Matematisk Statistik: Modelbaseret Inferens

Gamle eksamensopgaver

Jens Ledet Jensen

Forår 2019 opgave 1

Efter det sidste Europavalg i 2014 var 5 af de valgte medlemmer fra Danmark kvinder, og 8 var mænd; i Sverige var 11 ud af 20 medlemmer kvinder og i Finland var det 7 ud af 13. Undersøg ved et test, om kvindernes chancer for at blive medlem i Europaparlamentet er den samme i de tre lande. Giv korte svar:

(a) Hvilket test bruger du, og hvilken fordeling antages for teststørrelsen under nul hypotesen?

(b) Angiv teststørrelsen og p-værdien.

(c) Hvad er den faglige konklusion?

Besvarelse

Lad K_{DK}, K_S og K_F være antal valgte kvinder i de tre lande. Jeg bruger modellen

$$K_{DK} \sim \text{binom}(13, p_{DK})$$
$$K_S \sim \text{binom}(20, p_S)$$
$$K_F \sim \text{binom}(13, p_F)$$

og vil teste hypotesen $p_{DK} = p_S = p_F$

For at bruge Resultat 1.4 betragter jeg multinomialfordelinger:

$$(K_{DK}, 13 - K_{DK}) \sim \text{binom}(13, (p_{DK}, 1 - p_{DK}))$$

med tilsvarende for de to andre lande

(a) Som teststørrelse bruger jeg G fra Resultat 1.4, og den approksimative fordeling er $\chi^2((3-1)(2-1)) = \chi^2(2)$ under forudsætning om at alle forventede er ≥ 5

Besvarelse

(b) Her kommer R-kørsel:

```
obs=rbind(c(5,8),c(11,9),c(7,6))
ex=outer(rowSums(obs),colSums(obs))/sum(obs)
gTest=2*sum(obs*log(obs/ex))
pval=1-pchisq(gTest,(dim(obs)[1]-1)*(dim(obs)[2]-1))
list(Forventede=ex,G=gTest,Pvaerdi=pval)

$Forventede
     [,1] [,2]
[1,]  6.5  6.5
[2,] 10.0 10.0
[3,]  6.5  6.5

$G
[1] 0.975921

$Pvaerdi
[1] 0.6138771
```

Besvarelse

(b) Teststørrelsen er $G = 0.976$ og p-værdien er 0.62

(c) Data strider ikke mod at antage samme andel af kinder i de tre lande

Besvarelse

Bonusspørgsmål: Angiv, under antagelse af samme sandsynlighed for kvinde i de tre lande et 95%-konfidensinterval for sandsynligheden

Mode $K \sim \text{binom}(46, p)$, K er antal kvinder i de tre lande, observeret til 23
Formel for konfidensintervallet står i afsnit 1.2 (skjulte punkt) i webbogen og kan i R beregnes som

```
prop.test(23,46)$conf.int
```

```
[1] 0.3611894 0.6388106
attr(,"conf.level")
[1] 0.95
```

Konfidensintervallet er således $[0.36, 0.64]$. Intervallet er bredt, da vi kun har observeret $n = 46$. Intervallet indeholder værdien 0.5 svarende til at der er lige stor andel af mænd som kvinder.

Sceneskift

Reeksamen 2018, opgave 4

Wafers er siliciumskiver, der danner råmateriale til elektroniske chips. Wafers produceres i renrum, da hver mikropartikel, der lander på overfladen, mindsker udbyttet af chippene. Det kan dog ikke helt undgås, at wafers kontamineres af partikler. Når man betragter antallet af partikler på en wafer, antages tit, at denne er Poisson fordelt. Er denne antagelse altid rimelig? Tabellen nedenunder viser observerede partikelantal for 100 små kvadratiske (2cm x 2cm) udsnit på wafers, samt det forventede antal beregnet fra en Poisson fordeling fittet til data.

	antal partikler per kvadratisk udsnit										
	0	1	2	3	4	5	6	7	8	9	>=10
observeret	25	24	19	16	6	4	1	2	2	1	0
forventet	13.67	27.20	27.07	17.95	8.93	3.55	1.18	0.34	0.08	0.02	0.00

(a) Gør rede for, hvordan det forventede antal udsnit uden partikler er blevet estimeret til 13.67.
(b) Lav et goodness of fit test for hypotesen, at antal partikler per kvadrat er Poisson fordelt. Benyt signifikansniveauet $\alpha = 0.05$.
(c) Fortolk resultatet af hypotesetestet. Tyder data på, at partiklerne sætter sig på overfladen af en wafer uafængigt af hinanden? Hvis nej, hvordan afviger observationerne fra det, man ville forvente?

