Kapitel 6

Poisson modeller

Meget af det som står i dette kapitel ligner overfladisk set gentagelser af resultater vedrørende polynomialfordelingsmodeller. Grunden til dette er, at man ved betingning med totalsummen i en model for uafhængige Poissonfordelte variable får (jvf. opgave 3.6.3 i Ssr.) en polynomialfordelingsmodel. Men fortolkningen af modellerne, deres parametriseringer og de relevante hypoteser ser lidt anderledes ud. Poissonfordelingsmodellerne er på mange måder simplere, fordi de opererer med uafhængige variable.

6.1. Én Poissonfordelt variabel.

Antag at observationen $x \in \mathbf{N}_0$ er fremkommet som værdien af en stokastisk variabel X, der er Poissonfordelt med ukendt parameter $\lambda > 0$. Likelihoodfunktionen for denne meget lille model er

$$L(\lambda) = \frac{1}{x!}\lambda^x e^{-\lambda} ,$$

og logaritmen til den er, idet vi ignorerer den konstante faktor $\frac{1}{x!}$,

$$l(\lambda) = x \log \lambda - \lambda .$$

Ved differentiation m.h.t. λ fås

$$l'(\lambda) = \frac{x}{\lambda} - 1 ,$$

og da denne funktion for $x > 0$ aftager fra $+\infty$ til -1 når λ gennemløber intervallet $]0, +\infty[$, følger det umiddelbart at ML–estimatoren er veldefineret og givet ved

$$\hat{\lambda} = x .$$

Opgave 6.1.1. Gør rede for at der også for $x = 0$ gælder $\hat{\lambda} = x$, når parameterområdet udvides til $[0, +\infty[$.

6.2. Test for simpel hypotese.

Antag nu, at vi i denne model ønsker at teste hypotesen $\lambda = \lambda_0$ for en given fast værdi $\lambda_0 > 0$. Kvotientteststørrelsen bliver

$$-2 \log q = 2(l(x) - l(\lambda_0)) = 2((x \log x - x) - (x \log \lambda_0 - \lambda_0))$$

Ved rækkeudvikling til og med 2. orden af log–likelihooden som funktion af x omkring punktet $x = \lambda_0$ (idet x midlertidigt opfattes som en kontinuert variabel) ser man, at der med god approksimation for x nær ved λ_0 gælder

$$-2\log q \approx \frac{(x-\lambda_0)^2}{\lambda_0} \; .$$

Højre side kaldes (igen) Pearson's teststørrelse. Den approksimative χ^2–fordeling af denne størrelse er nem at bevise. Poissonfordelingen er jo, ifølge den centrale grænseværdisætning og Poissonfordelingens foldningsegenskab, for store værdier af λ_0 godt approksimeret ved en normalfordeling med middelværdi λ_0 og varians λ_0. Pearsons teststørrelse ses derfor umiddelbart at være approksimativt χ^2–fordelt med 1 frihedsgrad, som kvadratet på den approksimativt normeret normalfordelte stokastiske variabel $(X-\lambda_0)/\sqrt{\lambda_0}$. Heraf følger igen — p.g.a. ovennævnte approksimation — at kvotientteststørrelsen er approksimativt χ^2–fordelt med 1 frihedsgrad.

For små værdier af λ_0 (typisk $\lambda_0 < 5$) bør testet udføres eksakt, ved at den (eller de) tilsvarende halesandsynlighed(er) i Poissonfordelingen beregnes eller findes ved tabelopslag.

Tests for simple hypoteser af formen $\lambda = \lambda_0$ i Poissonfordelingen er sjældent relevante i praksis. Som eksempel er vi nødt til at tage en lidt fiktiv situation:

Eksempel 6.1. (frit efter Inge Henningsen: Statistik, Institut for Matematisk Statistik, Københavns Universitet) Et eksempel i forbindelse med Poissonfordelingens udledning i noterne i sandsynlighedsregning handlede om en cyklist C, som punkterede i gennemsnit 3 gange pr. 1000 km. Man kan med en vis ret spørge sig selv, hvor hun ved det fra. I virkeligheden kan hun vel allerhøjst vide, hvor langt hun alt i alt har kørt, og hvor mange gange hun alt i alt er punkteret. Tallet 3 må derfor fortolkes som et estimat, som hun har beregnet ved at dividere det samlede antal punkteringer med den samlede distance hun har kørt, regnet i enheden 1000 km. Men hvis vi forestiller os, at hun er en overordentlig erfaren cyklist, vil dette estimat være ret nøjagtigt. Vi vil derfor tillade os fortolke værdien 3 bogstaveligt som den nøjagtige værdi af parameteren i den Poissonfordeling, der beskriver antallet af punkteringer på en hvilken som helst 1000 km strækning hun kører, forudsat at der ikke ændres på omstændigheder som kan have indflydelse på punkteringshyppigheden.

