle resultater fra lineær algebra

De næste sider præsenterer nogle resultater fra lineær algebra, nærmere bestemt fra teorien for endeligdimensionale reelle vektorrum med indre produkt.

Notation og definitioner

Vektorrummet betegnes typisk V. *Underrum* betegnes $L, L_1, L_2, \ldots$ *Vektorer* betegnes normalt med fede bogstaver ($\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{v}$ osv.). Nulvektoren er $\mathbf{0}$. Når vi repræsenterer vektorerne ved hjælp af deres koordinatsæt i forhold til en valgt basis, skriver vi koordinatsættene som *søjlematricer*.

V L

$\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}$ $\mathbf{0}$

Lineære afbildninger [og deres matricer] betegnes ofte med bogstaver som A og B; den *transponerede* til A betegnes A'. *Nulrummet* for A betegnes $\mathcal{N}(A)$ og *billedrummet* betegnes $\mathcal{R}(A)$; *rangen* af A er tallet $\dim \mathcal{R}(A)$. Den identiske afbildning [enhedsmatricen] betegnes I.

A

A' $\mathcal{N}(A)$

$\mathcal{R}(A)$

I

Skalarproduktet eller det indre produkt af $\boldsymbol{u}$ og $\boldsymbol{v}$ skrives $\langle \boldsymbol{u}, \boldsymbol{v} \rangle$ og i matrixnotation $\boldsymbol{u}'\boldsymbol{v}$. *Længden* af $\boldsymbol{u}$ skrives $\|\boldsymbol{u}\|$.

$\langle \boldsymbol{u}, \boldsymbol{v} \rangle$

$\boldsymbol{u}'\boldsymbol{v}$ $\|\boldsymbol{u}\|$

To vektorer $\boldsymbol{u}$ og $\boldsymbol{v}$ er *ortogonale*, kort $\boldsymbol{u} \perp \boldsymbol{v}$, hvis $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0$. To underrum L_1 og L_2 er ortogonale, kort $L_1 \perp L_2$, hvis enhver vektor i L_1 er ortogonal på enhver vektor i L_2. I så fald defineres den ortogonale direkte sum af L_1 og L_2 som underrummet

$\boldsymbol{u} \perp \boldsymbol{v}$

$L_1 \perp L_2$

$L_1 \oplus L_2$

$$L_1 \oplus L_2 = \{\boldsymbol{u}_1 + \boldsymbol{u}_2 : \boldsymbol{u}_1 \in L_1, \boldsymbol{u}_2 \in L_2\}.$$

Hvis $\boldsymbol{v} \in L_1 \oplus L_2$, så har $\boldsymbol{v}$ en entydig opspaltning som $\boldsymbol{v} = \boldsymbol{u}_1 + \boldsymbol{u}_2$ hvor $\boldsymbol{u}_1 \in L_1$ og $\boldsymbol{u}_2 \in L_2$. Der gælder at $\dim L_1 \oplus L_2 = \dim L_1 + \dim L_2$.

Det *ortogonale komplement* til underrummet L betegnes $L^\perp$. Der gælder at $V = L \oplus L^\perp$. *Ortogonalprojektionen* af V på underrummet L er den lineære afbildning $\mathfrak{p} : V \to V$ for hvilken $\mathfrak{p}\boldsymbol{x} \in L$ og $\boldsymbol{x} - \mathfrak{p}\boldsymbol{x} \in L^\perp$ for alle $\boldsymbol{x} \in V$.

$L^\perp$

$\mathfrak{p}$

En symmetrisk lineær afbildning [en symmetrisk matrix] A er *positivt semidefinit* hvis $\langle \boldsymbol{x}, A\boldsymbol{x} \rangle \geq 0$ [eller $\boldsymbol{x}'A\boldsymbol{x} \geq 0$] for alle $\boldsymbol{x}$; den er *positivt definit* hvis ulighedstegnet er skarpt for alle $\boldsymbol{x} \neq \mathbf{0}$. Vi vil undertiden bruge skrivemåderne $A \geq 0$ og $A > 0$ til at angive at A er positivt semidefinit hhv. positivt definit.

$A \geq 0$

$A > 0$

Forskellige resultater

Denne sætning turde være velkendt fra lineær algebra:

Sætning B.1
Lad A være en lineær afbildning af $\mathbb{R}^p$ ind i $\mathbb{R}^n$. Da gælder at $\mathcal{R}(A)$ og $\mathcal{N}(A')$ er hinandens ortogonale komplementer (i $\mathbb{R}^n$), eller kort $\mathbb{R}^n = \mathcal{R}(A) \oplus \mathcal{N}(A')$.

Korollar B.2
Lad A være en symmetrisk lineær afbildning af $\mathbb{R}^n$ ind i sig selv. Da gælder at $\mathcal{R}(A)$ og $\mathcal{N}(A)$ er hinandens ortogonale komplementer (i $\mathbb{R}^n$).

Korollar B.3
$\mathcal{R}(A) = \mathcal{R}(AA')$.

Bevis for korollar B.3
Ifølge sætningen kan vi vise korollaret ved at vise at $\mathcal{N}(A') = \mathcal{N}(AA')$, altså at $A'\boldsymbol{u} = \mathbf{0} \Leftrightarrow AA'\boldsymbol{u} = \mathbf{0}$.

