STATISTIKNOTER

Simple Poissonfordelingsmodeller

Jørgen Larsen

IMFUFA
Roskilde Universitetscenter

Februar 1999

IMFUFA, Roskilde Universitetscenter, Postboks 260, DK-4000 Roskilde.
Jørgen Larsen: **STATISTIKNOTER: Simple Poissonfordelingsmodeller**
IMFUFA tekst nr. 304c/1995 51 sider ISSN 0106-6242

Dette hæfte er en del af undervisningsmaterialet til et kursus i statistik og statistiske modeller. Undervisningsmaterialet omfatter blandt andet følgende titler:

a. Simple binomialfordelingsmodeller
b. Simple normalfordelingsmodeller
c. Simple Poissonfordelingsmodeller
d. Simple multinomialfordelingsmodeller
e. Mindre matematisk-statistisk opslagsværk, indeholdende bl.a. ordforklaringer, resuméer og tabeller

•

Om kurset og kursusmaterialet kan blandt andet siges at

- når det er et gennemgående tema at påpege at likelihoodmetoden kan benyttes som et overordnet princip for valg af estimatorer og teststørrelser, er det blandt andet begrundet i *at* likelihoodmetoden har mange egenskaber der fra et matematisk-statistisk synspunkt anses for ønskelige, *at* likelihoodmetoden er meget udbredt og nyder stor anerkendelse (ikke mindst i Danmark), og *at* det i al almindelighed er værd at gøre opmærksom på at man også inden for faget statistik har overordnede og strukturerende begreber og metoder;
- når kursusmaterialet er skrevet på dansk (og ikke for eksempel på 'scientific English'), er det for at bidrage til at vedligeholde traditionerne for *hvordan* og *at* man kan tale om slige emner på dansk, og så sandelig også fordi dansk er det sprog som forfatteren – og vel også den forventede læser – er bedst til;
- når hæfterne foruden de sædvanlige simple modeller, metoder og eksempler også indeholder eksempler der er væsentligt sværere, er det for at antyde nogle af de retninger man kan arbejde videre i, og for at der kan være lidt udfordringer til den krævende læser.

Indhold

1 Poissonfordelingen

Dette kapitel introducerer Poissonfordelingen der er opkaldt efter den franske matematiker og fysiker S.-D. Poisson (1781-1840). Poissonfordelingsmodeller kan blandt andet komme på tale når man har at gøre med *antalsobservationer* der angiver hvor mange gange et bestemt fænomen optræder i et vist tidsrum og/eller et vist geografisk område eller lignende (det kunne for eksempel være trafikulykker på et år på en bestemt vejstrækning).

Det gennemgående eksempel i dette kapitel kan måske i første omgang forekomme lidt kuriøst, men det er meget berømt idet det optræder i næsten alle lærebøger i statistik.[1]

Eksempel 1.1 (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. 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 fire. Ved klassificeringen bliver der derfor fem klasser svarende til 0, 1, 2, 3 og 4 døde pr. år. Tabel 1.1 viser hvordan de faktiske tal blev.

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. Der kan således være fornuft i at beskæftige sig med denne modelbygningsopgave: *Find et forslag til en matematisk model der kan levere sandsynligheder for at have netop y døde i et bestemt regiment, $y = 0, 1, 2, \ldots$.*

1.1 Udledning

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

[1] Eksemplet stammer fra L. von Bortkiewicz (1898): *Das Gesetz der kleinen Zahlen*, Leipzig: Teubner.

Tabel 1.1 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

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

2. Der er et »grundeksperiment« der består i at man i et vist tidsinterval (af længde 1 år) holder øje med hvor mange gange en bestemt type begivenhed (dødsfald ved hestespark) indtræffer.

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

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

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

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

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

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

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

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

$j = 1, 2, \ldots, n$.

Hvis Δt er meget lille, så er det samlede antal Y af begivenheder i intervallet $]t_0, t_1]$ ifølge betragtningerne i punkt 5 cirka lig med $I_1 + I_2 + \ldots + I_n$.

7. Antag at der i alle $n = n(\Delta t)$ delintervaller er den *samme* sandsynlighed $p = p(\Delta t)$ for at der sker en begivenhed. (Der bliver altså ikke i løbet af perioden indført nye sikkerhedsforanstaltninger der nedsætter chancen for at blive sparket til døde af en hest. Og antallet af soldater og af heste i regimentet er stort set konstant året igennem.)

   Antag også at det der sker i ét interval er *stokastisk uafhængigt* af det der sker i andre intervaller. (Hvis der *tilfældigvis* var to soldater der i begyndelsen af året blev sparket til døde af heste, så tager de øvrige soldater i regimentet *ikke* i den anledning ekstra forholdsregler i resten af året.)

8. Da $I_1, I_2, \ldots, I_n$ således er uafhængige og identisk fordelte 01-variable, er $\sum_{j=1}^{n} I_j$ binomialfordelt med parametre $n = n(\Delta t)$ og $p = p(\Delta t)$, og da totalantallet Y af begivenheder i $]t_0, t_1]$ cirka er lig med $\sum_{j=1}^{n} I_j$, er Y således cirka binomialfordelt med parametre n og p.

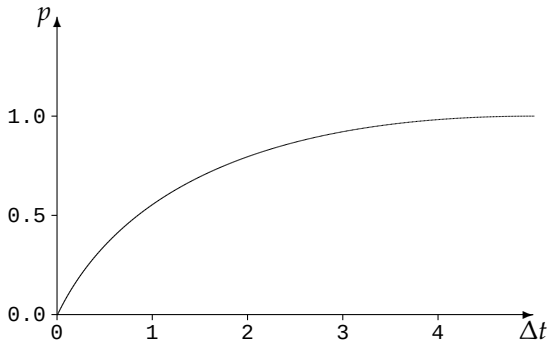
   Forbeholdet »cirka« bortfalder når Δt bliver tilstrækkelig lille, dvs. vi skal på et senere stadium lade Δt gå mod nul.

9. Den måde hvorpå n afhænger af Δt er simpel idet som tidligere anført

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

   Derimod mangler vi at overveje hvordan p afhænger af Δt.

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



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

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

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

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

$$\mathrm{P}(Y = y) \quad \approx \quad \binom{n}{y} p^y (1-p)^{n-y} \tag{1.1}$$

hvor »≈« bliver til »=« når $\Delta t \to 0$. Derfor må det næste skridt være at bestemme grænseværdien

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

under den grænseovergang hvor $\Delta t \to 0$ og dermed $n = \dfrac{t_1 - t_0}{\Delta t} \to \infty$. I punkt 10 vedtog vi at der under denne grænseovergang skal gælde at $\dfrac{p}{\Delta t} = \dfrac{p(\Delta t)}{\Delta t} \to \lambda$, og derfor vil

$$np = \frac{(t_1 - t_0) \cdot p}{\Delta t} \quad \to \quad \lambda \cdot (t_1 - t_0). \tag{1.2}$$

12. Vi omskriver binomialsandsynligheden på følgende måde:

$$\begin{aligned}
&\binom{n}{y} p^y (1-p)^{n-y} \\
&= \frac{n}{1} \frac{n-1}{2} \cdots \frac{n-y+1}{y} \cdot p^y \cdot (1-p)^{-y} \cdot (1-p)^n \\
&= \underbrace{1\,(1-\tfrac{1}{n}) \ldots (1-\tfrac{y-1}{n})}_{\text{(a)}} \cdot \underbrace{\frac{(np)^y}{y!}}_{\text{(b)}} \cdot \underbrace{(1-p)^{-y}}_{\text{(c)}} \cdot \underbrace{(1-p)^n}_{\text{(d)}}.
\end{aligned}$$

13. Under grænseovergangen vil de forskellige faktorer opføre sig på forskellige måder:

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

(b) $\dfrac{(np)^y}{y!} \to \dfrac{\big(\lambda \cdot (t_1 - t_0)\big)^y}{y!}.$

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

(d) $(1-p)^n \to \exp\big(-\lambda \cdot (t_1 - t_0)\big)$ hvilket indses således:

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

$$\frac{\ln(1+h)}{h} = \frac{\ln(1+h) - \ln 1}{h} \quad \to \quad 1\,.$$

ii. Ved at benytte dette samt formel (1.2) fås

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

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

$$(1-p)^n \quad \to \quad \exp\big(-\lambda\cdot(t_1-t_0)\big)$$

som ønsket.

Alt i alt vil binomialsandsynligheden i formel (1.1) konvergere mod

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

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

$$\mathrm{P}(Y=y) \quad = \quad \frac{\big(\lambda\cdot(t_1-t_0)\big)^y}{y!}\,\exp\big(-\lambda\cdot(t_1-t_0)\big)\,, \tag{1.3}$$

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

I formel (1.3) optræder den ukendte parameter λ der i punkt 10 blev indført som værende cirka »sandsynligheden for en begivenhed i et meget kort tidsinterval divideret med tidsintervallets længde«. Størrelsen λ har derfor dimensionen tid^{-1}, dvs. λ angives i f.eks. dag^{-1} eller år^{-1}.

Jo større λ er, jo tilbøjeligere er begivenhederne til at indtræffe; λ er en såkaldt *intensitet* (der i hestesparkeksemplet specielt kunne kaldes for en *ulykkesintensitet* eller en *dødsintensitet*).[2]

1.2 Definition og egenskaber

Man definerer Poissonfordelingen således:

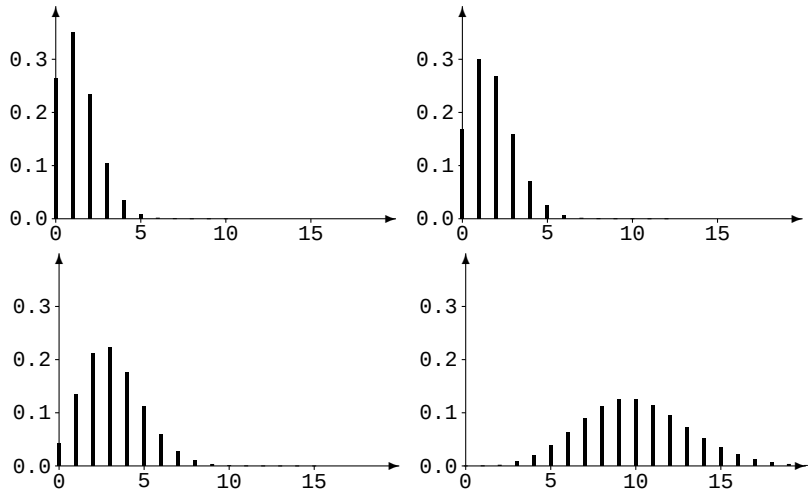
Definition 1.1 (Poissonfordeling)
Poissonfordelingen med parameter $\mu \geq 0$ er den sandsynlighedsfordeling på udfaldsrummet $\mathcal{X}=\{0,1,2,\ldots\}$ som har sandsynlighedsfunktion

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

Figur 1.1 viser nogle Poissonfordelinger.

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

[2] Antagelsen i punkt 7 går ud på at begivenhederne indtræffer med *samme* tilbøjelighed, med samme intensitet, overalt på (den betragtede del af) tidsaksen. Der er imidlertid ikke noget i vejen for at konstruere mere indviklede modeller hvor intensiteten er tidsafhængig, dvs. $\lambda=\lambda(t)$.



Figur 1.1 Poissonfordelinger med middelværdier 1.33, 1.78, 3.16 og 10.

Bemærkning
Strengt taget bør den givne definition af Poissonfordelingen følges op af en redegørelse for at $f(y;\mu)$ faktisk *er* en sandsynlighedsfunktion, dvs. at der er tale om ikke-negative tal der summerer til 1. Det er klart at f-erne er ikke-negative; at de summerer til 1 følger af eksponentialfunktionens rækkeudvikling $\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ som ikke vil blive bevist her. □

En egenskab ved Poissonfordelingen er at dens middelværdi er lig dens varians: *Hvis den stokastiske variabel Y er Poissonfordelt med parameter μ, så er*

$$\begin{aligned} \mathrm{E}\, Y &= \mu \\ \mathrm{Var}\, Y &= \mu \end{aligned}$$

dvs. både middelværdien og variansen af Y er lig μ.