Det er imidlertid netop det, vi forestiller os at hun gør nu. For at afprøve en ny type dæk kører hun 1000 km med de nye dæk. På denne tur punkterer hun 9 gange. Hvad kan vi, på dette grundlag, sige om disse dæk?

Før vi udregner kvotienttest m.m. kan vi rent intuitivt argumentere således. Umiddelbart ser det jo ud til, at man punkterer oftere med de nye dæk. Spørgsmålet er om den øgede frekvens (9 punkteringer, kun 3 forventet) er signifikant. Sandsynligheden for at få udfaldet 9 eller mere i en Poissonfordeling med parameter 3 er

$$e^{-3}\left(\frac{1}{9!}3^9+\frac{1}{10!}3^{10}+\frac{1}{11!}3^{11}+\dots\right)$$

$$=0.00270+0.00081+0.000022+0.00006+0.00001+\cdots=0.00380\,.$$

Denne halesandsynlighed på ca. 0.4 % kan vi så vælge at gange med 2, fordi testet formelt bør foretages tosidet. Men konklusionen er under alle omstændigheder, at de nye dæk ser stærkt ud til at være dårligere end de gamle.

Kvotienttestet giver

$$-2\log q=2((9\log 9-9)-(9\log 3-3))=7.775$$

medens Pearsons teststørrelse bliver

$$\frac{(9-3)^2}{3}=12.$$

Den ret store forskel mellem disse to størrelser advarer os om, at approksimationen med en χ^2–fordeling ikke er god nok, og vi får da også lidt forskellige konklusioner: 7.775 ligger tæt på 99.5 % fraktilen i χ^2–fordelingen med 1 frihedsgrad, medens 12 er noget større end 99.9 % fraktilen. Men de tilsvarende halesandsynligheder 0.005 og 0.001 er trods alt af samme størrelsesorden som den eksakte P–værdi 0.0038.

6.3. Modeller for uafhængige Poissonfordelte variable.

Lad observationen $(x_1,\dots,x_k)$ være fremkommet som den observerede værdi af $(X_1,\dots,X_k)$, hvor $X_1,\dots,X_k$ er uafhængige, Poissonfordelte, med parametre $\lambda_1,\dots,\lambda_k$. Ved den *fulde model* forstår vi i denne forbindelse modellen givet ved $(\lambda_1,\dots,\lambda_k)\in]0,+\infty[^k$, altså modellen hvor hver observation har sin egen frie parameter. Ved en *glat model af dimension* $d_0<k$ forstår vi en model givet på formen $(\lambda_1,\dots,\lambda_k)\in\Theta_0$, hvor Θ_0 er en glat delmængde af $]0,+\infty[^k$ af dimension d_0. Vi har tidligere diskuteret begrebet glat delmængde, og vil ikke uddybe det yderligere her.

Maksimaliseringsestimaterne for Poisson parametrene under den fulde model bliver åbenbart (da likelihood'en på indlysende måde splitter op som et produkt af funktioner, der hver for sig er likelihood for en enkelt Poisson variabel)

$$\hat\lambda_j=x_j.$$

Lad $\hat\lambda_{0j}$, $j=1,\dots,k$, betegne ML–estimatorerne (som vi antager eksisterer) under en glat model af dimension $d_0<k$. Da gælder

SÆTNING 6.1. *Kvotientteststørrelsen for test mod den fulde model er*

$$-2\log q = 2\sum_{j=1}^{k}\left(x_j \log\frac{x_j}{\hat{\lambda}_{0j}} + \hat{\lambda}_{0j} - x_j\right).$$

Under hypotesen gælder, for store værdier af parametrene λ_j, at denne størrelse er approksimativt χ^2–fordelt med $k - d_0$ frihedsgrader. Endvidere gælder approksimationsformlen

$$-2\log q \approx \sum_{j=1}^{k}\frac{(x_j - \hat{\lambda}_{0j})^2}{\hat{\lambda}_{0j}}.$$

Højre side af den sidste formel kalder vi — som sædvanlig — *Pearsons teststørrelse.* For brugbarheden af χ^2–approksimationen gælder, som tommelfingerregel, at alle estimaterne $\hat{\lambda}_{0j}$ skal være ≥ 5, idet approksimationen dog kan bruges med passende forbehold selv når nogle af disse størrelser er helt nede i nærheden af 1.

