Kapitel 4

Estimation og testning i en polynomialfordeling

Vi betragter følgende model: $x = (x_1, \ldots, x_k)$ antages at være den observerede værdi af en stokastisk variabel $X = (X_1, \ldots, X_k)$, som er polynomialfordelt af orden k med (kendt) antalsparameter n og ukendte sandsynlighedsparametre $(p_1, \ldots, p_k)$.
Observationsrummet betegnes

$$D(n,k) = \{(x_1, \ldots, x_k) \in \mathbf{N}_0^k \mid x_1 + \cdots + x_k = n\},$$

og parameterrummet

$$\Delta_k = \{(p_1, \ldots, p_k) \in \mathbf{R}_+^k \mid p_1 + \cdots + p_k = 1\}.$$

Dette kapitel handler om, hvordan man i denne model estimerer sandsynlighedsparametrene og tester hypoteser af formen

$$(p_1, \ldots, p_k) \in \Theta_0$$

for passende delmængder $\Theta_0 \subset \Delta_k$.

4.1. Estimation.

Likelihoodfunktionen bliver åbenbart

$$L(p_1, \ldots, p_k) = \binom{n}{x_1 \ldots x_k} p_1^{x_1} \ldots p_k^{x_k},$$

og log–likelihood'en (idet logaritmen til polynomialkoefficienten ignoreres)

$$l(p_1, \ldots, p_k) = x_1 \log p_1 + \cdots + x_k \log p_k$$

På grund af parameterrummets specielle udseende kan vi ikke maksimere denne funktion bare ved at differentiere efter hver af de k parametre og sætte resultatet lig med 0. Da parameterrummet Δ_k jo er defineret som mængden af sæt $(p_1, \ldots, p_k)$ af positive tal, der opfylder betingelsen

$$p_1 + \cdots + p_k = 1,$$

er der tale om et maksimeringsproblem *under bibetingelser,* som kan løses ved hjælp af *Lagrange multiplikatorer* på følgende måde. Betragt ligningssystemet

$$\frac{d}{dp_j}\left(l(p_1, \ldots, p_k) - \lambda(p_1 + \cdots + p_k)\right) = 0, \; j = 1, \ldots, k,$$

$$p_1 + \cdots + p_k = 1.$$

Hvis vi kan løse disse $k+1$ ligninger med hensyn til de $k+1$ ubekendte $p_1, \ldots, p_k$ og λ, får vi et sæt parametre der med lidt held maksimerer funktionen $l(p_1, \ldots, p_k) - \lambda(p_1 + \cdots + p_k)$ på R_+^k som funktion af $(p_1, \ldots, p_k)$. Hvis dette faktisk er tilfældet (det skal naturligvis undersøges), vil dette sæt af parametre også maksimere funktionen under bibetingelsen. Og da sidste led $-\lambda(p_1 + \cdots + p_k)$ er konstant under bibetingelsen, følger det, at dette sæt af parametre også maksimerer første led $l(p_1, \ldots, p_k)$ under bibetingelsen.

Ligningssystemet bliver i dette tilfælde

$$\frac{x_j}{p_j} - \lambda = 0,\ j = 1, \ldots, k,$$

$$p_1 + \cdots + p_k = 1.$$

De k første ligninger kan også skrives

$$x_j = p_j \lambda,\ j = 1, \ldots, k,$$

og ved at addere disse ligninger og indsætte bibetingelsen får vi, at der må gælde $\lambda = x_1 + \cdots + x_k = n$. Herefter bliver de k første ligninger til

$$x_j = p_j n,\ j = 1, \ldots, k,$$

med løsning

$$p_j = \frac{x_j}{n}.$$

Og det er ikke svært at indse, at funktionen

$$x_1 \log p_1 + \cdots + x_k \log p_k - n(p_1 + \cdots + p_k)$$

som funktion af $(p_1, \ldots, p_k) \in \mathbf{R}_+^k$ faktisk har maksimum i dette punkt, idet den kan skrives som sum af k funktioner $x_j \log p_j - np_j$ af hver sit argument, som er lette at undersøge. Heraf følger, at den også som funktion på Δ_k har maksimum i dette punkt, dvs. at maksimum–likelihood estimatoren i polynomialfordelingen er

$$\hat{p}_j = \frac{x_j}{n},$$

hvilket må siges at være intuitivt rimeligt når man tænker på polynomialfordelingens definition og fortolkningen af parametrene $(p_1, \ldots, p_k)$.