Det er klart at $A'\mathbf{0} \Rightarrow AA'\boldsymbol{u} = \mathbf{0}$. Vi behøver således kun vise at $AA'\boldsymbol{u} = \mathbf{0} \Rightarrow A'\boldsymbol{u} = \mathbf{0}$. Antag derfor at $AA'\boldsymbol{u} = \mathbf{0}$. Da $A(A'\boldsymbol{u}) = \mathbf{0}$, er $A'\boldsymbol{u} \in \mathcal{N}(A) = \mathcal{R}(A')^\perp$, og da automatisk $A'\boldsymbol{u} \in \mathcal{R}(A')$, er $A'\boldsymbol{u} \in \mathcal{R}(A')^\perp \cap \mathcal{R}(A') = \{\mathbf{0}\}$. □

Sætning B.4
Lad A og B være injektive lineære afbildninger af $\mathbb{R}^p$ ind i $\mathbb{R}^n$ således at $AA' = BB'$. Så findes en isometri C af $\mathbb{R}^p$ ind i sig selv med den egenskab at $A = BC$.

Bevis
Sæt $L = \mathcal{R}(A)$. Ifølge antagelsen og korollar B.3 er $L = \mathcal{R}(A) = \mathcal{R}(AA') = \mathcal{R}(BB') = \mathcal{R}(B)$.

Da A er injektiv, er $\mathcal{N}(A) = \{\mathbf{0}\}$; ifølge sætning B.1 er da $\mathcal{R}(A') = \{\mathbf{0}\}^\perp = \mathbb{R}^p$, dvs. til et givet $\boldsymbol{u} \in \mathbb{R}^p$ har ligningen $A'\boldsymbol{x} = \boldsymbol{u}$ mindst en løsning $\boldsymbol{x} \in \mathbb{R}^n$, og da $\mathcal{N}(A') = \mathcal{R}(A)^\perp = L^\perp$, er der i L præcis én løsning $\boldsymbol{x}(\boldsymbol{u})$ til $A'\boldsymbol{x} = \boldsymbol{u}$.

Vi vil vise at den herved definerede afbildning $\boldsymbol{u} \mapsto \boldsymbol{x}(\boldsymbol{u})$ af $\mathbb{R}^p$ ind i L er lineær. Lad $\boldsymbol{u}_1$ og $\boldsymbol{u}_2$ være to vektorer i $\mathbb{R}^p$ og α_1 og α_2 to skalarer. Da er

$$\begin{aligned} A'\Big(\boldsymbol{x}(\alpha_1\boldsymbol{u}_1 + \alpha_2\boldsymbol{u}_2)\Big) &= \alpha_1\boldsymbol{u}_1 + \alpha_2\boldsymbol{u}_2 \\ &= \alpha_1 A'\boldsymbol{x}(\boldsymbol{u}_1) + \alpha_2 A'\boldsymbol{x}(\boldsymbol{u}_1) \\ &= A'\Big(\alpha_1\boldsymbol{x}(\boldsymbol{u}_1) + \alpha_2\boldsymbol{x}(\boldsymbol{u}_2)\Big), \end{aligned}$$

og da ligningen $A'\boldsymbol{x} = \boldsymbol{u}$ som nævnt har en entydig løsning $\boldsymbol{x} \in L$, følger heraf at $\boldsymbol{x}(\alpha_1\boldsymbol{u}_1 + \alpha_2\boldsymbol{u}_2) = \alpha_1\boldsymbol{x}(\boldsymbol{u}_1) + \alpha_2\boldsymbol{x}(\boldsymbol{u}_2)$, hvilket skulle vises.

Vi kan derfor definere en lineær afbildning $C : \mathbb{R}^p \to \mathbb{R}^p$ givet ved $C\boldsymbol{u} = B'\boldsymbol{x}(\boldsymbol{u})$. For $\boldsymbol{u} \in \mathbb{R}^p$ er da $BC\boldsymbol{u} = BB'\boldsymbol{x}(\boldsymbol{u}) = AA'\boldsymbol{x}(\boldsymbol{u}) = A\boldsymbol{u}$, altså $A = BC$.

Afslutningsvis vil vi vise at C bevarer indre produkt, hvoraf specielt følger at den er en isometri: For $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^p$ er $\langle C\boldsymbol{u}_1, C\boldsymbol{u}_2\rangle = \langle B'\boldsymbol{x}(\boldsymbol{u}_1), B'\boldsymbol{x}(\boldsymbol{u}_2)\rangle = \langle \boldsymbol{x}(\boldsymbol{u}_1), BB'\boldsymbol{x}(\boldsymbol{u}_2)\rangle = \langle \boldsymbol{x}(\boldsymbol{u}_1), AA'\boldsymbol{x}(\boldsymbol{u}_2)\rangle = \langle A'\boldsymbol{x}(\boldsymbol{u}_1), A'\boldsymbol{x}(\boldsymbol{u}_2)\rangle = \langle \boldsymbol{u}_1, \boldsymbol{u}_2\rangle$. □

Sætning B.5
Antag at A er en symmetrisk og positiv semidefinit lineær afbildning af $\mathbb{R}^n$ ind i sig selv. Der findes netop én symmetrisk og positivt semidefinit lineær afbildning $A^{1/2}$ af $\mathbb{R}^n$ ind i sig selv med den egenskab at $A^{1/2}A^{1/2} = A$.