Bemærkning. For langt de fleste af de modeller man betragter gælder, at

$$\sum_{j=1}^{k}\hat{\lambda}_{0j} = \sum_{j=1}^{k}x_j.$$

Dette gælder, mere præcist, for modeller bestemt ved delmængder Θ_0 af $]0,+\infty[^k$ som er invariante under multiplikation med positive skalarer, i den forstand at

$$(\lambda_1,\ldots,\lambda_k)\in\Theta_0, \beta>0 \implies (\beta\lambda_1,\ldots,\beta\lambda_k)\in\Theta_0.$$

I så fald har vi åbenbart den simplere formel

$$-2\log q = 2\sum_{j=1}^{k}x_j\log\frac{x_j}{\hat{\lambda}_{0j}}$$

for kvotientteststørrelsen.

I analogi med resultaterne for glatte polynomialfordelingsmodeller har vi også følgende resultat om successiv testning:

Betragt to glatte hypoteser, givet ved delmængder $\Theta_{00} \subset \Theta_0 \subset]0,+\infty[^k$ af dimensioner $d_{00} < d_0 < k$. Med indlysende betegnelser for ML–estimaterne under de to hypoteser gælder så

SÆTNING 6.2. *Kvotientteststørrelsen for test af hypotesen givet ved Θ_{00} mod hypotesen givet ved Θ_0 er*

$$-2\log q = 2\sum_{j=1}^{k}\left(x_j\log\frac{\hat{\lambda}_{0j}}{\hat{\lambda}_{00j}} + \hat{\lambda}_{00j} - \hat{\lambda}_{0j}\right).$$

Under hypotesen $(\lambda_1, \ldots, \lambda_k) \in \Theta_{00}$ *gælder, for store værdier af parametrene* λ_j*, at denne størrelse er approksimativt* χ^2*–fordelt med* $d_0 - d_{00}$ *frihedsgrader. Endvidere gælder approksimationsformlen*

$$-2 \log q \approx \sum_{j=1}^{k} \frac{(\hat{\lambda}_{0j} - \hat{\lambda}_{00j})^2}{\hat{\lambda}_{00j}}.$$

Højre side af den sidste formel kalder vi — endnu engang — Pearsons teststørrelse. Hovedreglen for brugbarhed af approksimationen er at alle estimaterne $\hat{\lambda}_{00j}$ skal være ≥ 5.

Bemærk, at hvis

$$\sum_{j=1}^{k} \hat{\lambda}_{00j} = \sum_{j=1}^{k} \hat{\lambda}_{0j} = x.$$

(dvs. hvis begge de betragtede modeller er invariante under multiplikation med positive skalarer) så har vi den simplere formel

$$-2 \log q = 2 \sum_{j=1}^{k} x_j \log \frac{\hat{\lambda}_{0j}}{\hat{\lambda}_{00j}}$$

for kvotientteststørrelsen.

6.4. Test for homogenitet.

Lad $x_1, \ldots, x_k$ være observerede værdier af k uafhængige Poissonfordelte variable $X_1, \ldots, X_k$ med parametre $\lambda_1, \ldots, \lambda_k$. Betragt *homogenitetshypotesen*

$$\lambda_1 = \lambda_2 = \cdots = \lambda_k = \lambda.$$

Under denne hypotese ser likelihooden sådan ud:

$$L(\lambda) = \frac{1}{x_1!}\lambda^{x_1} e^{-\lambda} \ldots \frac{1}{x_k!}\lambda^{x_k} e^{-\lambda} = \text{const.} \times \lambda^{x\cdot} e^{-k\lambda}$$

så log likelihooden bliver (idet vi ignorerer konstantleddet)

$$l(\lambda) = x. \log \lambda - k\lambda = k(\bar{x} \log \lambda - \lambda) \ ,$$

hvor $\bar{x} = x./k$ betegner gennemsnittet af de k observationer. Da denne funktion, på nær faktoren k, har præcis samme form som likelihooden for en enkelt Poissonfordelt variabel, får vi umiddelbart at ML–estimatoren er veldefineret for $x. > 0$ og givet ved

$$\hat{\lambda} = \bar{x}\,.$$

Kvotientteststørrelsen for hypotesen bliver således

$$\begin{aligned}-2\log q &= 2((x_1\log x_1 - x_1) + \cdots + (x_k\log x_k - x_k) - (x.\log\bar{x} - k\bar{x}))\\ &= 2(x_1\log x_1 + \cdots + x_k\log x_k - x. - x.(\log x. - \log k) + x.)\\ &= 2(x_1\log x_1 + \cdots + x_k\log x_k - x.\log x. + x.\log k)\,.\end{aligned}$$

Dette er — bortset fra lidt ændring af notation — præcis samme størrelse, som vi i kapitel 4 (side 26) udledte som kvotientteststørrelsen for hypotesen om homogenitet i en polynomialfordeling. Da antallet af frihedsgrader også er det samme $(k-1)$ skal teststørrelsen vurderes i samme fordeling, så konklusionen af testet bliver den samme.