Man kan i øvrigt også indse dette uden brug af Lagrange multiplikatorer. I stedet for at parametrisere ved $(p_1, \ldots, p_k)$ kan vi erstatte $p_2, \ldots, p_k$ med

$$\pi_2 = \frac{p_2}{p_2 + \cdots + p_k},$$

$$\vdots$$
$$\pi_k = \frac{p_k}{p_2 + \cdots + p_k},$$

De nye parametre $\pi_2, \ldots, \pi_k$ kan fortolkes som betingede sandsynligheder for, at det enkelte udfald (jvf. polynomialfordelingens fortolkning som fordelingen af et sæt af antal) falder i kategorierne $2, \ldots, k$, givet at det falder i en af disse. Det er let at se, at vi herved får en ny parametrisering af den samme model ved parametrene $p_1 \in]0, 1[$ og $(\pi_2, \ldots, \pi_k) \in \Delta_{k-1}$. Likelihooden bliver, som funktion af de nye parametre (med udeladelse af polynomialkoefficienten)

$$p_1^{x_1}(p_2 + \cdots + p_k)^{x_2+\cdots+x_k} \pi_2^{x_2} \ldots \pi_k^{x_k}$$
$$= p_1^{x_1}(1 - p_1)^{n-x_1} \pi_2^{x_2} \ldots \pi_k^{x_k}$$

Hermed har vi fået omskrevet likelihoodfunktionen til et produkt af

$$p_1^{x_1}(1 - p_1)^{n-x_1},$$

der som funktion af p_1 er en almindelig binomial–likelihood, og

$$\pi_2^{x_2} \ldots \pi_k^{x_k}$$

som er en polynomial likelihood af orden $k - 1$. Da disse to led kan maksimeres hver for sig, har vi hermed, ved at reducere problemet til et problem af samme slags af orden $k - 1$ i stedet for k, lagt op til et induktionsargument, der umiddelbart fører til det påståede resultat.

Hvis man vil angive en usikkerhed på det enkelte estimat $\hat{p}_j$, kan man benytte at dette som stokastisk variabel $\hat{p}_j = \frac{X_j}{n}$ har varians $\frac{1}{n}p_j(1-p_j)$, idet X_j jo er binomialfordelt (n, p_j). Man kan så indsætte $\hat{p}_j$ for p_j i dette udtryk, og i øvrigt gå frem som i kapitel 2 (side 7).

4.2. Test for dimensionsreducerende hypotese.

Ved en *dimensionsreducerende hypotese* forstår vi en hypotese af formen

$$(p_1, \ldots, p_k) \in \Theta_0,$$

hvor Θ_0 er en delmængde af Δ_k af dimension $d_0 < d = k - 1$, som opfylder passende "glathedsbetingelser". En fuldstændig matematisk præcisering af disse glathedsbetingelser må vi afstå fra her. Vi nøjes med, ved eksempler og løs snak, at forklare hvad vi forstår ved en glat delmængde af Δ_k og ved dimensionen af sådan en.

Vi bemærker, at det fulde parameterrum Δ_k, som åben delmængde af hyperplanen $p_1 + \cdots + p_k = 1$ i $\mathbf{R}^k$, naturligt tillægges dimensionen

$d = k - 1$. Den fulde polynomialfordelingsmodel kan jo parametriseres injektivt ved de $k-1$ parametre $p_1, \ldots, p_{k-1}$, som varierer i det åbne område givet ved $p_1, \ldots, p_{k-1} > 0$ og $p_1 + \cdots + p_{k-1} < 1$.

For $k = 2$ (hvor polynomialfordelingen blot er en lettere omfortolket binomialfordeling) er således $d = 2 - 1 = 1$, dvs. parameterrummet er af dimension 1, i overenstemmelse med at vi i dette tilfælde kan parametrisere med en enkelt parameter p_1 (da $p_2 = 1 - p_1$ er givet ved p_1). En delmængde af lavere dimension må således være af dimension 0, dvs. bestå af et enkelt punkt (eller to eller tre ... — men det ser vi bort fra her). Den teori, vi skal gennemgå i det følgende, reducerer i tilfældet $k = 2$ til det der er sagt i kapitel 2 om test for hypoteser af formen $p = p_0$ i binomialfordelingen.

