Introducerende Statistik og Dataanalyse med R

Homogenitetstest

Jens Ledet Jensen

I DAG

Sammenligne to (eller flere) sæt multinomialfordelte data

Dagens spørgsmål: er der lige mange der dropper ud af studiet på matematik som på matematik-økonomi og som på datavidenskab ?

Søndagsavisen

Fedmetallene

Periode	Antal på session	Antal med BMI > 30	procent
1. halvdel 2003	11527	796	6.9%
2. halvdel 2003	12259	825	6.7%

Har søndagsavisen ret ?

Umiddelbare formulering: er der forskel mellem de to halvår ?

Formulering

Hvad mener vi med spørgsmålet: er der forskel mellem de to halvår ?

Vi har procentvis observeret færre med højt BMI i andet halvår,

men er dette blot en tilfældighed ?

eller: er dette udtryk for forskel i bagvedliggende populationer ?

Hvad betyder egentligt "bagvedliggende population"?

Som et tankeeksperiment kan vi godt forestille os de 12000 unge skiftet ud med 12000 andre unge:

forskelle i genetik og opvækst vil give et andet antal med BMI>30

Vender nu spørgsmålet om:

er data i overensstemmelse med en hypotese om ingen forskel ?

hypotese: der er samme andel med højt BMI i de to halvår

Statistisk model

Data:

	Kasse 1	Kasse 2	Total
Gruppe 1	a_1	$n_1 - a_1$	n_1
Gruppe 2	a_2	$n_2 - a_2$	n_2
Sum	$a_\bullet$	$n - a_\bullet$	n

Statistisk Model (model M_0):

$$A_1 \sim \text{binomial}(n_1, p_1), \qquad A_2 \sim \text{binomial}(n_2, p_2)$$

Hypotese: $p_1 = p_2$

Svarer til model M_1: $A_1 \sim \text{binomial}(n_1, p)$, $A_2 \sim \text{binomial}(n_2, p)$

Parameterskøn under M_0: $\hat{p}_1 = \frac{A_1}{n_1}$, $\hat{p}_2 = \frac{A_2}{n_2}$

Parameterskøn under M_1: $\hat{p} = \frac{A_\bullet}{n}$ $\quad A_\bullet = A_1 + A_2$

Likelihoodratio test

Likelihoodfunktion: $L(p_1,p_2) = \binom{n_1}{A_1} p_1^{A_1}(1-p_1)^{n_1-A_1} \binom{n_2}{A_2} p_2^{A_2}(1-p_2)^{n_2-A_2}$

$$Q = \frac{\max_p L(p,p)}{\max_{p_1,p_2} L(p_1,p_2)} = \frac{\hat{p}^{A_1}(1-\hat{p})^{n_1-A_1}\hat{p}^{A_2}(1-\hat{p})^{n_2-A_2}}{(\frac{A_1}{n_1})^{A_1}(1-\frac{A_1}{n_1})^{n_1-A_1}(\frac{A_2}{n_2})^{A_2}(1-\frac{A_2}{n_2})^{n_2-A_2}}$$

$$= \frac{1}{(\frac{A_1}{n_1\hat{p}})^{A_1}(\frac{n_1-A_1}{n_1(1-\hat{p})})^{n_1-A_1}(\frac{A_2}{n_2\hat{p}})^{A_2}(\frac{1-A_2}{n_2(1-\hat{p})})^{n_2-A_2}}$$

Teststørrelse: $G = -2\log Q$:

$$G = 2\sum_{4 \text{ celler}} \text{observeret} \cdot \log\left(\frac{\text{observeret}}{\text{forventet}}\right)$$

Forventede: $e_{11} = n_1\hat{p}$, $e_{12} = n_1(1-\hat{p})$, $e_{21} = n_2\hat{p}$, $e_{22} = n_2(1-\hat{p})$

Kritiske værdier: Q lille eller $G = -2\log(Q)$ stor
Approksimation: $G \approx \chi^2(1)$

Sessionstal

Observeret

	BMI $<$ 30	BMI $>$ 30	Total
2003-1	11527	796	12323
2003-2	12259	825	13084
Sum	23786	1621	25407