Den matematiske forklaring på denne ækvivalens er, at hvis man i Poissonmodellen betinger med hændelsen $X. = x.$, så får man netop en polynomialfordelingsmodel, jvf. Ssr. opgave 3.6.3, hvor sættet $(X_1, \ldots, X_k)$ beskrives som polynomialfordelt af orden k med antalsparameter $n = x.$ og sandsynlighedsparametre $p_j = \lambda_j/\lambda.$. Og hypotesen om homogenitet i Poissonmodellen svarer netop til hypotesen om homogenitet i polynomialfordelingsmodellen.

Eksempel 6.2. Betragt tabellen på side 26 i kapitel 4, der angiver antal levendefødte børn i 1988 fordelt på fødselsmåned. Vi analyserede disse tal ved hjælp af en polynomialmodel, hvor det samlede antal 58844 af fødsler i 1988 blev opfattet som fast. Det kan forekomme lidt unaturligt. En lidt mere rimelig model går ud på at opfatte de 12 antal som uafhængige, binomialfordelte, med en fælles antalsparameter der kan fortolkes som noget i retning af det samlede antal kvinder i den fødedygtige alder. Hver af de 12 sandsynlighedsparametre skal så fortolkes som sandsynligheden for, at en bestemt kvinde blandt disse netop føder i den tilsvarende måned. Da en binomialfordeling med meget stor antalsparameter og lille sandsynlighedsparameter er godt approksimeret ved en Poissonfordeling, fører dette umiddelbart til en model, hvor antallene af fødsler pr. måned er uafhængige og Poissonfordelte. Konklusionen af ovenstående er, at testet for homogenitet i de to modeller fører til præcis det samme.

For $k = 2$ er testet for $\lambda_1 = \lambda_2$ specielt sammenfaldende med testet for $p = \frac{1}{2}$ i binomialfordelingen. Her udfører man ofte testet som eksakt binomialtest, svarende til at man betinger med (den for hypotesen ret irrelevante størrelse) $x. = x_1 + x_2$. Formelt går testet altså ud på at udregne sandsynlighederne

$$P(X_1 \geq x_1) \text{ og } P(X_1 \leq x_1)$$

under antagelse af at X_1 er binomialfordelt med antalsparameter $x_1 + x_2$ og sandsynlighedsparameter $\frac{1}{2}$. Det mindste af disse to tal, ganget med 2 (af hensyn til "tosidigheden"), er testets P–værdi.

Eksempel 6.1 fortsat. Antag nu (selvom det sikkert gør ondt) om vores cyklist C, at hun, i modsætning til hvad vi har troet indtil nu, overhovedet ikke er nogen særligt erfaren cyklist. Hendes påstand om at hun “plejer” at punktere 3 gange pr. 1000 km er faktisk baseret på, at hun sidste sommer kørte 1000 km og netop punkterede 3 gange.

Vurderingen af den nye dæktype — hvor vi stadig antager, at hun har oplevet 9 punkteringer med de nye dæk på en 1000 km strækning — bliver nu væsentlig ændret. Vi skal nu teste $\lambda_1 = \lambda_2$ i en model for to uafhængige, Poissonfordelte variable, hvor $x_1 = 3$ og $x_2 = 9$ er observeret. Vi kan direkte argumentere for det eksakte binomialtest på følgende måde. Hvis de to dæktyper var lige gode, ville vi forvente at hendes i alt 12 punkteringer var tilfældigt placeret på de to strækninger efter en almindelig binomialfordeling med antalsparameter 12 og sandsynlighedsparameter $\frac{1}{2}$. Sandsynligheden for, at en binomialfordelt variabel med disse parametre er ≥ 9, kan udregnes til 7.3 %. Det er jo ikke særligt ekstremt, og den sandsynlighed skal oven i købet ganges med 2 hvis vi forestiller os et test mod tosidet alternativ, som også tager i betragtning at de nye dæk kunne være *bedre* end de gamle.

Forklaringen på, at konklusionen i så høj grad har ændret sig i forhold til hvad vi tidligere nåede frem til, er naturligvis at angivelsen “3 punkteringer pr. 1000 km” — som vi før tog helt bogstaveligt som 3.0000 punkteringer i gennemsnit pr. 1000 km — nu er blevet forsynet med en meget stor usikkerhed.

6.5. Proportionalitet med en given baggrundsvariabel.

I mange anvendelser af Poissonmodeller er observationerne antal af hændelser, som er indtruffet i grupper af forskellige størrelser eller i tidsintervaller af forskellige længder. I så fald er den naturlige hypotese om “homogenitet” ikke, at de Poissonfordelinger der beskriver antallene har samme parameter, men at parametrene følger gruppernes størrelser eller tidsintervallernes længder proportionalt.