Besvarelse

Lad $(A_1, \ldots, A_{11})$ være antallene i de 11 kasser. Jeg bruger modellen (M_0)

$(A_1, \ldots, A_{11}) \sim \text{multinom}(100, (\pi_1, ldots, \pi_{11})$

hvor $\pi_j \geq 0$ og $\pi_1 + \cdots + \pi_{11} = 1$.

Jeg vil teste hypotesen $\pi_j = \text{dpois}(j-1, \lambda)$, $j = 1, \ldots, 10$,

$\pi_{11} = 1 - \pi_1 - \cdots - \pi_{10}$, $\lambda \geq 0$.

(a) Som skøn over λ benytter jeg gennemsnittet af de 100 værdier (MSRR porposition 6.1.2). Denne beregnes nedenfor til $\hat{\lambda} = 1.99$. Det forventede antal i kasse 1 er derfor $100 \cdot \text{dpois}(0, 1.99) = 13.6695$

Besvarelse

(b) For at lave goodness of fit test, og have alle de forventede ≥ 5, slår jeg de sidste 6 kasser sammen. Den observerede værdi for disse bliver nu 10 og den forventede værdi bliver 5.18.

G-teststørrelsen fra Resultat 1.1 beregnes til 15.44, og p-værdien fra en $\chi^2(6-1-1)$-fordeling er 0.0039.

Da pværdien er langt under signifikansniveauet forkaster jeg hypotesen om at data stammer fra en poissonfordeling.

(c) Hvis partiklerne sætter sig tilfældigt uafhængigt af hinanden burde antallet i kvadraterne være poissonfordelt. Da vi forkaster poissonfordelingen holder denne antagelse ikke.

Den mest markante afvigelse er at der er for mange områder uden nogen partikler i forhold til hvad vi forventer. Der ser også ud til at være lidt for mange områder med rigtig mange partikler.

Besvarelse

```
a=c(25, 24, 19, 16, 6, 4, 1, 2, 2, 1, 0)
n=sum(a)
lambdahat=sum(a*c(0:10))/n

forventet=n*dpois(c(0:9),lambdahat)
forventet=c(forventet,n-sum(forventet))

forventet
 [1] 13.669542545 27.202389664 27.066377715 17.954030551  8.932130199
 [6]  3.554987819  1.179070960  0.335193030  0.083379266  0.018436082
[11]  0.004462168

lambdahat
[1] 1.99
```

Besvarelse

```
a1=c(a[1:5],sum(a[6:11]))
ex1=c(forventet[1:5],sum(forventet[6:11]))

rbind(a1,round(ex1,2))
    [,1] [,2]  [,3]  [,4] [,5]  [,6]
a1 25.00 24.0 19.00 16.00 6.00 10.00
   13.67 27.2 27.07 17.95 8.93  5.18

G=2*sum(a1*log(a1/ex1))

c(G1-pchisq(G,6-1-1))
[1] 15.437703709  0.003874401
```

Sceneskift

Forår 2020 opgave 3

Data i tabellen nedenfor viser for 200 kvinder, hvor mange der er blevet gravide efter første, andet og tredje forsøg med ivf-behandling, og hvor mange der ikke er blevet gravide i de tre forsøg (data er opdigtede).

