Peatükk 6

Kruskal-Wallise test

Kruskal-Wallise test on Wilcoxoni astaksummatesti üldistus enam kui kahe valimi jaoks.

Oletame, et meil on k sõltumatut valimit, igaühes neist on tehtud n_i mõõtmist, kokku N vaatlust. Paigutame vaatlused taas ühisesse variatsioonritta ja nõnda saab i. valimist pärit j. vaatlus endale järjekorranumbri ehk astaku R_{ij}. Iga valimi jaoks leiame tema vaatluste astakute keskmise

$$\overline{R_i} = \frac{1}{N} \sum_{j=1}^{n_i} R_{ij}.$$

Ühe vaatluse keskmine astak on $1/N \sum_{i=1}^{n} i = N(N+1)/(2N) = (N+1)/2$ ja nullhüpoteesi kehtides (kõik töötlused on võrdsed) on i. grupi oodatavaks keskmiseks astakuks $(N+1)/2$. Kruskal-Wallise teststatistikuks on

$$K = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i (\overline{R_i} - (N+1)/2)^2.$$

Alternatiivne statistiku K arvutusvalem on

$$K = \frac{12}{N(N+1)} \sum_{i=1}^{k} R_i^2/n_i - 3(N+1),$$

kus

$$R_i = \sum_{j=1}^{n_i} R_{ij}.$$

Nullhüpoteesi kehtides peaksid $K > 0$ väärtused olema väiksed, nullilähedased. Alternatiivse hüpoteesi kehtides (kui mõnest populatsioonist pärit

vaatlused kipuvad olema teistest suuremad) aga peaks K väärtus tulema suur (sest erinevus tegeliku keskmise astaku ja H_0 puhul oodatava puhul läheb suureks).

Kuidas leida Kruskal-Wallise testi olulisustõenäosust? Esmalt peaksime leidma, mitu erinevat võimalust on N astakut jagada gruppidesse suurusega $n_1, n_2, \ldots, n_k$. Selleks on muidugi

$$\left(\begin{array}{c} N \\ n_1 \end{array} \right) \cdot \left(\begin{array}{c} N-n_1 \\ n_2 \end{array} \right) \cdot \ldots \cdot \left(\begin{array}{c} N-n_1-\ldots-n_{k-1} \\ n_k \end{array} \right)$$

erinevat võimalust. Kõik mainitud võimalused astakute jagunemiseks k grupi vahel on H_0 kehtides võrdtõenäosed. Seejärel tuleks kokku lugeda, kui paljud neist võimalustest on sellised, mille puhul statistiku K väärtus tuleb suurem (või samasuur) kui meie valimis nähtud K väärtus. Jagades sobivate võimaluste arvu kõigi võimaluste arvuga saamegi teada, millise tõenäosusega H_0 kehtides võiks näha sama ekstreemset või veelgi ekstreemsemat statistiku väärtust ehk oleme leidnud testi olulisustõenäosuse.

Teoorias on kõik ilus, aga praktikas läheb erinevate astakute konfiguratsioonide arv ruttu astronoomiliseks. Kui näiteks meie valimi suurused on $n_1 = 10, n_2 = 11, n_3 = 8$, siis võimalikke viise astakuid jagada neisse kolme gruppi oleks

$$\left(\begin{array}{c} 29 \\ 10 \end{array} \right) \cdot \left(\begin{array}{c} 19 \\ 11 \end{array} \right) \cdot 1 = 20030010 \cdot 75582 \approx 1,5 \cdot 10^{12}.$$

Nii paljusid eri võimalusi läbi vaadata on isegi arvutilgi raske. Mis siis veel saab, kui valimimahud hakkavad muutuma "mõistlikult" suureks?

Lahenduseks on muidugi asümptootilise jaotuse kasutamine. Selgub, et asümptootiliselt (ja nullhüpoteesi kehtides) on K jaotuseks $\chi^2_{df=k-1}$.

Proovime seda tulemust ka pool-tunnetuslikult tõestada. Tähistame X-ga järgmise vektori:

$$X := \left(\sqrt{\frac{12n_1}{N(N+1)}}(\overline{R_1} - (N+1)/2), \ldots, \sqrt{\frac{12n_k}{N(N+1)}}(\overline{R_k} - (N+1)/2) \right)$$

ja paneme tähele, et $K = X^T X$. Kui juhusliku vektori X jaotuseks oleks normaaljaotus, $X \sim N(0, \Sigma)$, siis vektori $X^T X$ jaotuseks on χ^2-jaotus (siis ja ainult) siis, kui Σ on idempotentne, $\Sigma\Sigma = \Sigma$. Üritame aru saada, milline on meie poolt defineeritud X-i jaotus. Kuna $\sqrt{n}(\overline{Y} - \mu)$ jaotus on tsentraalse piirteoreemi tõttu normaaljaotus, siis võime oletada, et meie poolt defineeritud vektori kõigi elementide asümptootiliseks jaotuseks on normaaljaotus.