Stiller man sig for eksempel ved en landevej og tæller, hvor mange biler der kommer forbi i forskellige tidsintervaller, så er den naturlige hypotese om “konstant trafikintensitet” givet ved, at disse antal har middelværdier som er proportionale med længderne af de tilsvarende tidsintervaller. En model der beskriver disse antal som Poissonfordelte med *samme* parameter er kun relevant, hvis tidsintervallerne er lige lange.

Et andet eksempel har vi i de netop omtalte antal børnefødsler per måned i 1988. I kapitel 4 (side 27) stillede vi en opgave, som gik ud på at opstille en model der tog hensyn til at årets 12 måneder ikke er lige lange. I den tilsvarende Poisson model ville det være naturligt at tage højde for dette ved at opstille en model, hvor de 12 Poisson parametre er proportionale med månedernes længder.

EKSEMPEL 6.3. (fra Inge Henningsen: Statistik, Institut for Matematisk Statistik, Københavns Universitet). Følgende tabel angiver antallet af tilfælde af testikelkræft i perioden 1974–78 i fire Nordsjællandske kommuner, samt det gennemsnitlige antal mænd i disse kommuner i samme periode.

Kommune	Tilfælde af testikelkræft	Antal mænd
Birkerød	4	10900
Frederikssund	12	7336
Helsingør	9	27671
Hillerød	9	15499
I alt	34	61406

Den helt rigtige model ville her være en binomialfordelingsmodel, der for eksempel beskriver antal tilfælde i Birkerød Kommune som den observerede værdi af en binomialfordelt variabel med antalsparameter 10900 og en sandsynlighedsparameter p_1, der kan fortolkes som sandsynligheden for at blive ramt af testikelkræft i en fire års periode for en mand der bor i Birkerød. Den naturlige hypotese i denne model ville så være homogenitetshypotesen $p_1 = \cdots = p_4$, som siger at de fire kommuner ikke adskiller sig fra hinanden på dette punkt. Men da sandsynlighedsparametrene er meget små og antalsparametrene meget store kan vi uden videre benytte approksimationen af binomialfordelingen med en Poissonfordeling. Den tilsvarende Poissonmodel siger så, at de fire antal 4, 12, 9 og 9 er fremkommet som værdier af uafhængige Poissonfordelte variable. I første omgang (den fulde model) lader vi parametrene $\lambda_1, \ldots, \lambda_4$ for disse fordelinger variere frit. Den hypotese, som svarer til hypotesen $p_1 = \cdots = p_4$ ovenfor, bliver så (jvf. sammenhængen $\lambda = np$ mellem parametrene i en binomialfordeling og parameteren i den approksimerende Poissonfordeling)

$$\frac{\lambda_1}{10900} = \frac{\lambda_2}{7336} = \frac{\lambda_3}{27671} = \frac{\lambda_4}{15499}$$

eller

$$\lambda_i = \beta n_i\,, \quad i = 1, 2, 3, 4,$$

hvor $n_1, \ldots, n_4$ betegner tallene i tabellens sidste søjle og parameteren β kan fortolkes som det forventede antal tilfælde af testikelkræft pr. mandlig indbygger over en periode på fire år.

Log–likelihooden under denne hypotese bliver åbenbart

$$4\log(10900\beta) - 10900\beta \;+\; 12\log(7336\beta) - 7336\beta$$

$$+9\log(27671\beta) - 27671\beta \;+\; 9\log(15499\beta) - 15499\beta$$

$$= \text{const.} + 34 \log \beta - 61406\beta = \text{const.} + 34 \log(61406\beta) - 61406\beta.$$

Som funktion af 61406β har denne funktion samme form som log likelihooden i modellen for en enkelt Poisson variabel, hvor $x = 34$ er observeret. Heraf følger umiddelbart at

$$\hat{\beta} = \frac{34}{61406} = 0.0005537$$

hvilket forekommer helt rimeligt efter ovennævnte fortolkning af β. Estimaterne for de fire Poisson parametre under hypotesen — også kaldet de fittede værdier — er således

$$\begin{array}{rcrcr}
\hat{\lambda}_1 & = & 10900\frac{34}{61406} & = & 6.035 \\
\hat{\lambda}_2 & = & 7336\frac{34}{61406} & = & 4.062 \\
\hat{\lambda}_3 & = & 27671\frac{34}{61406} & = & 15.321 \\
\hat{\lambda}_4 & = & 15499\frac{34}{61406} & = & 8.582
\end{array}$$

Da modellen åbenbart er invariant under multiplikation med positive skalarer bliver kvotientteststørrelsen for test af proportionalitetshypotesen mod den fulde model

$$-2 \log q = 2\left(4 \log \frac{4}{6.035} + \cdots + 9 \log \frac{9}{8.582}\right) = 13.99$$

Da dimensionen af den fulde model er 4, medens dimensionen af proportionalitetsmodellen åbenbart er 1, skal denne størrelse vurderes i en χ^2–fordeling med 3 frihedsgrader. Da 99.5 % procent fraktilen i denne fordeling er 12.838 må hypotesen om proportionalitet forkastes.