Bevis
Det vises på følgende måde der kræver lidt kendskab til uendelige rækker:

Middelværdien af Y er pr. definition $\mathrm{E}Y = \sum_{y=0}^{\infty} y \cdot f(y)$, og der gælder:

$$\begin{aligned} \mathrm{E}\, Y &= \sum_{y=0}^{\infty} y \frac{\mu^y}{y!} \exp(-\mu) \\ &= \sum_{y=1}^{\infty} y \frac{\mu^y}{y!} \exp(-\mu) \\ &= \mu \left(\sum_{y=1}^{\infty} \frac{\mu^{y-1}}{(y-1)!} \right) \exp(-\mu) \\ &= \mu \left(\sum_{n=0}^{\infty} \frac{\mu^n}{n!} \right) \exp(-\mu) \\ &= \mu. \end{aligned}$$

Variansen af Y er $\mathrm{Var}(Y) = \mathrm{E}\left((Y-\mathrm{E}Y)^2\right) = \mathrm{E}(Y^2) - (\mathrm{E}Y)^2$. Vi kender $\mathrm{E}Y$, men $\mathrm{E}(Y^2)$ er besværlig at regne ud; det er smart at benytte omskrivningen $\mathrm{E}(Y^2) = \mathrm{E}(Y(Y-1)+Y) = \mathrm{E}(Y(Y-1)) + \mathrm{E}Y$. Nu er

$$\begin{aligned}
\mathrm{E}(Y(Y-1)) &= \sum_{y=0}^{\infty} y(y-1)\,\frac{\mu^y}{y!}\exp(-\mu) \\
&= \sum_{y=2}^{\infty} y(y-1)\,\frac{\mu^y}{y!}\exp(-\mu) \\
&= \mu^2\left(\sum_{y=2}^{\infty}\frac{\mu^{y-2}}{(y-2)!}\right)\exp(-\mu) \\
&= \mu^2\left(\sum_{n=0}^{\infty}\frac{\mu^n}{n!}\right)\exp(-\mu) \\
&= \mu^2,
\end{aligned}$$

så $\mathrm{Var}(Y) = \mathrm{E}(Y(Y-1)) + \mathrm{E}Y - (\mathrm{E}Y)^2 = \mu^2 + \mu - \mu^2 = \mu$. □

1.3 Afrunding

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

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

1.4 Opgaver

Opgave 1.1
I næste kapitel vil det vise sig, at det i hestesparkeksemplet er fornuftigt at estimere λ ved $\hat{\lambda} = 0.61$ dødsfald pr. år.

Lav et pindediagram[3] der viser hvordan Poissonfordelingen med parameter 0.61 ser ud.

TIP: Når man skal udregne $f(y;\mu) = \frac{\mu^y}{y!}\exp(-\mu)$ for en hel masse y-værdier, kan det være smart at gøre det rekursivt:

$$\begin{aligned} f(0;\mu) &= \exp(-\mu) \\ f(y;\mu) &= \frac{\mu}{y} \cdot f(y-1;\mu), \quad y = 1,2,3,\ldots \end{aligned}$$

Opgave 1.2 (Raunkiær-cirklinger)
Inden for planteøkologi bestemmer man (i Danmark) ofte planters skudtæthed ved hjælp af en metode der kaldes Raunkiær-cirklinger. I sin simpleste form er metoden som følger (tænk på at det handler om at undersøge planter på en mark): Man anbringer et tilfældigt sted på prøvearealet en cirkel med areal a og ser efter om den planteart man undersøger, findes inden for cirklen eller ej; dette gentages n gange (idet man sørger for at de n cirkler ikke overlapper). Typisk er $a = 0.1\text{m}^2$ og $n = 10$.

Antag for eksempel at man i 10 cirklinger med en 0.1m^2 cirkel fik netop 7 tilfælde hvor planten blev fundet inden for cirklen.

Man ønsker som nævnt at bestemme skudtætheden λ (der måles i antal/m^2). Man må derfor gøre en antagelse om at en bestemt slags sandsynlighedsmodel har placeret skuddene ud over marken. Den simpleste antagelse er at skuddene er placeret efter en Poisson-proces, hvilket betyder at antal skud i et delområde med areal a er Poissonfordelt med parameter $\lambda \cdot a$, og at antal skud i disjunkte delområder er stokastisk uafhængige.

1. Hvad er sandsynligheden for at man ved én cirkling oplever at der er netop k skud inde i cirklen?

2. Hvad er sandsynligheden for at man ved én cirkling oplever at der er mindst et skud inde i cirklen?

   TIP: »mindst et« er det modsatte af »ingen«.

3. Hvis man udfører $n = 10$ cirklinger, hvad er da sandsynligheden for at der i netop $y = 7$ tilfælde findes mindst et skud inde i cirklen?

(Raunkiær-cirklinger genoptages i Opgave 2.3.)

[3] Et pindediagram er sådan noget som Figur 1.1 indeholder fire eksempler på.

2 En- og flerstikprøveproblemer i Poissonfordelingen

I Kapitel 1 nåede vi ved teoretiske overvejelser frem til at antallet af dødsfald pr. regiment pr. år måtte være Poissonfordelt med parameter $\mu = \lambda \cdot 1$ år, men stemmer det overhovedet med virkeligheden, og hvordan estimerer man intensiteten λ?

Vi skal i dette kapitel beskæftige os med estimation af parametre og test af hypoteser om parametre i Poissonfordelinger, og med spørgsmålet om kontrol af modellen.

2.1 Enstikprøveproblemet

I hestespark-eksemplet fra Kapitel 1 er situationen den at der er $n = 200$ uafhængige observationer $y_1, y_2, \ldots, y_n$ fra Poissonfordelingen med parameter $\mu = \lambda \cdot 1$ år. Det er et eksempel på et »enstikprøveproblem« fordi der er tale om et antal observationer, en *stikprøve*, fra en og samme fordeling.

Generelt har man at gøre med uafhængige observationer $y_1, y_2, \ldots, y_n$ fra en Poissonfordeling med parameter μ, svarende til at *modelfunktionen* er

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

hvor $y_\bullet$ er statistikerens sædvanlige korte skrivemåde for $y_1 + y_2 + \ldots + y_n$.

Estimation af parameteren

Poissonparameteren μ estimeres ved likelihoodmetoden. *Likelihoodfunktionen* svarende til observationerne $y_1, y_2, \ldots, y_n$ er

$$L(\mu) = \frac{\mu^{y_\bullet}}{\text{konstant}} \exp(-n\mu),$$

så at

$$\ln L(\mu) = \text{konstant} + y_\bullet \ln \mu - n\mu.$$

Ifølge de sædvanlige principper er det bedste estimat over μ den værdi $\hat{\mu}$ der maksimaliserer L eller $\ln L$. For at bestemme denne værdi løser vi ligningen $\frac{d}{d\mu} \ln L = 0$. Man finder at

$$\frac{d}{d\mu} \ln L(\mu) \quad = \quad \frac{y_\bullet}{\mu} - n$$

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

$$\frac{d^2}{d\mu^2} \ln L(\mu) \quad = \quad -\frac{y_\bullet}{\mu^2}$$

altid er negativ, er $\overline{y}$ et maksimumspunkt. Dermed er vist at maksimaliseringsestimatet for μ er gennemsnittet af observationerne:[1]

$$\hat{\mu} \quad = \quad \overline{y} \; = \; \frac{1}{n} \sum_{j=1}^{n} y_j \, .$$

I taleksemplet får man

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

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

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

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

Modelkontrol

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

For at få et fingerpeg om hvor god denne »bedste beskrivelse« er, udregner vi nogle »forventede« antal under forudsætning af at modellen er rigtig: Ifølge modellen er sandsynligheden for at der i et bestemt regiment-år er netop y dødsfald,

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

Ud af de 200 regiment-år skulle man derfor forvente ca. $200 \cdot f(0; \hat{\lambda})$ tilfælde med 0 dødsfald, ca. $200 \cdot f(1; \hat{\lambda})$ tilfælde med 1 dødsfald, ca. $200 \cdot f(2; \hat{\lambda})$ tilfælde med 2 dødsfald, osv. Disse forventede tal udregnes, og man får Tabel 2.1. Det ses at de »forventede« antal stemmer fint overens med de observerede, og det må vi tage som tegn på at Poissonmodellen ikke er helt hen i vejret.

[1] Det er i udledningen forudsat at $y_\bullet > 0$. Hvis $y_\bullet = 0$, så er log-likelihoodfunktionen lig $-n\mu +$ konstant, og den antager sit maksimum når $\mu = 0$.

Tabel 2.1 Hestespark-eksemplet: De observerede antal år med y dødsfald sammenlignet med de »forventede« antal år med y dødsfald beregnet ud fra Poissonmodellen.

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

Dispersionstestet

Undertiden vil man gerne udføre et numerisk test for rimeligheden af at antage at et sæt observationer $y_1, y_2, \ldots, y_n$ er en stikprøve fra en Poissonfordeling. Vi vil omtale en nem metode hertil.

Som nævnt side 10 har Poissonfordelingen den egenskab at middelværdi og varians er ens. Man kunne derfor udregne den empiriske middelværdi

$$\overline{y} \quad = \quad \frac{1}{n}\sum_{i=1}^{n} y_i$$

og den empiriske varians

$$s^2 \quad = \quad \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2,$$

og så se efter om de to er nogenlunde ens. Det viser sig hensigtsmæssigt at gøre dette på den måde at man udregner størrelsen

$$d \quad = \quad \frac{s^2}{\overline{y}}.$$

Hvis modelantagelsen er rigtig, skal d være tæt på 1, så man vil forkaste modellen hvis *enten* d_{obs} er så meget større end 1 at der kun er lille sandsynlighed (for eksempel 0.025) for at få en større værdi, *eller* d_{obs} er så meget mindre end 1 at der kun er lille sandsynlighed (for eksempel 0.025) for at få en mindre værdi. – Man kan bevise at når modellen er rigtig (og Poissonparameteren ikke er alt for lille), så vil d med god tilnærmelse følge en såkaldt χ^2/f-fordeling med $f = n - 1$ frihedsgrader.

I hestespark-eksemplet fandt vi tidligere at $\overline{y} = 0.61$. Videre er

$$\begin{aligned}
&\sum_{i=1}^{n}(y_i - \overline{y})^2 \\
&= 109 \cdot (0 - 0.61)^2 + 65 \cdot (1 - 0.61)^2 + \\
&\quad 22 \cdot (2 - 0.61)^2 + 3 \cdot (3 - 0.61)^2 + 1 \cdot (4 - 0.61)^2 \\
&= 121.58,
\end{aligned}$$

så $s^2 = 121.58/199 = 0.611$ og dermed $d_{\text{obs}} = 0.611/0.61 = 1.002$. Den fundne d_{obs}-værdi ligger meget tæt på 1, også målt i forhold til χ^2/f-fordelingen med 199 frihedsgrader, og det numeriske test bekræfter dermed indtrykket af at Poissonfordelingen giver en god beskrivelse af tallene.

2.2 Sammenligning af to Poissonfordelinger

Vi vil diskutere spørgsmålet om sammenligning af to Poissonfordelinger ud fra følgende eksempel.

Eksempel 2.1 (Ultralydsscanning)
Det er meget udbredt at foretage ultralydsscanning af gravide kvinder. Det menes/frygtes imidlertid at fostrene kan lide skade derved idet der måske sker kromosomforandringer. For at undersøge dette nærmere har man udført en række laboratorieforsøg med mus.[2]

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

I dette eksempel (der kun behandler en del af forsøgets talmateriale) optræder to grupper à tre mus: en behandlingsgruppe og en kontrolgruppe. Behandlingsgruppen har fået ultralyd, hvorefter man har ladet gå 18 timer inden musen blev dræbt og prøverne udtaget. Kontrolgruppen er behandlet på samme måde, på nær at der denne gang ikke blev tændt for ultralydapparatet. Fra hver mus udtog man otte prøver; i alt undersøgte man for hver mus ca. 2000 celler og afgjorde om det var en mikrokærnecelle eller ej. Derved fremkom resultaterne i Tabel 2.2.

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

Modelopstilling

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

I den foreliggende problemstilling er det der er genstand for den grundlæggende interesse formentlig chancen for at en celle omdannes til en mikrokærnecelle. I den forbindelse er det uinteressant at søge at opstille en model der kan påtage sig at beskrive variationen i antal optalte celler pr. mus. De indgående tider er de samme for alle forsøgsdyr; derfor behøver vi ikke indbygge tidsafhængigheder i modellen. Derimod skal vi søge at formulere en model der kan beskrive variationen i antallet af mikrokærneceller i en prøve af en given størrelse, udtaget fra en mus der har fået en given behandling. I modellen skal r-erne derfor indgå som givne konstanter og y-erne som udfald af stokastiske variable.

