



Statistisk modellering for biologer og bioteknologer, ST2304

11. august, 2012

Løsningsforslag

Oppgave 1

- a) X er hypergeometrisk fordelt med parametere $m = 40$ (antall hanner i populasjonen) $n = 60$ (antall hunner) og $k = 20$ (antall individ i utvalget). Uttrykket
`rhyper(1, 40, 60, 20)`
vil simulere én realisasjon av X .
- b) `dhyper(12, 40, 60, 20)`
- c) Hvis det skal være *flere* hanner enn hunner i utvalget må antall hanner være *minst* 11, vi skal med andre ord finne $P(X \geq 11) = P(X > 10) = 1 - P(X \leq 10)$ som kan beregnes ett av uttrykkene
`phyper(10, 40, 60, 20, lower.tail=FALSE)`
`1-phyper(10, 40, 60, 20)`
- d) Det er spurt etter 50%-kvantilen også kjent som medianen til X som kan beregnes ved hjelp av uttrykket
`qhyper(.5, 40, 60, 20)`

Oppgave 2

- a) I følge modellen er antall guttebarn Y_i i hver familie $i = 1, 2, \dots, m$ binomisk fordelt med parametere n_i og p_i hvor

$$\text{logit}(p_i) = \beta_0 + \beta_1 \text{inntekt}_i, \quad (1)$$

og inntekt_i er inntekten i den i 'te familien. Modellen antar at sannsynligheten for å få barn av ett gitt kjønn varierer mellom familier som over og at kjønnene til hvert barn innen hver familie er uavhengig slik at totalt antall av det en kjønn blir binomisk fordelt. Valget av logit link-funksjon sikrer at sannsynlighetene p_i blir liggende på intervallet fra 0 til 1 for alle inntekter og parameterverdier.

- b) Vi har testet $H_0 : \beta_1 = 0$ mot $H_1 : \beta_1 \neq 0$, altså nullhypotesen at inntekt har ingen effekt på sannsynlighet for gutt i forskjellige familier mot alternativ hypotese at det er en effekt. Testens konklusjon er at vi ikke kan forkaste nullhypotesen til fordel alternativ hypotese; dette innebærer ikke at inntekt ikke har noen effekt; vi kan kun konkludere med at dataene ikke gir støtte for H_1 . Dette kan f.eks. skyldes liten utvalgsstørrelse og dermed teststyrke.

- c) Vi løser ligning (1) for p og finner at for inntekt=600000 blir sannsynlighet for gutt

$$\begin{aligned}\hat{p} &= \frac{1}{1 + \exp(-(\beta_0 + \beta_1 \text{inntekt}))} \\ &= \frac{1}{1 + \exp(-(1.963 - 3.592 \cdot 10^{-6} \cdot 600000))} = 0.452.\end{aligned}\quad (2)$$

- d) Under nullhypotesen om ingen overdispersjon er deviansen χ^2 -fordelt med $n - p$ frihetsgrader, her $16 - 2 = 14$. Det er vanskelig å tenke seg mekanismer som kan genererer underdispersjon i dette tilfelle og det kan derfor være naturlig å la alternativ hypotese være positiv overdispersjon. Vi får dermed en en-sidig test hvor vi forkaster for store devianser. Kritisk verdi blir øvre 0.05-kuantil i en χ^2 -fordeling med 14 frihetsgrader, d.v.s. 23.68. Siden observert devians ("residual deviance") er større enn dette kan vi konkludere med at det er overdispersjon i dataene.

Positive kovarians mellom kjønnet til barn innen en gitt familie, utelatte kovariater som påvirker p_i og eventuell feil link-funksjon vil kunne gi overdispersjon i dette tilfelle.

Oppgave 3

- a) Antallene av hver genotype X_1, X_2, X_3 vil være multinomisk fordelt med parametre lik genotypfrekvensene i populasjonen, la oss si P_{AA}, P_{Aa}, P_{aa} og $n = 100$.
- b) SME av frekvensen av genvarianten A blir $\hat{p}_A = (2 \cdot 10 + 65)/(2 \cdot 100) = 0.425$. SME av genotypfrekvensene blir $\hat{P}_{AA} = \hat{p}_A^2 = 0.180625$, $\hat{P}_{Aa} = 2\hat{p}_A(1 - \hat{p}_A) = 0.488750$ og $\hat{P}_{aa} = (1 - \hat{p}_A)^2 = 0.330625$.
- c) Vi har multinomiske data med $k = 3$ kategorier. De tre multinomiske sannsynlighetene (genotypfrekvensene) kan alle skrives som funksjoner av $s = 1$ parameter (frekvensen p_A av genvarianten A). Dermed er (i følge setning i notat 2) testobservatoren

$$D = \sum_{i \in \{AA, Aa, aa\}} \frac{(X_i - n\hat{P}_i)^2}{n\hat{P}_i} \quad (3)$$

kji-kvadrat fordelt med $k - 1 - s = 3 - 1 - 1 = 1$ frihetsgrad. Vi forkaster for store D (store forskjeller mellom observerte og forventede antall). Kritisk verdi blir dermed øvre 0.05-kvantil i en kji-kvadrat fordelingen med 1 frihetsgrad, 3.84.

Setter vi inn for X_i og hver $\hat{P}_i$ finner vi at den observerte verdien blir $D = 10.885$. Vi kan dermed forkaste nullhypotesen at genotypefrekvensene i populasjonen er i Hardy-Weinberg-likevekt.