Matematisk Statistik: Modelbaseret Inferens

Uafhængighedstest

Jens Ledet Jensen

I DAG

G-test kontra C-test

Fishers eksakte test

Test for uafhængighed

Eksamensopgaver

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 webbog 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
$P(\|)$	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: $X_1 \sim \text{binom}(n_1, p)$, $\;X_2 \sim \text{binom}(n_2, p)$, 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

Sceneskift

Næste: Fishers eksakte test for 2×3 tabel

Betinge i 2×3 tabel

Model:

$(X_{11}, X_{12}, X_{13}) \sim \text{multinom}(n_1, \pi)$, $\quad (X_{21}, X_{22}, X_{23}) \sim \text{multinom}(n_2, \pi)$,
betinge med $(X_{11} + X_{21}, X_{12} + X_{22}, X_{13} + X_{23})$

$$P(X_{11} = x, X_{12} = y | X_{11} + X_{21} = a, X_{12} + X_{22} = b, X_{13} + X_{23} = c)$$

$$= \binom{n_1}{x, y, n_1 - x - y} \binom{n_2}{a - x, b - y, n_2 - a - b + x + y} \Big/ \binom{n}{k}$$

```
res=c()
for (x in 1:n1){
 for (y in 0:(n1-x)){
  x13=n1-x-y
  x21=a-x; x22=b-y; x23=n2-a-b+x+y
  if ((x21>=0)&(x22>=0)&(x23>=0)){
   pr=choose(n1,x)*choose(n1-x,y)*choose(n2,x21)*choose(n2-x21,x22)/
   (choose(n,a)*choose(n-a,b))
  res=rbind(res,c(x,y,pr))
  }
 }
```

Sceneskift

Fishers eksakte test er omtalt

Næste: lidt om uafhængighed

Eksempel

Skema angiver sandsynlighed for at en tilfældig voksen kvinde i alderen 20-30 år er ryger og om vedkommende er kaffedrikker:

	ryger	ikke ryger	sum
kaffe	0.10	0.30	0.40
ikke kaffe	0.15	0.45	0.60
sum	0.25	0.75	1.00

Klikker: Er rygevaner uafhængig af kaffedrikning?

Eksempel

Skema angiver sandsynlighed for at en tilfældig stoppet bilist jævnligt taler i mobiltelefon under kørsel, og om vedkommende har været indblandet i trafikuheld indenfor de seneste to år:

	uheld	ikke uheld	sum
mobil	0.1	0.1	0.2
ikke mobil	0.1	0.7	0.8
sum	0.2	0.8	1.0

Uafhængighed

	ryger	ikke ryger	sum
kaffe	αp	$(1-\alpha)p$	p
ikke kaffe	$\beta(1-p)$	$(1-\beta)(1-p)$	$(1-p)$
sum	–	–	1

Uafhængighed: $\frac{\alpha p}{(1-\alpha)p} = \frac{\beta(1-p)}{(1-\beta)(1-p)}$ eller $\frac{\alpha}{1-\alpha} = \frac{\beta}{1-\beta}$

Eller: $\alpha = \beta$

	ryger	ikke ryger	sum
kaffe	αp	$(1-\alpha)p$	p
ikke kaffe	$\alpha(1-p)$	$(1-\alpha)(1-p)$	$(1-p)$
sum	α	$1-\alpha$	1

Produkt!

Uafhængighed

Definition: X og Y er uafhængige hvis

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$$

	ryger	ikke ryger	sum
kaffe	αp	$(1 - \alpha)p$	p
ikke kaffe	$\alpha(1 - p)$	$(1 - \alpha)(1 - p)$	$(1 - p)$
sum	α	$1 - \alpha$	1

Cereals

Datasæt fra MSRR: cereals.csv

> $n = 43$ morgenmadsprodukter klassificeret efter
>
> Rettet mod "adult" eller "children" (H)
>
> Placering på hylde: bottom, middle, top (M)

```
   ID      Age  Shelf  Sodiumgram Proteingram
1   1    adult bottom 0.007000000  0.10000000
2   2 children bottom 0.006666667  0.06666667
...
42 42    adult    top 0.000000000  0.09259259
43 43    adult    top 0.002452830  0.09433962
```