Bevis
Lad $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ være en ortonormalbasis af egenvektorer for A, og lad de tilhørende egenværdier være $\lambda_1, \lambda_2, \ldots, \lambda_n \geq 0$. Hvis vi definerer $A^{1/2}$ som den lineære afbildning der afbilder $\boldsymbol{e}_j$ over i $\lambda_j^{1/2}\boldsymbol{e}_j$, $j = 1, 2, \ldots, n$, har vi en afbildning med den ønskede egenskab.

Antag at $A^\star$ er en løsning, dvs. $A^\star$ er symmetrisk og positivt semidefinit og $A^\star A^\star = A$. Der findes en ortonormalbasis $\boldsymbol{e}_1^\star, \boldsymbol{e}_2^\star, \ldots, \boldsymbol{e}_n^\star$ af egenvektorer for $A^\star$; lad de tilhørende egenværdier være $\mu_1, \mu_2, \ldots, \mu_n \geq 0$. Da $A\boldsymbol{e}_j^\star = A^\star A^\star \boldsymbol{e}_j^\star = \mu_j^2 \boldsymbol{e}_j^\star$, er $\boldsymbol{e}_j^\star$ en A-egenvektor med tilhørende egenværdi μ_j^2. Deraf følger at $A^\star$ nødvendigvis må have de samme egenrum som A og med egenværdier som er de ikke-negative kvadratrødder af de tilsvarende A-egenværdier, altså $A^\star = A^{1/2}$. □

Sætning B.6
Antag at A er en symmetrisk og positiv semidefinit lineær afbildning af $\mathbb{R}^n$ ind i sig selv, og sæt $p = \dim \mathcal{R}(A)$.

Da findes en injektiv lineær afbildning B af $\mathbb{R}^p$ ind i $\mathbb{R}^n$ med den egenskab at $BB' = A$, og B er entydigt bestemt på nær isometri.

Bevis
Lad $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ være en ortonormalbasis af egenvektorer for A med tilhørende egenvektorer $\lambda_1, \lambda_2, \ldots, \lambda_n$ hvor $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ og $\lambda_{p+1} = \lambda_{p+2} = \ldots = \lambda_n = 0$.

Som B kan vi bruge den lineære afbildning hvis matrixrepræsentation i forhold til standardbasen i $\mathbb{R}^p$ og basen $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ i $\mathbb{R}^n$ er den $n \times p$-matrix som på den (i, j)-te plads har $\lambda_i^{1/2}$ hvis $i = j \leq p$ og 0 ellers. Entydigheden af B følger f.eks. af sætning B.4. □

C Tabeller

I FØR-COMPUTER-ÆRAEN var tabelværker over blandt andet fordelingsfunktioner og inverse fordelingsfunktioner for de gængse fordelinger et uundværligt arbejdsredskab for den praktisk arbejdende statistiker. De følgende sider indeholder nogle få mindre eksempler på sådanne statistiske tabeller.

Til orientering: For en sandsynlighedsfordeling med strengt voksende og kontinuert fordelingsfunktion F defineres *α-fraktilen* x_α som løsningen til ligningen $F(x) = \alpha$; her er $0 < \alpha < 1$. (Hvis F ikke vides at være strengt voksende og kontinuert, kan man definere mængden af α-fraktiler som det afsluttede interval med endepunkter $\sup\{x : F(x) < \alpha\}$ og $\inf\{x : F(x) > \alpha\}$.)

Fordelingsfunktionen $\Phi(u)$ for standardnormalfordelingen

u	$u+5$	$\Phi(u)$	u	$u+5$	$\Phi(u)$
			0.00	5.00	0.5000
−3.75	1.25	0.0001	0.10	5.10	0.5398
−3.50	1.50	0.0002	0.20	5.20	0.5793
−3.25	1.75	0.0006	0.30	5.30	0.6179
−3.00	2.00	0.0013	0.40	5.40	0.6554
−2.80	2.20	0.0026	0.50	5.50	0.6915
−2.60	2.40	0.0047	0.60	5.60	0.7257
−2.40	2.60	0.0082	0.70	5.70	0.7580
−2.20	2.80	0.0139	0.80	5.80	0.7881
−2.00	3.00	0.0228	0.90	5.90	0.8159
−1.90	3.10	0.0287	1.00	6.00	0.8413
−1.80	3.20	0.0359	1.10	6.10	0.8643
−1.70	3.30	0.0446	1.20	6.20	0.8849
−1.60	3.40	0.0548	1.30	6.30	0.9032
−1.50	3.50	0.0668	1.40	6.40	0.9192
−1.40	3.60	0.0808	1.50	6.50	0.9332
−1.30	3.70	0.0968	1.60	6.60	0.9452
−1.20	3.80	0.1151	1.70	6.70	0.9554
−1.10	3.90	0.1357	1.80	6.80	0.9641
−1.00	4.00	0.1587	1.90	6.90	0.9713
−0.90	4.10	0.1841	2.00	7.00	0.9772
−0.80	4.20	0.2119	2.20	7.20	0.9861
−0.70	4.30	0.2420	2.40	7.40	0.9918
−0.60	4.40	0.2743	2.60	7.60	0.9953
−0.50	4.50	0.3085	2.80	7.80	0.9974
−0.40	4.60	0.3446	3.00	8.00	0.9987
−0.30	4.70	0.3821	3.25	8.25	0.9994
−0.20	4.80	0.4207	3.50	8.50	0.9998
−0.10	4.90	0.4602	3.75	8.75	0.9999