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

Tabel 2.2 Resultater af mikrokærnetællinger.

1. Behandlingsgruppen

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

2. Kontrolgruppen

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

Da der for en enkelt mus optælles et meget stort antal celler der hver især har en meget lille chance for at være blevet omdannet til en mikrokærnecelle, kan vi antage (jf. trykfejlseksemplet side 11) at antal mikrokærneceller i en prøve med r celler er Poissonfordelt med parameter $\mu = \lambda r$, hvor λ er en »omdannelsesintensitet«, nemlig sandsynligheden for at en optalt celle er en mikrokærnecelle. Den systematiske forskel mellem behandlingsgrupperne skal beskrives ved hjælp af modellens parametre, så derfor skal mus med samme behandling have samme intensitet λ, hvorimod behandlingsgruppen og kontrolgruppen skal have hver sit λ.

Vi indfører lidt notation for at kunne formulere modellen præcist:

$$\begin{aligned} r_{ij} &= \text{antal optalte celler fra mus nr. } j \text{ i gruppe } i, \\ y_{ij} &= \text{antal mikrokærneceller fra mus nr. } j \text{ i gruppe } i, \end{aligned}$$

hvor $i = 1$ svarer til behandlingsgruppen og $i = 2$ til kontrolgruppen. Det vil sige at Tabel 2.2 skematisk ser således ud:

	$i = 1$	
1	r_{11}	y_{11}
2	r_{12}	y_{12}
3	r_{13}	y_{13}
	$r_{1\cdot}$	$y_{1\cdot}$

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

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

$$\prod_{i=1}^{2}\prod_{j=1}^{3}\frac{(\lambda_i r_{ij})^{y_{ij}}}{y_{ij}!}\exp(-\lambda_i r_{ij})\,. \tag{2.1}$$

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

Estimation af parametre

Maksimaliseringsestimaterne over λ_1 og λ_2 skal bestemmes på grundlag af *likelihoodfunktionen*. Ud fra modelfunktionen (2.1) får vi

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

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

$$\hat{\lambda}_i = \frac{y_{i\bullet}}{r_{i\bullet}},$$

nemlig det totale antal observerede mikrokærneceller i gruppe i divideret med det totale antal optalte celler i gruppe i, hvilket også er et estimat der virker umiddelbart rimeligt.

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

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

som har maksimum i

$$\hat{\lambda} = \frac{y_{\bullet\bullet}}{r_{\bullet\bullet}}.$$

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

I eksemplet bliver estimaterne

$$\begin{aligned} \hat{\lambda}_{\text{behandling}} &= \frac{18}{6320} &&= 2.8 \cdot 10^{-3} \\ &&&\sim \text{knap 3 mikrokærneceller pr. 1000 celler} \\ \hat{\lambda}_{\text{kontrol}} &= \frac{10}{6288} &&= 1.6 \cdot 10^{-3} \\ &&&\sim \text{godt 1.5 mikrokærneceller pr. 1000 celler} \\ \hat{\lambda}_{\text{fælles}} &= \frac{28}{12608} &&= 2.2 \cdot 10^{-3} \\ &&&\sim \text{godt 2 mikrokærneceller pr. 1000 celler.} \end{aligned}$$

Man kan spørge hvor stor tiltro man nu kan have til disse tal. Det er ikke i statistikerens magt at udtale noget fornuftigt om diverse eksterne fejlkilder der eventuelt måtte have været i spil (det véd eksperimentator bedre). Statistikeren kan udtale sig om dén tilfældige variation der beskrives af den statistiske model, for eksempel konkretiseret til *middelfejlene på estimatorerne*. Lad os derfor bestemme middelfejlen (standardafvigelsen) på $\hat{\lambda}_i$ i det foreliggende eksempel: Da $\hat{\lambda}_i = \frac{Y_{i\bullet}}{r_{i\bullet}}$, er den søgte størrelse $\sqrt{\text{Var}\hat{\lambda}_i} = \sqrt{\text{Var}\left(\frac{Y_{i\bullet}}{r_{i\bullet}}\right)}$. Ifølge regnereglerne for varianser og kvadratrødder er

$$\sqrt{\text{Var}\left(\frac{Y_{i\bullet}}{r_{i\bullet}}\right)} = \sqrt{\frac{\text{Var}(Y_{i\bullet})}{r_{i\bullet}^2}} = \frac{\sqrt{\text{Var}(Y_{i\bullet})}}{r_{i\bullet}}.$$

Da $Y_{i\bullet}$ er Poissonfordelt med parameter $\lambda_i r_{i\bullet}$, er $\text{Var}(Y_{i\bullet}) = \lambda_i r_{i\bullet}$ (se side 10), så middelfejlen på $\hat{\lambda}_i$ er

$$\frac{\sqrt{\text{Var}(Y_{i\bullet})}}{r_{i\bullet}} = \sqrt{\frac{\lambda_i}{r_{i\bullet}}}.$$

Da vi ikke kender λ_i, men kun et estimat $\hat{\lambda}_i = y_{i\bullet}/r_{i\bullet}$, kan vi kun udregne en talværdi for *den estimerede middelfejl* på λ_i, og den bliver

$$\sqrt{\frac{\hat{\lambda}_i}{r_{i\bullet}}} = \sqrt{\frac{y_{i\bullet}/r_{i\bullet}}{r_{i\bullet}}} = \frac{\sqrt{y_{i\bullet}}}{r_{i\bullet}}.$$

Man finder de estimerede middelfejl på $\hat{\lambda}_{\text{behandling}}$, $\hat{\lambda}_{\text{kontrol}}$ og $\hat{\lambda}_{\text{fælles}}$ til $0.67 \cdot 10^{-3}$, $0.50 \cdot 10^{-3}$ og $0.42 \cdot 10^{-3}$.

Som læseren vil have bemærket, benytter vi ved beregningen af de forskellige estimater slet ikke de individuelle værdier af r og y for de enkelte mus, vi benytter kun totalerne for hver gruppe. Er det da lige meget hvad værdierne for de enkelte mus er? Ja, det er det faktisk, *sålænge der ikke er tvivl om Poissonmodellens brugbarhed*. Men hvis vi er på udkig efter indicier for (eller imod) anvendeligheden af Poissonmodellen, så er det i høj grad påkrævet at kende de enkelte værdier. For den statistiske model skal jo beskrive enkeltobservationernes tilfældige variation omkring et bestemt niveau, og hvis man vil vurdere antagelsen om at den tilfældige variation kan beskrives ved netop en Poissonfordeling, så skal man se på enkeltobservationernes faktiske variation og vurdere om den ligner den fittede Poissonfordeling.

Hypoteseprøvning

Som nævnt skal vi teste den statistiske hypotese $H_0 : \lambda_1 = \lambda_2$. Det gøres på traditionel vis med et kvotienttest. Vi udregner kvotienttesstørrelsen

$$\begin{aligned} Q &= \frac{L(\hat{\lambda}, \hat{\lambda})}{L(\hat{\lambda}_1, \hat{\lambda}_2)} \\ &= \frac{\hat{\lambda}^{y_{1\cdot}} \hat{\lambda}^{y_{2\cdot}} \exp(-\hat{\lambda} r_{1\cdot} - \hat{\lambda} r_{2\cdot})}{\hat{\lambda}_1^{y_{1\cdot}} \hat{\lambda}_2^{y_{2\cdot}} \exp(-\hat{\lambda}_1 r_{1\cdot} - \hat{\lambda}_2 r_{2\cdot})} \\ &= \left(\frac{\hat{\lambda}}{\hat{\lambda}_1}\right)^{y_{1\cdot}} \left(\frac{\hat{\lambda}}{\hat{\lambda}_2}\right)^{y_{2\cdot}} \frac{\exp(-y_{1\cdot} - y_{2\cdot})}{\exp(-y_{1\cdot} - y_{2\cdot})} \\ &= \left(\frac{\hat{\lambda} r_{1\cdot}}{y_{1\cdot}}\right)^{y_{1\cdot}} \left(\frac{\hat{\lambda} r_{2\cdot}}{y_{2\cdot}}\right)^{y_{2\cdot}} \\ &= \left(\frac{\hat{y}_{1\cdot}}{y_{1\cdot}}\right)^{y_{1\cdot}} \left(\frac{\hat{y}_{2\cdot}}{y_{2\cdot}}\right)^{y_{2\cdot}} , \end{aligned}$$

hvor $\hat{y}_{i\cdot} = \hat{\lambda} r_{i\cdot}$ er det »forventede« antal mikrokærneceller i gruppe i, forudsat at H_0 er rigtig.

Derfor er

$$-2 \ln Q = 2 \left(y_{1\cdot} \ln \frac{y_{1\cdot}}{\hat{y}_{1\cdot}} + y_{2\cdot} \ln \frac{y_{2\cdot}}{\hat{y}_{2\cdot}} \right) .$$

Små værdier af Q, dvs. store værdier af $-2 \ln Q$, er *signifikante*, dvs. de er tegn på at hypotesen H_0 *ikke* er forenelig med de foreliggende data.

For at vurdere om $-2 \ln Q_{\text{obs}}$ er signifikant stor, skal man bestemme testsandsynligheden

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

altså sandsynligheden under H_0 for at få et mindst lige så afvigende observationssæt som det foreliggende. Ved beregningen af ε kan man udnytte at når

H_0 er rigtig, så er $-2\ln Q$ med god tilnærmelse[3] χ^2-fordelt med $f = 2-1$ frihedsgrader[4], således at ε med god tilnærmelse kan udregnes som sandsynligheden for at få en værdi større end eller lig med $-2\ln Q_{\text{obs}}$ i χ^2-fordelingen med 1 frihedsgrad:

$$\varepsilon \;=\; \mathrm{P}\left(\chi^2_1 \geq -2\ln Q_{\text{obs}}\right).$$

I taleksemplet er $\hat{y}_{1\bullet} = 14.0$ og $\hat{y}_{2\bullet} = 14.0$, så

$$\begin{aligned} -2\ln Q &= 2\left(18\,\ln\frac{18}{14.0} + 10\,\ln\frac{10}{14.0}\right) \\ &= 2.32\,. \end{aligned}$$

I χ^2-fordelingen med 1 frihedsgrad er 80%-fraktilen 1.64 og 90%-fraktilen 2.71, så den fundne $-2\ln Q$-værdi svarer til et ε på mellem 10% og 20%. Man vil almindeligvis sige at en sådan ε-værdi ikke er lille nok til at man vil forkaste H_0. Vi kan dermed konkludere at de foreliggende tal *ikke* giver statistisk belæg for at mene at ultralyd er skadeligt. (På den anden side giver de næppe heller belæg for at mene at ultralyd *ikke* er skadeligt.)

2.3 Et sværere eksempel

I dette afsnit gennemgås et eksempel hvor Poissonfordelingen søges anvendt; det viser sig imidlertid at den model der foreslås i første omgang, ikke passer særlig godt; derfor må man finde på en anden model.

Præsentation af eksemplet

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). Tabel 2.3 viser 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.[5]

Lad y_i betegne antal ulykker som kvinde nr. i kommer ud for; y_i tænkes at være en observation af en stokastisk variabel Y_i, $i = 1, 2, \ldots, n$. Vi går ud fra at de stokastiske variable $Y_1, Y_2, \ldots, Y_n$ er indbyrdes uafhængige (men det er måske en lidt diskutabel antagelse).

[3] χ^2-approksimationen kan anvendes når de forventede antal $\hat{y}_{i\bullet}$ er mindst fem.

[4] antal parametre i grundmodellen er 2; antal parametre under H_0 er 1; antal frihedsgrader er derfor $f = 2-1$.

[5] Eksemplet stammer fra M. Greenwood & G.U. Yule (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**, 255-279. – Her i landet er eksemplet især kendt via sin forekomst i hvad der i mere end en menneskealder har været en toneangivende lærebog i statistik, nemlig A. Hald (1948, 1968): *Statistiske Metoder*, Akademisk Forlag.