Pearson's test giver

$$\frac{(4 - 6.035)^2}{6.035} + \cdots + \frac{(9 - 8.582)^2}{8.582} = 18.83.$$

Her er konklusionen endnu mere overbevisende. Selv om den store forskel fra kvotienttestet antyder, at approksimation med en χ^2–fordeling er lidt tvivlsom, kan man godt tillade sig at konkludere, at proportionalitetshypotesen må forkastes med en P–værdi af størrelsesorden 1 %.

Denne konklusion svækkes imidlertid lidt af forhistorien, som vi ikke har fortalt endnu. Baggrunden for undersøgelsen var en mistanke om, at forekomsten af testikelkræft i Frederikssund var usædvanligt høj. Man foretog så en nøjagtig registrering i Frederikssund og tre sammenlignelige Nordsjællandske kommuner. Ser vi på antal tilfælde af testikelkræft pr.

10000 mandlige indbyggere,

Kommune	Tilfælde af testikelkræft pr. 10000
Birkerød	3.67
Frederikssund	16.36
Helsingør	3.25
Hillerød	5.81
Samlet	5.54

er det da også tydeligt, at det er Frederikssund der skiller sig ud. Men man må være opmærksom på, at når man på denne måde udvælger Frederikssund Kommune *fordi* den har så usædvanligt mange tilfælde af testikelkræft, så lyder man ikke særligt troværdig når man bagefter forbløffet udbryder "det var da utroligt så mange tilfælde af testikelkræft der er i Frederikssund". En af landets mange kommuner må jo nødvendigvis have rekorden, og en sammenligning af denne med nogle få gennemsnitskommuner vil med stor sandsynlighed føre til formel signifikans, også selv om proportionalitetsmodellen i virkeligheden er god nok. Hvis det er rigtigt, at Frederikssund er udvalgt på denne måde, skal der en P–værdi af størrelsesorden 0.0001 til, før man for alvor kan konkludere noget (hvorfor?).

Opgave 6.5.1. For børnefødselstallene (kapitel 4, side 26), opstil en model hvor antal fødsler i årets 12 måneder er uafhængige, Poissonfordelte med parametre som er proportionale med månedernes længder, og test denne model imod den fulde model. Vis, at betingning i denne model med totalsummen ($x. = 58844$) netop fører til den polynomialfordelingsmodel, der er beskrevet i opgave 4.3.1 på side 27, samt at testet for denne model imod den fulde model giver samme resultat.

Opgave 6.5.2. For data i eksempel 6.3. (testikelkræfttilfælde i fire kommuner), formuler og test en hypotese, der svarer til at det kun er Frederikssund som skiller sig ud, medens risikoen for testikelkræft er den samme i de tre andre kommuner.

Hvis denne hypotese bliver godkendt (og det bliver den faktisk), kan man forsøge at teste videre herfra til hypotesen om at risikoen er den samme i alle fire byer. Hvad bliver konklusionen nu?

6.6. Multiplikative Poissonmodeller.

Det simpleste eksempel på en multiplikativ Poissonmodel — og det eneste vi vil se på her — er følgende. Lad der være givet en $R \times S$–antalstabel

(x_{rs}) af hele ikke–negative tal. Antag at disse er fremkommet som værdier af uafhængige Poissonfordelte variable (X_{rs}). Lad λ_{rs} betegne parameteren for X_{rs}'s fordeling. Ved hypotesen om *multiplikativitet* forstås i denne forbindelse hypotesen

$$\lambda_{rs} = \alpha_r \beta_s$$

om at parameteren for den rs'te observation kan skrives som produkt af en rækkeparameter α_r og en søjleparameter β_s.