Fraktiler i standardnormalfordelingen

α	$u_\alpha = \Phi^{-1}(\alpha)$	$u_\alpha + 5$	α	$u_\alpha = \Phi^{-1}(\alpha)$	$u_\alpha + 5$
			0.500	0.000	5.000
0.001	−3.090	1.910	0.520	0.050	5.050
0.002	−2.878	2.122	0.540	0.100	5.100
0.005	−2.576	2.424	0.560	0.151	5.151
0.010	−2.326	2.674	0.580	0.202	5.202
0.015	−2.170	2.830	0.600	0.253	5.253
0.020	−2.054	2.946	0.620	0.305	5.305
0.025	−1.960	3.040	0.640	0.358	5.358
0.030	−1.881	3.119	0.660	0.412	5.412
0.035	−1.812	3.188	0.680	0.468	5.468
0.040	−1.751	3.249	0.700	0.524	5.524
0.045	−1.695	3.305	0.720	0.583	5.583
0.050	−1.645	3.355	0.740	0.643	5.643
0.055	−1.598	3.402	0.750	0.674	5.674
0.060	−1.555	3.445	0.760	0.706	5.706
0.070	−1.476	3.524	0.780	0.772	5.772
0.080	−1.405	3.595	0.800	0.842	5.842
0.090	−1.341	3.659	0.825	0.935	5.935
0.100	−1.282	3.718	0.850	1.036	6.036
0.125	−1.150	3.850	0.875	1.150	6.150
0.150	−1.036	3.964	0.900	1.282	6.282
0.175	−0.935	4.065	0.910	1.341	6.341
0.200	−0.842	4.158	0.920	1.405	6.405
0.220	−0.772	4.228	0.930	1.476	6.476
0.240	−0.706	4.294	0.940	1.555	6.555
0.250	−0.674	4.326	0.945	1.598	6.598
0.260	−0.643	4.357	0.950	1.645	6.645
0.280	−0.583	4.417	0.955	1.695	6.695
0.300	−0.524	4.476	0.960	1.751	6.751
0.320	−0.468	4.532	0.965	1.812	6.812
0.340	−0.412	4.588	0.970	1.881	6.881
0.360	−0.358	4.642	0.975	1.960	6.960
0.380	−0.305	4.695	0.980	2.054	7.054
0.400	−0.253	4.747	0.985	2.170	7.170
0.420	−0.202	4.798	0.990	2.326	7.326
0.440	−0.151	4.849	0.995	2.576	7.576
0.460	−0.100	4.900	0.998	2.878	7.878
0.480	−0.050	4.950	0.999	3.090	8.090

Fraktiler i χ^2-fordelingen med f frihedsgrader

	Sandsynlighed i procent						
f	50	90	95	97.5	99	99.5	99.9
1	0.45	2.71	3.84	5.02	6.63	7.88	10.83
2	1.39	4.61	5.99	7.38	9.21	10.60	13.82
3	2.37	6.25	7.81	9.35	11.34	12.84	16.27
4	3.36	7.78	9.49	11.14	13.28	14.86	18.47
5	4.35	9.24	11.07	12.83	15.09	16.75	20.52
6	5.35	10.64	12.59	14.45	16.81	18.55	22.46
7	6.35	12.02	14.07	16.01	18.48	20.28	24.32
8	7.34	13.36	15.51	17.53	20.09	21.95	26.12
9	8.34	14.68	16.92	19.02	21.67	23.59	27.88
10	9.34	15.99	18.31	20.48	23.21	25.19	29.59
11	10.34	17.28	19.68	21.92	24.72	26.76	31.26
12	11.34	18.55	21.03	23.34	26.22	28.30	32.91
13	12.34	19.81	22.36	24.74	27.69	29.82	34.53
14	13.34	21.06	23.68	26.12	29.14	31.32	36.12
15	14.34	22.31	25.00	27.49	30.58	32.80	37.70
16	15.34	23.54	26.30	28.85	32.00	34.27	39.25
17	16.34	24.77	27.59	30.19	33.41	35.72	40.79
18	17.34	25.99	28.87	31.53	34.81	37.16	42.31
19	18.34	27.20	30.14	32.85	36.19	38.58	43.82
20	19.34	28.41	31.41	34.17	37.57	40.00	45.31
21	20.34	29.62	32.67	35.48	38.93	41.40	46.80
22	21.34	30.81	33.92	36.78	40.29	42.80	48.27
23	22.34	32.01	35.17	38.08	41.64	44.18	49.73
24	23.34	33.20	36.42	39.36	42.98	45.56	51.18
25	24.34	34.38	37.65	40.65	44.31	46.93	52.62
26	25.34	35.56	38.89	41.92	45.64	48.29	54.05
27	26.34	36.74	40.11	43.19	46.96	49.64	55.48
28	27.34	37.92	41.34	44.46	48.28	50.99	56.89
29	28.34	39.09	42.56	45.72	49.59	52.34	58.30
30	29.34	40.26	43.77	46.98	50.89	53.67	59.70
31	30.34	41.42	44.99	48.23	52.19	55.00	61.10
32	31.34	42.58	46.19	49.48	53.49	56.33	62.49
33	32.34	43.75	47.40	50.73	54.78	57.65	63.87
34	33.34	44.90	48.60	51.97	56.06	58.96	65.25
35	34.34	46.06	49.80	53.20	57.34	60.27	66.62
36	35.34	47.21	51.00	54.44	58.62	61.58	67.99
37	36.34	48.36	52.19	55.67	59.89	62.88	69.35
38	37.34	49.51	53.38	56.90	61.16	64.18	70.70
39	38.34	50.66	54.57	58.12	62.43	65.48	72.05
40	39.34	51.81	55.76	59.34	63.69	66.77	73.40