	Første	Andet	Tredje	Ikke-gravide
Antal	29	21	21	129

(a) Opskriv en statistisk model for antallet af kvinder i de fire kategorier Første, Andet, Tredje og Ikke-gravide.
(b) Betragt hypotesen (fast-rate-hypotesen), at sandsynligheden for at blive gravid er den samme i hvert forsøg. Hvis θ er sandsynligheden for at blive gravid i det enkelte forsøg, er sandsynligehden for at blive gravid i det første forsøg ?, gravid i det andet forsøg $(1-\theta)\theta$, gravid i det tredje forsøg $(1-\theta)^2\theta$, og sandsynligheden for ikke at blive gravid $(1-\theta)^3$. Opstil likelihoodfunktionen under hypotesen og find et udtryk for $\hat{\theta}$. Eftervis, at for data i tabellen ovenfor er $\hat{\theta} = 0.1363$.

(c) Lav et test for fast-rate-hypotesen.

Besvarelse

(a) Lad (a_1, a_2, a_3, a_4) være antallene i de fire kasser. Vi benytter modellen

$$(A_1, A_2, A_3, A_4) \sim \text{multinom}(200, (\pi_1, \pi_2, \pi_3, \pi_4))$$
$$\pi_j \geq 0, \ \pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$$

(b) Likelihoodfunktionen under hypotesen er

$$L(\theta) = \binom{200}{a} \theta^{a_1} ((1-\theta)\theta)^{a_2} ((1-\theta)^2\theta)^{a_3} ((1-\theta)^3)^{a_4}$$
$$= \binom{200}{a} \theta^{a_1+a_2+a_3} (1-\theta)^{a_2+2a_3+3a_4}$$

Denne har samme strukur som likelihoodfunktionen i en binomialmodel og fra bevis af Propsition 6.1.1 i MSRR får vi

$$\hat{\theta} = \frac{a_1+a_2+a_3}{(a_1+a_2+a_3)+(a_2+2a_3+3a_4)}$$

Med data indsat giver dette $\hat{\theta} = 0.1362764$ (se R-kørsel)

Besvarelse

(c) Vi benytter G-testet fra Resultat 1.1 i webbogen. Fra R-kørsel får vi

forventede: $27.26, 23.54, 20.33, 128.87$, alle ≥ 5, så vi kan bruge $\chi^2(4-1-1)$-fordelingen til beregning af p-værdi

$G = 0.4159$ og p-værdien $= 0.8123$

Da p-værdien er langt over 0.05 siger vi, at data ikke strider mod hypotesen om fast rate.

Besvarelse

```
a=c(29,21,21,129)
thetahat=sum(a[-4])/sum(a*c(1,2,3,3))
thetahat
[1] 0.1362764

ex=200*c(thetahat,thetahat*(1-thetahat),thetahat*(1-thetahat)^2,(1-t
round(ex,2)
[1]  27.26  23.54  20.33 128.87

G=2*sum(a*log(a/ex))
c(G,1-pchisq(G,4-1-1))
[1] 0.4158731 0.8122586
```

Sceneskift

MSRR 10.5

I general survey (CSS2002) er personer spurgt om hvor glade de er og hvem de har stemt på til præsidentvalg i 2000.

	Bush	Gore	Nader
Not too happy	29	46	5
Pretty happy	245	233	17
Very happy	164	123	7

a) State the hypothesis of interest.
b) Perform the test using the chisq.test command in R.
c) R returns a warning message. Compute the expected counts for each cell to see why.
(JLJ b+c) Gennemfør G-test for hypotesen
d) Perform a permutation test of independence.
e) Perform the test using Fisher's exact test.
f) Compare the P-values for the three approaches. Would you come to the same conclusion regardless of which test you used?