Forventet

	BMI $<$ 30	BMI $>$ 30
2003-1	11536.78	786.22
2003-2	12249.22	834.78

$$e_{11} = 12323 \cdot \frac{23786}{25407} = 11536.78$$

$$\begin{aligned} G = 2\Big[& 11527 \cdot \log\left(\frac{11527}{11536.78}\right) + 796 \cdot \log\left(\frac{796}{786.2236}\right) \\ & + 12259 \cdot \log\left(\frac{12259}{12249.22}\right) + 825 \cdot \log\left(\frac{825}{834.7764}\right)\Big] = 0.2520978 \end{aligned}$$

Prøv webbogen

Gå til den sidste "Beregning i R" i afsnit 1.6

Erstat Obs=rbind(c(6,28,38),c(13,56,10)) med

```
Obs=rbind(c(11527,796),c(12259,825))
```

Sceneskift

Homogenitetstest er vist for 2×2 tabel

Næste: Homogenitetstest for generel $r \times k$ tabel

Generel formulering

A_{ij}: Antal i kasse j i den i'te række

	kasse				
	1	2	$\cdots$	k	sum
række 1	a_{11}	a_{12}	$\cdots$	a_{1k}	n_1
række 2	a_{21}	a_{22}	$\cdots$	a_{2k}	n_2
$\vdots$					
række r	a_{r1}	a_{r2}	$\cdots$	a_{rk}	n_r
sum	$a_{\bullet 1}$	$a_{\bullet 2}$	$\cdots$	$a_{\bullet k}$	n

Række i: $\quad a_{i*} = (a_{i1}, a_{i2}, \ldots, a_{ik})$

Søjlesum: $\quad a_{\bullet j} = a_{1j} + a_{2j} + \cdots + a_{rj}$

Homogenitetshypotesen

r grupper eller rækker, $\quad a_{i*} = (a_{i1}, \ldots, a_{ik}), \quad \pi_{i*} = (\pi_{i1}, \ldots, \pi_{ik})$

Model M_0 :

$$\begin{aligned} A_{1*} &\sim \mathrm{multinomial}(n_1, \pi_{1*}) \\ A_{2*} &\sim \mathrm{multinomial}(n_2, \pi_{2*}) \\ &\vdots \\ A_{r*} &\sim \mathrm{multinomial}(n_r, \pi_{r*}) \end{aligned}$$

Rækker er uafhængige, sandsynligheder er vilkårlige i hver række

Homogenitetshypotesen: alle grupperne (eller rækkerne) har det samme sæt sandsynligheder $(\pi_1, \pi_2, \ldots, \pi_k)$ (model M_1):

Eller: $\pi_{1*} = \pi_{2*} = \cdots = \pi_{r*}$

Homogenitetshypotesen

Fulde model (M_0):

	kasse				
	1	2	$\cdots$	k	sum
række 1	π_{11}	π_{12}	$\cdots$	π_{1k}	1
række 2	π_{21}	π_{22}	$\cdots$	π_{2k}	1
$\vdots$					
række r	π_{r1}	π_{r2}	$\cdots$	π_{rk}	1

Hypotese (model M_1):

	kasse				
	1	2	$\cdots$	k	sum
række 1	π_1	π_2	$\cdots$	π_k	1
række 2	π_1	π_2	$\cdots$	π_k	1
$\vdots$					
række r	π_1	π_2	$\cdots$	π_k	1

Estimation

Vi kender estimaterne i en multinomialmodel med frie parametre: under M_0: $\hat{\pi}_{ij} = \frac{a_{ij}}{n_i}$

Under model M_1 er likelihoodfunktionen

$$L_{M_1}(\pi_1, \ldots \pi_k) = \prod_i \binom{n_i}{a_{i*}} \pi_1^{a_{i1}} \cdots \pi_k^{a_{ik}}$$

$$= \frac{\prod_i \binom{n_i}{a_{i*}}}{\binom{n}{a_{\bullet 1} \ldots, a_{\bullet k}}} \binom{n}{a_{\bullet 1} \ldots, a_{\bullet k}} \pi_1^{a_{\bullet 1}} \cdots \pi_k^{a_{\bullet k}}$$

Estimation af $(\pi_1 \ldots \pi_k)$ under model M_1 svarer til at bruge

$$(A_{\bullet 1} \ldots, A_{\bullet k}) \sim \text{multinom}(n, (\pi_1 \ldots \pi_k))$$

