STATISTIKNOTER

Simple multinomialfordelingsmodeller

Jørgen Larsen

IMFUFA
Roskilde Universitetscenter

Februar 1999

IMFUFA, Roskilde Universitetscenter, Postboks 260, DK-4000 Roskilde.
Jørgen Larsen: **STATISTIKNOTER: Simple multinomialfordelingsmodeller**
IMFUFA tekst nr. 304d/1995 35 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 Multinomialfordelingen

Multinomialfordelingen kan ses som en naturlig generalisation af binomialfordelingen:

- I situationer hvor man har at gøre med n gentagelser af et elementarforsøg der kan resultere i et af *to* mulige udfald, vil antallet af gange man får den ene slags udfald, blive *binomialfordelt*.
- I situationer hvor man har at gøre med n gentagelser af et elementarforsøg der kan resultere i et af r mulige udfald, vil man et vist antal gange, y_1, få det første udfald, et vist antal gange, y_2, det andet udfald, ... , og et vist antal gange, y_r, det r-te udfald; talsættet $(y_1, y_2, \ldots, y_r)$ bliver *multinomialfordelt*.

Eksempel 1.1
En simpel form for politisk meningsmålingsundersøgelse kunne bestå i at man tilfældigt udvælger n personer og spørger dem hvilket af de r politiske partier de ville stemme på hvis der var folketingsvalg i morgen.

Her består elementarforsøget i at spørge én person og notere den pågældendes svar ned. Den samlede undersøgelse resulterer i at et vist antal (y_1) svarer det første parti, et vist antal (y_2) svarer det andet parti, ... , og et vist antal (y_r) svarer det r-te parti. Da der i alt er spurgt n personer, vil der gælde at $y_1 + y_2 + \ldots + y_r = n$, forudsat at alle de adspurgte faktisk svarer.

Den multinomialfordelingsmodel vi i det følgende vil diskutere, svarer til at når man vælger en tilfældig person, så vil denne med en vis sandsynlighed (p_1) svare parti nr. 1, med en vis sandsynlighed (p_2) svare parti nr. 2, ... , og med en vis sandsynlighed (p_r) svare parti nr. r. Da vi forudsætter at alle adspurgte giver et af de r mulige svar, er $p_1 + p_2 + \ldots + p_r = 1$.

1.1 Den grundlæggende multinomialfordelingsmodel

Antag at vi har klassificeret n individer i r klasser; i den generelle diskussion kaldes klasserne $A_1, A_2, \ldots, A_r$, i en konkret modelsituation har de ofte nogle mere sigende betegnelser. Skematisk er situationen som vist i Figur 1.1.

Vi går ud fra at de n individer stammer fra en og samme »population« således at hver gang man tilfældigt udvælger et individ, er der sandsynligheden p_1 for at individet tilhører klassen A_1, sandsynligheden p_2 for at individet tilhører klassen A_2, osv. Sandsynlighederne $p_1, p_2, \ldots, p_r$ (der summerer til 1) er *ukendte parametre* der er karakteristiske for populationen.

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

Figur 1.1 Multinomialfordelingssituationen, skematisk.

Hermed har vi sådan set beskrevet den statistiske model for ét individ. Når der er et større antal individer, plejer man ikke at angive hvilken klasse hvert enkelt individ viser sig at tilhøre, man nøjes med at angive hvor mange individer der er i hver klasse, dvs. man angiver de observerede værdier af de stokastiske variable $Y_1, Y_2, \ldots, Y_r$ defineret som

$$Y_i = \text{antal individer der viser sig at tilhøre klassen } A_i \quad (i = 1, 2, \ldots, r).$$

Den statistiske model vi skal nå frem til, skal specificere sandsynlighedsfordelingen for sættet $(Y_1, Y_2, \ldots, Y_r)$ af stokastiske variable, eller sagt på en anden måde, vi skal fastlægge $\mathrm{P}(Y_1 = y_1,\ Y_2 = y_2,\ \ldots,\ Y_r = y_r)$.

Hvis der kun er *to* klasser, så er der tale om et binomialfordelingsproblem. For at løse problemet med r klasser går vi frem på en måde der er stærkt inspireret af udledningen af binomialfordelingen.

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

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

Hvis vi har n klassenavne $x_1, x_2, \ldots, x_n$, og hvis det er sådan at netop y_i af x-erne er et A_i, $i = 1, 2, \ldots, r$, så er

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

Den søgte sandsynlighed $\mathrm{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_r = y_r)$ fås nu ved at summere disse sandsynligheder over alle mulige n-tupler $(x_1, x_2, \ldots, x_n)$ be-

stående af y_1 A_1-er, y_2 A_2-er, ... , y_r A_r-er:

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

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

$$\binom{n}{y_1\ y_2\ \ldots\ y_r}$$

der kaldes en *multinomialkoefficient* (eller *polynomialkoefficient*). Den fundne sandsynlighedsfunktion

$$\mathrm{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_r = y_r) \quad = \quad \binom{n}{y_1\ y_2\ \ldots\ y_r} p_1^{y_1} p_2^{y_2} \ldots p_r^{y_r}$$

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

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

Multinomialkoefficienter

Definition 1.1 (Multinomialkoefficient)
Multinomialkoefficienten

$$\binom{n}{y_1\ y_2\ \ldots\ y_r}$$

betegner antallet af forskellige måder hvorpå man kan placere r symboler $A_1, A_2, \ldots, A_r$ på n pladser således at symbolet A_1 kommer på y_1 af pladserne, symbolet A_2 kommer på y_2 af pladserne, ... , symbolet A_r kommer på y_r af pladserne.

Man kan let udlede formler der gør det muligt at udregne multinomialkoefficienter. Vi illustrerer fremgangsmåden med et eksempel, hvor vi vil bestemme talværdien af $\binom{7}{2\ 3\ 2}$:

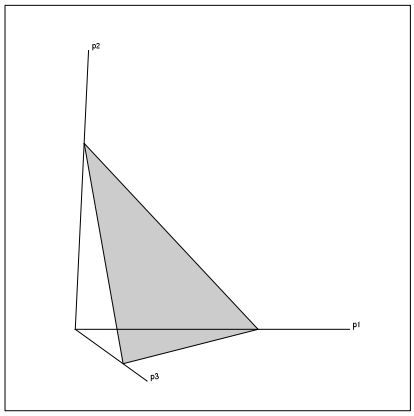
1. Det søgte tal er pr. definition antallet af placeringer af symbolerne A_1, A_2 og A_3 på syv pladser således at to af pladserne får et A_1, tre af pladserne et A_2 og to af pladserne et A_3. – En mulig placering er $A_1, A_3, A_1, A_2, A_2, A_2, A_3$.