For $k = 3$ er $\Delta_k \subset \mathbf{R}^3$ den åbne trekant med hjørner (1,0,0), (0,1,0) og (0,0,1) (heraf navnet "Δ"). En glat delmængde af dimension 1 vil typisk fremkomme som trekantens fællesmængde med en kurve i hyperplanen $p_1 + p_2 + p_3 = 1$.

Eksempel 4.1. I visse genetiske anvendelser ser man på polynomialfordelinger af orden 3, hvor sandsynlighedsparametrene er givet på formen

$$p_1 = \pi^2,$$

$$p_2 = 2\pi(1-\pi),$$

$$p_3 = (1-\pi)^2.$$

Det svarer til at antage $(p_1, p_2, p_3) \in \Theta_0$, hvor Θ_0 er den 1–dimensionale glatte delmængde (eller kurve)

$$\Theta_0 = \{(\pi^2, 2\pi(1-\pi), (1-\pi)^2) \mid \pi \in]0,1[\}$$

i Δ_3.

Tilsvarende kan man, i det generelle tilfælde, tænke på en glat d_0–dimensional delmængde af Δ_k som en mængde af formen

$$\Theta_0 = \{t(\pi_1, \ldots, \pi_{d_0}) \mid (\pi_1, \ldots, \pi_{d_0}) \in \Pi\},$$

hvor Π er en åben delmængde af $\mathbf{R}^{d_0}$ og t er en *injektiv, kontinuert differentiabel* afbildning $t : \Pi \to \Delta_k$. Dette kommer tæt på den korrekte definition. Der mangler blot en række besværlige antagelser til sikring af, at Θ_0 ikke opfører sig underligt, for eksempel ved at skære sin egen rand, oscillere vildt eller ende brat inde midt i Δ_k. Vi må indskrænke os til at bemærke, at når man i praksis, ved hjælp af almindelige regnearter og matematiske standardfunktioner, opskriver en hypotese om polynomialfordelingens sandsynlighedsparametre, enten

— som formler for $(p_1, \ldots, p_k)$, udtrykt ved hjælp af et mindre antal (typisk d_0) parametre, eller

— i form af et antal (typisk $d - d_0 = k - 1 - d_0$) bånd på parametrene $(p_1, \ldots, p_k)$,

så vil dette i almindelighed definere en glat delmængde Θ_0. *Dimensionen* d_0 af denne mængde er det *mindste* antal parametre, man kan parametrisere den med.

Hovedresultatet om test i polynomialfordelingen er nu følgende.

SÆTNING 4.1. *Lad Θ_0 være en glat delmængde af Δ_k af dimension $d_0 < d = k-1$. Lad $\hat{p}_j = \frac{x_j}{n}$, $j = 1, \ldots, k$, betegne maksimaliseringsestimatorerne for sandsynlighedsparametrene under den fulde model bestemt ved Δ_k, og lad tilsvarende $\hat{p}_{01}, \ldots, \hat{p}_{0k}$ betegne maksimaliseringsestimatorerne for sandsynlighedsparametrene i den reducerede model bestemt ved Θ_0. Så er, for $(p_1, \ldots, p_k) \in \Theta_0$,*

$$-2 \log Q = 2 \left(x_1 \log \frac{\hat{p}_1}{\hat{p}_{01}} + \cdots + x_k \log \frac{\hat{p}_k}{\hat{p}_{0k}} \right)$$

approksimativt χ^2–fordelt med $d - d_0$ frihedsgrader. Endvidere gælder approksimationsformlen

$$-2 \log Q \approx \sum_{j=1}^{k} \frac{\left(x_j - n\hat{p}_{0j}\right)^2}{n\hat{p}_{0j}}.$$

Højre side af den sidste relation kaldes Pearson's teststørrelse.

Vi vil ikke forsøge at bevise sætningen (vi har jo ikke engang formuleret den helt præcist). I kapitel 2 og 3 har vi delvis godtgjort resultatet i nogle simple specialtilfælde.

Den helt præcise formulering af sætningen er i form af et asymptotisk resultat, der siger at fordelingen af $-2 \log Q$ *konvergerer* i passende forstand imod en χ^2–fordeling for $n \to \infty$. Bemærk, at der i sætningen indgår en forudsætning om at maksimaliseringsestimatorne $\hat{p}_{01}, \ldots, \hat{p}_{0k}$ under hypotesen er veldefinerede. Det vil de i almindelighed være, i det mindste med meget stor sandsynlighed når n er stor, men i den helt korrekte formulering af sætningen indgår naturligvis også en præcisering af dette.