Forventede

Under hypotesen estimeres den fælles π ved de observerede frekvenser for de k kasser

$$\hat{\pi}_j = \frac{\sum_i x_{ij}}{n} = \frac{x_{\bullet j}}{n}, \; j = 1, \ldots, k, \qquad n = n1 + \cdots + n_k$$

De forventede i kasse j for den i'te gruppe: $\qquad e_{ij} = n_i \hat{\pi}_j = \frac{n_i x_{\bullet j}}{n}$

antal i gruppe i gange skøn over sandsynlighed for kasse j

rækkesum gange søjlesum divideret med samlede antal

Likelihood ratio

Likelihoodratio teststørrelse Q:

$$Q = \frac{\max_{M_1} L}{\max_{M_0} L} = \frac{\prod_i \binom{n_i}{a_{i*}} \hat{\pi}_1^{a_{ij}} \cdots \hat{\pi}_k^{a_{ij}}}{\prod_i \binom{n_i}{a_{i*}} \hat{\pi}_{i1}^{a_{ij}} \cdots \hat{\pi}_{ik}^{a_{ik}}}$$

$$= \prod_i \prod_j \frac{\hat{\pi}_j^{a_{ij}}}{\hat{\pi}_{ij}^{a_{ij}}} = \prod_i \prod_j \frac{1}{\left(\frac{a_{ij}}{(n_i a_{\bullet j})/n_\bullet}\right)^{a_{ij}}}$$

Små værdier af Q er det samme som store værdier af $G = -2\log(Q)$

$$G = 2\sum_{i=1}^{r}\sum_{j=1}^{k} a_{ij} \log\left(\frac{a_{ij}}{e_{ij}}\right), \quad e_{ij} = \frac{n_i a_{\bullet j}}{n}$$

P-værdi

Teststørrelse: $G = 2 \cdot \sum_{i=1}^{r} \sum_{j=1}^{k} a_{ij} \log\left(\frac{a_{ij}}{e_{ij}}\right)$

Approksimativ p-værdi: Hvis $e_{ij} \geq 5$, for alle i, j:

$$p\text{-værdi} \approx 1 - \chi^2_{\mathrm{cdf}}(G, (r-1)(k-1)) = 1 - \mathrm{pchisq}(G, (r-1)(k-1))$$

$df = (r-1)(k-1)$: antal rækker minus 1 gange antal søjler minus 1

Nakkesmerter ved brug af smartphone

Observeret antal:

	Nakkesmerter		
Tidsforbrug	Nej	Ja	Total
Lav	16	4	20
Medium	15	19	24
Høj	7	30	37
sum	38	53	91

$\text{Pers}_{ij} \quad i = L, M, H$ (lav,medium,høj), $j = \text{Nej,Ja}$

Model:

$$(\text{Pers}_{\text{L,Nej}}, \text{Pers}_{\text{L,Ja}}) \sim \text{multinom}(20, (\pi_{\text{L,Nej}}, \pi_{\text{L,Ja}}))$$
$$\pi_{\text{L},j} \geq 0, \ \pi_{\text{L,Nej}} + \pi_{\text{L,Ja}} = 1$$
$$(\text{Pers}_{\text{M,Nej}}, \text{Pers}_{\text{M,Ja}}) \sim \text{multinom}(24, (\pi_{\text{M,Nej}}, \pi_{\text{M,Ja}}))$$
$$\pi_{\text{M},j} \geq 0, \ \pi_{\text{M,Nej}} + \pi_{\text{M,Ja}} = 1$$
$$(\text{Pers}_{\text{H,Nej}}, \text{Pers}_{\text{H,Ja}}) \sim \text{multinom}(37, (\pi_{\text{H,Nej}}, \pi_{\text{H,Ja}}))$$
$$\pi_{\text{H},j} \geq 0, \ \pi_{\text{H,Nej}} + \pi_{\text{H,Ja}} = 1$$

Hypotese: $(\pi_{\text{L,Nej}}, \pi_{\text{L,Ja}})) = (\pi_{\text{M,Nej}}, \pi_{\text{M,Ja}})) = (\pi_{\text{H,Nej}}, \pi_{\text{H,Ja}}))$