Fulde data

n individer inddeles efter to inddelingskriterier, H og M

$$\text{data: } (H_1, M_1), (H_2, M_2), \ldots, (H_n, M_n)$$

H har r niveauer, M har k niveauer

$$P(H = i) = \alpha_i, \quad P(M = j | H = i) = \gamma_{ij}$$

Samlet sandsynlighed: $P(\text{Data}) = \prod_{u=1}^{n} \alpha_{h_u} \gamma_{h_u, m_u}$

Hypotese om uafhængighed: $\gamma_{ij} = \beta_j$

$$P(\text{Data}) = \prod_u \alpha_{h_u} \beta_{m_u}$$

Fordeling af $(H_1, \ldots, H_n)$ er irrelevant for spørgsmålet om uafhængighed

Første betingning

Betinger med $H_1, \dots, H_n$

$$P(\text{Data}|H) = \frac{\prod_u \alpha_{h_u}\gamma_{h_u,m_u}}{\prod_u \alpha_{h_u}} = \prod_u \gamma_{h_u,m_u}$$

hypotese om uafhængighed: $P(\text{Data}|H) = \prod_u \beta_{m_u}$

Fordeling af M $(\beta_1, \dots, \beta_k)$er irrelevant for spørgsmålet om uafhængighed

betinge med hvor mange der er på hvert niveau

$B_j =$ antal M_u-er med værdien j, $j = 1, \dots, k$, $B_j = \sum_u 1(M_u = j)$

$(B_1, \dots, B_k) \sim \text{multinom}(n, (\beta_1, \dots, \beta_k))$

Anden betingning

$$P(\text{Data}|H,B) = \frac{\prod_u \beta_{m_u}}{\binom{n}{b_1,\ldots,b_k}\beta_1^{b_1}\cdots\beta_k^{b_k}} = \frac{1}{\binom{n}{b_1,\ldots,b_k}}$$

I den betingede fordeling er alle kombinationer af $(m_1, m_2, \ldots, m_n)$ som opfylder $b_j = \sum_u 1(m_u = j)$ lige sandsynlige

Vi kan simulere betingede fordeling ved at lave tilfældige permutationer af dataværdierne $m_1, \ldots, m_n$

R: sample(m) m er en vektor med $m_1, \ldots, m_n$

Næste slide: grundliggende dele af program

Simulere

Vi kan simulere p-værdi for test i den dobbelt-betingede fordeling baseret på en teststørrelse der beregnes i *test*:

```
nSim=10000
tval=rep(0,nSim)

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

tobs=test(cbind(h,m))
pval=(1+sum(tval>=tobs))/(1+nSim)
```

Sceneskift

Princip i simulering af betingede fordeling er indført

Næste: vælge teststørrelse

Teststørrelse

Vi mangler at specificere en teststørrelse

Fra fulde data til $r \times k$ tabel:

$$A_{ij} = \sum_{u=1}^{n} 1(H_u = i, M_u = j),\ i = 1, \dots, r,\ j = 1, \dots, k$$

Model M_{I0}: $(A_{1,1}, \dots, A_{r,k}) \sim \text{multinom}(n, (\pi_{11}, \dots, \pi_{rk}))$

$$\sum_{ij} \pi_{ij} = 1 \text{ eller } \pi_{ij} = \alpha_i \gamma_{ij},\ \sum_j \gamma_{ij} = 1$$

Notation:

$A_{i\bullet}$: i'te rækkesum; $A_{\bullet j}$: j'te søjlesum
$A_{i*} = (A_{i1}, \dots, A_{ik})$, den i'te række
$A_{\bullet *} = (A_{\bullet 1}, \dots, A_{\bullet k})$, vektor med søjlesummer
$A_{*\bullet} = (A_{1\bullet}, \dots, A_{r\bullet})$, vektor med rækkesummer