2. Vi kan bestemme en placering ved først at bestemme hvilke to pladser der skal have et A_1, dernæst hvilke tre pladser der skal have et A_2, og så endelig placere et A_3 på de to tiloversblevne pladser.
   (a) Der er $\binom{7}{2} = 21$ forskellige placeringer af de to A_1-er (ifølge definitionen af binomialkoefficienter).
   (b) Hver gang vi har placerer de to A_1-er, er der fem pladser tilbage, og på de fem pladser skal vi fordele tre A_2-er og to A_3-er; dette kan gøres på $\binom{5}{3} = 10$ forskellige måder. Hver gang vi har en af de $\binom{7}{2}$ placeringer af A_1, er der altså $\binom{5}{3}$ placeringer af A_2 og A_3.
3. I alt er der derfor $\binom{7}{2} \cdot \binom{5}{3}$ forskellige placeringer af A-erne så
$$\binom{7}{2\,3\,2} = \binom{7}{2} \cdot \binom{5}{3} = 21 \cdot 10 = 210.$$
4. Vi kan benytte formlen $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$ og få
$$\binom{7}{2\,3\,2} = \binom{7}{2} \cdot \binom{5}{3} = \frac{7!}{2!\,5!} \cdot \frac{5!}{3!\,2!} = \frac{7!}{2!\,3!\,2!}.$$

Et generelt udtryk for multinomialkoefficienter fås på ganske tilsvarende måde. Man skal placere y_1 A_1-er, y_2 A_2-er, ..., og y_r A_r-er på n pladser ($n = y_1 + y_2 + \ldots + y_r$). Først kan A_1-erne placeres på $\binom{n}{y_1}$ forskellige måder; dernæst kan A_2-erne placeres på de de resterende $n - y_1$ pladser på $\binom{n-y_1}{y_2}$ forskellige måder; dernæst kan A_3-erne placeres på de resterende $n - y_1 - y_2$ pladser på $\binom{n-y_1-y_2}{y_3}$, osv. Slutresultatet bliver at
$$\binom{n}{y_1\ y_2\ \ldots\ y_r} = \frac{n!}{y_1!\,y_2!\,\ldots\,y_r!}$$
når $y_1 + y_2 + \ldots + y_r = n$.

Definition af multinomialfordelingen

Definition 1.2 (Multinomialfordeling)
At den r-dimensionale stokastiske variabel $(Y_1, Y_2, \ldots, Y_r)$ er multinomialfordelt med antalsparameter n og sandsynlighedsparameter
$$\boldsymbol{p} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_r \end{pmatrix}$$
betyder at
$$\mathrm{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_r = y_r) = \binom{n}{y_1\ y_2\ \ldots\ y_r} p_1^{y_1} p_2^{y_2} \ldots p_r^{y_r} \quad (1.1)$$
når $y_1, y_2, \ldots, y_r$ er ikke-negative heltal der summerer til n.



Figur 1.2 Sandsynlighedssimplexet i det tredimensionale rum.

Estimation af parametrene

I den generelle situation er modelfunktionen givet ved formel (1.1), og likelihoodfunktionen er dermed

$$L(\boldsymbol{p}) \quad = \quad \text{konstant} \cdot p_1^{y_1} p_2^{y_2} \dots p_r^{y_r}.$$

Spørgsmålet er nu hvordan man estimerer parameteren $\boldsymbol{p}$.

De almene principper for analyse af statistiske modeller påbyder at estimere $\boldsymbol{p}$ ved det r-dimensionale talsæt $\hat{\boldsymbol{p}}$ der maksimaliserer likelihoodfunktionen. Likelihoodfunktionen er en funktion af $\boldsymbol{p}$, dvs. af r variable $p_1, p_2, \dots, p_r$; disse kan ikke variere frit, men opfylder »bibetingelserne«

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

I specialtilfældet $r = 3$ kan vi anskueliggøre mulighedsområdet, dvs. mængden af $\boldsymbol{p}$-er der opfylder bibetingelserne, som et trekantet område, det såkaldte sandsynlighedssimplex, i det tredimensionale rum, se Figur 1.2.

Opgaven er at bestemme det punkt

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

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

Sætning 1.1
Antag at $a_1, a_2, \ldots, a_r$ er givne ikke-negative tal, og betragt funktionen

$$f : (p_1, p_2, \ldots, p_r) \quad \mapsto \quad p_1^{a_1}\, p_2^{a_2}\, \ldots\, p_r^{a_r}$$

defineret på mængden af ikke-negative talsæt $(p_1, p_2, \ldots, p_r)$ der summerer til 1. Vi sætter $a_\bullet = a_1 + a_2 + \ldots + a_r$ og $\hat{p}_i = a_i/a_\bullet$, $i = 1, 2, \ldots, r$.

Da har f et entydigt maksimumspunkt, nemlig $(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_r)$.

Bevis
Vi vil sammenligne funktionsværdierne $f(p_1, p_2, \ldots, p_r)$ og $f(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_r)$ ved at se på størrelsen $\ln \dfrac{f(p_1, p_2, \ldots, p_r)}{f(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_r)}$ som er negativ hvis og kun hvis $f(p_1, p_2, \ldots, p_r) < f(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_r)$.

Der gælder først at

$$\ln \frac{f(p_1, p_2, \ldots, p_r)}{f(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_r)} \quad = \quad \sum_{i=1}^{r} a_i \ln \frac{p_i}{\hat{p}_i}.$$

Nu benyttes en egenskab ved logaritmefunktionen, nemlig at $\ln t \leq t - 1$ for alle $t > 0$, og med lighedstegn hvis og kun hvis $t = 1$. Derfor er

$$\begin{aligned} \sum_{i=1}^{r} a_i \ln \frac{p_i}{\hat{p}_i} &\leq \sum_{i=1}^{r} a_i \left(\frac{p_i}{\hat{p}_i} - 1 \right) \\ &= \sum_{i=1}^{r} \left(\frac{a_i p_i}{a_i/a_\bullet} - a_i \right) \\ &= \sum_{i=1}^{r} p_i a_\bullet - \sum_{i=1}^{r} a_i \\ &= a_\bullet - a_\bullet \\ &= 0, \end{aligned}$$

hvor »mindre end eller lig med« bliver »lig med« hvis og kun hvis alle tallene $p_i/\hat{p}_i$ er lig 1, dvs. hvis og kun hvis $p_i = \hat{p}_i$ for alle i. □

Anvendt på funktionen $(p_1, p_2, \ldots, p_r) \mapsto p_1^{y_1} p_2^{y_2} \ldots p_r^{y_r}$ fortæller sætningen at likelihoodfunktionen L antager sit maksimum i det entydigt bestemte punkt $(y_1/n,\ y_2/n,\ \ldots,\ y_r/n)$. Altså er maksimaliseringsestimatet $\hat{\boldsymbol{p}}$ for $\boldsymbol{p}$ givet ved

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

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

1.2 Sammenligning af multinomialfordelinger

Man har undertiden brug for at kunne sammenligne forskellige multinomialfordelinger for at afgøre om de har samme sandsynlighedsparameter. Her er et eksempel; det vil blive analyseret mere indgående i Kapitel 2:

Eksempel 1.2 (Torsk i Østersøen)
Den 6. marts 1961 fangede nogle havbiologer 69 torsk ved Lolland og undersøgte arten af blodets hæmoglobin i hver enkelt torsk. Senere på året fangede man også nogle torsk ved Bornholm og ved Ålandsøerne og bestemte deres genotype.[1]

Man mener at hæmoglobin-arten bestemmes af ét enkelt gen, og det som biologerne bestemte, var torskenes genotype for så vidt angår dette gen. Genet kan optræde i to udgaver som traditionen tro kaldes for A og a, og de mulige genotyper er da AA, Aa og aa. Tabel 1.1 viser den fundne fordeling på genotyper for hver af de tre lokaliteter. I dette afsnit vil vi udelukkende opfatte symbolerne AA, Aa og aa som *navne* på klasser man klassificerer torskene i. I Kapitel 2 vil vi smugle lidt genetik ind i en mere udbygget statistisk model for tallene.