Beregning i R

```
obs=rbind(c(16,4),c(15,19),c(7,30))

ex=outer(rowSums(obs),colSums(obs))/sum(obs)

obs1=ifelse(obs==0,1,obs)
G=2*sum(obs*log(obs/ex))

pval=1-pchisq(G,(dim(obs)[1]-1)*(dim(obs)[2]-1))

list(Forventede=ex,G=G,Pvaerdi=pval)
```

rowSums: rækkesummer;   colSums: søjlesummer

outer(x,y): 3×2 matriks indgange $x_i y_j$

R-leg

	Nakkesmerter		
Tidsforbrug	Nej	Ja	Total
Lav	16	4	20
Medium	15	19	24
Høj	7	30	37
sum	38	53	91

1. Prøv: chisq.test(obs) (obs er matriks med data)

2. Lav 95%-konfidensinterval for sandsynlighed for "Ja" for hver række

Benyt: prop.test(x,n)$conf.int

3. Lav eventuelt figur med de tre konfidensintervaller (errorbar)

Simuleringseksperiment

Simulere fordeling af G-teststørrelse for 3×2 tabel

```
dataObs=rbind(c(16,4),c(15,19),c(7,30))
p0=(16+15+7)/(20+34+37)

simFct=function(p){
 x=rbinom(3,c(20,34,37),p)
 return(cbind(x,c(20,34,37)-x))
 }

testFct=function(obs){
 ex=outer(rowSums(obs),colSums(obs))/sum(obs)
 obs1=ifelse(obs==0,1,obs)
 gTest=2*sum(obs*log(obs1/ex))
 return(gTest)
}

Gobs=testFct(dataObs)
```

Simuleringseksperiment

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

for (i in 1:nSim){
 dataSim=simFct(p0)
 gTest[i]=testFct(dataSim)
}

c(100*sum(gTest>=qchisq(0.95,2))/nSim,
100*(1+sum(gTest>=Gobs))/(1+nSim))

hist(gTest,probability=TRUE,ylim=c(0,0.04))
curve(dchisq(x,df=2),from=0,to=20,add=TRUE)
abline(v=qchisq(0.95,2),col=2)

hist(log(gTest),probability=TRUE)
curve(dchisq(exp(x),df=2)*exp(x),from=-6,to=3,add=TRUE)
abline(v=log(qchisq(0.95,2)),col=2)
```

Baggrund: Antal frihedsgrader i χ^2-fordeling generelt

Data $\sim$ model M_1

Hypotese: Data $\sim$ model M_2

Test: $$Q = \frac{\max_{M_2} L}{\max_{M_1} L}, \qquad G = -2\log(Q)$$

$$pværdi \approx 1 - \chi^2_{\text{cdf}}(G, df)$$

df = antal frie parametre i M_1 - antal frie parametre i M_2

Goodness of fit: $df = (k-1) - d$

Homogenitetstest: $df = r(k-1) - (k-1) = (r-1)(k-1)$

Baggrund

Hvorfor er $G \approx \chi^2(df)$?

Svar: Taylorudvikling + centrale grænseværdisætning

Sceneskift

Vi har indført test for homogenitet af r multinomialfordelinger

Næste: Simpsons paradox

Berkeley: kønsdiskriminering

Data:

Køn	Optaget	Ikke optaget	n
mænd	3714 (44%)	4728	8442
kvinder	1512 (35%)	2809	4321
sum	5226	7537	12763

Model: $(3714, 4728) \sim\sim \text{multinomial}(8442, (\pi_{11}, \pi_{12}))$
$(1512, 2809) \sim\sim \text{multinomial}(4321, (\pi_{21}, \pi_{22}))$

Hypotese: samme procentdel optages af kvinder og mænd:
$\pi_{11} = \pi_{21}$ eller $\pi_{\text{mænd,optaget}} = \pi_{\text{kvinder,optaget}}$

Kørsel i R:

```
Obs=rbind(c(3714,4728),c(1512,2809))
...
```

Berkeley: kønsdiskriminering

```
$Forventede
         [,1]     [,2]
[1,] 3456.702 4985.298
[2,] 1769.298 2551.702