Teststørrelse

Teste uafhængighedshypotesen: der eksisterer et sæt sandsynligheder $(\alpha_1 \ldots, \alpha_r)$ og et andet sæt sandsynligheder $(\beta_1 \ldots \beta_k)$ således at

$$\pi_{ij} = \alpha_i \beta_j \text{ for alle } i, j \text{ (model } M_{I1})$$

Næste slide: udregne likelihoodratio teststørrelsen Q for reduktion fra model M_{I0} til model M_{I1}

Alternativ til simulerings-pværdi: approksimative χ^2-fordeling for $-2\log(Q)$

Likelihoodratio Test

$$Q = \frac{\max_{M_{I1}} L}{\max_{M_{I0}} L} = \frac{\binom{n}{a}\prod_{ij}(\hat{\alpha}_i\hat{\beta}_j)^{A_{ij}}}{\binom{n}{a}\prod_{ij}(\hat{\pi}_{ij}(M_{I0})^{A_{ij}}}$$

$$\hat{\pi}_{ij}(M_{I0}) = \frac{A_{ij}}{n}$$

$$L_{M_{I1}}(\alpha, \beta) = \prod_{ij}(\alpha_i\beta_j)^{A_{ij}} = \prod_i \alpha_i^{A_{i\bullet}} \prod_j \beta_j^{A_{\bullet j}}$$

$$\hat{\alpha}_i = \frac{A_{i\bullet}}{n}, \; \hat{\beta}_j = \frac{A_{\bullet j}}{n}$$

forventede under M_{I1}: $e_{ij} = \frac{A_{i\bullet}A_{\bullet j}}{n}$

$$Q = \frac{\prod_{ij}\left(\frac{A_{i\bullet}}{n}\frac{A_{\bullet j}}{n}\right)^{A_{ij}}}{\prod_{ij}\left(\frac{A_{ij}}{n}\right)^{A_{ij}}} = \prod_{ij}\frac{1}{\left(\frac{A_{ij}}{A_{i\bullet}A_{\bullet j}/n}\right)^{A_{ij}}} = \prod_{ij}\frac{1}{\left(\frac{A_{ij}}{e_{ij}}\right)^{A_{ij}}}$$

Likelihoodratio Test

G-teststørrelse:

$$G = -2\log(Q) = \sum_{ij} A_{ij} \log\left(\frac{A_{ij}}{e_{ij}}\right)$$

Denne teststørrelse vil vi bruge i simulering af p-værdien

Prøv selv:

afsnit 1.8 i webbog, erstat obs=.. med cereal-data:

obs=rbind(c(2,1,14),c(7,18,1))

R-kørsel

Vise og forklare program i R

R ved ikke at $0 \cdot \log(0) = 0$

Simuleret p-værdi når der bruges teststørrelsen $C = \sum_{ij}(A_{ij} - e_{ij})^2/e_{ij}$:

R: `chisq.test(obs,simulate.p.value=TRUE,B=9999)`

Sceneskift

Næste: G-test for uafhængighed

$= G$-test for homogenitet

Forbindelse til homogenitetstest

G-teststørrelse fra uafhængighedstest $=$ G fra homogenitetstest

hvorfor ?

Model M_{I0}: $(A_{1,1},\ldots,A_{r,k}) \sim \text{multinom}(n,(\pi_{11},\ldots,\pi_{rk})),\ \pi_{ij} = \alpha_i\gamma_{ij}$

$\alpha_1 + \cdots + \alpha_r = 1,\ \gamma_{i1} + \gamma_{i2} + \cdots + \gamma_{ik} = 1,\ i = 1,\ldots,r$