Fraktiler i χ^2-fordelingen med f frihedsgrader

	Sandsynlighed i procent						
f	50	90	95	97.5	99	99.5	99.9
41	40.34	52.95	56.94	60.56	64.95	68.05	74.74
42	41.34	54.09	58.12	61.78	66.21	69.34	76.08
43	42.34	55.23	59.30	62.99	67.46	70.62	77.42
44	43.34	56.37	60.48	64.20	68.71	71.89	78.75
45	44.34	57.51	61.66	65.41	69.96	73.17	80.08
46	45.34	58.64	62.83	66.62	71.20	74.44	81.40
47	46.34	59.77	64.00	67.82	72.44	75.70	82.72
48	47.34	60.91	65.17	69.02	73.68	76.97	84.04
49	48.33	62.04	66.34	70.22	74.92	78.23	85.35
50	49.33	63.17	67.50	71.42	76.15	79.49	86.66
51	50.33	64.30	68.67	72.62	77.39	80.75	87.97
52	51.33	65.42	69.83	73.81	78.62	82.00	89.27
53	52.33	66.55	70.99	75.00	79.84	83.25	90.57
54	53.33	67.67	72.15	76.19	81.07	84.50	91.87
55	54.33	68.80	73.31	77.38	82.29	85.75	93.17
56	55.33	69.92	74.47	78.57	83.51	86.99	94.46
57	56.33	71.04	75.62	79.75	84.73	88.24	95.75
58	57.33	72.16	76.78	80.94	85.95	89.48	97.04
59	58.33	73.28	77.93	82.12	87.17	90.72	98.32
60	59.33	74.40	79.08	83.30	88.38	91.95	99.61
61	60.33	75.51	80.23	84.48	89.59	93.19	100.89
62	61.33	76.63	81.38	85.65	90.80	94.42	102.17
63	62.33	77.75	82.53	86.83	92.01	95.65	103.44
64	63.33	78.86	83.68	88.00	93.22	96.88	104.72
65	64.33	79.97	84.82	89.18	94.42	98.11	105.99
66	65.33	81.09	85.96	90.35	95.63	99.33	107.26
67	66.33	82.20	87.11	91.52	96.83	100.55	108.53
68	67.33	83.31	88.25	92.69	98.03	101.78	109.79
69	68.33	84.42	89.39	93.86	99.23	103.00	111.06
70	69.33	85.53	90.53	95.02	100.43	104.21	112.32
71	70.33	86.64	91.67	96.19	101.62	105.43	113.58
72	71.33	87.74	92.81	97.35	102.82	106.65	114.84
73	72.33	88.85	93.95	98.52	104.01	107.86	116.09
74	73.33	89.96	95.08	99.68	105.20	109.07	117.35
75	74.33	91.06	96.22	100.84	106.39	110.29	118.60
76	75.33	92.17	97.35	102.00	107.58	111.50	119.85
77	76.33	93.27	98.48	103.16	108.77	112.70	121.10
78	77.33	94.37	99.62	104.32	109.96	113.91	122.35
79	78.33	95.48	100.75	105.47	111.14	115.12	123.59
80	79.33	96.58	101.88	106.63	112.33	116.32	124.84

90% fraktiler i F-fordelingen.

f_1 er antal frihedsgrader for tælleren, f_2 er antal frihedsgrader for nævneren.

	f_1										
f_2	1	2	3	4	5	6	7	8	9	10	15
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	61.22
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.42
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.20
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.87
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.24
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.87
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.63
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.46
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.34
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.24
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.17
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.10
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.05
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.01
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	1.97
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.94
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.91
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.89
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.86
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.84
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.81
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.78
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.76
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.74
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.72
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.66
50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73	1.63
75	2.77	2.37	2.16	2.02	1.93	1.85	1.80	1.75	1.72	1.69	1.58
100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66	1.56
150	2.74	2.34	2.12	1.98	1.89	1.81	1.76	1.71	1.67	1.64	1.53
200	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66	1.63	1.52
300	2.72	2.32	2.10	1.96	1.87	1.79	1.74	1.69	1.65	1.62	1.51
400	2.72	2.32	2.10	1.96	1.86	1.79	1.73	1.69	1.65	1.61	1.50
500	2.72	2.31	2.09	1.96	1.86	1.79	1.73	1.68	1.64	1.61	1.50