På hver geografisk lokalitet er der sket det at man har klassificeret et antal torsk i tre mulige klasser, så derfor kan man sige at der på hver lokalitet er tale om en multinomialfordelingssituation (når der er tre klasser, taler man også om en *tri*nomialfordeling). Det kunne måske være af interesse at undersøge om genotypefordelingen er den samme på de tre lokaliteter, altså om sandsynligheden for at en torsk har en bestemt genotype, er den samme for alle tre lokaliteters vedkommende. (Skønt når man ser på tallene virker denne formodning lidet plausibel.)

Den generelle model

I den generelle model antages det at vi har klassificeret nogle individer i r forskellige klasser $A_1, A_2, \ldots, A_r$. Individerne er på forhånd delt op i grupper, og der er s forskellige grupper med hhv. $n_1, n_2, \ldots, n_s$ individer. Det har vist sig at i gruppe j hører y_{1j} af individerne til gruppen A_1, y_{2j} af individerne til gruppen A_2, $y_{3,j}$ af individerne til gruppen A_3, osv. Skematisk ser situationen ud som vist i Figur 1.3.

[1] K. Sick (1965): Haemoglobin polymorphism of cod in the Baltic and the Danish Belt Sea. *Hereditas* **54**, 19-48.

Tabel 1.1 Genotypefordeling af torsk fra tre lokaliteter i Østersøen.

	lokalitet		
genotype	Lolland	Bornholm	Ålandsøerne
AA	27	14	0
Aa	30	20	5
aa	12	52	75
i alt	69	86	80

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

Figur 1.3 Sammenligning af multinomialfordelinger, generelt. y_{ij} betegner antallet af individer fra gruppe j der tilhører klassen A_i.

I torskeeksemplet er der $s = 3$ grupper svarende til de tre geografiske lokaliteter og $r = 3$ klasser svarende til de tre forskellige genotyper.

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

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

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

  som en observeret værdi af en r-dimensional stokastisk variabel

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

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

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

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

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

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

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

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

eller mere udførligt

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

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

$$f(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_s; \boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_s) \quad = \quad \prod_{j=1}^{s} \binom{n_j}{y_{1j}\ y_{2j}\ \cdots\ y_{rj}} p_{1j}^{y_{1j}} p_{2j}^{y_{2j}} \cdots p_{rj}^{y_{rj}} .$$

Likelihoodfunktionen er dermed

$$L(\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_s) \quad = \quad \text{konstant} \cdot \prod_{j=1}^{s} p_{1j}^{y_{1j}} p_{2j}^{y_{2j}} \cdots p_{rj}^{y_{rj}} \tag{1.2}$$

hvor konstanten er produktet af de s multinomialkoefficienter.

I torskeeksemplet er likelihoodfunktionen

$$L(\boldsymbol{p}_{\mathrm{L}}, \boldsymbol{p}_{\mathrm{B}}, \boldsymbol{p}_{\mathrm{Å}}) \quad = \quad \text{konstant} \cdot p_{1\mathrm{L}}^{27}\, p_{2\mathrm{L}}^{30}\, p_{3\mathrm{L}}^{12} \cdot p_{1\mathrm{B}}^{14}\, p_{2\mathrm{B}}^{20}\, p_{3\mathrm{B}}^{52} \cdot p_{1\mathrm{Å}}^{0}\, p_{2\mathrm{Å}}^{5}\, p_{3\mathrm{Å}}^{75} .$$

Likelihoodfunktionen er sandsynligheden for at observere det faktisk observerede, betragtet som funktion af det ukendte sæt parametre. Som sædvanlig udnævner vi de værdier der maksimaliserer likelihoodfunktionen (eller log-likelihoodfunktionen) til at være de bedste estimater over de ukendte parametre. I

den foreliggende model er likelihoodfunktionen et produkt af s del-likelihoodfunktioner der hver især vedrører én enkelt gruppe og ét enkelt $\boldsymbol{p}_j$. Når vi skal maksimalisere L mht. $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_s$, kan det derfor ske ved at maksimalisere hver del-likelihoodfunktion for sig. Det j-te delproblem er en simpel multinomialfordelingsmodel, så derfor følger det uden videre af resultatet på side 11 at

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

I taleksemplet er specielt

$$\hat{\boldsymbol{p}}_{\mathrm{L}} = \begin{pmatrix} \hat{p}_{1\mathrm{L}} \\ \hat{p}_{2\mathrm{L}} \\ \hat{p}_{3\mathrm{L}} \end{pmatrix} \quad = \quad \begin{pmatrix} 27/69 \\ 30/69 \\ 12/69 \end{pmatrix} = \begin{pmatrix} 0.39 \\ 0.43 \\ 0.17 \end{pmatrix} ,$$

$$\hat{\boldsymbol{p}}_{\mathrm{B}} = \begin{pmatrix} \hat{p}_{1\mathrm{B}} \\ \hat{p}_{2\mathrm{B}} \\ \hat{p}_{3\mathrm{B}} \end{pmatrix} \quad = \quad \begin{pmatrix} 14/86 \\ 20/86 \\ 52/86 \end{pmatrix} = \begin{pmatrix} 0.16 \\ 0.23 \\ 0.60 \end{pmatrix} ,$$

$$\hat{\boldsymbol{p}}_{\mathrm{Å}} = \begin{pmatrix} \hat{p}_{1\mathrm{Å}} \\ \hat{p}_{2\mathrm{Å}} \\ \hat{p}_{3\mathrm{Å}} \end{pmatrix} \quad = \quad \begin{pmatrix} 0/80 \\ 5/80 \\ 75/80 \end{pmatrix} = \begin{pmatrix} 0.00 \\ 0.06 \\ 0.94 \end{pmatrix} .$$

Hypoteseprøvning

Vi skal herefter undersøge om det er rimeligt at antage at hypotesen

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

om ens sandsynlighedsparametre holder. Under H_0 er der ingen forskel på de s grupper, så da kan vi lige så godt slå dem sammen til én stor gruppe bestående af $n_\bullet = n_1 + n_2 + \ldots + n_s$ individer der fordeler sig med

$$\begin{array}{rcccl}
y_{1\bullet} & = & y_{11} + y_{12} + \ldots + y_{1s} & = & \displaystyle\sum_{j=1}^{s} y_{ij} \quad \text{i klassen } A_1 \\
y_{2\bullet} & = & y_{21} + y_{22} + \ldots + y_{2s} & = & \displaystyle\sum_{j=1}^{s} y_{2j} \quad \text{i klassen } A_2 \\
\vdots & & \vdots & & \vdots \\
y_{i\bullet} & = & y_{i1} + y_{i2} + \ldots + y_{is} & = & \displaystyle\sum_{j=1}^{s} y_{ij} \quad \text{i klassen } A_i \\
\vdots & & \vdots & & \vdots \\
y_{r\bullet} & = & y_{r1} + y_{r2} + \ldots + y_{rs} & = & \displaystyle\sum_{j=1}^{s} y_{rj} \quad \text{i klassen } A_r
\end{array}$$

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

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

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

I likelihoodfunktionen (1.2) erstatter vi alle $\boldsymbol{p}_j$-erne med $\boldsymbol{p}$ og får derved *likelihoodfunktionen under* H_0:

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

Det valg af $p_1, p_2, \ldots, p_r$ der maksimaliserer denne likelihoodfunktion, er ifølge sætningen på side 10 netop $\hat{p}_i = y_{i\bullet}/n_\bullet$ som formodet.