Da middelværdien af en Poissonfordelt variabel er det samme som fordelingens parameter kan hypotesen om multiplikativitet fortolkes sådan, at tallene i tabellen på nær stokastisk variation omkring deres middelværdi opfylder en proportionalitetsbetingelse, som man efter behag kan udtrykke ved at sige at rækkerne er proportionale eller at søjlerne er proportionale. I denne henseende ligner multiplikativitetshypotesen til forveksling hypotesen om uafhængighed i en polynomialfordelingsmodel. Det gælder da også, som man let kan overbevise sig om, at i den polynomialfordelingsmodel man får ved at betinge med totalsummen $x_{..}$ er der uafhængighed hvis og kun hvis Poissonmodellens parametre opfylder multiplikativitetsbetingelsen. Desuden er kvotienttestet for multiplikativitet i Poissonmodellen identisk (beregningsmæssigt) med kvotienttestet for uafhængighed i polynomialfordelingsmodellen, som vi skal se i det følgende.

Antallet af parametre i den multiplikative model er umiddelbart $R + S$, idet der jo er R α'er og S β'er. Men modellen er overparametriseret, idet forskellige sæt af α'er og β'er kan give anledning til de samme Poisson parametre $\lambda_{rs} = \alpha_r \beta_s$. Som man umiddelbart kan se vil multiplikation af alle α'erne med en positiv konstant og division af alle β'erne med den samme konstant ikke ændre på λ'erne. Det betyder, at der i realiteten er en parameter for meget i brug. Modellens virkelige dimension er derfor $R + S - 1$.

Hvis man har brug for det kan en entydig parametrisering opnås ved at sætte en af parametrene — f.eks. α_1 — til 1. Det er forholdsvis let at indse, at man herved får en injektiv parametrisering, og heraf følger også, at der ikke er andre "skjulte overparametriseringer" end den vi lige har omtalt.

Men det er ikke særlig naturligt at indføre betingelser som $\alpha_1 = 1$, og det er faktisk heller ikke altid nødvendigt. Likelihoodfunktionen afhænger nemlig kun af parametrene α_r og β_s gennem Poissonparametrene $\lambda_{rs} = \alpha_r \beta_s$. Så hvis vi blot kan finde et sæt af λ'er der kan skrives på denne form, og som giver likelihooden sin maksimale værdi under multiplikativitetshypotesen, så har vi både løst estimationsproblemet (i den forstand at vi har fundet estimaterne for selve Poissonparametrene under hypotesen) og testproblemet (i den forstand at vi kan udregne den maksimale værdi af likelihooden under hypotesen, og dermed udføre kvotienttestet mod den fulde model).

Likelihooden under hypotesen er (med udeladelse af de konstante faktorer $\frac{1}{x_{rs}!}$)

$$L(\alpha_1, \ldots, \alpha_R, \beta_1, \ldots, \beta_S) = \prod_{r=1}^{R} \prod_{s=1}^{S} e^{-\alpha_r \beta_s} (\alpha_r \beta_s)^{x_{rs}}$$

og log–likelihooden

$$l(\alpha_1, \ldots, \alpha_R, \beta_1, \ldots, \beta_S) = \sum_{r=1}^{R} \sum_{s=1}^{S} (-\alpha_r \beta_s + x_{rs}(\log \alpha_r + \log \beta_s))$$

$$= -\alpha.\beta. + \sum_{r=1}^{R} x_{r\cdot} \log \alpha_r + \sum_{s=1}^{S} x_{\cdot s} \log \beta_s$$

Vi opstiller likelihoodligningerne (altså de ligninger der sætter log–likelihoodens partielle afledede lig med 0):

$$-\beta. + \frac{x_{r\cdot}}{\alpha_r} = 0 \quad , \qquad -\alpha. + \frac{x_{\cdot s}}{\beta_s} = 0 \; .$$

Man ser umiddelbart, at enhver løsning til likelihoodligningerne nødvendigvis må opfylde

$$\alpha_r = \frac{x_{r\cdot}}{\beta.} \quad , \qquad \beta_s = \frac{x_{\cdot s}}{\alpha.} \; .$$

Ved summation af udtrykket for α_r over r (eller summation af udtrykket for β_s over s) fås

$$\alpha.\beta. = x.. \, ,$$

og for selve Poissonparametrenes vedkommende betyder dette at der må gælde

$$\lambda_{rs} = \alpha_r \beta_s = \frac{x_{r\cdot}}{\beta.} \frac{x_{\cdot s}}{\alpha.} = \frac{x_{r\cdot} x_{\cdot s}}{\beta.\alpha.} = \frac{x_{r\cdot} x_{\cdot s}}{x..} .$$

Hermed har vi, uden at tage stilling til præcis hvordan α'er og β'er skal normeres, vist at enhver løsning af likelihoodligningerne fører til estimatorerne

$$\hat{\lambda}_{rs} = \frac{x_{r\cdot} x_{\cdot s}}{x..}$$

for selve Poisson parametrene. Vi mangler at gennemføre en egentlig funktionsundersøgelse, som godtgør at løsningen svarer til et maksimum for likelihooden, men det vil vi springe over i denne omgang.