Besvarelse

869 personer er blevet spurgt om deres grad af happiness (Not too happy/Pretty happy/Very happy = 1/2/3) og om hvem de har stemt på (Bush/Gore/Nader = 1/2/3). Lad

a_{ij} være antal blandt de 869 med svarene (i, j), $i, j = 1, 2, 3$

Model: $(A_{11}, A_{12}, A_{13}, A_{21}, A_{22}, A_{23}, A_{31}, A_{32}, A_{33}) \sim$ multinom$(137, (\pi_{11}, \pi_{12}, \pi_{13}, \pi_{21}, \pi_{22}, \pi_{23}, \pi_{31}, \pi_{32}), \pi_{33}))$,
$\pi_{ij} \geq 0$ $\sum_{ij} \pi_{ij} = 1$

(a) Vi ønsker at teste om svar på de to spørgsmål er uafhængige. Hypotesen er, at der eksisterer α, β således at

$$\pi_{ij} = \alpha_i \beta_j \ i, j = 1, 2, 3$$

hvor $\alpha_1 + \alpha_2 + \alpha_3 = 1$ og $\beta_1 + \beta_2 + \beta_3 = 1$

Besvarelse

(b+c)

Fra R-output set at C-teststørrelsen er 11.3 med approksimativ p-værdi på 0.023. Vi er dog usikre på denne p-værdi da en af cellerne har en forventet værdi på 2.7 under hypotesen om uafhængighed.

Vi vil derfor også være usikre på den approksimative pværdi for G-testet. G-teststørrelsen fra Resultat 1.4 er 11.15 og den approksimative p-værdi fra en $\chi^2(4)$-fordeling er 0.025

Besvarelse

```
obs=rbind(c(29,46,5),c(245,233,17),c(164,123,7))

# C-test
chisq.test(obs)
data:  obs
X-squared = 11.3, df = 4, p-value = 0.02339
Advarselsbesked:
I chisq.test(obs) : Chi-squared approximation may be incorrect

chisq.test(obs)$expected
          [,1]      [,2]      [,3]
[1,]  40.32221  37.00806  2.669735
[2,] 249.49367 228.98734 16.518987
[3,] 148.18412 136.00460  9.811277
```

Besvarelse

```
# G-test
ex=outer(rowSums(obs),colSums(obs))/sum(obs)
gTest=2*sum(obs*log(obs/ex))
pval=1-pchisq(gTest,(dim(obs)[1]-1)*(dim(obs)[2]-1))
list(Forventede=ex,G=gTest,Pvaerdi=pval)

$Forventede
          [,1]      [,2]      [,3]
[1,]  40.32221  37.00806  2.669735
[2,] 249.49367 228.98734 16.518987
[3,] 148.18412 136.00460  9.811277

$G
[1] 11.14512

$Pvaerdi
[1] 0.02498055
```

Besvarelse

(d) Vi bruger kode fra afsnit 1.8 i webbogen til beregning af simulerede (betingede) p-værdi. Denne bliver 0.030.

(e) Fra R-kørsel ses at Fishers eksakte test giver en p-værdi på 0.01934

(f) P-værdien fra Fishers eksakte test ligger lidt under den simulerede værdi på 0.030 (95%-konfideninterval 0.027-0.034) (bygger på samme betingede fordeling, men bruger forskellig teststørrelse). De to approksimative værdier ligner hinanden og er nærmest midt mellem den eksakte og den simulerede værdi. Alle værdier fortæller cirka den samme historie.

Med et signifikansniveau på 0.05 vil de forskellige test give samme resultat for data her.

Besvarelse

```
obs=rbind(c(29,46,5),c(245,233,17),c(164,123,7))

r=dim(obs)[1]; k=dim(obs)[2]
rs=rowSums(obs)
cs=colSums(obs)
h=rep(c(1:r),rs)
m=rep(c(1:k),cs)

Gfct=function(Amat){
ex=outer(rowSums(Amat),colSums(Amat))/sum(Amat)
A1=ifelse(Amat==0,1,Amat)
return(2*sum(Amat*log(A1/ex)))
}
Gobs=Gfct(obs)

nSim=10^4-1
tval=rep(0,nSim)
```

Besvarelse

```
for (i in 1:nSim){
 msamp=sample(m)
 tabelsamp=table(h,msamp)
 tval[i]=Gfct(tabelsamp)
}

pval=(1+sum(tval>=Gobs))/(1+nSim)
c(G=Gobs,pværdi=pval)
       G   pværdi
11.14512  0.03000

# Fishers test
fisher.test(obs)

data:  obs
p-value = 0.01934
```

Sceneskift

Opgave