STATISTIKKURSUS
NAT-BAS, efterår 2007

Jørgen Larsen
10. december 2007

Besvarelse af opgave 1 i Afleveringsopgavesæt 3

Opgave 1. *Et nedbørsdøgn er et døgn med nedbør på 0.1 mm eller mere. Nedbøren i et døgn angives ved den højde hvortil vandet ville stige hvis det ikke kunne løbe væk eller fordampe, beregnet som et gennemsnit over 600 stationer. I 1988 og 1989 var der henholdsvis 173 og 146 nedbørsdøgn, og deres fordelinger på kvartaler var*

	januar	*april*	*juli*	*oktober*
1988	*58*	*25*	*48*	*42*
1989	*46*	*26*	*34*	*40*

1. *Undersøg om der er signifikant forskel på årene 1988 og 1989 hvad angår den kvartalsvise fordeling af nedbørsdøgn.*
2. *De såkaldte normalværdier for det kvartalsvise antal nedbørsdøgn beregnes som et gennemsnit af tallene for årene 1931-60. Ud fra disse er beregnet følgende sandsynligheder for den kvartalsvise fordeling af nedbørsdøgn:*

januar	*april*	*juli*	*oktober*
0.239	*0.207*	*0.258*	*0.296*

Undersøg (idet vi opfatter disse sandsynligheder som kendte tal) om fordelingen af nedbørsdøgn i årene 1988 og 1989 adskiller sig signifikant fra fordelingen beregnet ud fra normalværdierne.

Delopgave 1:

I hvert af de to år 1988 og 1989 har man fordelt et vist antal nedbørsdøgn på kvartaler, dvs. for hvert af årene er der tale om en multinomialfordelingssituation med fire klasser. Den første delopgave er derfor en »sammenligning af to multinomialfordelinger«.

Vi kan indføre lidt matematisk notation: det observerede antal nedbørsdøgn i kvartal i og år j vil vi betegne y_{ij}; her antager j værdierne 1988 og 1989, og i værdierne januar, april, juli og oktober. Med den sædvanlige punktnotation (jf. bogens afsnit 14.2) er $y_{\cdot j} = y_{\text{januar},j} + y_{\text{april},j} + y_{\text{juli},j} + y_{\text{oktober},j}$ årstotalen for år j ($j = 1988, 1989$), og $y_{i\cdot} = y_{i,1988} + y_{i,1989}$ er kvartalstotalerne for toårsperioden (i = januar, april, juli, oktober). Desuden er $n = y_{\cdot\cdot}$ det samlede antal nedbørsdøgn i toårsperioden. Under antagelsen om at de to år har samme fordeling af nedbørsdøgn, er de »forventede« antal nedbørsdøgn

Tabel 1 Observerede og forventede antal nedbørsdøgn

observerede værdier (y_{ij})				forventede værdier ($\widehat{y}_{ij}$)			
	1988	1989	i alt ($y_{i\cdot}$)		1988	1989	i alt
januar	58	46	104	januar	56.4	47.6	104
april	25	26	51	april	27.7	23.3	51
juli	48	34	82	juli	44.5	37.5	82
oktober	42	40	82	oktober	44.5	37.5	82
i alt ($y_{\cdot j}$)	173	146	319	i alt	173	146	319

tallene $\widehat{y}_{ij} = y_{i\cdot}y_{\cdot j}/n$. I tabel 1 ses de observerede og de forventede værdier. Hypotesen om samme fordeling de to år, testes med teststørrelsen

$$-2\ln Q = 2\sum \text{obs.antal} \cdot \ln\frac{\text{obs.antal}}{\text{forv.antal}} = 2\sum_{j=1988}^{1989}\sum_{i=\text{januar}}^{\text{oktober}} y_{ij}\ln\frac{y_{ij}}{\widehat{y}_{ij}}$$

Ved indsættelse af de faktiske værdier finder man $-2\ln Q = 1.57$; denne værdi skal sammenlignes med χ^2-fordelingen med $(4-1)(2-1) = 3$ frihedsgrader; testsandsynligheden bliver dermed ca. 0.67, således at hypotesen om samme fordeling for de to år stemmer fint overens med de foreliggende observationer.

Delopgave 2:

Nu skal vi sammenligne data for årene 1988 og 1989 med et givet sæt normalværdier for perioden 1931-60. Problemstillingen er en generalisation af bogens afsnit 2.2 til multinomialfordelingen (i stedet for binomialfordelingen): Vi har de to observerede talsæt $\begin{pmatrix} y_{1,1988} \\ y_{2,1988} \\ y_{3,1988} \\ y_{4,1988} \end{pmatrix} = \begin{pmatrix} 58 \\ 25 \\ 48 \\ 42 \end{pmatrix}$ og $\begin{pmatrix} y_{1,1989} \\ y_{2,1989} \\ y_{3,1989} \\ y_{4,1989} \end{pmatrix} = \begin{pmatrix} 46 \\ 26 \\ 34 \\ 40 \end{pmatrix}$ der ifølge delopgave 1 stammer fra to multinomialfordelinger der har samme sandsynlighedsparameter $\boldsymbol{p} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix}$. Vi skal teste hypotesen $\boldsymbol{p} = \boldsymbol{p}_0$ hvor $\boldsymbol{p}_0 = \begin{pmatrix} p_{01} \\ p_{02} \\ p_{03} \\ p_{04} \end{pmatrix} = \begin{pmatrix} 0.239 \\ 0.207 \\ 0.258 \\ 0.296 \end{pmatrix}$. – Med resultatet fra delopgave 1 in mente kunne man alternativt opfatte situationen på den måde at der foreligger ét observeret talsæt $\begin{pmatrix} y_{1\cdot} \\ y_{2\cdot} \\ y_{3\cdot} \\ y_{4\cdot} \end{pmatrix} = \begin{pmatrix} 104 \\ 51 \\ 82 \\ 82 \end{pmatrix}$ fra en multinomialfordeling med sandsynlighedsparameter $\boldsymbol{p}$, og at vi i denne model skal teste hypotesen $\boldsymbol{p} = \boldsymbol{p}_0$.