Approksimationen af de to teststørrelsers fordeling med en χ^2–fordeling anses — som en tommelfingerregel — for at være god nok til praktiske formål, når alle de estimerede forventede værdier $\mathrm{E}X_j = n\hat{p}_{0j}$ under hypotesen er ≥ 5. Ofte bruger man, i mangel af bedre, disse tests i situationer hvor nogle af disse størrelser er helt nede i nærheden af 1. I så fald skal man ikke tage χ^2–approksimationen alt for bogstaveligt, men størrelsesordenen af de således udregnede P–værdier kan man så nogenlunde stole på.

EKSEMPEL 4.2. Et helt banalt eksempel af nogenlunde samme relevans som møntkasteksemplet er følgende. En terning kastes 100 gange. Det resulterer i

11 ettere,
16 toere,
15 treere,
17 firere,
12 femmere og
29 seksere.

Opfører denne terning sig, som terninger skal?

Den oplagte model går ud på at (11,16,15,17,12,29) er en observation fra en polynomialfordeling af orden 6 med antalsparameter 100 og (i første omgang) ukendte sandsynlighedsparametre $p_1, \ldots, p_6$. ML–estimatorerne er

$$\hat{p}_1 = 0.11,\ \hat{p}_2 = 0.16,\ \hat{p}_3 = 0.15,\ \hat{p}_4 = 0.17,\ \hat{p}_5 = 0.12,\ \hat{p}_6 = 0.29.$$

Hypotesen om at "terningen opfører sig som den skal" skal naturligvis forstås som

$$p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6}.$$

Da der er tale om en "glat" hypotese $\Theta_0 = \{(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})\}$ (selvom et punkt ikke umiddelbart virker særligt glat) af dimension 0, kan vi teste den ved at udregne enten $-2\log Q$ eller Pearson's teststørrelse, og undersøge om denne størrelse er ekstremt stor i en χ^2–fordeling med 5 frihedsgrader. Her forsøger vi naturligvis begge dele:

$$-2\log Q = 2\left(11\log\frac{0.11}{1/6} + \cdots + 29\log\frac{0.29}{1/6}\right) = 11.31,$$

$$\text{“Pearson”} = \frac{(11 - 100/6)^2}{100/6} + \cdots + \frac{(29 - 100/6)^2}{100/6} = 12.56.$$

Begge disse størrelser ligger mellem 95%–fraktilen og 99% fraktilen i χ^2–fordelingen med 5 frihedsgrader. Der er altså ikke tale om nogen ekstrem grad af signifikans, og hvis det drejer sig om en ganske almindelig terning, som man i øvrigt ikke har nogen mistanke til, ville det være urimeligt at reagere på dette ved at kassere terningen, klage til producenten eller lignende.

4.3. Test for homogenitet.

Det test vi har udført i eksemplet ovenfor kan fortolkes som et specialtilfælde af følgende:

Lad $(x_1, \ldots, x_k)$ være en observation fra en polynomialfordeling af orden k med antalsparameter n og sandsynlighedsparametre $p_1, \ldots, p_k$. Undertiden (omend ikke særligt ofte) er man interesseret i at teste hypotesen om *homogenitet,* dvs. den hypotese at de k sandsynlighedparametre er ens,

$$p_1 = \cdots = p_k = \frac{1}{k}.$$

For $k = 2$ reducerer dette til testet for $p = \frac{1}{2}$ i binomialfordelingen, som vi har studeret så umådeligt grundigt i kapitel 2.

Da estimatorerne i den fulde model er de relative hyppigheder $\hat{p}_j = x_j/n$, og estimatorerne under homogenitetshypotesen naturligvis er $\hat{p}_{0j} = 1/k$, kan vi umiddelbart opskrive kvotientteststørrelsen:

$$Q = \frac{\left(\frac{1}{k}\right)^{x_1} \cdots \left(\frac{1}{k}\right)^{x_k}}{\left(\frac{x_1}{n}\right)^{x_1} \cdots \left(\frac{x_k}{n}\right)^{x_k}} ,$$

altså

$$-2 \log Q = 2 \left(x_1 \log x_1 + \cdots + x_k \log x_k - n \log n + n \log k\right) .$$

Kvotienttestet foretages således ved vurdering af denne størrelse i en χ^2–fordeling med $(k-1) - 0 = k - 1$ frihedsgrader.