I taleksemplet bliver $\hat{\boldsymbol{p}} = \begin{pmatrix} 41/235 \\ 55/235 \\ 139/235 \end{pmatrix} = \begin{pmatrix} 0.17 \\ 0.23 \\ 0.59 \end{pmatrix}$.

Når man vil vurdere hvor godt det faktisk observerede beskrives under H_0 i forhold til den aktuelle grundsmodels beskrivelse, skal man udregne *kvotientteststørrelsen*

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

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

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

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

hvor $\hat{y}_{ij} = \hat{p}_i n_j = y_{i\bullet} n_j / n_\bullet$ er det »forventede« antal individer fra gruppe j der klassificeres som A_i.

For at bestemme $-2 \ln Q$ i taleksemplet udregnes først de forventede antal, se Tabel 1.2. Dermed er

$$\begin{aligned} -2 \ln Q_{\text{obs}} &= 2 \Big(27 \ln \frac{27}{12.0} + 14 \ln \frac{14}{15.0} + 0 \ln \frac{0}{14.0} \\ &\qquad + 30 \ln \frac{30}{16.1} + 20 \ln \frac{20}{20.1} + 5 \ln \frac{5}{18.7} \\ &\qquad + 12 \ln \frac{12}{40.8} + 52 \ln \frac{52}{50.9} + 75 \ln \frac{75}{47.3} \Big) \\ &= 107.8 \end{aligned}$$

Tabel 1.2 Genotypefordeling hos torsk fra tre lokaliteter i Østersøen: forventede antal under antagelse af ens fordelinger på de tre lokaliteter.

	lokalitet		
genotype	Lolland	Bornholm	Ålandsøerne
AA	12.0	15.0	14.0
Aa	16.1	20.1	18.7
aa	40.8	50.9	47.3
i alt	68.9	86.0	80.0

For at afgøre om en opnået $-2\ln Q_{\text{obs}}$-værdi (som f.eks. 107.8) nu er tæt på 0 eller ej, skal man sammenligne den med alle de andre $-2\ln Q$-værdier man også kunne have fået ifølge den aktuelle model når H_0 er rigtig. Vi skal derfor finde *testsandsynligheden* ε, dvs. sandsynligheden for at få en værre (større) $-2\ln Q$-værdi end den observerede, under forudsætning af at H_0 er rigtig:

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

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

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

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

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

Bemærk at χ^2-fordelingen kun er en approksimation; for at man skal kunne bruge den, *skal alle de »forventede« antal* $\hat{y}_{ij}=\hat{p}_i n_j = y_{i\cdot} n_j/n$. *være mindst fem*. Hvis denne betingelse ikke er opfyldt, kan man måske opnå at den bliver opfyldt ved at man udelader nogle grupper eller klasser eller slår nogle grupper eller klasser sammen.

I det gennemgående taleksempel er der ingen problemer med at de »forventede« antal er for små. Vi kan derfor uden videre sammenligne $-2\ln Q_{\text{obs}} = 107.8$ med χ^2-fordelingen med $(3-1)(3-1)=4$ frihedsgrader. Da 99.9%-fraktilen i denne fordeling er 18.47, er testsandsynligheden ε mindre end 0.1%.

Da det således er temmelig usandsynligt at få en værre værdi af teststørrelsen $-2\ln Q$ end 107.8, er teststørrelsen *signifikant* stor, og vi forkaster H_0. Man må altså sige at der er en signifikant forskel på genotypen af torsk på de tre geografiske lokaliteter. – Denne konklusion er ikke overraskende hvis man man sammenligner Tabel 1.1 og 1.2.

1.3 Opgaver

Opgave 1.1 (Medarbejderaktier)
Det er blevet almindeligt at firmaer indfører ordninger med medarbejderaktier; derved skulle medarbejderne komme til at føle større medansvar og forpligtelse over for deres arbejdsplads. Det er dog ikke altid at firmaets opfordring til medarbejderne om at blive aktionærer opfattes på samme måde af alle medarbejdergrupper. For at danne sig et indtryk af medarbejderes motiver til at erhverve sig aktier har man foretaget et rundspørge blandt medarbejderne på en bestemt virksomhed (som har en medarbejderaktie-ordning) og bedt dem nævne deres motiver for at gå med i aktieordningen. Svarmulighederne var »for at bevare jobbet«, »som en investering« og »tror på idéen med medarbejderaktier«.

Nedenstående tabel viser respondenternes fordeling på motiv og medarbejderkategori. Hvad kan man på denne baggrund sige om en eventuel sammenhæng mellem medarbejdernes motiver for at deltage i ordningen og arten af deres arbejde?

	arbejdere	funktionærer	mellemledere	topledere
for at bevare jobbet	77	25	11	8
som en investering	37	13	8	4
tror på idéen	35	14	7	11

Opgave 1.2 (Test af simpel hypotese)
Antag at $(Y_1, Y_2, \ldots, Y_r)$ er multinomialfordelt med parametre n og $\boldsymbol{p}$, og lad

$$\boldsymbol{p}_0 = \begin{pmatrix} p_{01} \\ p_{02} \\ \vdots \\ p_{0r} \end{pmatrix}$$

være et sæt *kendte* ikke-negative tal der summerer til 1. Man ønsker at teste hypotesen $H_0 : \boldsymbol{p} = \boldsymbol{p}_0$ (eller altså $p_i = p_{0i}$ for alle i).

1. Udled $-2\ln Q$-størrelsen for denne hypotese.
2. Der gælder at når H_0 er rigtig, så er $-2\ln Q$ asymptotisk χ^2-fordelt med et antal frihedsgrader der kan udregnes som ændringen i antal frie parametre.

Hvad er antallet af frihedsgrader for $-2 \ln Q$?

2 Et større eksempel: Torsk i Østersøen

I dette kapitel vil vi tage et tidligere omtalt eksempel op til nærmere behandling. Eksemplet er blandt andet et eksempel på at man kan indbygge noget teori i den statistiske model, og et eksempel der viser nytten af maximum likelihood metoden til parameterestimation.

2.1 Præsentation af eksemplet

Den 6. marts 1961 fangede nogle havbiologer 69 torsk ved Lolland og undersøgte arten af blodets hæmoglobin i hver enkelt torsk. Senere på året fangede man desuden nogle torsk ved Bornholm og ved Ålandsøerne og bestemte deres genotype.[1]

Man mener at hæmoglobin-arten bestemmes af ét enkelt gen, og det som biologerne bestemte, var torskenes genotype for så vidt angår dette gen. Genet kan optræde i to udgaver som traditionen tro kaldes for A og a, og de mulige genotyper er da AA, Aa og aa. Tabel 2.1 viser den fundne fordeling på genotyper for hver af de tre lokaliteter.