Tabel 2.3 Fordelingen af $n = 647$ kvinder efter antallet y af ulykkestilfælde i en fem ugers periode.

y	f_y = antal kvinder med y ulykker
0	447
1	132
2	42
3	21
4	3
5	2
6+	0
	647

Tabel 2.4 Model 1: Observerede antal f_y og forventede antal $\hat{f}_y$.

y	f_y	$\hat{f}_y$
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

Vi indfører 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 da lig $0f_0 + 1f + 2f_2 + \ldots = \sum_{y=0}^{\infty} y f_y = 301$.

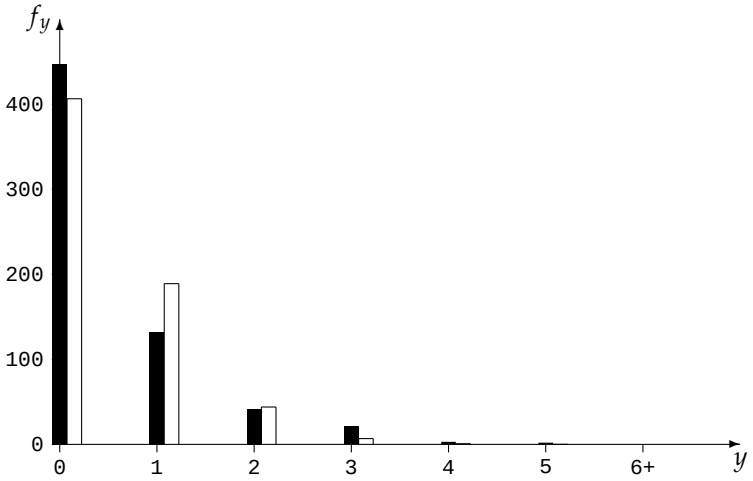
Model 1

I første omgang kan man forsøge sig med en model hvor $Y_1, Y_2, \ldots, Y_n$ er uafhængige og identisk Poissonfordelte med parameter μ, dvs.

$$\mathrm{P}(Y_i = y) \quad = \quad \frac{\mu^y}{y!} \exp(-\mu).$$

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«.

I denne model estimeres μ ved $\hat{\mu} = \overline{y} = 301/647 = 0.465$ (der sker 0.465 ulykker pr. kvinde pr. fem uger). Det forventede antal kvinder med y ulykker er $\hat{f}_y = n \dfrac{\hat{\mu}^y}{y!} \exp(-\hat{\mu})$; værdierne heraf vises i Tabel 2.4 og i Figur 2.1. Det ses at der ikke er nogen særlig god overensstemmelse mellem de observerede og de



Figur 2.1 Model 1: Observerede antal (sorte søjler) og forventede antal (hvide søjler) fra Tabel 2.4.

forventede antal. Man kan udregne variansen til $s^2 = 0.692$, og det er næsten halvanden gange middelværdien, hvilket er endnu et tegn på at Poissonmodellen er dårlig. Man kan derfor give sig til at overveje en anden model.

Model 2

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, dvs. Y_i er 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 få et perfekt fit (med $\hat{\mu}_i = y_i$, $i = 1, 2, \ldots, n$). 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 fordeling med en tæthedsfunktion af formen

$$g(\mu) \quad = \quad \frac{1}{\Gamma(\kappa)\beta^\kappa}\,\mu^{\kappa-1}\exp(-\mu/\beta), \quad \mu > 0.$$

(Symbolet $\Gamma(\kappa)$ betegner den såkaldte Gammafunktion, udregnet i κ. Pr. definition er $\Gamma(\kappa) \;=\; \int_0^{+\infty} t^{\kappa-1}\exp(-t)\,dt$. Hvis m er et naturligt tal, så er $\Gamma(m+1) = m!$.[6])

Fordelingen med denne tæthedsfunktion g er en *gammafordeling* med formparameter $\kappa > 0$ og skalaparameter $\beta > 0$.

[6]Gammafunktionen kommer ind i billedet fordi tæthedsfunktionen g skal integrere til 1, og det gør den da også, hvilket ses ved at foretage substitutionen $t = \mu/\beta$.

- Sandsynligheden for at en kvinde kommer ud for y ulykker ville nu være lig med $\dfrac{\mu^y}{y!}\exp(-\mu)$, hvis vi altså kendte værdien af μ for den pågældende kvinde. Men da vi kun véd at μ følger fordelingen med tæthedsfunktion g, bliver den faktiske sandsynlighed for y ulykker et vægtet middeltal af værdierne $\dfrac{\mu^y}{y!}\exp(-\mu)$ med $g(\mu)$ som vægte, og det betyder at sandsynligheden for at en kvinde kommer ud for y ulykker alt i alt bliver

$$\begin{aligned}
\mathrm{P}(Y=y) &= \int_0^{+\infty} \frac{\mu^y}{y!}\exp(-\mu)\cdot g(\mu)\,d\mu \\
&= \int_0^{+\infty} \frac{\mu^y}{y!}\exp(-\mu)\,\frac{1}{\Gamma(\kappa)\beta^\kappa}\,\mu^{\kappa-1}\exp(-\mu/\beta)\,d\mu \\
&= \frac{\Gamma(y+\kappa)}{y!\,\Gamma(\kappa)}\left(\frac{1}{\beta+1}\right)^\kappa\left(\frac{\beta}{\beta+1}\right)^y \\
&= \frac{\Gamma(y+\kappa)}{y!\,\Gamma(\kappa)}p^\kappa(1-p)^y,
\end{aligned}$$

  hvor $p = 1/(\beta+1)$. Med betegnelsen $\binom{y+\kappa-1}{y} = \dfrac{\Gamma(y+\kappa)}{y!\,\Gamma(\kappa)}$ (som hvis m er et naturligt tal, blot er den sædvanlige definition af binomialkoefficient) er sandsynligheden for y ulykker

$$\mathrm{P}(Y=y) = \binom{y+\kappa-1}{y}p^\kappa(1-p)^y,\ \ y\in\{0,1,2,\dots\}.$$

  Denne fordeling af Y er den såkaldte *negative binomialfordeling* med formparameter κ og sandsynlighedsparameter $p = 1/(\beta+1)$.

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 middelværdi og varians af Y givet ved

$$\begin{aligned}
\mathrm{E}(Y) &= \kappa(1-p)/p \\
&= \kappa\beta
\end{aligned}$$

og

$$\begin{aligned}
\mathrm{Var}(Y) &= \kappa(1-p)/p^2 \\
&= \mathrm{E}(Y)/p \\
&= \kappa\beta(\beta+1).
\end{aligned}$$

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 med likelihoodmetoden til estimation af de ukendte parametre. Likelihoodfunktionen er

$$\begin{aligned} L(\kappa, p) &= \prod_{i=1}^{n} \binom{y_i + \kappa - 1}{y_i} p^{\kappa} (1-p)^{y_i} \\ &= p^{n\kappa} (1-p)^{y_1 + y_2 + \ldots + y_n} \prod_{i=1}^{n} \binom{y_i + \kappa - 1}{y_i} \\ &= \text{konstant} \cdot p^{n\kappa} (1-p)^{y_\cdot} \prod_{k=1}^{\infty} (\kappa + k - 1)^{\sum_{j=k}^{\infty} f_j}, \end{aligned}$$

hvor f_k stadig betegner antal observationer som har værdien k. Logaritmen til likelihoodfunktionen bliver derfor (pånær en konstant)

$$\ln L(\kappa, p) = n\kappa \ln p + y_\cdot \ln(1-p) + \sum_{k=1}^{\infty} \left(\sum_{j=k}^{\infty} f_j \right) \ln(\kappa + k - 1)$$

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

$$\begin{aligned} \kappa\beta &= 0.464 \\ \kappa\beta(\beta + 1) &= 0.692. \end{aligned}$$

Man finder (idet $p = 1/(\beta + 1)$)

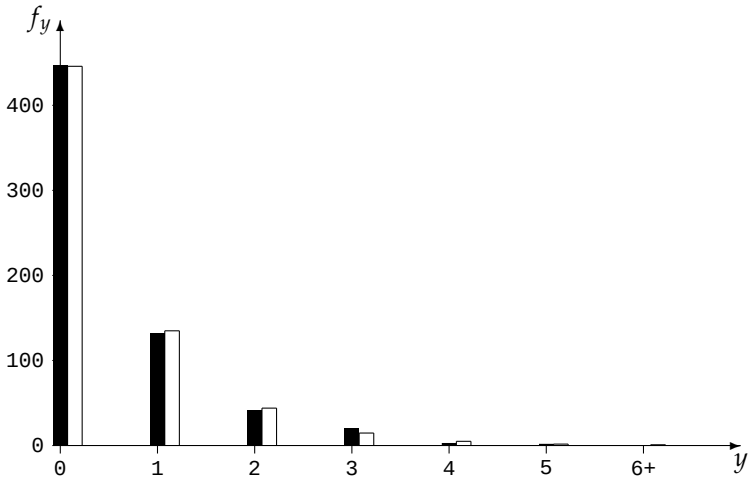
$$\begin{aligned} \tilde{\beta} &= 0.4882 \\ \tilde{\kappa} &= 0.9525 \\ \tilde{p} &= 0.6720. \end{aligned}$$

Disse værdier benyttes som startværdier i en iteration der leder frem til likelihoodfunktionens maksimumspunkt som er

$$\begin{aligned} \hat{\beta} &= 0.5378 \\ \hat{\kappa} &= 0.8651 \\ \hat{p} &= 0.6503. \end{aligned}$$

Tabel 2.5 Model 2: Observerede antal f_y og forventede antal $\hat{\hat{f}}_y$ i den negative binomialfordelingsmodel.

y	f_y	$\hat{\hat{f}}_y$
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



Figur 2.2 Model 2: Observerede antal (sorte søjler) og forventede antal (hvide søjler) fra Tabel 2.5.

Tabel 2.5 og Figur 2.2 viser de tilsvarende forventede antal

$$\hat{\hat{f}}_y = n \binom{y + \hat{\kappa} - 1}{y} \hat{p}^{\hat{\kappa}} (1 - \hat{p})^y$$

beregnet ud fra den estimerede negative binomialfordeling. På baggrund heraf tillader vi os at konkludere at den negative binomialfordelingsmodel beskriver observationerne godt nok.

2.4 Opgaver

Opgave 2.1 (Udsendelse af α-partikler)
I et berømt eksperiment har Rutherford og Geiger[7] talt op hvor mange α-partikler der udsendes fra en bestemt portion af det radioaktive stof Polonium i et

[7] E. Rutherford and H. Geiger (1910): The Probability Variations in the Distribution of α Particles. *Philosophical Magazine* **xx**, 698-707, genoptrykt med mindre rettelser i *The Collected Papers of Lord Rutherford of Nelson*, Vol. 2, side 203-211, London, 1963.

Tabel 2.6 Opgave 2.1: Antal tidsintervaller f_y hvor der udsendes netop y α-partikler.

y	f_y	y	f_y
0	57	8	45
1	203	9	27
2	383	10	10
3	525	11	4
4	532	12	0
5	408	13	1
6	273	14	1
7	139		

tidsinterval af længde 7.5 sekund; man har foretaget optællingen for i alt 2608 sådanne tidsintervaller. Resultaterne fremgår af Tabel 2.6.

Det formodes at antal α-partikler udsendt i et tidsinterval af længde t (som er meget mindre end stoffets halveringstid) kan opfattes som en observation af en Poissonfordelt stokastisk variabel med parameter $\lambda \cdot t$, hvor λ er en slags strålingsintensitet.

1. Gør rede for rimeligheden af Poissonfordelingsantagelsen, og præcisér den statistiske model.
2. Estimér λ ud fra de givne observationer.
3. Hvad kan dispersionstestet fortælle om rimeligheden af den foreslåede model?

Opgave 2.2

Tabel 2.7 indeholder 20 stikprøver $y_1, y_2, \ldots, y_{10}$ fra en Poissonfordeling med $\mu = 3.14$.

1. Udregn $\hat{\mu}$ for hver stikprøve. Hvordan fordeler $\hat{\mu}$ sig omkring μ?
2. Udregn dispersionsteststørrelsen d for hver stikprøve. Hvordan ligger værdierne i forhold til χ^2/f-fordelingen?
3. Man kan bevise at en sum af uafhængige Poissonfordelte størrelser er Poissonfordelt med en parameter der er lig summen af parametrene. Derfor kan man opfatte de 20 værdier i y.-søjlen som 20 observationer fra en Poissonfordeling med parameter 10μ ($= 31.4$).