Eksempel 4.3. (Kilde: Opgaver i Statistik, Inge Henningsen, Københavns Universitet, Institut for Matematisk Statistik 1996). I noterne i sandsynlighedsregning (side 7) udregnes sandsynligheden for, at der i en klasse med 25 børn ikke findes to med samme fødselsdag, til 0.43. Denne udregning forudsætter, at børns fødselsdage er jævnt fordelt over året. Følgende tabel angiver antal levendefødte børn i Danmark 1988 fordelt på måneder:

Januar	4500
Februar	4471
Marts	5201
April	5061
Maj	5169
Juni	4989
Juli	5160
August	5288
September	5158
Oktober	4716
November	4474
December	4657
I alt	58844

Hvis vi opfatter det samlede antal børnefødsler i 1988 som givet, er det naturligt her at opstille en polynomialfordelingsmodel, der beskriver observationen $(4500, \ldots, 4657)$ som polynomialfordelt af orden 12 med antalsparameter 58844. Kvotienttestet for homogenitet giver her

$$-2\log Q = 2(4500\log 4500 + \cdots + 4657\log 4657$$

$$-58844\log 58844 + 58844\log 12) = 225.98.$$

Da den fulde model har 11 (=12–1) parametre, og hypotesen har 0 parametre, skal denne størrelse vurderes i en χ^2–fordeling med 11 frihedsgrader. Da 99.99%–fraktilen i denne fordeling er 37.367, kan vi med meget stor overbevisning sige, at fødsler (i hvert fald i 1988) *ikke* fordeler sig jævnt over årets 12 måneder. Inhomogeniteten kan groft beskrives (som man ser ved hjælp af en simpel tegning) ved at der fødes flere børn i månederne marts–september end i oktober–februar. Så sandsynligheden 0.43 for "ingen dobbeltfødselsdage" er kun en approksimation. Det rigtige resultat bliver lidt mindre, fordi fødselsdagene har en svag tendens til at klumpe sammen i marts–september, hvorved sammenfald naturligvis bliver lidt hyppigere end ved en helt jævn fordeling.

Opgave 4.3.1. I eksempel 4.3 (børnefødslers fordeling over årets 12 måneder) er det ikke helt naturligt at teste for homogenitet i den simple forstand vi har defineret det her. Da årets 12 måneder ikke er lige lange, ville det være mere rimeligt at teste hypotesen

$$p_1 = \frac{31}{366},\ p_2 = \frac{29}{366},\ \ldots,\ p_{12} = \frac{31}{366}$$

som (idet vi husker at 1988 var et skudår) svarer til at fødslerne fordeler sig ligeligt over årets 366 *dage*. Udfør dette test.

4.4. Successiv testning.

Ofte interesserer man sig i en statistisk analyse for flere hypoteser. Typisk begynder man med en stor "grundmodel" — i tilfælde af polynomialfordelingen sædvanligvis den fulde model med parameterrum Δ_k — som man så simplificerer trin for trin, indtil man ikke kan komme videre.

For at forstå de væsentligste begreber i denne forbindelse er det nok at se på situationen, hvor vi har to hypoteser, svarende til glatte delmængder

$$\Theta_{00} \subset \Theta_0 \subset \Delta_k$$

af dimensioner henholdsvis d_{00} og d_0, hvor

$$d_{00} < d_0 < d = k - 1.$$

Antag, at vi ved vurdering af kvotientteststørrelsen

$$Q_0 = \frac{L(\hat{p}_{01}, \ldots, \hat{p}_{0k})}{L(\hat{p}_1, \ldots, \hat{p}_k)}$$

har godkendt hypotesen $(p_1, \ldots, p_k) \in \Theta_0$. Hvis vi herefter vil teste hypotesen $(p_1, \ldots, p_k) \in \Theta_{00}$ kan det i princippet gøres ved et test af samme slags, altså et *test imod den fulde model,* der forkaster hvis

$$Q_{00} = \frac{L(\hat{p}_{001}, \ldots, \hat{p}_{00k})}{L(\hat{p}_1, \ldots, \hat{p}_k)}$$

er ekstremt lille i sin fordeling. Det forekommer imidlertid mere naturligt at udnytte den viden vi har, om at hypotesen $(p_1, \ldots, p_k) \in \Theta_0$ kan accepteres, og erstatte den oprindelige grundmodel med modellen givet ved Θ_0, dvs. benytte et *test imod den reducerede model,* som forkaster hvis

$$Q_{0,00} = \frac{L(\hat{p}_{001}, \ldots, \hat{p}_{00k})}{L(\hat{p}_{01}, \ldots, \hat{p}_{0k})}$$

er ekstremt lille i sin fordeling. Ved således at udnytte den viden vi har i forvejen, får vi et test der er mere følsomt overfor afvigelser fra den yderligere reducerede model.