$G
[1] 96.69686

$Pvaerdi
[1] 0
```

Bemærk: Alle forventede er ≥ 5

$p$værdi $= 1 - \chi^2_{\mathrm{cdf}}(96.70, 1) = 8.07 \cdot 10^{-23}$

Data peger tydeligt på en forskel i optagelsesprocent for mænd og kvinder, og dog:

Berkeley: ingen kønsdiskriminering

	mænd		kvinder		
Afdeling	antal	optaget	antal	optaget	pval
A	825	62%	108	82%	0.00
B	560	63%	25	68%	0.61
C	325	37%	593	34%	0.39
D	417	33%	375	35%	0.59
E	191	28%	393	24%	0.32
F	272	6%	341	7%	0.56

Simpsons paradox!

Årsag: mænd og kvinder vælger forskellige studiefag og der er forskel i adgangskrav mellem studier

Simulere Simpsons paradoks

```
n11=50; n21=10
n12=10; n22=50
p1=0.8; p2=0.2
x1=rbinom(1,n11,p1)+rbinom(1,n12,p2)
x2=rbinom(1,n21,p1)+rbinom(1,n22,p2)

obs=rbind(c(x1,(n11+n12)-x1),c(x2,(n21+n22)-x2))
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)
```

Simpson's paradoks

Sceneskift

Simpsons paradoks er illustreret

Næste: Cochrans regel

Cochrans regel

Contingency tables with more than 1 d.f. If relatively few expectations are less than 5 (say in 1 cell out of 5 or more, or 2 cells out of 10 or more), a minimum expectation of 1 is allowable in computing X^2. (Cochran 1954, p. 420).

Alle forventede skal være større end eller lig med 1

højst 20 procent må være under 5

Lad os vende tilbage til GOF fra sidste forelæsning

Cochrans regel

Index	1	2	3	4	5
Værdi	0	1	2	3	≥ 4
A_j	144	91	32	11	2
e_j	139.04	97.33	34.07	7.95	1.61

Slår to sidste kasser sammen: $G = 1.84$, p-værdi $= 0.399$

Slår ikke sammen: $G = 1.86$, p-værdi $= 0.603$

Kasse	1	2	3	4	5	6	7	8	9	10	11	12
a_j	1	2	9	23	34	47	37	24	12	9	0	1
e_j	1.0	3.2	9.5	20.9	34.6	42.6	39.2	27.0	13.8	5.3	1.5	0.4

Slår to første og tre sidste kasser sammen: $G = 2.73$, p-værdi $= 0.84$

Slår to første og to sidste kasser sammen: $G = 3.57$, p-værdi $= 0.83$

Sceneskift

Cochrans regel er omtalt

Næste: G-test eller χ^2-test (MSRR: C-test)

χ^2-test

I webbogen bruges teststørrelsen: $G = 2\sum \text{observeret} \cdot \log\left(\frac{\text{observeret}}{\text{forventet}}\right)$

Ude i verden (og i MSRR) bruges ofte af historiske grunde:

$$C = \sum \frac{(\text{observeret} - \text{forventet})^2}{\text{forventet}}$$

For store datasæt er der ikke forskel. For små datasæt kan χ^2-approksimationen være lidt bedre for G end for C

Gå til webbo afsnit 1.6 og tilføj i sidste skjulte punkt inde i list:

C=chisq.test(obs)

Sammenlige G og C ved simulering

Teste $p = p_0$ i binomialmodel ($=$ multinomialmodel)

```
nsim=1000000
p0=0.3
n=100

x1=rbinom(nsim,n,p0)
x2=n-x1
phat=rep(p0,nsim)
e1=n*phat
e2=n*(1-phat)
x11=ifelse(x1==0,1,x1)
x22=ifelse(x2==0,1,x2)