   Udregn parameterestimatet og dispersionsteststørrelsen for disse 20 observationer.

Opgave 2.3 (fortsættelse af Opgave 1.2)

1. Antag at der er udført n cirklinger, og at i netop y tilfælde fandtes der et skud inde i cirklen.

   Hvordan skal man på denne baggrund estimere λ?

Tabel 2.7 20 eksempler på udfald af stokastiske variable $Y_1, Y_2, \ldots, Y_{10}$ frembragt af en Poissonfordelings-tilfældighedsmekanisme med $\mu = 3.14$.

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	$y_\bullet$	$\overline{y}$	s^2
2	4	2	2	0	5	2	3	4	2	26	2.60	2.04
3	3	3	0	4	3	3	3	3	5	30	3.00	1.56
4	2	5	1	0	6	5	6	1	2	32	3.20	5.07
3	1	4	3	3	2	0	2	3	3	24	2.40	1.38
3	4	5	4	4	2	3	6	2	1	34	3.40	2.27
2	5	6	6	5	3	4	2	4	2	39	3.90	2.54
2	2	6	2	4	2	4	1	1	6	30	3.00	3.56
4	1	2	0	2	3	7	4	4	2	29	2.90	3.88
1	3	4	2	1	2	2	2	3	3	23	2.30	0.90
3	3	2	2	5	4	2	3	6	4	34	3.40	1.82
6	5	5	3	3	2	3	3	3	2	35	3.50	1.83
3	4	1	4	3	4	4	3	3	4	33	3.30	0.90
1	2	3	2	1	2	1	2	5	5	24	2.40	2.27
4	2	3	1	5	8	5	5	2	1	36	3.60	4.93
3	4	3	4	3	1	5	2	3	5	33	3.30	1.57
2	2	2	6	4	5	3	2	2	0	28	2.80	3.07
3	4	3	2	3	2	3	2	0	1	23	2.30	1.34
2	0	1	2	2	5	6	4	2	2	26	2.60	3.38
1	2	4	2	3	3	4	0	3	4	26	2.60	1.82
6	0	7	3	0	6	3	4	3	4	36	3.60	5.60

2. Hvis man skal kunne opdage sjældne plantearter med denne metode, skal man nok bruge mere end 10 cirklinger.

   Antag at man stadig bruger cirkler med areal $a = 0.1\text{m}^2$. Hvis en plante vokser med en tæthed på ca. en pr. 5m^2 (dvs. $\lambda = 0.2\text{m}^{-2}$), hvor mange cirklinger skal man da foretage for at være 90% sikker på at opdage planten?

   TIP: Opskriv først sandsynligheden for at man i n cirklinger *ikke* opdager planten.

Opgave 2.4 (Fluor i drikkevandet)
Det menes at fluor i drikkevandet kan modvirke huller i tænderne. I 1960-erne foretog man en undersøgelse af børns »tandstatus« og sammenholdt den med koncentrationen af fluor-ioner i drikkevandet fra det lokale vandværk. Tabel 2.8 viser data fra to vandværksdistrikter i Næstved. Man har bestemt antal DMF-tænder, dvs. tænder med huller efter caries samt udtrukne og plomberede tænder, hos de 12-årige drenge i de to distrikter. (Det kan i øvrigt nævnes at F^--koncentrationen ved det gamle vandværk var 1.9 ppm og ved hjælpevandværket 1.2 ppm.)

Undersøg ved hjælp af en Poissonfordelingsmodel om der er en signifikant forskel på forekomsten af DMF-tænder i de to vandværksdistrikter.

Tabel 2.8 Opgave 2.4: Fordelingen af drenge fra to vandværksdistrikter efter antal DMF-tænder.

y	antal drenge med y DMF-tænder	
	gamle vv.	hjælpe-vv.
0	1	
1	1	
2	8	6
3	3	5
4	13	6
5	7	3
6	8	7
7	2	4
8		3
9	2	4
10		1
11		
12		
13		1

3 Multiplikative Poissonmodeller

I dette kapitel vil vi gennemgå et eksempel på en såkaldt multiplikativ Poissonmodel. Modellen er ganske vist en smule mere indviklet end hvad der hidtil er blevet præsenteret, men på den anden side er det en type modeller der benyttes en del. Derudover 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.

3.1 Det gennemgående eksempel: Lungekræft i Fredericia

I midten af 1970-erne var der en større debat om hvorvidt der var særlig stor risiko for at få lungekræft når man boede i byen Fredericia. Grunden til at der kunne være en større risiko, var at der i Fredericia var en betydelig mængde luftforurenende industri som tilmed lå midt inde i byen. For at kunne afgøre spørgsmålet indsamlede man data om lungekræfthyppigheden i perioden 1968-71, dels i Fredericia, dels i byerne Horsens, Kolding og Vejle. De tre sidste byer skulle tjene som sammenligningsgrundlag idet det var byer af nogenlunde samme art som Fredericia, pånær den mistænkte industri.

Lungekræft opstår tit som et resultat af 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, og det er under alle omstændigheder tilfældet at lungekræft optræder med meget forskellig hyppighed i forskellige aldersklasser. Det er derfor ikke nok at se på totalantallene af lungekræfttilfælde, man skal se på antallene af tilfælde i forskellige aldersklasser. De foreliggende tal er vist i Tabel 3.1. Da antallene af lungekræfttilfælde i sig selv ikke siger noget sålænge man ikke kender risikogruppernes størrelse, må man også rapportere antal indbyggere i de forskellige aldersklasser og byer, se Tabel 3.2.[1]

Det der nu er statistikerens opgave, er at beskrive tallene i Tabel 3.1 ved hjælp af en statistisk model hvori der indgår nogle 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.

[1] Tallene i Tabel 3.1 og 3.2 er citeret efter E.B.Andersen (1977): Multiplicative Poisson models

Tabel 3.1 Lungekræfttilfælde i fire byer fordelt på aldersklasser.

aldersklasse	Fredericia	Horsens	Kolding	Vejle	i alt
40-54	11	13	4	5	»33
55-59	11	6	8	7	32
60-64	11	15	7	10	43
65-69	10	10	11	14	45
70-74	11	12	9	8	40
75+	10	2	12	7	31
i alt	64	58	51	51	224

Tabel 3.2 Antal indbyggere i de forskellige aldersklasser i de fire byer.

aldersklasse	Fredericia	Horsens	Kolding	Vejle	i alt
40-54	3059	2879	3142	2520	11600
55-59	800	1083	1050	878	3811
60-64	710	923	895	839	3367
65-69	581	834	702	631	2748
70-74	509	634	535	539	2217
75+	605	782	659	619	2665
i alt	6264	7135	6983	6026	26408

3.2 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 variabel der skal specificeres af den statistiske model.

Vi indfører noget notation:

$$\begin{aligned} y_{ij} &= \text{antal tilfælde i aldersgruppe } i \text{ i by } j, \\ r_{ij} &= \text{antal personer i aldersgruppe } i \text{ i by } j, \end{aligned}$$

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}.

Inspireret af Kapitel 1 kunne man foreslå at Y_{ij} skulle 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 $\mu_{ij} = \lambda_{ij} \cdot r_{ij}$, 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

with unequal cell rates. *Scand. J. Statist.* **4**, 153-8.

gå ud fra at de enkelte Y_{ij}-er er stokastisk uafhængige. *Grundmodellen* bliver således

> 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 nok at estimere parametrene i grundmodellen. Eksempelvis estimeres intensiteten λ_{21} for 55-59–årige i Fredericia til $11/800 = 0.014$ (dvs. 0.014 tilfælde pr. person pr. 4 år). Den generelle opskrift er $\hat{\lambda}_{ij} = y_{ij}/r_{ij}$.

Nu var det jo tanken at vi gerne ville kunne komme til at sammenligne byerne efter at vi havde 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 hvori λ_{ij} er 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 α_1, α_2, α_3, α_4, α_5, α_6, β_1, β_2, β_3, β_4 er ukendte parametre. (Mere udførligt lyder hypotesen: Der findes parametre α_1, α_2, α_3, α_4, α_5, α_6, β_1, β_2, β_3, β_4 således at der for by j og aldersgruppe i gælder at lungekræftrisikoen λ_{ij} fås som $\lambda_{ij} = \alpha_i \beta_j$. – Hypotesen H_0 specificerer 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. At en parametrisering er *injektiv* betyder at forskellige parametersæt giver forskellige udgaver af modellen.

De 10 parametre $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_1, \beta_2, \beta_3, \beta_4$ indgår udelukkende i modellen via produkterne $\alpha_i \beta_j$ ($= \lambda_{ij}$). Antag nu at to parametersæt

$$(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_1, \beta_2, \beta_3, \beta_4)$$

og

$$(\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*, \alpha_5^*, \alpha_6^*, \beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*)$$

giver anledning til de samme produkter, dvs. antag at

$$\alpha_i \beta_j \quad = \quad \alpha_i^* \beta_j^* \tag{3.1}$$

for alle i og j. Så gælder også

$$\alpha_i / \alpha_i^* \quad = \quad \beta_j^* / \beta_j \tag{3.2}$$

for alle i og j. Da højresiden af formel (3.2) ikke involverer i, så kan venstresiden heller ikke afhænge af i, det vil sige der findes en konstant c således at $\alpha_i/\alpha_i^* = c$ og dermed

$$\alpha_i^* = \frac{\alpha_i}{c}$$

for alle i. Videre er $\beta_j^*/\beta_j = \alpha_i/\alpha_i^* = c$, det vil sige

$$\beta_j^* = c\beta_j$$

for alle j. Parametersættet $(\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*, \alpha_5^*, \alpha_6^*, \beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*)$ må altså nødvendigvis være af formen

$$\begin{aligned}&(\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*, \alpha_5^*, \alpha_6^*, \beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*)\\ &= (\frac{\alpha_1}{c}, \frac{\alpha_2}{c}, \frac{\alpha_3}{c}, \frac{\alpha_4}{c}, \frac{\alpha_5}{c}, \frac{\alpha_6}{c}, c\beta_1, c\beta_2, c\beta_3, c\beta_4)\end{aligned} \tag{3.3}$$

hvor c er en positiv konstant. Omvendt gælder også at hvis det stjernede parametersæt er defineret ved formel (3.3), så vil formel (3.1) være opfyldt. Hermed har vi fået klarlagt dels at parametriseringen ikke er injektiv, dels hvilke parametersæt der giver den samme model.

De 10 parametre skal pålægges ét bånd for at få en injektiv parametrisering. Et sådant bånd kan være at $\alpha_1 = 1$, eller at $\alpha_1 + \alpha_2 + \ldots + \alpha_6 = 1$, 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 definerer at parameteren for Fredericia skal være lig 1. Med denne betingelse er parametriseringen injektiv, for hvis både β_1 og $\beta_1^* = c\beta_1$ skal være 1, så må c nødvendigvis være lig 1.

Samtidig noterer vi at der er $10 - 1 = 9$ forskellige parametre at estimere.

3.3 Estimation i den multiplikative model

I den multiplikative model lader det sig ikke gøre at opskrive simple udtryk for estimaterne, man er henvist til at benytte numeriske metoder for at bestemme talværdierne i de konkrete tilfælde. En datamat med noget ordentligt statistikprogrammel vil uden videre kunne levere estimaterne, men hvis man foretrækker at gå ud fra de grundlæggende principper og foretage udregningerne med håndkraft/lommeregner, er det faktisk heller ikke særlig besværligt. Det skal vi se på de næste sider.