$(A_{1\bullet},\ldots,A_{r\bullet}) \sim \text{multinom}(n,(\alpha_1,\ldots,\alpha_r))$

uafhængighed: $\gamma_{ij} = \beta_j$, afhænger ikke af i

Næste slide: fra M_{I0} til M_0 via betingning

Betinge med rækkesummer

$$P(A = a | A_{*\bullet} = a_{*\bullet}) = \frac{\binom{n}{a} \prod_{ij} (\alpha_i \gamma_{ij})^{a_{ij}}}{\binom{n}{a_{*\bullet}} \prod_i \alpha_i^{a_{i\bullet}}}$$

$$= \prod_i \binom{a_{i\bullet}}{a_{i*}} \gamma_{i1}^{a_{i1}} \cdots \gamma_{ik}^{a_{ik}}$$

Dette er model M_0 fra homogenitetstest: r multinomialfordelinger

uafhængighedshypotesen = homogenitetshypotesen

Derfor: samme G og skal vurderes i samme χ^2-fordeling

Frihedsgrader: $(rs - 1) - \{(r-1) + (k-1)\} = (r-1)(k-1)$

Udpensling

$$L_{M_{I0}}(\{\pi_{ij}\}) = L_{M_{I0}}(\{\alpha_i\gamma_{ij}\}) = L_{A_{*\bullet}}(\alpha)L_{A|A_{*\bullet}}(\{\gamma_{ij}\})$$

$$= L_{A_{*\bullet}}(\alpha)L_{M_0}(\{\gamma_{ij}\}) \quad (M_0 \text{ fra homogenitetstest})$$

$$Q_I = \frac{\max_{\alpha,\beta} L_{M_{I0}}(\{\alpha_i\beta_j\})}{\max_{\alpha,\gamma} L_{M_{I0}}(\{\alpha_i\gamma_{ij}\})} = \frac{\max_\alpha L_{A_{*\bullet}}(\alpha) \max_\beta L_{A|A_{*\bullet}}(\beta,\ldots\beta)}{\max_\alpha L_{A_{*\bullet}}(\alpha) \max_\gamma L_{A|A_{*\bullet}}(\{\gamma_{ij}\})}$$

$$= \frac{\max_\beta L_{M_0}(\beta,\ldots\beta)}{\max_\gamma L_{M_0}(\{\gamma_{ij}\})} = Q_{\text{Hom}}$$

Generelt: model med parameter (θ, ξ) og $L(\theta, \xi) = L_1(\theta)L_2(\xi)$:

Likelihoodratio for hypotese om θ vedrører kun $L_1(\theta)$

Rækkesummer og søjlesummer

Under hypotesen om uafhængighed

$$L_A(\alpha, \beta) = L_{A_{*\bullet}}(\alpha) L_{A|A_{*\bullet}}(\beta) = L_{A_{*\bullet}}(\alpha) L_{A_{\bullet *}|A_{*\bullet}}(\beta) L_{A|A_{*\bullet}, A_{\bullet *}}()$$

idet vi fra før har

$$L_{A|A_{*\bullet}}(\beta) = \prod_i L_{A_{i*}|A_{i\bullet}}(\beta)$$

I ord: rækkerne i A er uafhængige givet rækkesummerne

$$\text{række } i\text{: } A_{i*}|A_{i\bullet} \sim \text{multinom}(a_{i\bullet}, \beta)$$

Summen af rækkerne er derfor også multinomialfordelt:

$$A_{\bullet *}|A_{*\bullet} \sim \text{multinom}(n, \beta)$$

Rækkesummer og søjlesummer

Da dette udtryk ikke afhænger af rækkesummerne har vi at søjlesummer og rækkesummer er uafhængige

$$L_{A_{\bullet *}|A_{*\bullet}}(\beta) = L_{A_{\bullet *}}(\beta)$$

og

$$L_{A|A_{*\bullet},A_{\bullet *}} = \frac{\prod_i \binom{a_{i\bullet}}{a_{i*}} \prod_j \beta_j^{a_{ij}}}{\binom{n}{a_{\bullet *}} \prod_j \beta_j^{a_{\bullet j}}} = \frac{\prod_i \binom{a_{i\bullet}}{a_{i*}}}{\binom{n}{a_{\bullet *}}}$$