På hver geografisk lokalitet er der sket det at man har klassificeret et antal torsk i tre mulige klasser, så på hver lokalitet er der tale om en multinomialfordelingssituation (når der er tre klasser, taler man også om en *tri*nomialfordeling). Som grundmodel benytter vi derfor den model der siger, at de tre observations»vektorer«

$$\begin{aligned}
\boldsymbol{y}_{\mathrm{L}} &= \begin{pmatrix} y_{1\mathrm{L}} \\ y_{2\mathrm{L}} \\ y_{3\mathrm{L}} \end{pmatrix} = \begin{pmatrix} 27 \\ 30 \\ 12 \end{pmatrix}, \\
\boldsymbol{y}_{\mathrm{B}} &= \begin{pmatrix} y_{1\mathrm{B}} \\ y_{2\mathrm{B}} \\ y_{3\mathrm{B}} \end{pmatrix} = \begin{pmatrix} 14 \\ 20 \\ 52 \end{pmatrix}, \\
\boldsymbol{y}_{\mathrm{Å}} &= \begin{pmatrix} y_{1\mathrm{Å}} \\ y_{2\mathrm{Å}} \\ y_{3\mathrm{Å}} \end{pmatrix} = \begin{pmatrix} 0 \\ 5 \\ 75 \end{pmatrix}.
\end{aligned}$$

stammer fra hver sin multinomialfordeling med antalsparametre $n_{\mathrm{L}} = 69$,

[1] K. Sick (1965): Haemoglobin polymorphism of cod in the Baltic and the Danish Belt Sea. *Hereditas* **54**, 19-48.

$n_{\mathrm{B}} = 86$ og $n_{\mathrm{Å}} = 80$ og med sandsynlighedsparametre

$$\begin{aligned} \boldsymbol{p}_{\mathrm{L}} &= \begin{pmatrix} p_{1\mathrm{L}} \\ p_{2\mathrm{L}} \\ p_{3\mathrm{L}} \end{pmatrix}, \\ \boldsymbol{p}_{\mathrm{B}} &= \begin{pmatrix} p_{1\mathrm{B}} \\ p_{2\mathrm{B}} \\ p_{3\mathrm{B}} \end{pmatrix}, \\ \boldsymbol{p}_{\mathrm{Å}} &= \begin{pmatrix} p_{1\mathrm{Å}} \\ p_{2\mathrm{Å}} \\ p_{3\mathrm{Å}} \end{pmatrix}. \end{aligned}$$

2.2 Hardy-Weinberg ligevægt

Grundmodellen er at hver geografisk lokalitet har sin egen multinomialfordeling, og at hver multinomialfordeling har en sandsynlighedsparameter

$$\boldsymbol{p} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix}$$

hvor p_1, p_2 og p_3 kan være hvilkesomhelst tre ikke-negative tal der summerer til 1. Imidlertid kan man argumentere for at der under visse omstændigheder må være en bestemt sammenhæng mellem de tre p-er.

Lad os antage at i en bestemt torskegeneration optræder de tre genotyper AA, Aa og aa med hyppighederne p_1, p_2 og p_3 (hvor $p_1 + p_2 + p_3 = 1$). Lad os desuden antage at næste generation fremstilles ved »tilfældig parring« således at hvert af en torskeunges to hæmoglobin-gener vælges uafhængigt af hinanden på følgende måde: først vælges et tilfældigt forældre-individ, dernæst vælges et tilfældigt af dette individs hæmoglobin-gener. Sandsynligheden for at vælge A er da $p_1 + {}^1\!/_2\, p_2$ hvilket vi kalder β, og sandsynligheden for at vælge a er ${}^1\!/_2\, p_2 + p_3 = 1 - \beta$. I den nye generation bliver genotypefordelingen derfor

$$\begin{aligned} \mathsf{AA}&: \quad \beta^2 \\ \mathsf{Aa}&: \quad 2\beta(1-\beta) \\ \mathsf{aa}&: \quad (1-\beta)^2 \end{aligned}$$

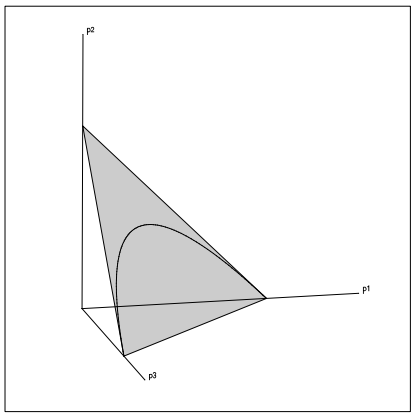
(bemærk at $\beta^2 + 2\beta(1-\beta) + (1-\beta)^2 = \big(\beta + (1-\beta)\big)^2 = 1$). Heraf kan vi se at genotypefordelingen i den nye generation ikke kan være hvadsomhelst, men at der er en vis sammenhæng mellem de tre sandsynligheder, styret af størrelsen β.

Vi kan prøve at se hvad der sker hvis der er en tilsvarende sammenhæng mellem sandsynlighederne i forældregenerationen. Lad os sige at i forældregenerationen er

$$\begin{aligned} \mathsf{AA}&: \quad p_1 = \alpha^2 \\ \mathsf{Aa}&: \quad p_2 = 2\alpha(1-\alpha) \\ \mathsf{aa}&: \quad p_3 = (1-\alpha)^2. \end{aligned}$$

Tabel 2.1 (= Tabel 1.1) Genotypefordeling af torsk fra tre lokaliteter i Østersøen.

genotype	Lolland	Bornholm	Ålandsøerne
AA	27	14	0
Aa	30	20	5
aa	12	52	75
i alt	69	86	80



Figur 2.1 Det tonede område er sandsynlighedssimplexet, dvs. mængden af tripler $\boldsymbol{p} = (p_1, p_2, p_3)$ af ikke-negative tal der summerer til 1. Kurven består af de $\boldsymbol{p}$ der kan optræde hvis der er Hardy-Weinberg ligevægt.

Så bliver $\beta = p_1 + {}^1\!/_2\, p_2 = \alpha^2 + {}^1\!/_2\, 2\alpha(1-\alpha) = \alpha$, dvs. *sandsynlighederne er uforandrede fra den ene generation til den anden*.

Man siger at populationen er i **Hardy-Weinberg ligevægt** hvis det er sådan at de tre genotyper optræder i forholdet

$$\begin{array}{rl} \mathsf{AA}\text{:} & p_1 = \beta^2 \\ \mathsf{Aa}\text{:} & p_2 = 2\beta(1-\beta) \\ \mathsf{aa}\text{:} & p_3 = (1-\beta)^2 \end{array}$$

for en eller anden værdi af $\beta \in [0,1]$. Hvis der er Hardy-Weinberg ligevægt, er det altså kun ganske særlige sandsynlighedstripler (p_1, p_2, p_3) der kan komme på tale, se Figur 2.1

2.3 Hypotesen om Hardy-Weinberg ligevægt

Vi vil undersøge om der er Hardy-Weinberg ligevægt på hver af de tre lokaliteter. Vi begynder med Lolland.