95% fraktiler i F-fordelingen.

f_1 er antal frihedsgrader for tælleren, f_2 er antal frihedsgrader for nævneren.

	f_1										
f_2	1	2	3	4	5	6	7	8	9	10	15
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	245.95
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87
75	3.97	3.12	2.73	2.49	2.34	2.22	2.13	2.06	2.01	1.96	1.80
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.73
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.72
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.70
400	3.86	3.02	2.63	2.39	2.24	2.12	2.03	1.96	1.90	1.85	1.69
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.69

97.5% fraktiler i F-fordelingen.

f_1 er antal frihedsgrader for tælleren, f_2 er antal frihedsgrader for nævneren.

	f_1										
f_2	1	2	3	4	5	6	7	8	9	10	15
1	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63	984.87
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.43
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.25
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.66
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.27
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.57
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.10
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.77
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.52
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.33
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.18
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.05
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	2.95
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.86
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.79
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.72
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.67
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.62
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.57
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.50
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.44
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.39
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.34
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.31
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.18
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.11
75	5.23	3.88	3.30	2.96	2.74	2.58	2.46	2.37	2.29	2.22	2.01
100	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	1.97
150	5.13	3.78	3.20	2.87	2.65	2.49	2.37	2.28	2.20	2.13	1.92
200	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18	2.11	1.90
300	5.07	3.73	3.16	2.83	2.61	2.45	2.33	2.23	2.16	2.09	1.88
400	5.06	3.72	3.15	2.82	2.60	2.44	2.32	2.22	2.15	2.08	1.87
500	5.05	3.72	3.14	2.81	2.59	2.43	2.31	2.22	2.14	2.07	1.86

99% fraktiler i F-fordelingen.

f_1 er antal frihedsgrader for tælleren, f_2 er antal frihedsgrader for nævneren.

	f_1										
f_2	1	2	3	4	5	6	7	8	9	10	15
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6157.28
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.43
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	26.87
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.20
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.72
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.25
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.82
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.66
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.52
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.41
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.31
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.23
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.15
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	2.98
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	2.89
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.81
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.75
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.70
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.52
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.42
75	6.99	4.90	4.05	3.58	3.27	3.05	2.89	2.76	2.65	2.57	2.29
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.22
150	6.81	4.75	3.91	3.45	3.14	2.92	2.76	2.63	2.53	2.44	2.16
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.13
300	6.72	4.68	3.85	3.38	3.08	2.86	2.70	2.57	2.47	2.38	2.10
400	6.70	4.66	3.83	3.37	3.06	2.85	2.68	2.56	2.45	2.37	2.08
500	6.69	4.65	3.82	3.36	3.05	2.84	2.68	2.55	2.44	2.36	2.07

Fraktiler i t-fordelingen med f frihedsgrader

	Sandsynlighed i procent						
f	90	95	97.5	99	99.5	99.9	f
1	3.078	6.314	12.706	31.821	63.657	318.309	1
2	1.886	2.920	4.303	6.965	9.925	22.327	2
3	1.638	2.353	3.182	4.541	5.841	10.215	3
4	1.533	2.132	2.776	3.747	4.604	7.173	4
5	1.476	2.015	2.571	3.365	4.032	5.893	5
6	1.440	1.943	2.447	3.143	3.707	5.208	6
7	1.415	1.895	2.365	2.998	3.499	4.785	7
8	1.397	1.860	2.306	2.896	3.355	4.501	8
9	1.383	1.833	2.262	2.821	3.250	4.297	9
10	1.372	1.812	2.228	2.764	3.169	4.144	10
11	1.363	1.796	2.201	2.718	3.106	4.025	11
12	1.356	1.782	2.179	2.681	3.055	3.930	12
13	1.350	1.771	2.160	2.650	3.012	3.852	13
14	1.345	1.761	2.145	2.624	2.977	3.787	14
15	1.341	1.753	2.131	2.602	2.947	3.733	15
16	1.337	1.746	2.120	2.583	2.921	3.686	16
17	1.333	1.740	2.110	2.567	2.898	3.646	17
18	1.330	1.734	2.101	2.552	2.878	3.610	18
19	1.328	1.729	2.093	2.539	2.861	3.579	19
20	1.325	1.725	2.086	2.528	2.845	3.552	20
21	1.323	1.721	2.080	2.518	2.831	3.527	21
22	1.321	1.717	2.074	2.508	2.819	3.505	22
23	1.319	1.714	2.069	2.500	2.807	3.485	23
24	1.318	1.711	2.064	2.492	2.797	3.467	24
25	1.316	1.708	2.060	2.485	2.787	3.450	25
30	1.310	1.697	2.042	2.457	2.750	3.385	30
50	1.299	1.676	2.009	2.403	2.678	3.261	50
75	1.293	1.665	1.992	2.377	2.643	3.202	75
100	1.290	1.660	1.984	2.364	2.626	3.174	100
150	1.287	1.655	1.976	2.351	2.609	3.145	150
200	1.286	1.653	1.972	2.345	2.601	3.131	200
400	1.284	1.649	1.966	2.336	2.588	3.111	400

