

Eksamen juni 2001, Opgave 3

(a)

Vi fortolker de fem antal rygere (1517, ..., 287) som observationer af uafhængige binomialfordelte variable med hver sin ukendte sandsynlighedsparameter ($p_1, \dots, p_5$ i det følgende) og det totale antal personer i aldersgruppen (2406, ..., 625) som antalsparameter.

Estimater for de fem sandsynlighedsparametre:

$$\hat{p}_1 = \frac{1517}{2406} = 0.631$$

$$\hat{p}_2 = \frac{2309}{3283} = 0.703$$

$$\hat{p}_3 = \frac{3178}{4886} = 0.650$$

$$\hat{p}_4 = \frac{1702}{2980} = 0.571$$

$$\hat{p}_5 = \frac{287}{625} = 0.459$$

Approksimative 95% sikkerhedsgrænser udregnes efter formlen

$$\hat{p}_j \pm 1.96 \sqrt{\frac{1}{n_j} \hat{p}_j (1 - \hat{p}_j)}, \quad j = 1, \dots, 5.$$

Det giver

Gruppe 1: 0.631 ± 0.019 , sikkerhedsinterval [0.611 , 0.650]

Gruppe 2: 0.703 ± 0.016 , sikkerhedsinterval [0.688 , 0.719]

Gruppe 3: 0.650 ± 0.013 , sikkerhedsinterval [0.637 , 0.663]

Gruppe 4: 0.571 ± 0.018 , sikkerhedsinterval [0.553 , 0.589]

Gruppe 5: 0.459 ± 0.039 , sikkerhedsinterval [0.420 , 0.498]

(*det er ikke absolut nødvendigt at angive både usikkerheden på formen " $\pm \dots$ " og det egentlige sikkerhedsinterval, en af delene må være nok*).

(b)

Testet for "ingen forskel mellem rygevaner i aldersgrupper" er ækvivalent med det sædvanlige test for uafhængighed i den givne 5×2 antalstabel.

Vi får

$$-2 \log q = 2(1517 \log 1517 + \dots + 338 \log 338$$

$$-2406 \log 2406 - \dots - 625 \log 625$$

$$-8993 \log 8993 - 5187 \log 5187 + 14180 \log 14180) = 204.4450.$$

Denne teststørrelse er ekstremt stor, vurderet i en χ^2 -fordeling med $(5-1)(2-1)=4$ frihedsgrader (99.99% fraktilen i denne fordeling er 23.513), så vi får en helt klar forkastelse af hypotesen. Der er altså forskel på aldersgruppernes rygevaner.

Den tilsvarende Pearson teststørrelse, hvis nogle foretrækker at regne den ud, er 206.8847. Konklusionen er den samme.

(c)

Her vil det være rimeligt at vise estimaternes placering i forhold til hinanden ved hjælp af en tegning, enten en tegning af rygehyppigheden som funktion af alderen (med aldersintervallernes midtpunkter som de naturlige x -værdier, sikkerhedsintervaller tegnet som lodrette liniestykker; valget af "midtpunkter" for de to yderintervaller må naturligvis foretages lidt arbitrært), eller med intervallerne vandret indtegnet på fem linier under hinanden. Tegningerne gengives ikke her.

Man ser, at rygehyppigheden er størst i aldersgruppen 41–50. Dem under 40 ryger mindre, og efter 50 falder rygningen med alderen. Da sikkerhedsintervallerne for nabo-aldersgrupper er tydeligt disjunkte, må man gå ud fra at disse tendenser er signifikante. Den eneste reduktion, som man måske ville kunne få godkendt, går ud på at slå dem under 40 sammen med dem der er mellem 50 og 60. Men denne modelreduktion må siges at være lidt unaturlig. Hvis man kan få den godkendt siger den lige så meget om valget af aldersgruppering som om folks rygevaner. Det væsentlige er at rygehyppigheden ser ud til at toppe et sted mellem 40 og 50.

Sikkerhedsintervaller