Dermed har vi vist: $L_A(\alpha,\beta) = L_{A_{*\bullet}}(\alpha) L_{A_{\bullet *}}(\beta) L_{A|A_{*\bullet},A_{\bullet *}}()$

Konklusion: under uafhængighedshypotesen baseres inferens om α på rækkesummerne og inferens for β baseres på søjlesummerne

Leddet $L_{A|A_{*\bullet},A_{\bullet *}}()$ bruges i Fishers eksakte test

Sceneskift

Næste: simuleret p-værdi

$$= p\text{-værdi fra Fishers eksakte test (næsten)}$$

Samme betingede fordeling, men forskellig teststørrelse

Fra simulering til Fisher

$(M_1, \ldots, M_)|(H_1, \ldots, H_n, B_1, \ldots, B_k)$ samme som

$$(M_1, \ldots, M_n)|(H_1, \ldots, H_n, A_{\bullet 1}, \ldots, A_{\bullet k})$$

da j'te søjlesum netop er $B_j = \sum_u 1(M_u = j)$

$$\text{sandsynlighed} = \frac{1}{\binom{n}{A_{\bullet 1}, \ldots, A_{\bullet k}}}$$

Bemærk: når vi betinger med $(H_1, \ldots, H_n)$ så har vi betinget med rækkesummerne $A_{1\bullet}, \ldots, A_{r\bullet}$

Alle muligheder har samme sandsynlighed. For at få sandsynlighed for tabel $\{A_{ij}\}$ givet rækkesummer og søjlesummer, skal vi tælle antal muligheder for $(M_1, \ldots, M_n)$

Tælle op

Hvis vi peger på alle dem i række 1, alle med $H_u = 1$, så skal vi vælge A_{11} ud som vi giver M-værdien 1, vælge A_{12} ud som vi giver M-værdien 2, og så videre. Antallet af måder er

$$\binom{A_{1\bullet}}{A_{11},\ldots,A_{1k}}$$

Tilsvarende med række 2 op til række r, i alt:

$$\binom{A_{1\bullet}}{A_{11},\ldots,A_{1k}} \cdot \binom{A_{2\bullet}}{A_{21},\ldots,A_{2k}} \cdots \binom{A_{r\bullet}}{A_{r1},\ldots,A_{rk}}$$

Betingede sandsynlighed for tabel er denne divideret med $\binom{n}{A_{\bullet 1},\ldots,A_{\bullet k}}$

Fishers eksakte

I Fishers eksakte test bruges den betingede sandsynlighed:

$$\frac{\binom{n}{A_{11},\ldots,A_{rk}}}{\binom{n}{A_{\bullet 1},\ldots,A_{\bullet k}}\binom{n}{A_{1\bullet},\ldots,A_{r\bullet}}} = \frac{\binom{A_{1\bullet}}{A_{11},\ldots,A_{1k}}\cdot\binom{A_{2\bullet}}{A_{21},\ldots,A_{2k}}\cdots\binom{A_{r\bullet}}{A_{r1},\ldots,A_{rk}}}{\binom{n}{A_{\bullet 1},\ldots,A_{\bullet k}}}$$

som er den samme som vi fandt ovenfor

Fisher eller simulere

Hvorfor bruger vi ikke altid Fishers eksakte test i stedet for at simulere ?

fisher.test:
"can get too large for the exact test in which case an error is signalled. Apart from increasing workspace sufficiently, which then may lead to very long running times, using simulate.p.value=TRUE may then often be sufficient and hence advisable."

Opsummering

Se på data for at afgøre om tabel er

- en stor multinomialfordeling (to inddelingskriterier)
- eller r multinomialfordelinger (antal i rækker er "design")

Lav G-teststørrelse

- hvis forventede er store: brug χ^2-approksimation
- hvis forventede ikke er store nok:
  - brug fisher.test hvis tabel ikke er for stor
  - ellers brug simulering

Opgaver

Regne gamle eksamensopgaver