Parametersættet $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_1, \beta_2, \beta_3, \beta_4)$ (med den bibetingelse at $\beta_1 = 1$) skal ifølge de sædvanlige principper bestemmes så det maksimaliserer likelihoodfunktionen. I grundmodellen er likelihoodfunktionen

$$\begin{aligned}L &= \prod_{i=1}^{6}\prod_{j=1}^{4} \frac{(\lambda_{ij}r_{ij})^{y_{ij}}}{y_{ij}!} \exp(-\lambda_{ij}r_{ij})\\ &= \text{konstant} \cdot \prod_{i=1}^{6}\prod_{j=1}^{4} \lambda_{ij}^{y_{ij}} \exp(-\lambda_{ij}r_{ij}).\end{aligned}$$

Når vi her erstatter λ_{ij} med $\alpha_i\beta_j$, får vi likelihoodfunktionen under H_0:

$$\begin{aligned} L_0 &= \text{konstant} \cdot \prod_{i=1}^{6}\prod_{j=1}^{4} \alpha_i^{y_{ij}} \beta_j^{y_{ij}} \exp(-\alpha_i\beta_j r_{ij}) \\ &= \text{konstant} \cdot \left(\prod_{i=1}^{6} \alpha_i^{y_{i\bullet}}\right) \left(\prod_{j=1}^{4} \beta_j^{y_{\bullet j}}\right) \exp\left(-\sum_{i=1}^{6}\sum_{j=1}^{4} \alpha_i\beta_j r_{ij}\right). \end{aligned}$$

Vi skal bestemme det parametersæt $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\alpha}_5, \hat{\alpha}_6, 1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$ der maksimaliserer L_0. Vi vil dog benytte *log-likelihoodfunktionen* $\ln L_0$ i stedet:

$$\ln L_0 = \text{konstant} + \left(\sum_{i=1}^{6} y_{i\bullet} \ln \alpha_i\right) + \left(\sum_{j=1}^{4} y_{\bullet j} \ln \beta_j\right) - \sum_{i=1}^{6}\sum_{j=1}^{4} \alpha_i\beta_j r_{ij}.$$

Opgaven at maksimalisere $\ln L_0$ lader sig ikke løse sådan lige uden videre, og man må derfor inddrage den generelle matematiske teori for hvordan man maksimaliserer en funktion af mange variable. Hvis $\ln L_0$ havde været en funktion af én variabel θ, så kunne man (under visse omstændigheder) bestemme maksimumspunktet $\hat{\theta}$ som løsningen til »likelihoodligningen«

$$\tfrac{d}{d\theta} \ln L_0 = 0.$$

Tilsvarende kan maksimumspunktet $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\alpha}_5, \hat{\alpha}_6, 1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$ for den aktuelle log-likelihoodfunktion bestemmes som løsning til »likelihoodligningerne«

$$\left.\begin{aligned} \tfrac{\partial}{\partial\alpha_1} \ln L_0 &= 0, \\ \tfrac{\partial}{\partial\alpha_2} \ln L_0 &= 0, \\ \tfrac{\partial}{\partial\alpha_3} \ln L_0 &= 0, \\ \tfrac{\partial}{\partial\alpha_4} \ln L_0 &= 0, \\ \tfrac{\partial}{\partial\alpha_5} \ln L_0 &= 0, \\ \tfrac{\partial}{\partial\alpha_6} \ln L_0 &= 0, \\ \tfrac{\partial}{\partial\beta_1} \ln L_0 &= 0, \\ \tfrac{\partial}{\partial\beta_2} \ln L_0 &= 0, \\ \tfrac{\partial}{\partial\beta_3} \ln L_0 &= 0, \\ \tfrac{\partial}{\partial\beta_4} \ln L_0 &= 0. \end{aligned}\right\} \tag{3.4}$$

Her er eksempelvis $\frac{\partial}{\partial\alpha_2} \ln L_0$ den såkaldte *partielle afledede af* $\ln L_0$ *med hensyn til* α_2, dvs. den afledede af $\ln L_0$ med hensyn til α_2 når de øvrige variable holdes fast. – Man finder at

$$\tfrac{\partial}{\partial\alpha_2} \ln L_0 = \frac{y_{2\bullet}}{\alpha_2} - \sum_{j=1}^{4} \beta_j r_{2j}.$$

Deraf ses at ligningen $\frac{\partial}{\partial\alpha_2} \ln L_0 = 0$ er ensbetydende med at

$$\alpha_2 = \frac{y_{2\bullet}}{\sum_{j=1}^{4} \beta_j r_{2j}}.$$

Hvis man på samme måde løser alle de øvrige ligninger i (3.4), får man at følgende relationer skal være opfyldt:

$$\alpha_i = \frac{y_{i\cdot}}{\sum_{j=1}^{4} \beta_j r_{ij}}, \quad i \in \{1,2,3,4,5,6\} \tag{3.5}$$

$$\beta_j = \frac{y_{\cdot j}}{\sum_{i=1}^{6} \alpha_i r_{ij}}, \quad j \in \{1,2,3,4\} \tag{3.6}$$

og stadig

$$\beta_1 = 1.$$

I ligningerne (3.5) og (3.6) er y-erne og r-erne kendte tal og α-erne og β-erne de ubekendte. Man kan ikke løse ligningerne eksplicit, dvs. man kan ikke opskrive en løsning af formen »estimaterne over α og β = en kendt funktion af y-erne og r-erne«. I stedet er man henvist til at bestemme en numerisk løsning iterativt, for eksempel ved brug af følgende algoritme:

1. Vælg startværdier for β_2, β_3 og β_4 ($\beta_1 = 1$).
2. Indsæt β-værdierne i (3.5) og få derved $(\alpha_1, \alpha_2, \ldots, \alpha_6)$.
3. Indsæt α-værdierne i (3.6) og få derved $(\beta_1, \beta_2, \beta_3, \beta_4)$.
4. Justér β-værdierne så $\beta_1 = 1$, dvs. erstat de beregnede værdier med $\left(\frac{\beta_1}{\beta_1} = 1, \frac{\beta_2}{\beta_1}, \frac{\beta_3}{\beta_1}, \frac{\beta_4}{\beta_1}\right)$.
5. Du har nu gennemført et iterationstrin. Du kan så enten gå tilbage til punkt 2, eller du kan slutte.

Man vil almindeligvis vælge at slutte hvis enten det samlede antal iterationstrin er blevet for stort, eller parameterestimaterne næsten ikke har ændret sig siden forrige iterationstrin.

Den metode prøver vi: Som startværdi vælges $(\beta_1, \beta_2, \beta_3, \beta_4) = (1,1,1,1)$ svarende til at byerne er ens. Skridt 2 leverer værdierne

$$\begin{aligned}
\alpha_1 &= 0.0028 \\
\alpha_2 &= 0.0084 \\
\alpha_3 &= 0.0128 \\
\alpha_4 &= 0.0164 \\
\alpha_5 &= 0.0180 \\
\alpha_6 &= 0.0116.
\end{aligned}$$

Skridt 3 leverer et sæt nye β-er:

$$\begin{aligned}
\beta_1 &= 1.2779 \\
\beta_2 &= 0.9187 \\
\beta_3 &= 0.8814 \\
\beta_4 &= 0.9733.
\end{aligned}$$

Skridt 4 består i at dividere hver af de fundne β-værdier med 1.2779 hvorved fås

$$\begin{aligned}\beta_1 &= 1\\ \beta_2 &= 0.7189\\ \beta_3 &= 0.6897\\ \beta_4 &= 0.7616.\end{aligned}$$

Således fortsættes et par gange indtil værdierne (med ca. tre betydende cifre) har stabiliseret sig. De på denne måde bestemte estimater er

$$\begin{aligned}\hat{\alpha}_1 &= 0.0036\\ \hat{\alpha}_2 &= 0.0108\\ \hat{\alpha}_3 &= 0.0164\\ \hat{\alpha}_4 &= 0.0210\\ \hat{\alpha}_5 &= 0.0229\\ \hat{\alpha}_6 &= 0.0148\\ \beta_1 &= 1\\ \hat{\beta}_2 &= 0.719\\ \hat{\beta}_3 &= 0.690\\ \hat{\beta}_4 &= 0.762.\end{aligned}$$

3.4 Den multiplikative models beskrivelse af data

Efter at have bestemt de bedste estimater over α-erne og β-erne skal vi nu beskæftige os med hvor god en beskrivelse de faktisk giver af datamaterialet.

Formelt består opgaven i at *teste* multiplikativitetshypotesen H_0, og dette gøres som sædvanlig med et kvotienttest: Man udregner $-2\ln Q$ hvor

$$Q = \frac{L_0(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\alpha}_5, \hat{\alpha}_6, 1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)}{L(\hat{\lambda}_{11}, \hat{\lambda}_{12}, \ldots, \hat{\lambda}_{63}, \hat{\lambda}_{64})}.$$

Små værdier af Q eller store værdier af $-2\ln Q$ er signifikante, dvs. de tyder på at H_0 *ikke* giver en tilstrækkelig god beskrivelse af data. For at afgøre om $-2\ln Q_{\text{obs}}$ er signifikant stor skal vi se på testsandsynligheden ε, altså sandsynligheden for at få en værre $-2\ln Q$-værdi forudsat at H_0 er rigtig:

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

Når H_0 er rigtig, er $-2\ln Q$ med god tilnærmelse χ^2-fordelt med $f = 24 - 9 = 15$ frihedsgrader.[2] Det betyder at testsandsynligheden kan bestemmes som

$$\varepsilon = \mathrm{P}\Big(\chi^2_{15} \geq -2\ln Q_{\text{obs}}\Big).$$

[2] Tilnærmelsen er som sædvanlig god nok når de forventede antal alle er mindst fem.

Udtrykket for kvotienttesstørrelsen Q kan omformes således:

$$\begin{aligned} Q &= \frac{\prod_{i=1}^{6}\prod_{j=1}^{4}(\hat{\alpha}_i\hat{\beta}_j)^{y_{ij}}\exp(-\hat{\alpha}_i\hat{\beta}_j r_{ij})}{\prod_{i=1}^{6}\prod_{j=1}^{4}\hat{\lambda}_{ij}^{y_{ij}}\exp(-\hat{\lambda}_{ij}r_{ij})} \\ &= \prod_{i=1}^{6}\prod_{j=1}^{4}\left(\frac{\hat{\alpha}_i\hat{\beta}_j}{\hat{\lambda}_{ij}}\right)^{y_{ij}}\cdot\exp\left(-\sum_{i=1}^{6}\sum_{j=1}^{4}\hat{\alpha}_i\hat{\beta}_j r_{ij}+\sum_{i=1}^{6}\sum_{j=1}^{4}\hat{\lambda}_{ij}r_{ij}\right) \\ &= \prod_{i=1}^{6}\prod_{j=1}^{4}\left(\frac{\hat{\alpha}_i\hat{\beta}_j r_{ij}}{y_{ij}}\right)^{y_{ij}}\cdot\exp\left(-\sum_{i=1}^{6}\sum_{j=1}^{4}\hat{\alpha}_i\hat{\beta}_j r_{ij}+\sum_{i=1}^{6}\sum_{j=1}^{4}y_{ij}\right) \\ &= \prod_{i=1}^{6}\prod_{j=1}^{4}\left(\frac{\hat{y}_{ij}}{y_{ij}}\right)^{y_{ij}}\cdot\exp\left(-\sum_{i=1}^{6}\sum_{j=1}^{4}\hat{y}_{ij}+\sum_{i=1}^{6}\sum_{j=1}^{4}y_{ij}\right), \end{aligned}$$

hvor $\hat{y}_{ij} = \hat{\alpha}_i\hat{\beta}_j r_{ij}$ er det forventede antal lungekræfttilfælde i aldersklasse i i by j.