G=2*(x1*log(x11/e1)+x2*log(x22/e2))
Ctest=(x1-e1)^2/e1+(x2-e2)^2/e2

100*c(sum(G>3.8415),sum(Ctest>3.8415))/nsim
```

Sceneskift

χ^2-test er omtalt

Næste: Fishers eksakte test

Tea party

En person bliver givet 8 kopper med te og mælk blandet sammen. De 4 af kopperne er lavet ved at der først er hældt te i koppen og dernæst mælk, og de 4 andre kopper er lavet ved at mælk er hældt i først og dernæst te. Personen ved ikke noget om hvordan de 8 kopper er fordelt på de to typer.

Hypotese: Person kan ikke kende forskel (= vælger tilfældigt)

X_1 antal gange der siges "te først" blandt de 4 med te først
X_2 antal gange der siges "mælk først" blandt de 4 med mælk først
Model: $X_1 \sim \text{binom}(4, p_1), \quad X_2 \sim \text{binom}(4, p_2)$

Hypotese $p_1 = p_2$

Tea party

	Observerede	
	Siger te	Siger mælk
Te først	4	0
Mælk først	0	4

	Forventede	
	Siger te	Siger mælk
Te først	2	2
Mælk først	2	2

Kan ikke bruge Cochran regel

Der er 25 kombinationer af x_1 og x_2 med $n_1 = n_2 = 4$ men kun 7 forskellige værdier af G

G	0.00	0.54	1.53	2.09	3.45	6.09	11.09
$P,\ p = 0.5$	0.273	0.375	0.062	0.125	0.094	0.062	0.008
$P,\ p = 0.1$	0.518	0.029	0.383	0.002	0.064	0.005	0.000
$\chi^2(1),\ \geq$	1.000	0.462	0.216	0.148	0.063	0.014	0.001

Fishers eksakte test

Hvis $p_1 = p_2$ så er $X_1 + X_2 \sim \text{binom}(n_1 + n_2, p)$, $\hat{p} = \frac{X_1+X_2}{n_1+n_2}$

vi er ikke interesseret i p, kun i spørgsmålet $p_1 = p_2$

Betinge med værdien af $X_1 + X_2$:

$$\begin{array}{cc|c} 0 & 4 & 4 \\ 4 & 0 & 4 \\ \hline 4 & 4 & 8 \end{array} \qquad \begin{array}{cc|c} 1 & 3 & 4 \\ 3 & 1 & 4 \\ \hline 4 & 4 & 8 \end{array} \qquad \begin{array}{cc|c} 2 & 2 & 4 \\ 2 & 2 & 4 \\ \hline 4 & 4 & 8 \end{array} \qquad \begin{array}{cc|c} 3 & 1 & 4 \\ 1 & 3 & 4 \\ \hline 4 & 4 & 8 \end{array} \qquad \begin{array}{cc|c} 4 & 0 & 4 \\ 0 & 4 & 4 \\ \hline 4 & 4 & 8 \end{array}$$

G	11.09	2.09	0.00	2.09	11.09
$P(\vert)$	0.014	0.229	0.514	0.229	0.014

P-værdi fra betingede test $= 0.014 + 0.014 = 0.028$

Betinge i 2×2 tabel

Model: $X_1 \sim \text{binom}(n_1, p), \quad X_2 \sim \text{binom}(n_2, p)$, betinge med $X_1 + X_2$

$$P(X_1 = x_1, X_2 = x_2 | X_1 + X_2 = k)$$

$$= \frac{\binom{n_1}{x_1} p^{x_1} (1-p)^{n_1 - x_1} \binom{n_2}{k - x_1} p^{k - x_1} (1-p)^{n_2 - (k - x_1)}}{\binom{n}{k} p^k (1-p)^{n-k}}, \quad x_1 + x_2 = k, \; n_1 + n_2 = n$$

$$= \frac{\binom{n_1}{x_1}\binom{n_2}{k - x_1}}{\binom{n}{k}}$$

Afhænger IKKE af p derfor "eksakt" test

Beregning af betingede sandsynlighed i R: dhyper(x,b,c,d)

x	b-x	b
d-x	c-d+x	c
d	n-d	n=b+c

Kritisk område

Hvad er "mere kritisk" i betingede fordeling?

R, fisher.test: alle udfald hvor betingede sandsynlighed er $\leq$ sandsynlighed for faktiske observation

Alternativ: alle udfald i betingede fordeling med $G(x) \geq G(x_{\text{obs}})$

Se eksempel i afsnit 1.8.1

Generelt: I skal blot bruge fisher.test medmindre jeg beder jer om andet

Betinge i 2×3 tabel

Sceneskift

Slut for i dag