At der er Hardy-Weinberg ligevægt ved Lolland kan formuleres som den statistiske hypotese

$$H_{\mathrm{L}} : \begin{pmatrix} p_{1\mathrm{L}} \\ p_{2\mathrm{L}} \\ p_{3\mathrm{L}} \end{pmatrix} = \begin{pmatrix} \beta_{\mathrm{L}}^2 \\ 2\beta_{\mathrm{L}}(1-\beta_{\mathrm{L}}) \\ (1-\beta_{\mathrm{L}})^2 \end{pmatrix}.$$

I grundmodellen er likelihoodfunktionen

$$L(p_{1\mathrm{L}}, p_{2\mathrm{L}}, p_{3\mathrm{L}}) \quad = \quad \text{konstant} \cdot p_{1\mathrm{L}}^{27}\, p_{2\mathrm{L}}^{30}\, p_{3\mathrm{L}}^{12}$$

der har maksimum i

$$\hat{\boldsymbol{p}}_L \quad = \quad \begin{pmatrix} 27/69 \\ 30/69 \\ 12/69 \end{pmatrix}.$$

Under H_{L} er likelihoodfunktionen

$$\begin{aligned} L_{\mathrm{L}}(\beta_L) &= L\left(\beta_{\mathrm{L}}^2, 2\beta_{\mathrm{L}}(1-\beta_{\mathrm{L}}), (1-\beta_{\mathrm{L}})^2\right) \\ &= \text{konstant} \cdot \left(\beta_{\mathrm{L}}^2\right)^{27} \left(2\beta_{\mathrm{L}}(1-\beta_{\mathrm{L}})\right)^{30} \left((1-\beta_{\mathrm{L}})^2\right)^{12} \\ &= \text{konstant} \cdot \beta_{\mathrm{L}}^{2\cdot 27+30}(1-\beta_{\mathrm{L}})^{30+2\cdot 12}, \end{aligned}$$

som har maksimum i $\hat{\beta}_{\mathrm{L}} = \frac{2\cdot 27+30}{2\cdot 69} = \frac{84}{138} = 0.609$ (dvs. det observerede antal A divideret med det samlede antal gener).

Man tester hypotesen ved brug af den sædvanlige kvotienttesstørrelse $Q = L(\hat{\beta}_{\mathrm{L}}^2, 2\hat{\beta}_{\mathrm{L}}(1-\hat{\beta}_{\mathrm{L}}), (1-\hat{\beta}_{\mathrm{L}})^2) / L(\hat{p}_{1\mathrm{L}}, \hat{p}_{2\mathrm{L}}, \hat{p}_{3\mathrm{L}})$ eller $-2\ln Q$; sidstnævnte kan udtrykkes som

$$-2\ln Q \quad = \quad 2\sum_{i=1}^{r} y_i \ln \frac{y_i}{\hat{y}_i}$$

hvor $(\hat{y}_1, \hat{y}_2, \hat{y}_3) = (n_{\mathrm{L}}\hat{\beta}_{\mathrm{L}}^2, n_{\mathrm{L}}2\hat{\beta}_{\mathrm{L}}(1-\hat{\beta}_{\mathrm{L}}), n_{\mathrm{L}}(1-\hat{\beta}_{\mathrm{L}})^2)$ er de »forventede« antal under H_{L}.

Man finder at $-2\ln Q = 0.52$ med $(3-1)-1 = 1$ frihedgrader, svarende til en testsandsynlighed på ca. 47%, så man kan sagtens antage at torskebestanden ved Lolland er i Hardy-Weinberg ligevægt.

Noget tilsvarende kan gøres med de to andre lokaliteter. Man får maksimaliseringsestimaterne $\hat{\beta}_{\mathrm{B}} = 0.279$ og $\hat{\beta}_{\mathrm{Å}} = 0.031$. Tabel 2.2 viser de forventede

Tabel 2.2 Forventede antal $\hat{y}$ under forudsætning af Hardy-Weinberg ligevægt på hver lokalitet.

genotype	Lolland	Bornholm	Ålandsøerne
AA	25.6	6.7	0.1
Aa	32.9	34.6	4.8
aa	10.6	44.7	75.1
i alt	69	86	80

antal $\hat{y}$ hvert sted. Ved Ålandsøerne kan man oplagt antage Hardy-Weinberg ligevægt.[2] Ved Bornholm er der større uoverensstemmelse mellem de observerede og de forventede antal, og teststørrelsen er her $-2\ln Q = 14.4$, svarende til en testsandsynlighed af størrelsesorden 10^{-4}.

2.4 En samlet model

Man kan sige at hypotesen om Hardy-Weinberg ligevægt er sådan en »pæn« hypotese fordi man kan »forstå« (dvs. levere en simpel forklaring på) den. Derfor er det ærgerligt at Bornholm tilsyneladende falder uden for det pæne billede. For at reparere på tingene kunne man forsøge sig med en modificeret hypotese H_1 gående ud på at

- ved Lolland er der Hardy-Weinberg ligevægt med parameter β_{L},
- ved Ålandsøerne er der Hardy-Weinberg ligevægt med parameter $\beta_{\mathring{\mathrm{A}}}$,
- ved Bornholm er populationen en blanding af Lollandstorsk og Ålandstorsk i forholdet $\alpha : (1-\alpha)$ hvor $\alpha \in]0,1[$ er ukendt parameter.

Mere præcist går H_1 altså ud på at der findes værdier af β_{L}, $\beta_{\mathring{\mathrm{A}}}$ og α så

$$
\begin{aligned}
\boldsymbol{p}_{\mathrm{L}} &= \begin{pmatrix} \beta_{\mathrm{L}}^2 \\ 2\beta_{\mathrm{L}}(1-\beta_{\mathrm{L}}) \\ (1-\beta_{\mathrm{L}})^2 \end{pmatrix}, \\
\boldsymbol{p}_{\mathring{\mathrm{A}}} &= \begin{pmatrix} \beta_{\mathring{\mathrm{A}}}^2 \\ 2\beta_{\mathring{\mathrm{A}}}(1-\beta_{\mathring{\mathrm{A}}}) \\ (1-\beta_{\mathring{\mathrm{A}}})^2 \end{pmatrix}, \\
\boldsymbol{p}_{\mathrm{B}} &= \alpha\boldsymbol{p}_{\mathrm{L}} + (1-\alpha)\boldsymbol{p}_{\mathring{\mathrm{A}}} \\
&= \begin{pmatrix} \alpha\beta_{\mathrm{L}}^2 + (1-\alpha)\beta_{\mathring{\mathrm{A}}}^2 \\ \alpha 2\beta_{\mathrm{L}}(1-\beta_{\mathrm{L}}) + (1-\alpha)2\beta_{\mathring{\mathrm{A}}}(1-\beta_{\mathring{\mathrm{A}}}) \\ \alpha(1-\beta_{\mathrm{L}})^2 + (1-\alpha)(1-\beta_{\mathring{\mathrm{A}}})^2 \end{pmatrix}.
\end{aligned}
$$

[2] Man kan ikke benytte χ^2-approksimationen til $-2\ln Q$ fordi et af de forventede antal er alt for lille. Til gengæld reproducerer modellen jo observationerne særdeles fint.

Tabel 2.3 Forventede antal $\hat{y}$ i blandingsmodellen.

genotype	Lolland	Bornholm	Ålandsøerne
AA	25.7	13.7	0.1
Aa	32.8	20.3	4.8
aa	10.4	52.0	75.1
i alt	69	86	80

Bemærk at der nu er tale om én samlet model for alle tre lokaliteter.