Der gælder at det totale forventede antal $\sum_{i=1}^{6}\sum_{j=1}^{4}\hat{y}_{ij}$ er lig det totale observerede antal $y_{\cdot\cdot} = \sum_{i=1}^{6}\sum_{j=1}^{4}y_{ij}$, fordi da estimaterne opfylder ligningerne (3.5) og (3.6) på side 36, så er

$$\hat{y}_{ij} = \hat{\alpha}_i\hat{\beta}_j r_{ij} = \frac{y_{i\cdot}}{\sum_{l=1}^{4}\hat{\beta}_l r_{il}}\hat{\beta}_j r_{ij} = y_{i\cdot}\frac{\hat{\beta}_j r_{ij}}{\sum_{l=1}^{4}\hat{\beta}_l r_{il}}$$

hvoraf

$$\sum_{j=1}^{4}\hat{y}_{ij} = y_{i\cdot}\sum_{j=1}^{4}\frac{\hat{\beta}_j r_{ij}}{\sum_{l=1}^{4}\hat{\beta}_l r_{il}} = y_{i\cdot}$$

og dermed

$$\sum_{i=1}^{6}\sum_{j=1}^{4}\hat{y}_{ij} = \sum_{i=1}^{6}y_{i\cdot} = y_{\cdot\cdot}.$$

Det betyder at det fundne udtryk for Q reduceres til

$$Q = \prod_{i=1}^{6}\prod_{j=1}^{4}\left(\frac{\hat{y}_{ij}}{y_{ij}}\right)^{y_{ij}},$$

og dermed er

$$-2\ln Q = 2\sum_{i=1}^{6}\sum_{j=1}^{4}y_{ij}\ln\frac{y_{ij}}{\hat{y}_{ij}}.$$

Tabel 3.3 Estimerede alders- og byspecifikke lungekræftintensiteter i perioden 1986-71 under forudsætning af den multiplikative Poissonmodel. Værdierne er antal pr. 1000 indbyggere pr. 4 år.

aldersklasse	Fredericia	Horsens	Kolding	Vejle
40-54	3.6	2.6	2.5	2.7
55-59	10.8	7.8	7.5	8.2
60-64	16.4	11.8	11.3	12.5
65-69	21.0	15.1	14.5	16.0
70-74	22.9	16.5	15.8	17.4
75+	14.8	10.6	10.2	11.3

Tabel 3.4 De forventede antal $\hat{y}_{ij}$ af lungekræfttilfælde under den multiplikative Poissonmodel.

aldersklasse	Fredericia	Horsens	Kolding	Vejle	i alt
40-54	11.01	7.45	7.80	6.91	»33.17
55-59	8.64	8.41	7.82	7.23	32.10
60-64	11.64	10.88	10.13	10.48	43.13
65-69	12.20	12.59	10.17	10.10	45.06
70-74	11.66	10.44	8.45	9.41	39.96
75+	8.95	8.32	6.73	6.98	30.98
i alt	64.10	58.09	51.10	51.11	224.40

Som led i beregningerne af $\hat{y}_{ij}$ udregnes de estimerede alders- og by-specifikke lungekræftintensiteter $\hat{\alpha}_i\hat{\beta}_j$. Værdierne af $1000\hat{\alpha}_i\hat{\beta}_j$, dvs. de forventede antal tilfælde pr. 1000 indbyggere, ses i Tabel 3.3. Selve de forventede antal $\hat{y}_{ij}$ i de forskellige byer og aldersklasser ses i Tabel 3.4. Indsættes tallene fra Tabel 3.1 og Tabel 3.4 i udtrykket for $-2\ln Q$, får man $-2\ln Q_{\text{obs}} = 22.6$ (men se dog også Afsnit 3.9, blandt andet Figur 3.1!). I χ^2-fordelingen med $f = 24 - 9 = 15$ frihedsgrader er 90%-fraktilen 22.3 og 95%-fraktilen 25.0. Den opnåede værdi $-2\ln Q_{\text{obs}} = 22.6$ svarer altså til en testsandsynlighed ε på godt 5%, og der er dermed ikke alvorlig evidens imod modellens brugbarhed. Vi tillader os at gå ud fra at modellen faktisk *er* anvendelig, dvs. at *lungekræftrisikoen afhænger multiplikativt af by og alder*.

Hermed er vi nået frem til en statistisk model der beskriver data ved hjælp af nogle by-parametre og nogle alders-parametre, men 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.

3.5 Ens byer?

Det hele går ud på at vurdere om der er nogen signifikant forskel på byerne. Hvis der ikke er nogen forskel, så må byparametrene være ens, dvs. $\beta_1 = \beta_2 = \beta_3 = \beta_4$, og da $\beta_1 = 1$, må den fælles værdi være 1. Derfor vil vi teste den statistiske hypotese

$$H_1 : \ \beta_1 = \beta_2 = \beta_3 = \beta_4 = 1.$$

Hypotesen skal testes i forhold til den aktuelle grundmodel H_0, så teststørrelsen bliver

$$Q \quad = \quad \frac{L_1(\hat{\hat{\alpha}}_1, \hat{\hat{\alpha}}_2, \hat{\hat{\alpha}}_3, \hat{\hat{\alpha}}_4, \hat{\hat{\alpha}}_5, \hat{\hat{\alpha}}_6)}{L_0(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\alpha}_5, \hat{\alpha}_6, 1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)}$$

hvor

$$L_1(\alpha_1, \alpha_2, \ldots, \alpha_6) \quad = \quad L_0(\alpha_1, \alpha_2, \ldots, \alpha_6, 1, 1, 1, 1)$$

er likelihoodfunktionen under H_1, og hvor $\hat{\hat{\alpha}}_1, \hat{\hat{\alpha}}_2, \ldots, \hat{\hat{\alpha}}_6$ er estimaterne over $\alpha_1, \alpha_2, \ldots, \alpha_6$ under H_1, dvs. $\hat{\hat{\alpha}}_1, \hat{\hat{\alpha}}_2, \ldots, \hat{\hat{\alpha}}_6$ maksimaliserer L_1.

Likelihoodfunktionen L_1 kan omskrives til et produkt af seks funktioner, hver med sit α:

$$\begin{aligned} L_1(\alpha_1, \alpha_2, \ldots, \alpha_6) &= \text{konstant} \cdot \prod_{i=1}^{6} \prod_{j=1}^{4} \alpha_i^{y_{ij}} \exp(-\alpha_i r_{ij}) \\ &= \text{konstant} \cdot \prod_{i=1}^{6} \alpha_i^{y_{i\bullet}} \exp(-\alpha_i r_{i\bullet}) \, . \end{aligned}$$

Maksimaliseringsestimatet findes derfor til $\hat{\hat{\alpha}}_i = \frac{y_{i\bullet}}{r_{i\bullet}}$. Talværdierne bliver

$$\begin{array}{rcrcl} \hat{\hat{\alpha}}_1 & = & 33/11600 & = & 0.002845 \\ \hat{\hat{\alpha}}_2 & = & 32/3811 & = & 0.00840 \\ \hat{\hat{\alpha}}_3 & = & 43/3367 & = & 0.0128 \\ \hat{\hat{\alpha}}_4 & = & 45/2748 & = & 0.0164 \\ \hat{\hat{\alpha}}_5 & = & 40/2217 & = & 0.0180 \\ \hat{\hat{\alpha}}_6 & = & 31/2665 & = & 0.0116. \end{array}$$

Kvotienttteststørrelsen omskrives således:

$$\begin{aligned} Q &= \frac{L_1(\hat{\hat{\alpha}}_1, \hat{\hat{\alpha}}_2, \hat{\hat{\alpha}}_3, \hat{\hat{\alpha}}_4, \hat{\hat{\alpha}}_5, \hat{\hat{\alpha}}_6)}{L_0(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\alpha}_5, \hat{\alpha}_6, 1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)} \\ &= \frac{\prod_{i=1}^{6} \prod_{j=1}^{4} \hat{\hat{\alpha}}_i^{y_{ij}} \exp(-\hat{\hat{\alpha}}_i r_{ij})}{\prod_{i=1}^{6} \prod_{j=1}^{4} (\hat{\alpha}_i \hat{\beta}_j)^{y_{ij}} \exp(-\hat{\alpha}_i \hat{\beta}_j r_{ij})} \\ &= \prod_{i=1}^{6} \prod_{j=1}^{4} \left(\frac{\hat{\hat{y}}_{ij}}{\hat{y}_{ij}} \right)^{y_{ij}} \cdot \exp\left(- \sum_{i=1}^{6} \sum_{j=1}^{4} \hat{\hat{y}}_{ij} + \sum_{i=1}^{6} \sum_{j=1}^{4} \hat{y}_{ij} \right), \end{aligned}$$

Tabel 3.5 De forventede antal $\hat{y}_{ij}$ 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

hvor $\hat{y}_{ij} = \hat{\alpha}_i\hat{\beta}_j r_{ij}$ (som hidtil) og $\hat{\hat{y}}_{ij} = \hat{\hat{\alpha}}_i r_{ij}$.

Da $\sum_{i=1}^{6}\sum_{j=1}^{4}\hat{y}_{ij} = \sum_{i=1}^{6}\sum_{j=1}^{4}\hat{\hat{y}}_{ij} = \sum_{i=1}^{6}\sum_{j=1}^{4} y_{ij}$, er

$$Q = \prod_{i=1}^{6}\prod_{j=1}^{4}\left(\frac{\hat{\hat{y}}_{ij}}{\hat{y}_{ij}}\right)^{y_{ij}}$$

og dermed

$$-2\ln Q = 2\sum_{i=1}^{6}\sum_{j=1}^{4} y_{ij}\ln\frac{\hat{y}_{ij}}{\hat{\hat{y}}_{ij}}\,.$$

Store værdier af $-2\ln Q$ er signifikante. Man skal sammenholde $-2\ln Q$ med χ^2-fordelingen med $f = 9 - 6 = 3$ frihedsgrader.

De forventede tal er vist i Tabel 3.5. Indsættes værdierne fra Tabel 3.1, Tabel 3.4 og Tabel 3.5 i udtrykket for $-2\ln Q$, fås $-2\ln Q_{\text{obs}} = 5.67$ (men se dog også Afsnit 3.9, blandt andet Figur 3.2!). I χ^2-fordelingen med $f = 9 - 6 = 3$ frihedsgrader er 80%-fraktilen 4.64 og 90%-fraktilen 6.25, således at testsandsynligheden ε er næsten 20%. De foreliggende observationer er altså fint forenelige med hypotesen H_1 om at der ikke er nogen forskel på byerne. Sagt på en anden måde, *der er ikke nogen signifikant forskel på byerne*.

3.6 En anden mulighed

Det er sjældent tilfældet at der er én bestemt måde at undersøge en praktisk problemstilling på ved hjælp af en statistisk model og en statistisk hypotese. Det aktuelle spørgsmål om der er en øget risiko for lungekræft ved at bo i Fredericia, blev i forrige afsnit belyst ved at vi 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 kan man imidlertid angribe problemet på en anden måde. Man kan sige at det hele drejer sig om at vurdere om det er farligere at bo i Fredericia end i en af de tre øvrige byer. Dermed er det indirekte forudsat at de tre øvrige byer

er stort set ens, hvilket man bør teste. Man kunne derfor anlægge følgende strategi for formulering og test af hypoteser:

1. Vi går stadig ud fra den multiplikative Poissonmodel H_0 som grundmodel.

2. Først undersøges om det kan antages at de tre byer Horsens, Kolding og Vejle er ens, dvs. vi vil teste hypotesen

$$H_2 : \beta_2 = \beta_3 = \beta_4$$

3. Hvis H_2 bliver accepteret, er der et fælles niveau kaldet β_0 for de tre »kontrolbyer«. Vi kan derefter sammenligne Fredericia med dette fælles niveau ved at teste om $\beta_1 = \beta_0$. Da β_1 pr. definition er lig 1, er den hypotese der skal testes,

$$H_3 : \beta_0 = 1.$$

Sammenligning af de tre kontrolbyer

Vi skal teste hypotesen $H_2 : \beta_2 = \beta_3 = \beta_4$ om ens kontrolbyer i forhold til den multiplikative model H_0. Det gøres med et kvotienttest

$$Q = \frac{L_2(\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_6, \tilde{\beta})}{L_0(\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_6, 1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)}$$

hvor

$$L_2(\alpha_1, \alpha_2, \ldots, \alpha_6, \beta) = L_0(\alpha_1, \alpha_2, \ldots, \alpha_6, 1, \beta, \beta, \beta)$$

er likelihoodfunktionen under H_2, og $\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_6, \tilde{\beta}$ er maksimaliseringsestimaterne under H_2.

Når H_2 er rigtig, er $-2 \ln Q$ med god tilnærmelse χ^2-fordelt med $f = 9 - 7 = 2$ frihedsgrader.