D Ordlister

Dansk-Engelsk

afbildning map
afkomstfordeling offspring distribution

betinget sandsynlighed conditional probability
binomialfordeling binomial distribution

Cauchyfordeling Cauchy distribution
central estimator unbiased estimator
Central Grænseværdisætning Central Limit Theorem

delmængde subset
Den Centrale Grænseværdisætning The Central Limit Theorem
disjunkt disjoint
diskret discrete

eksponentialfordeling exponential distribution
endelig finite
ensidet variansanalyse one-way analysis of variance
enstikprøveproblem one-sample problem
estimat estimate
estimation estimation
estimator estimator
etpunktsfordeling one-point distribution
etpunktsmængde singleton

flerdimensional fordeling multivariate distribution
fordeling distribution
fordelingsfunktion distribution function
foreningsmængde union
forgreningsproces branching process
forklarende variabel explanatory variable
formparameter shape parameter
forventet værdi expected value
fraktil quantile
frembringende funktion generating function
frihedsgrader degrees of freedom
fællesmængde intersection

gammafordeling gamma distribution
gennemsnit average, mean
gentagelse repetition, replicate
geometrisk fordeling geometric distribution

hypergeometrisk fordeling hypergeometric distribution
hypotese hypothesis
hypoteseprøvning hypothesis testing
hyppighed frequency
hændelse event

indikatorfunktion indicator function
khi i anden chi squared
klassedeling partition
kontinuert continuous
korrelation correlation
kovarians covariance
kvadratsum sum of squares
kvotienttesteststørrelse likelihood ratio test statistic

ligefordeling uniform distribution
likelihood likelihood
likelihoodfunktion likelihood function

maksimaliseringsestimat maximum likelihood estimate

maksimaliseringsestimator maximum likelihood estimator
marginal fordeling marginal distribution
middelværdi expected value; mean
multinomialfordeling multinomial distribution
mængde set

niveau level
normalfordeling normal distribution
nævner (i brøk) nominator

odds odds

parameter parameter
plat eller krone heads or tails
poissonfordeling Poisson distribution
produktrum product space
punktsandsynlighed point probability

regression regression
relativ hyppighed relative frequency
residual residual
residualkvadratsum residual sum of squares
række row

sandsynlighed probability
sandsynlighedsfunktion probability function
sandsynlighedsmål probability measure
sandsynlighedsregning probability theory
sandsynlighedsrum probability space
sandsynlighedstæthedsfunktion probability density function
signifikans significance
signifikansniveau level of significance
signifikant significant
simultan fordeling joint distribution
simultan tæthed joint density
skalaparameter scale parameter
skøn estimate
statistik statistics
statistisk model statistical model
stikprøve sample, random sample
stikprøvefunktion statistic
stikprøveudtagning sampling
stokastisk uafhængighed stochastic independence
stokastisk variabel random variable
Store Tals Lov Law of Large Numbers
Store Tals Stærke Lov Strong Law of Large Numbers
Store Tals Svage Lov Weak Law of Large Numbers
støtte support
sufficiens sufficiency
sufficient sufficient
systematisk variation systematic variation
søjle column

test test
teste test
testsandsynlighed test probability
teststørrelse test statistic
tilbagelægning (med/uden) replacement (with/without)
tilfældig variation random variation
tosidet variansanalyse two-way analysis of variance
tostikprøveproblem two-sample problem
totalgennemsnit grand mean
tæller (i brøk) denominator
tæthed density
tæthedsfunktion density function

uafhængig independent
uafhængige identisk fordelte independent identically distributed; i.i.d.
uafhængighed independence
udfald outcome
udfaldsrum sample space
udtage en stikprøve sample
uendelig infinite

variansanalyse analysis of variance
variation variation
vekselvirkning interaction

Engelsk-Dansk

analysis of variance variansanalyse
average gennemsnit

binomial distribution binomialfordeling
branching process forgreningsproces

Cauchy distribution Cauchyfordeling
Central Limit Theorem Den Centrale Grænseværdisætning
chi squared khi i anden
column søjle
conditional probability betinget sandsynlighed
continuous kontinuert
correlation korrelation
covariance kovarians

degrees of freedom frihedsgrader
denominator tæller
density tæthed
density function tæthedsfunktion
discrete diskret
disjoint disjunkt
distribution fordeling
distribution function fordelingsfunktion
estimate estimat, skøn
estimation estimation
estimator estimator
event hændelse
expected value middelværdi; forventet værdi
explanatory variable forklarende variabel
exponential distribution eksponentialfordeling

finite endelig
frequency hyppighed

gamma distribution gammafordeling
generating function frembringende funktion
geometric distribution geometrisk fordeling
grand mean totalgennemsnit

heads or tails plat eller krone
hypergeometric distribution hypergeometrisk fordeling
hypothesis hypotese
hypothesis testing hypoteseprøvning

i.i.d. = independent identically distributed
independence uafhængighed
independent uafhængig
indicator function indikatorfunktion
infinite uendelig
interaction vekselvirkning
intersection fællesmængde