Likelihoodfunktionen svarende til den første formulering er

$$L_1(\boldsymbol{p}) = \prod_{j=1}^{2} \binom{y_{\cdot j}}{y_{1j}\, y_{2j}\, y_{3j}\, y_{4j}} p_1^{y_{1j}} p_1^{y_{2j}} p_3^{y_{3j}} p_4^{y_{4j}} = \text{konstant}_1 \cdot p_1^{y_{1\cdot}} p_1^{y_{2\cdot}} p_3^{y_{3\cdot}} p_4^{y_{4\cdot}}$$

hvor konstanten er produktet af de to multinomialkoefficienter, og likelihoodfunktionen svarende til den anden formulering er

$$L_2(\boldsymbol{p}) = \text{konstant}_2 \cdot p_1^{y_{1\cdot}} p_1^{y_{2\cdot}} p_3^{y_{3\cdot}} p_4^{y_{4\cdot}}$$

hvor konstanten er en multinomialkoefficient. Det ses at de to likelihoodfunktioner er ens pånær en konstant faktor. Maksimaliseringsestimatet er det samme i begge tilfælde, nemlig

$$\widehat{\boldsymbol{p}} = \begin{pmatrix} \widehat{p}_1 \\ \widehat{p}_2 \\ \widehat{p}_3 \\ \widehat{p}_4 \end{pmatrix} = \begin{pmatrix} y_{1\cdot}/n \\ y_{2\cdot}/n \\ y_{3\cdot}/n \\ y_{4\cdot}/n \end{pmatrix}$$

Som teststørrelse for at teste hypotesen skal man bruge $-2\ln Q$-teststørrelsen der (såvel i den første som i den anden formulering) er

$$\begin{aligned}
-2\ln Q &= -2\ln \frac{L(\boldsymbol{p}_0)}{\max L(\boldsymbol{p})} \\
&= -2\ln \frac{\text{konstant } p_{01}^{y_{1\cdot}} p_{02}^{y_{2\cdot}} p_{03}^{y_{3\cdot}} p_{04}^{y_{4\cdot}}}{\text{konstant } (y_{1\cdot}/n)^{y_{1\cdot}} (y_{2\cdot}/n)^{y_{2\cdot}} (y_{3\cdot}/n)^{y_{3\cdot}} (y_{4\cdot}/n)^{y_{4\cdot}}} \\
&= -2\ln\left(\left(\frac{np_{01}}{y_{1\cdot}}\right)^{y_{1\cdot}} \left(\frac{np_{02}}{y_{2\cdot}}\right)^{y_{2\cdot}} \left(\frac{np_{03}}{y_{3\cdot}}\right)^{y_{3\cdot}} \left(\frac{np_{04}}{y_{4\cdot}}\right)^{y_{4\cdot}} \right) \\
&= 2\left(y_{1\cdot} \ln \frac{y_{1\cdot}}{np_{01}} + y_{2\cdot} \ln \frac{y_{2\cdot}}{np_{02}} + y_{3\cdot} \ln \frac{y_{3\cdot}}{np_{03}} + y_{4\cdot} \ln \frac{y_{4\cdot}}{np_{04}} \right)
\end{aligned}$$

altså den »sædvanlige« opskrift $-2\ln Q = 2\sum \text{obs.antal} \cdot \ln \frac{\text{obs.antal}}{\text{forv.antal}}$. Antallet af frihedsgrader for den χ^2-fordeling som $-2\ln Q$-størrelsen skal sammenlignes med, bestemmes som antal parametre i grundmodellen minus antal parametre under hypotesen, dvs. $(4-1)-0=3$. Man udregner $-2\ln Q$ til

$$\begin{aligned}
-2\ln Q &= 2\left(104 \ln \frac{104}{319 \cdot 0.239} + 51 \ln \frac{51}{319 \cdot 0.207} + 82 \ln \frac{82}{319 \cdot 0.258} + 82 \ln \frac{82}{319 \cdot 0.296} \right) \\
&= 14.5
\end{aligned}$$

svarende til en testsandsynlighed på ca. 0.2%, dvs. hypotesen må forkastes. Fordelingen af nedbørsdøgn i årene 1988-89 adskiller sig således signifikant fra normalværdierne fra 1931-60.