Modellen H_2 svarer til en multiplikativ Poissonmodel med *to* byer (nemlig Fredericia og resten) og seks aldersklasser, og der er derfor ingen principielt nye problemer forbundet med at estimere parametrene under H_2. Man finder

$$\begin{aligned}
\tilde{\alpha}_1 &= 0.00358 \\
\tilde{\alpha}_2 &= 0.0108 \\
\tilde{\alpha}_3 &= 0.0164 \\
\tilde{\alpha}_4 &= 0.0210 \\
\tilde{\alpha}_5 &= 0.0230 \\
\tilde{\alpha}_6 &= 0.0148 \\
\tilde{\beta}_1 &= 1 \\
\tilde{\beta}_0 &= 0.7220.
\end{aligned}$$

Tabel 3.6 De forventede antal $\tilde{y}_{ij}$ af lungekræfttilfælde under H_2.

aldersklasse	Fredericia	Horsens	Kolding	Vejle	i alt
40-54	10.95	7.44	8.12	»6.51	33.02
55-59	8.64	8.44	8.19	6.85	32.12
60-64	11.64	10.93	10.60	9.93	43.10
65-69	12.20	12.65	10.64	9.57	45.06
70-74	11.71	10.53	8.88	8.95	40.07
75+	8.95	8.36	7.04	6.61	30.96
i alt	64.09	58.35	53.47	48.42	224.33

Endvidere bliver

$$-2\ln Q \;=\; 2\sum_{i=1}^{6}\sum_{j=1}^{4} y_{ij}\ln\frac{\hat{y}_{ij}}{\tilde{y}_{ij}}$$

hvor $\hat{y}_{ij} = \hat{\alpha}_i\hat{\beta}_j r_{ij}$, se Tabel 3.4, og

$$\begin{aligned}\tilde{y}_{i1} &= \tilde{\alpha}_i r_{i1}\\ \tilde{y}_{ij} &= \tilde{\alpha}_i\tilde{\beta}_0\,, \quad j=2,3,4.\end{aligned}$$

De forventede antal $\tilde{y}_{ij}$ ses i Tabel 3.6. Når man indsætter værdierne fra Tabel 3.1, Tabel 3.4 og Tabel 3.6 i det netop fundne udtryk for $-2\ln Q$, fås $-2\ln Q_{\text{obs}} = 0.40$ der skal sammenholdes med χ^2-fordelingen med $f = 9-7 = 2$ frihedsgrader (men se dog også Afsnit 3.9, blandt andet Figur 3.3!). I χ^2-fordelingen med $f = 2$ frihedsgrader er 20%-fraktilen 0.446, så testsandsynligheden er altså godt 80%, og det betyder at H_2 er udmærket forenelig med de foreliggende data. Vi kan altså udmærket tillade os at gå ud fra at der ikke er nogen signifikant forskel mellem de tre byer.

Herefter kan vi gå over til at teste H_3, der går ud på 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 $\hat{\hat{\alpha}}_1, \hat{\hat{\alpha}}_2, \ldots, \hat{\hat{\alpha}}_6$ fra side 40.

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

$$Q \;=\; \frac{L_1(\hat{\hat{\alpha}}_1,\hat{\hat{\alpha}}_2,\ldots,\hat{\hat{\alpha}}_6)}{L_2(\tilde{\alpha}_1,\tilde{\alpha}_2,\ldots,\tilde{\alpha}_6,\tilde{\beta}_0)} \;=\; \frac{L_0(\hat{\hat{\alpha}}_1,\hat{\hat{\alpha}}_2,\ldots,\hat{\hat{\alpha}}_6,1,1,1,1)}{L_0(\tilde{\alpha}_1,\tilde{\alpha}_2,\ldots,\tilde{\alpha}_6,1,\tilde{\beta}_0,\tilde{\beta}_0,\tilde{\beta}_0)}$$

der let omformes til

$$Q \;=\; \prod_{i=1}^{6}\prod_{j=1}^{4}\left(\frac{\hat{y}_{ij}}{\tilde{y}_{ij}}\right)^{y_{ij}}$$

så at

$$-2\ln Q \;=\; 2\sum_{i=1}^{6}\sum_{j=1}^{4} y_{ij}\ln\frac{\tilde{y}_{ij}}{\hat{y}_{ij}}.$$

Store værdier af $-2\ln Q$ er signifikante. Når H_3 er rigtig, er $-2\ln Q$ med god tilnærmelse[3] χ^2-fordelt med $f = 7 - 6 = 1$ frihedsgrad.

Ved at indsætte værdierne fra Tabel 3.1, Tabel 3.5 og Tabel 3.6 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 $\tilde{\beta}_0 = 0.7$, så lungekræfthyppigheden i Fredericia er altså signifikant *større*.

Se det var jo en pæn og klar konklusion, der blot er stik modsat den vi nåede frem til på side 41!

3.7 Sammenligning af de to fremgangsmåder

Vi har benyttet to forskellige fremgangsmåder der kun var en smule forskellige, men gav helt forskellige 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, så har vi derved fået en ny aktuel grundmodel (nemlig den gamle med de simplifikationer som den accepterede hypotese giver).

   Fortsæt da med punkt 2

   (b) Hvis hypotesen forkastes, så slut. Data beskrives da ved den senest anvendte grundmodel.

Begge de anvendte fremgangsmåder tog udgangspunkt i den samme Poissonmodel, de adskiller sig udelukkende ved valgene af hypoteser i punkt 2. Tabel 3.7 og Tabel 3.8 giver oversigter over de to fremgangsmåder.

I den første fremgangsmåde tages skridtet fra den multiplikative model til »fire ens« på én gang, hvilket giver en teststørrelse på 5.67, som, da den kan fordeles på 3 frihedsgrader, ikke er signifikant. I den anden fremgangsmåde spalter vi op i

1. multiplikativitet → »tre ens«, og
2. »tre ens« → »fire ens«,

[3] forudsat at de indgående forventede antal er mindst fem.

Tabel 3.7 Oversigt over den første fremgangsmåde.

Model/Hypotese	$-2\ln Q$	f	ε
M: vilkårlige parametre H: multiplikativitet	22.65	24–9=15	godt 5%
M: multiplikativitet H: fire ens byer	5.67	9–6=»3	ca. 20

Tabel 3.8 Oversigt over den anden fremgangsmåde.

Model/Hypotese	$-2\ln Q$	f	ε
M: vilkårlige parametre H: multiplikativitet	22.65	24–9=15	godt 5%
M: multiplikativitet H: de tre byer ens	0.40	9–7=»2	godt 80%
M: de tre byer ens H: de fire byer ens	5.27	7–6=»1	ca. 2

og det viser sig så at de 5.67 med 3 frihedsgrader spaltes op i 0.40 med 2 frihedsgrader og 5.27 med 1 frihedsgrad, hvoraf den sidste er temmelig signifikant.

Det kan undertiden være hensigtsmæssigt at foretage en sådan trinvis testning. Man bør dog ikke stræbe efter at spalte op i så mange tests som muligt, men kun teste hypoteser der er *rimelige* i den foreliggende faglige sammenhæng.

3.8 Om teststørrelser

Læseren vil måske have bemærket visse fælles træk ved de $-2\ln Q$-udtryk der forekommer i dette kapitel. De er alle af formen

$$-2\ln Q \quad = \quad 2\sum \text{obs.antal} \cdot \ln \frac{\text{Modellens forventede antal}}{\text{Hypotesens forventede antal}}$$

og er (tilnærmelsesvis) χ^2-fordelt med et antal frihedsgrader som er »det reelle antal parametre under modellen« minus »det reelle antal parametre under hypotesen«. Dette gælder faktisk helt generelt[4] når man tester hypoteser om Poissonfordelte observationer.

3.9 Om beregninger

Værdierne af de forskellige estimater og teststørrelser i det foregående er alle udregnet med »håndkraft«, dvs. ved brug af papir og blyant og en alminde-

[4] under forudsætning af at summen af de forventede antal er lig summen af de observerede antal.

lig lommeregner. Undervejs er der i mellemregningerne foretaget afrundinger af tallene; i den multiplikative model er for eksempel $\hat{\alpha}$-erne afrundet til fire decimaler og $\hat{\beta}$-erne til tre decimaler efter kommaet (side 37), disse afrundede værdier er benyttet ved udregning af de forventede antal, og de forventede antal benyttes ved udregning af $-2\ln Q$. Afrundingsfejlene vil i et vist omfang forplante sig til teststørrelsen $-2\ln Q$.

Man kan forsøge sig med at lade en computer (med et passende statistikprogram) foretage udregningerne. Computeren laver ganske vist også afrundinger undervejs, men den regner med flere betydende cifre end den »typiske« person gider gøre, og ofte er den også programmeret til at foretage beregningerne på en sådan måde at afrundingsfejl får mindst mulig betydning.

Vi vil derfor også vise hvad der kommer ud af at lade et »typisk« statistikprogram (nemlig *StatUnit* (Tue Tjurs Turbo Pascal unit til statistisk analyse)) foretage udregningerne, se Figur 3.1 – 3.3. I udskrifterne vil estimater og forventede værdier blive udskrevet med samme antal decimaler som de tidligere resultater.

```
StatUnit.FitLogLinear

Dependent variable TILFÆLDE.
Offset variate LOGR.

Model terms
     BY
     ALDER

24 observations, 9 parameters estimated.
-2log(Likelihood) =              23.4475
Likelihood ratio test against full model
P[ ChiSquare(15) > -2log(Likelihood) ] =  0.075090
```

```
Estimerede parametre:
   Horsens:  0.719
   Kolding:  0.690
     Vejle:  0.762
Fredericia:  1.000
     40-54:  0.0036
     55-59:  0.0108
     60-64:  0.0164
     65-69:  0.0210
     70-74:  0.0229
      75+ :  0.0148
Forventede antal:
            Horsens     Kolding       Vejle  Fredericia
40-54:        10.95        7.41        7.76        6.87
55-59:         8.62        8.38        7.80        7.20
60-64:        11.61       10.85       10.09       10.45
65-69:        12.19       12.58       10.16       10.08
70-74:        11.67       10.45        8.46        9.41
 75+ :         8.96        8.33        6.73        6.98
```

Figur 3.1 Den multiplikative model: Udskrift fra *StatUnit*.
Estimaterne er de samme som tidligere (side 37), de forventede antal er *ikke* identiske med dem i Tabel 3.4, og $-2 \ln Q$ bliver nu 23.4475 mod tidligere 22.6.

```
StatUnit.FitLogLinear

Dependent variable TILFÆLDE.
Offset variate LOGR.

Model terms
     ALDER

24 observations, 6 parameters estimated.
-2log(Likelihood) =              28.3065
Likelihood ratio test against full model
P[ ChiSquare(18) > -2log(Likelihood) ] =  0.057541
```

```
Estimerede parametre:
     40-54:  0.002845
     55-59:  0.00840
     60-64:  0.0128
     65-69:  0.0164
     70-74:  0.0180
      75+ :  0.0116
Forventede antal:
            Horsens     Kolding       Vejle  Fredericia
40-54:        8.70        8.19        8.94        7.17
55-59:        6.72        9.09        8.82        7.37
60-64:        9.07       11.79       11.43       10.71
65-69:        9.51       13.66       11.50       10.33
70-74:        9.18       11.44        9.65        9.72
 75+ :        7.04        9.10        7.67        7.20
```

```
StatUnit.TestModelChange

3 parameters removed
-2log(Q) =               4.8590
P[ ChiSquare(3) > -2log(Q) ] =  0.182414
```

Figur 3.2 Hypotesen H_1 om ens byer: Udskrift fra *StatUnit*.
Estimaterne er de samme som tidligere (side 40), de forventede antal er *ikke* identiske med dem i Tabel 3.5, og $-2 \ln Q$ bliver nu 4.8590 mod tidligere 5.67.

```
StatUnit.FitLogLinear

Dependent variable TILFÆLDE.
Offset variate LOGR.

Model terms
     FRCIA
     ALDER

24 observations, 7 parameters estimated.
-2log(Likelihood) =              23.7001
Likelihood ratio test against full model
P[ ChiSquare(17) > -2log(Likelihood) ] =  0.127816
```

```
Estimerede parametre:
      40-54:  0.00358
      55-59:  0.0108
      60-64:  0.0164
      65-69:  0.0210
      70-74:  0.0230
       75+ :  0.0148
Øvrige byer:  0.7220
Forventede antal:
            Horsens     Kolding       Vejle  Fredericia
40-54:        10.94        7.44        8.11        6.51
55-59:         8.61        8.41        8.16        6.82
60-64:        11.62       10.90       10.57        9.91
65-69:        12.19       12.63       10.63        9.56
70-74:        11.69       10.51        8.87        8.94
 75+ :         8.96        8.37        7.05        6.62
```

```
StatUnit.TestModelChange

2 parameters removed
-2log(Q) =               0.2526
P[ ChiSquare(2) > -2log(Q) ] =  0.881352
```

Figur 3.3 Hypotesen H_2 om at de tre kontrolbyer er ens: Udskrift fra *StatUnit*. Estimaterne er de samme som tidligere (side 42), og de forventede antal er *ikke* identiske med dem i Tabel 3.6, og $-2 \ln Q$ bliver nu 0.2526 mod tidligere 0.40.

4 Stikord