Bemærk, at der gælder

$$Q_{0,00} = \frac{Q_{00}}{Q_0},$$

og dermed

$$-2 \log Q_{0,00} = -2 \log Q_{00} - (-2 \log Q_0)\,.$$

Hovedresultatet i forbindelse med successiv testning kan formuleres som følger.

SÆTNING 4.2. *Under hypotesen $(p_1, \ldots, p_k) \in \Theta_{00}$ er*

$$-2 \log Q_{0,00} = 2 \left(x_1 \log \frac{\hat{p}_{01}}{\hat{p}_{001}} + \cdots + x_k \log \frac{\hat{p}_{0k}}{\hat{p}_{00k}} \right)$$

approksimativt χ^2–fordelt med $d_0 - d_{00}$ frihedsgrader. Endvidere gælder approksimationsformlen

$$-2 \log Q_{0,00} \approx \sum_{j=1}^{k} \frac{(n\hat{p}_{0j} - n\hat{p}_{00j})^2}{n\hat{p}_{00j}},$$

hvor den sidste størrelse igen kaldes Pearson's teststørrelse.

Til dette resultat skal knyttes lignende forbehold som til sætningen om test imod den fulde model. Men dem vil vi ikke gentage her.

EKSEMPEL 4.2 fortsat. I forlængelse af eksemplet med de 100 kast med en terning, antag nu, at vi rent faktisk *har* en velunderbygget mistanke til denne terning. Fra sædvanligvis velunderrettet kilde har vi fået oplyst, at terningen muligvis stammer fra et parti falske terninger som, ved snedig indstøbning af en blyklump lige bag ved etter–prikken, er konstrueret sådan at de giver en sekser med sandsynlighed $p_6 > \frac{1}{6}$ og en etter med sandsynlighed $p_1 < \frac{1}{6}$, medens de fire andre sandsynligheder er ens ($p_2 = p_3 = p_4 = p_5$).

Det ændrer naturligvis situationen, og herefter vil vi måske være tilbøjlige til lægge meget større vægt på resultatet af det test, vi tidligere har foretaget.

Men vi kan faktisk gøre det endnu bedre nu. Ved testet mod den fulde polynomialfordelingsmodel retter mistanken sig jo imod en hvilken som helst form for skævhed, en terning kan have. Nu har vi en meget mere specifik mistanke, og denne mistanke passer unægteligt godt med data, idet terningen jo netop gav færrest ettere og flest seksere. Denne overvejelse kan vi formalisere på følgende måde:

Hvis mistanken er begrundet — og også selvom den ikke skulle være det — har vi

$$p_2 = p_3 = p_4 = p_5.$$

Disse tre relationer bestemmer en delmængde Θ_0 af Δ_6 af dimension 5–3=2. Vi kan parametrisere denne model ved to positive tal α og β med $\alpha + \beta < 1$ på formen

$$\begin{aligned} p_1 &= \alpha, \\ p_6 &= \beta, \\ p_2 = p_3 = p_4 = p_5 &= \frac{1-\alpha-\beta}{4}. \end{aligned}$$

Likelihooden under hypotesen bliver

$$L(\alpha, \beta) = \alpha^{11} \left(\frac{1-\alpha-\beta}{4} \right)^{16+15+17+12} \beta^{29}$$

som, på nær en proportionalitetsfaktor $(\frac{1}{4})^{16+15+17+12}$, er likelihood'en for en fuld polynomialfordelingsmodel af orden 3 med sandsynligheds-parametre α, $1-\alpha-\beta$ og β, hvor (11,60,29) er observeret. Vi får derfor

$$\hat{\alpha} = \frac{11}{100} \text{ og } \hat{\beta} = \frac{29}{100},$$

og dermed

$$\hat{p}_{01} = 0.11,\ \hat{p}_{02} = \hat{p}_{03} = \hat{p}_{04} = \hat{p}_{05} = \frac{60/100}{4} = 0.15 \text{ og } \hat{p}_{06} = 0.29.$$

Estimatet i denne model fremkommer altså helt naturligt ved at p_1 og p_6 — der varierer frit — estimeres præcis som i den fulde model, medens den resterende sandsynlighedsmasse 1–(0.11+0.29)=0.60 deles ligeligt ud på de fire sandsynlighedsparametre, som antages at være ens.