Den samlede likelihoodfunktion bliver produktet af de tre del-likelihoodfunktioner for de tre lokaliteter. Det er bekvemt at operere med *logaritmen* til likelihoodfunktionen, så den skriver vi op:

$$\begin{aligned}
&\ln L(\beta_{\mathrm{L}}, \beta_{\mathring{\mathrm{A}}}, \alpha) \\
&\quad = \quad 27 \ln p_{1\mathrm{L}} + 30 \ln p_{2\mathrm{L}} + 12 \ln p_{3\mathrm{L}} \\
&\qquad + 14 \ln p_{1\mathrm{B}} + 20 \ln p_{2\mathrm{B}} + 52 \ln p_{3\mathrm{B}} \\
&\qquad + 0 \ln p_{1\mathring{\mathrm{A}}} + 5 \ln p_{2\mathring{\mathrm{A}}} + 75 \ln p_{3\mathring{\mathrm{A}}} \\
&\quad = \quad \text{konstant} + 84 \ln \beta_{\mathrm{L}} + 54 \ln(1-\beta_{\mathrm{L}}) \\
&\qquad + 14 \ln \left(\alpha \beta_{\mathrm{L}}^2 + (1-\alpha) \beta_{\mathring{\mathrm{A}}}^2 \right) \\
&\qquad + 20 \ln \left(\alpha \beta_{\mathrm{L}} (1-\beta_{\mathrm{L}}) + (1-\alpha) \beta_{\mathring{\mathrm{A}}} (1-\beta_{\mathring{\mathrm{A}}}) \right) \\
&\qquad + 52 \ln \left(\alpha (1-\beta_{\mathrm{L}})^2 + (1-\alpha)(1-\beta_{\mathring{\mathrm{A}}})^2 \right) \\
&\qquad + 5 \ln \beta_{\mathring{\mathrm{A}}} + 155 \ln(1-\beta_{\mathring{\mathrm{A}}}).
\end{aligned}$$

Der synes ikke at være nogen praktisk anvendelig analytisk måde at maksimalisere denne funktion på, så man må benytte en iterationsmetode. Som startværdier til en sådan kan vi benytte de tidligere fundne estimater $\hat{\beta}_{\mathrm{L}} = 0.609$ og $\hat{\beta}_{\mathring{\mathrm{A}}} = 0.031$ og vælge α så det forventede antal Aa ved Bornholm er lig det observerede, dvs. ved at løse ligningen

$$\alpha \cdot 2\hat{\beta}_{\mathrm{L}}(1-\hat{\beta}_{\mathrm{L}}) + (1-\alpha) \cdot 2\hat{\beta}_{\mathring{\mathrm{A}}}(1-\hat{\beta}_{\mathring{\mathrm{A}}}) \quad = \quad 20/86,$$

hvilket giver $\alpha \approx 0.414$.

Man finder at $\ln L$ antager sit maksimum i punktet $(\hat{\hat{\beta}}_{\mathrm{L}}, \hat{\hat{\beta}}_{\mathring{\mathrm{A}}}, \hat{\hat{\alpha}}) = (0.611, 0.031, 0.425)$.[3] Herefter kan vi udregne den forventede genotypefordeling de tre steder, se Tabel 2.3. Det ses at der er langt bedre overensstemmelse mellem de observerede og de »forventede« værdier i denne model. Hvis man tester modellen i forhold til grundmodellen med en vilkårlig trinomialfordeling hvert sted, får man en $-2 \ln Q$-størrelse på 0.7 (– talværdien afhænger en

[3] Værdien af $\ln L$ i dette punkt er dog kun 0.02 større end værdien i det foreslåede udgangspunkt $(0.609, 0.031, 0.414)$, som derfor i sig selv er temmelig godt.

del af hvor mange cifre man har med i mellemregningerne), og selv om de forventede antal ikke alle er mindst 5, kan man jo alligevel godt skæve til χ^2-fordelingen med $3 \cdot (3-1) - 3 = 3$ frihedsgrader.

Alt i alt må man konkludere, at modellen med Hardy-Weinberg ligevægt ved Lolland og ved Ålandsøerne og med en blandingspopulation ved Bornholm giver en god beskrivelse af de foreliggende observationer.

3 Tosidede kontingenstabeller

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

I dette kapitel skal vi se hvorledes en bestemt art struktur i inddelingskriteriet kan afspejle sig i den statistiske model. Den pågældende struktur består i at der rent faktisk inddeles efter *to* kriterier på en gang.

Her er først en præsentation af det talmateriale der benyttes som gennemgående eksempel i dette kapitel.

Eksempel 3.1 (Hjernesvulstpatienter)
Man har klassificeret 141 hjernesvulstpatienter efter svulstens *art* (»godartet«, »ondartet« og »andet«) og *placering* i hjernevævet (»ved panden«, »ved tindingen« og »andre steder«). Resultaterne heraf fremgår af Tabel 3.1. Man er interesseret i at finde ud af om disse tal tyder på at der er en sammenhæng mellem svulstens art og dens placering.

Man kan sige at man har klassificeret $n = 141$ patienter som hørende til én af ni forskellige klasser, og at man derfor ifølge overvejelserne i Kapitel 1 kan betragte det observerede talsæt $(23, 21, \ldots, 17)$ som en observation af en multinomialfordelt stokastisk variabel. Imidlertid kan man også tænke på situationen på den måde at patienterne er klassificeret efter to kriterier på en gang, hvor hvert kriterium har tre niveauer.

3.1 Grundmodellen

Antag at vi har klassificeret n individer efter *to* kriterier. Det første kriterium har r niveauer og klasserne $A_1, A_2, \ldots, A_r$, og det andet har s niveauer og klas-

Tabel 3.1 141 hjernesvulstpatienter fordelt efter svulstens art og placering.

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

serne $B_1, B_2, \ldots, B_s$. Skematisk ser det sådan ud:

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

hvor

$$\begin{aligned} y_{ij} &= \text{antal individer i klassen } A_iB_j\ (= A_i \cap B_j), \\ y_{i\bullet} &= \sum_{j=1}^{s} y_{ij} = \text{ antal individer i klassen } A_i, \\ y_{\bullet j} &= \sum_{i=1}^{r} y_{ij} = \text{ antal individer i klassen } B_j. \end{aligned}$$

Da der er tale om at et antal individer er klassificeret i et antal klasser, benytter vi som grundmodel en *multinomialfordelingsmodel*: Den rs-dimensionale observation

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

er en observeret værdi af en rs-dimensional stokastisk variabel

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

som er multinomialfordelt med antalsparameter n og sandsynlighedsparameter

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

Størrelsen p_{ij} er sandsynligheden for at et individ udvalgt tilfældigt fra »populationen« vil tilhøre klassen A_iB_j, og den estimeres ved

$$\hat{p}_{ij} = y_{ij}/n\,. \tag{3.1}$$

3.2 Uafhængighedshypotesen

Den struktur der er i inddelingskriteriet (nemlig at der inddeles efter to kriterier på en gang) har foreløbig kun givet sig udslag i den måde de variable og parametrene er navngivet på (med index ij). Vi skal nu udlede en model der svarer til at der ikke er nogen sammenhæng mellem de to inddelingskriterier.

Den »sammenhæng« der skal være tale om, er ikke en årsagssammenhæng, men en statistisk sammenhæng. At der ikke er nogen sammenhæng mellem kriterium A og kriterium B skal betyde, at A og B i en vis forstand »virker« uafhængigt af hinanden, således at forstå, at en oplysning om, hvilken B-klasse et individ tilhører, ikke indeholder nogen information om, hvilken A-klasse individet tilhører, og omvendt. Det skal nu formaliseres i en matematisk model.