Kvotientteststørrelsen for test af multiplikativitet imod den fulde model bliver nu (idet $\hat{\lambda}_{\cdot\cdot} = x_{\cdot\cdot}$)

$$-2\log q = 2\sum\sum x_{rs} \log \frac{x_{rs}}{x_{r\cdot}x_{\cdot s}/x_{\cdot\cdot}}$$

$$= 2\left(\sum_{r=1}^{R}\sum_{s=1}^{S} x_{rs}\log x_{rs} - \sum_{r=1}^{R} x_{r\cdot}\log x_{r\cdot} - \sum_{s=1}^{S} x_{\cdot s}\log x_{\cdot s} + x_{\cdot\cdot}\log x_{\cdot\cdot}\right).$$

Dette er præcis samme teststørrelse, som vi i kapitel 5 udledte som kvotientteststørrelsen ved test for uafhængighed i en tosidet antalstabel. Den approksimerende χ^2–fordeling er også den samme, idet dimensionsfaldet fra den fulde model til den multiplikative model er

$$RS - (R + S - 1) = (R - 1)(S - 1).$$

Pearson's teststørrelse bliver i øvrigt også den samme som i polynomialfordelingsmodellen.

EKSEMPEL 6.4. Følgende tabel angiver antallet af konkurser i 1990, opdelt på landsdel og kvartal (kilde: Eksamensopgave i Statistik 0 vinter 93–94, Københavns Universitet):

	København	Øerne	Jylland	I alt
1. kvartal	244	155	321	720
2. kvartal	231	133	269	633
3. kvartal	204	121	278	603
4. kvartal	256	140	300	696
I alt	935	549	1168	2652

En naturlig model går her ud på at opfatte tallene i tabellen som observerede værdier af uafhængige, Poissonfordelte variable. I en given landsdel og for et givet kvartal har vi jo, ved kvartalets begyndelse, et stort antal eksisterende virksomheder, der hver for sig vil gå konkurs i løbet af det kommende kvartal med en ret lille sandsynlighed. Så argumentet for valg af Poissonfordelingen er igen, at der er tale om binomialfordelinger med store antalsparametre og små sandsynlighedsparametre.

Kvotientteststørrelsen for multiplikativitet bliver 3.071, den tilsvarende Pearson teststørrelse er 3.075. Begge ligger et godt stykke under 95%–fraktilen (12.6) i χ^2–fordelingen med 6 frihedsgrader, så hypotesen om multiplikativitet godkendes. Det betyder at variationen fra kvartal til kvartal (hvis der overhovedet er nogen) kan antages at være den samme i de tre regioner.

Man kan dernæst spørge om der er en forskel fra kvartal til kvartal. Det svarer til i den multiplikative model

$$\lambda_{kr} = \alpha_k \beta_r$$

($k =$ kvartal, $r =$ region) at opstille hypotesen

$$\alpha_1 = \cdots = \alpha_4 = 1 \quad \text{eller} \quad \lambda_{kr} = \beta_r \ .$$

Denne model og det tilhørende test — som vi ikke har omtalt i forbindelse med Poissonmodellerne — svarer helt til test for rækkehomogenitet, som omtalt i kapitel 5 (side 42), og det udføres på samme måde. Vi får

$$\begin{aligned}-2\log q = 2(&720\log 720 + \cdots + 696\log 696 \\ &-2652\log 2652 + 2652\log 4) = 13.36.\end{aligned}$$

Da antallet af parametre i den multiplikative model var 6, medens antal parametre i den reducerede model åbenbart er 3, skal denne teststørrelse vurderes i en χ^2–fordeling med 3 frihedsgrader. Da $13.36 > 12.838 = 99.5$ % fraktilen, må hypotesen om homogenitet afvises. Der er altså forskel på intensiteten af konkurser fra kvartal til kvartal. Åbenbart er de to vinterkvartaler værst, hvad dette angår. Hvilket vel også er hvad man på forhånd ville vente, i betragtning af at næsten alt i forretningslivet går langsommere i sommertiden på grund af ferie.

Opgave 6.6.1. I kapitel 5 (side 39–40) definerede vi de normerede residualer under uafhængighedsmodellen ved

$$u_{rs} = \frac{x_{rs} - \frac{x_{r\cdot}x_{\cdot s}}{x_{\cdot\cdot}}}{\sqrt{\frac{x_{r\cdot}x_{\cdot s}}{x_{\cdot\cdot}}}}$$

med nogle forblommede tilføjelser om, at dette var residualerne divideret med noget i retning af deres estimerede standardafvigelser. Gør rede for, at det er præcis hvad de er, når polynomialfordelingsmodellen med antagelse af uafhængighed erstattes med den multiplikative Poissonmodel.