Kvotienttestet for hypotesen givet ved Θ_0 imod den fulde model giver således

$$-2\log Q_0 = 2\left(0 + 16\log\frac{0.16}{0.15} + \cdots + 12\log\frac{0.12}{0.15} + 0\right) = 0.96,$$

som bestemt ikke er for stor til at kunne stamme fra en χ^2–fordeling med 3 frihedsgrader. Denne hypotese kan altså umiddelbart godkendes.

Med den således godkendte model som udgangspunkt vil vi herefter teste hypotesen $p_1 = \cdots = p_6$, altså hypotesen svarende til mængden Θ_{00} bestående af det ene punkt $(\frac{1}{6}, \ldots, \frac{1}{6})$. Vi får

$$-2\log Q_{0,00} = 2\left(11\log\frac{0.11}{1/6} + (16+15+17+12)\log\frac{0.15}{1/6} + 29\log\frac{0.29}{1/6}\right) = 10.34.$$

Til kontrol af udregningen kan man bemærke, at der, på nær afrundningsfejl, gælder relationen 11.31=0.96+10.34 mellem de tre $-2\log Q$–størrelser vi indtil nu har regnet ud. Da teststørrelsen 10.34 er større end 99%–fraktilen i χ^2–fordelingen med 2 frihedsgrader, får vi her en noget mere markant afvisning af hypotesen, end vi gjorde da vi testede direkte mod den fulde model. Det har naturligvis noget at gøre med, at vi udnytter vores viden om hvilken slags skævhed terningen er mistænkt for at have.

Faktisk kan man udnytte denne viden endnu bedre. Som udgangspunkt har vi her taget modellen, der antager $p_2 = p_3 = p_4 = p_5$, men vi har ikke udnyttet den viden der ligger i oplysningen om, at de falske terninger har $p_6 > p_1$. Hvis vi desuden antager, at de falske terninger i det væsentlige giver toere, treere, firere og femmere med samme sandsynlighed som almindelige terninger, kan vi argumentere (intutitivt) som følger. Antallene 16, …, 12 af "mellem–udfald" er ret irrelevante, det der først og fremmest kan afsløre den form for skævhed, vi har mistanke om, er et skævt forhold mellem antal seksere og antal ettere. Vi *betinger* derfor med hændelsen $X_1 + X_6 = 40$. Under denne betingelse er X_6 binomialfordelt med sandsynlighedsparameter $\frac{p_6}{p_1+p_6}$ og antalsparameter 40, jfv. opgave 3.4.3 (b1) og (b2) i Ssr. Det er derfor naturligt at bygge konklusionen på et ensidet test for hypotesen $p = \frac{1}{2}$ imod $p > \frac{1}{2}$ i en binomialfordeling med antalsparameter 40, hvor $X = 29$ er observeret. Det eksakte test giver her P–værdien 0.0032 — altså en endnu mere markant indikation af, at terningen faktisk er skæv.

Om man så vil tro på det eller ej må selvfølgelig afhænge af de konkrete omstændigheder. Strengt taget behøver man jo ikke at "tro" på noget som helst, bare fordi den formelle fremstilling af testteorien lægger op til det. Men hvis man var nysgerrig kunne man jo prøve at kaste terningen 1000 gange til. Eller smadre den med en hammer, for at se hvad der var indeni.

OPGAVE 4.4.1. Nedenfor er gengivet et diagram, som viser nogle hypoteser og tests vedrørende den famøse terning fra Eksempel 4.2. De (kvotient–) tests, som er foretaget i dette kapitel, er angivet på formen "teststørrelse/frihedsgrader".

En af hypoteserne (den næstsidste) har vi ikke set på før, bortset fra at den lige er nævnt på side 30. De tilsvarende teststørrelser er markeret med et spørgsmålstegn. Udregn disse, og diskuter konsekvenserne. Diskuter også relationen til det eksakte test, der er foretaget på side 30.