Vi indfører nogle hjælpevariable $\boldsymbol{X}_d = (X_{dA}, X_{dB})$, således at X_{dA} er navnet på den A-klasse som individ nr. d tilhører, og tilsvarende X_{dB} er navnet på den B-klasse som individ nr. d tilhører, det vil sige at

$$\boldsymbol{X}_d = (A_i, B_j) \qquad \text{betyder at} \qquad \begin{array}{l} \text{individ nr. } d \text{ tilhører} \\ A\text{-klassen } A_i \text{ og} \\ B\text{-klassen } B_j. \end{array}$$

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

$$\mathrm{P}(X_{dA} = A_i,\ X_{dB} = B_j) \quad = \quad \mathrm{P}(X_{dA} = A_i) \cdot \mathrm{P}(X_{dB} = B_j).$$

Nu er pr. definition $\mathrm{P}(X_{dA} = A_i,\ X_{dB} = B_j) = p_{ij}$, så at der ikke er nogen sammenhæng mellem A og B betyder altså at $p_{ij} = \alpha_i \beta_j$ hvor vi har sat $\alpha_i = \mathrm{P}(X_{dA} = A_i)$ og $\beta_j = \mathrm{P}(X_{dB} = B_j)$. Sammenfattende kan vi derfor sige at den matematiske formulering af antagelsen om at der ikke er nogen (statistisk) sammenhæng mellem kriterierne A og B, bliver at

$$p_{ij} \quad = \quad \alpha_i \beta_j$$

for alle i og j, hvor $\alpha_1, \alpha_2, \ldots, \alpha_r$ er ikke-negative tal der summerer til 1, og $\beta_1, \beta_2, \ldots, \beta_s$ er ikke-negative tal der summerer til 1. Udtrykt i ord går antagelsen ud på at sandsynligheden p_{ij} for på én gang at tilhøre både A_i og B_j er lig produktet af sandsynligheden α_i for at tilhøre A_i og sandsynligheden β_j for at tilhøre B_j.

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

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

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

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

Estimation af parametrene

Likelihoodfunktionen i grundmodellen er en almindelig multinomial-likelihoodfunktion:

$$L(\boldsymbol{p}) \quad = \quad \text{konstant} \cdot \prod_{i=1}^{r} \prod_{j=1}^{s} p_{ij}^{y_{ij}}$$

hvor konstanten er en multinomialkoefficient.

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

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

Det ses at L_0 er et produkt af en funktion af α-erne og en funktion af β-erne. Ifølge Sætning 1.1 antager disse to funktioner deres maksimumsværdier i hhv.

$$(\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_r) \quad = \quad \left(\frac{y_{1\bullet}}{n}, \frac{y_{2\bullet}}{n}, \ldots, \frac{y_{r\bullet}}{n} \right) \tag{3.2}$$

og

$$(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_s) \quad = \quad \left(\frac{y_{\bullet 1}}{n}, \frac{y_{\bullet 2}}{n}, \ldots, \frac{y_{\bullet s}}{n} \right) . \tag{3.3}$$

Dette er så maksimaliseringsestimaterne for parametrene. Resultatet er i øvrigt hvad man umiddelbart skulle forvente idet f.eks. sandsynligheden α_i for at tilhøre A-klassen A_i estimeres ved den observerede relative hyppighed $y_{i\bullet}/n$ af A_i.

I taleksemplet bliver $L = \text{konstant} \cdot p_{11}^{23} \, p_{12}^{21} \, p_{13}^{34} \, p_{21}^{9} \, p_{22}^{4} \, p_{23}^{24} \, p_{31}^{6} \, p_{32}^{3} \, p_{33}^{17}$. Ved at indsætte de aktuelle talværdier i (3.1), (3.2) og (3.3) fås estimaterne over de ukendte parametre, se Tabel 3.2.

Tabel 3.2 Estimaterne over grundmodellens parametre p_{ij} og uafhængighedsmodellens parametre α_i og β_j i hjernesvulsteksemplet. Tallene er sandsynligheder i procent.

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

Test for uafhængighed

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

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

hvor $\hat{y}_{ij} = n\hat{\alpha}_i\hat{\beta}_j = y_{i\cdot}y_{\cdot j}/n$ er det »forventede« antal individer i klassen A_iB_j under uafhængighedshypotesen. Dermed bliver

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

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

De »forventede« antal i hjernesvulsteksemplet er vist i Tabel 3.3; herudfra

Tabel 3.3 Den »forventede« fordeling af 141 hjernesvulstpatienter under forudsætning af uafhængighed mellem svulstens art og placering.

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

får man at

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

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

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

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

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

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

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

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

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

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

3.3 Jævnføring med andre tilsvarende modeller

Den læser der har studeret Afsnit 1.2 om sammenligning af multinomialfordelinger, vil måske have bemærket, at de dér præsenterede metoder har store ligheder med dem i indeværende kapitel. Vi kan opregne nogle af lighederne:

1. Der foreligger nogle observerede *antal* y_{ij} anbragt i et tosidet skema.
2. Man udregner nogle »forventede« antal $\hat{y}_{ij}$ efter opskriften *rækkesum* gange *søjlesum* divideret med *totalsum*.
3. Man udregner en teststørrelse $-2\ln Q_{\text{obs}} = \sum y \ln(y/\hat{y})$.
4. Man sammenligner $-2\ln Q_{\text{obs}}$ med χ^2-fordelingen med $(r-1)(s-1)$ frihedsgrader.

Selv om man *foretager sig* det samme i de to tilfælde, er det imidlertid på grundlag af to forskellige modeller:[1]

- I det ene tilfælde (dette kapitel) klassificerer man nogle individer efter *to* kriterier, og opgaven er da at undersøge om der er en sammenhæng mellem disse to kriterier.
- I det andet tilfælde (Afsnit 1.2) er individerne på forhånd delt ind i nogle grupper inden de klassificeres efter *et* kriterium. Opgaven er da at undersøge om der er forskel på grupperne (med hensyn til hvordan gruppernes individer fordeles på klasserne).

Om man skal benytte den ene eller den anden model, er således et spørgsmål om hvorledes man har designet det forsøg der har leveret talmaterialet. I eksemplet i dette kapitel sagde vi at det handlede om at man havde taget 141 hjernesvulstpatienter og klassificeret dem efter *to* kriterier; derved blev det et eksempel der illustrerede dette kapitels model og metoder. Hvis det derimod

[1] De to modeller er dog nært beslægtede; hvis man i dette kapitels model betinger med søjlesummerne, dvs. betinger med at $Y_{\bullet 1} = n_1, Y_{\bullet 2} = n_2, \ldots, Y_{\bullet s} = n_s$, så får man modellen i Afsnit 1.2, og uafhængighedshypotesen overføres til Afsnit 1.2s H_0.

havde handlet om at man havde taget 38 patienter med svulst i panden, 28 med svulst i tindingen og 75 hvor svulsten ikke var lokaliseret til pande eller tinding, og dernæst klassificeret disse patienter efter svulstens art, så havde det været et Afsnit 1.2-eksempel.

3.4 Opgaver

Opgave 3.1 (Hår- og øjenfarve)
Ved en sundhedsundersøgelse af 283 piger i St. Clement Street skole i Aberdeen blev hår- og øjenfarve observeret med et resultat som vist i nedenstående tabel. Viser dette materiale en sammenhæng mellem hårfarve og øjenfarve?

		Hårfarve			
		lys	rød	neutral	mørk
Øjenfarve	blå	30	4	27	6
	lys	30	5	28	11
	neutral	21	7	40	22
	mørk	6	3	23	20

4 Stikord