joint density simultan tæthed
joint distribution simultan fordeling

Law of Large Numbers Store Tals Lov
level niveau
level of significance signifikansniveau
likelihood likelihood
likelihood function likelihoodfunktion
likelihood ratio test statistic kvotienttest-størrelse

map afbildning
marginal distribution marginal fordeling
maximum likelihood estimate maksimaliseringsestimat
maximum likelihood estimator maksimaliseringsestimator
mean gennemsnit, middelværdi
multinomial distribution multinomialfordeling
multivariate distribution flerdimensional fordeling
nominator nævner
normal distribution normalfordeling

odds odds
offspring distribution afkomstfordeling
one-point distribution etpunktsfordeling
one-sample problem enstikprøveproblem
one-way analysis of variance ensidet variansanalyse
outcome udfald

parameter parameter
partition klassedeling
point probability punktsandsynlighed
Poisson distribution poissonfordeling
probability sandsynlighed
probability density function sandsynlighedstæthedsfunktion
probability function sandsynlighedsfunktion
probability measure sandsynlighedsmål
probability space sandsynlighedsrum
probability theory sandsynlighedsregning
product space produktrum

quantile fraktil

r.v. = random variable
random sample stikprøve
random variable stokastisk variabel
random variation tilfældig variation
regression regression
relative frequency relativ hyppighed
replacement (with/without) tilbagelægning (med/uden)
residual residual
residual sum of squares residualkvadratsum
row række

sample stikprøve; at udtage en stikprøve
sample space udfaldsrum
sampling stikprøveudtagning
scale parameter skalaparameter
set mængde
shape parameter formparameter
significance signifikans
significant signifikant
singleton etpunktsmængde
statistic stikprøvefunktion
statistical model statistisk model
statistics statistik
Strong Law of Large Numbers Store Tals Stærke Lov
subset delmængde
sufficiency sufficiens
sufficient sufficient
sum of squares kvadratsum
support støtte
systematic variation systematisk variation

test at teste; et test
test probability testsandsynlighed
test statistic teststørrelse
two-sample problem tostikprøveproblem
two-way analysis of variance tosidet variansanalyse

unbiased estimator central estimator
uniform distribution ligefordeling
union foreningsmængde
variation variation

Weak Law of Large Numbers Store Tals Svage Lov

Litteraturhenvisninger

Andersen, E. B. (1977). Multiplicative poisson models with unequal cell rates, *Scandinavian Journal of Statistics* **4**: 153–8.

Bachelier, L. (1900). Théorie de la spéculation, *Annales scientifiques de l'École Normale Supérieure*, 3^e *série* **17**: 21–86.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances, *Philosophical Transactions of the Royal Society of London* **53**: 370–418.

Bliss, C. I. and Fisher, R. A. (1953). Fitting the negative binomial distribution to biological data and Note on the efficient fitting of the negative binomial, *Biometrics* **9**: 176–200.

Bortkiewicz, L. (1898). *Das Gesetz der kleinen Zahlen*, Teubner, Leipzig.

Davin, E. P. (1975). *Blood pressure among residents of the Tambo Valley*, Master's thesis, The Pennsylvania State University.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society of London*, *Series A* **222**: 309–68.

Forbes, J. D. (1857). Further experiments and remarks on the measurement of heights by the boiling point of water, *Transactions of the Royal Society of Edinburgh* **21**: 135–43.

Gauß, K. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientum*, F. Perthes und I.H. Besser, Hamburg.

Greenwood, M. and Yule, G. U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents, *Journal of the Royal Statistical Society* **83**: 255–79.

Hald, A. (1948, 1968). *Statistiske Metoder*, Akademisk Forlag, København.

Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer, Berlin.

Kotz, S. and Johnson, N. L. (eds) (1992). *Breakthroughs in Statistics*, Vol. 1, Springer-Verlag, New York.

Larsen, J. (2006). *Basisstatistik, 2. udgave*, IMFUFA tekst nr 435, Roskilde Universitetscenter.
`http://akira.ruc.dk/~jl/tekster/nr435.pdf`

Lee, L. and Krutchkoff, R. G. (1980). Mean and variance of partially-truncated distributions, *Biometrics* **36**: 531–6.

Newcomb, S. (1891). Measures of the velocity of light made under the direction of the Secretary of the Navy during the years 1880-1882, *Astronomical Papers* **2**: 107–230.

Pack, S. E. and Morgan, B. J. T. (1990). A mixture model for interval-censored time-to-response quantal assay data, *Biometrics* **46**: 749–57.

Ryan, T. A., Joiner, B. L. and Ryan, B. F. (1976). *MINITAB Student Handbook*, Duxbury Press, North Scituate, Massachusetts.

Sick, K. (1965). Haemoglobin polymorphism of cod in the Baltic and the Danish Belt Sea, *Hereditas* **54**: 19–48.

Stigler, S. M. (1977). Do robust estimators work with *real* data?, *The Annals of Statistics* **5**: 1055–98.

Weisberg, S. (1980). *Applied Linear Regression*, Wiley series in Probability and Mathematical Statistics, John Wiley & Sons.

Wiener, N. (1976). *Collected Works*, Vol. 1, MIT Press, Cambridge, MA. Edited by P. Masani.

Alfabetisk register