Matematisk Statistik: Modelbaseret Inferens

Uafhængighedstest

Jens Ledet Jensen

I DAG

Relation mellem G-test for uafhængighed og for homogenitet

Relatoion mellem Fishers eksakte test og simulationsbaseret test

Eksamensopgaver

Sceneskift

Data

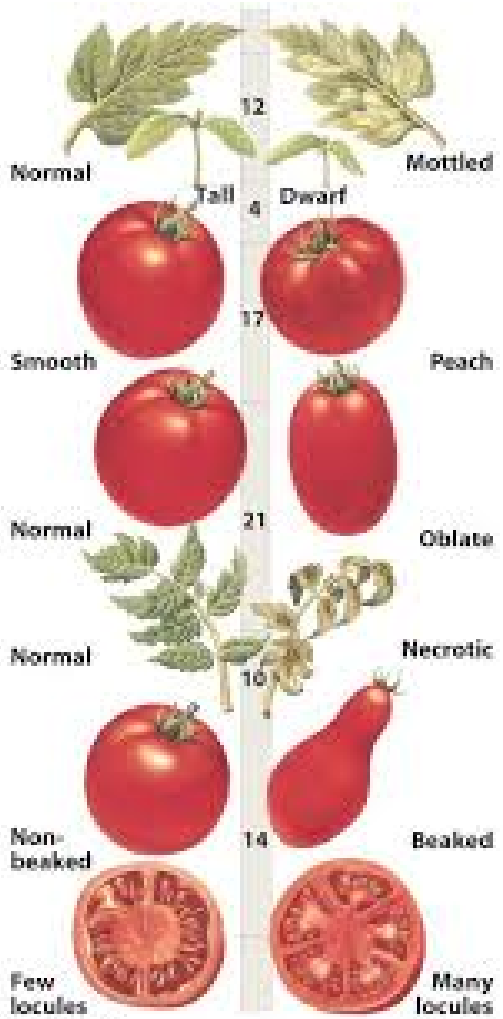

Figure 8-16
Biology of Plants, Seventh Edition

Data

Genetisk eksperiment: nedarves farve uafhængigt af form (tomater)

	nonbeaked	beaked	sum
rød	142	112	254
gul	113	133	246
sum	255	245	500

A_{ij}: antal i række i og søjle j, $i = 1, 2$, $j = 1, 2$

$n = A_{11} + A_{12} + A_{21} + A_{22} = 500$

Multinomialmodel M_0:

$(A_{11}, A_{12}, A_{21}, A_{22}) \sim \text{Multinom}(500, (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}))$

$pi_{ij} \geq 0, \ \pi_{11} + \pi_{12} + \pi_{21} + \pi_{22} = 1$

Uafhængighedshypotese (model M_1): $\pi_{ij} = \alpha_i \beta_j, \ \alpha_1 + \alpha_2 = 1, \ \beta_1 + \beta_2 = 1$

G-teststørrelse

Forventede under M_1: $e_{ij} = n\hat{\alpha}_i\hat{\beta}_j = \frac{Ai\bullet A_{\bullet j}}{n}$

G-test: $G = -2\ln\left(\frac{\max_{M_1} L}{\max_{M_0} L}\right) = 2\sum_{ij} a_{ij}\log\left(\frac{a_{ij}}{e_{ij}}\right)$

G-test

Vi kan nu anvende det generelle G-test:

	nonbeaked	beaked	sum
rød	142	112	254
gul	113	133	246
sum	255	245	500

	Forventede	
	nonbeaked	beaked
rød	129.54	124.46
gul	125.46	120.54

$$G = 2\Big\{142 \cdot \log(\tfrac{142}{129.54}) + 112 \cdot \log(\tfrac{112}{124.46}) + 113 \cdot \log(\tfrac{113}{125.46}) + 133 \cdot \log(\tfrac{133}{120.54})\Big\} = 4.98$$

Da alle forventede er ≥ 5 bruger vi:

$$p\text{værdi} = 1 - \chi^2_{\text{cdf}}(4.98, 3-2) = 0.026$$

Konklusion: data giver anledning til skepsis over for ufhængighed

Prøv i R

Gå til nederst i afsnit 1.6 i webbog erstat obs med:

```
obs=rbind( c(142,112), c(113,133) )
```

Bemærk: jeg beder jer om at gå hen i homogenitetstestafsnittet!

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

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

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

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

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

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-værdi fra Fishers eksakte test (næsten)

Samme betingede fordeling, men forskellig teststørrelse

Prøv i R

Webbog afsnit 1.8: Beregning i R simuleret p-værdi

```
obs=rbind(c(142,112),c(113,133))
```

Prøv dernæst:

```
fisher.test(obs)
```

Notation for simulering

Data bag tabel:

$(H_1, M_1), \ldots, (H_n, M_n)$

H_i: i'te tomats farve, M_i: i'te tomats form

B_j: antal tomater med j'te form

Simulering: $(M_1, \ldots, M_n)|(H_1, \ldots, H_n, B_1, \ldots, B_k)$

Fra simulering til Fisher

$(M_1, \ldots, M_n)|(H_1, \ldots, H_n, B_1, \ldots, B_k)$ 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 $B_j = \sum_u 1(M_u = j)$

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."

Husk: simulering ser på betingede fordelinge af G, fisher.test definerer "mere kritisk" på anden vis

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

Sceneskift

Slut med teori for i dag

Regne gamle eksamensopgaver