Saab näidata, et kogu vektor on mitmemõõtmelise normaaljaotusega (jätame tõestuse hetkel vahele). See, et X keskväärtus on (nullhüpoteesi kehtides) 0, on lihtne näha. Aga milline on Σ ja kas ta on ka idempotentne? Kõigepealt vaatame millised näevad välja Σ diagonaali elemendid ehk milline on $\mathrm{D}X_i$.

$$\begin{aligned}
\mathrm{D}X_i &= \mathrm{D}\left(\sqrt{\frac{12n_i}{N(N+1)}}(\overline{R_i} - (N+1)/2)\right) \\
&= \frac{12n_i}{N(N+1)}\left\{\mathrm{D}\left(\frac{1}{n_i}\sum_{j=1}^{n_i} R_{ij}\right) + \sum_{j\neq j'} Cov(R_{ij}, Rij')\right\} \\
&= \frac{12n_i}{N(N+1)}\left\{\frac{1}{n_i^2}n_i\frac{N^2-1}{12} + \frac{1}{n_i^2}n_i(n_i-1)\left(-\frac{N+1}{12}\right)\right\} \\
&= \frac{12n_i}{N(N+1)}\left\{\frac{N+1}{12n_i}(N-1-(n_i-1))\right\} \\
&= \frac{N-n_i}{N}
\end{aligned}$$

ja väljaspool peadiagonaali asuv element $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$ näeb välja selline:

$$\begin{aligned}
\mathrm{Cov}(X_i, X_j) &= \sqrt{n_i n_j}\frac{12}{N(N+1)}\mathrm{Cov}(\overline{R_i}, \overline{R_j}) \\
&= \sqrt{n_i n_j}\frac{12}{N(N+1)}\frac{1}{n_i n_j}\sum_{k=1\ldots n_i, l=1\ldots n_j}\mathrm{Cov}(R_{ik}, R_{jl}) \\
&= \sqrt{n_i n_j}\frac{12}{N(N+1)}\left(-\frac{N+1}{12}\right) \\
&= -\frac{\sqrt{n_i n_j}}{N}
\end{aligned}$$

Siit on juba lihtne näha, et Σ on idempotentne — Σ^2 diagonaali i element

on

$$\begin{aligned}
\Sigma_{ii}^2 &= \sum_{j=1} k\Sigma_{ij}\Sigma_{ji} \\
&= (\mathrm{D}X_i)^2 + \sum_{j\neq i} \mathrm{Cov}(X_i, X_j)^2 \\
&= \frac{(N-n_i)^2}{N^2} + \sum_{j\neq i} \frac{n_i n_j}{N^2} \\
&= \frac{(N-n_i)^2 + n_i(N-n_i)}{N^2} \\
&= \frac{(N-n_i)(N-n_i+n_i)}{N^2} \\
&= \frac{N-n_i}{N}
\end{aligned}$$

ja väljaspool diagonaali paiknev element Σ_{ij}^2 näeb välja ju selline:

$$\begin{aligned}
\Sigma_{ij}^2 &= \sum_{l=1}^{k} \Sigma_{il}\Sigma_{lj} \\
&= (\mathrm{D}X_i)\mathrm{Cov}(X_i, X_j) + (\mathrm{D}X_j)\mathrm{Cov}(X_i, X_j) + \sum_{l\neq i,j} \mathrm{Cov}(X_i, X_l)\mathrm{Cov}(X_l, X_j) \\
&= \frac{N-n_i}{N}\left(-\frac{\sqrt{n_i n_j}}{N}\right) + \frac{N-n_j}{N}\left(-\frac{\sqrt{n_i n_j}}{N}\right) + \sum_{l\neq i,j} -\frac{\sqrt{n_i n_l}}{N}\left(-\frac{\sqrt{n_l n_j}}{N}\right) \\
&= \frac{(N-n_i+N-n_j)\sqrt{n_i n_j}}{N^2} + \frac{\sqrt{n_i n_j}(N-n_i-n_j)}{N^2} \\
&= \frac{\sqrt{n_i n_j}(N-n_i-n_j-(N-n_i+N-n_j))}{N^2} \\
&= -\frac{\sqrt{n_i n_j}}{N}.
\end{aligned}$$

Seega on Σ idempotentne ja $X^T X$ on χ^2-jaotusega vabadusastmete arvuga $df = \mathrm{rank}(\Sigma) = k - 1$.

Veel mõningaid tulemusi. Post-hoc test — loe kaks töötlust/populatsiooni teineteisest erinevaks, kui

$$|\overline{R_i} - \overline{R_j}| \geq z_{\alpha/teste}\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$

Sarnaselt Wilcoxoni testi situatsioonile võime ka Kruskal-Wallise testi puhul asendada astakud mingi astakust sõltuva funktsiooniga, st i asemel summeerida/keskmistada arve $a(i)$. Näiteks võime (juhul kui kahtlustame, et vaatlused võiksid olla normaaljaotusest) kasutada $a(i) = \Phi^{-1}\left(\frac{i}{N+1}\right)$.

Sellisel juhul kasutame teststatistikuna statistikut

$$K_a = \frac{1}{s_a^2} \sum_{i=1}^{k} \frac{(\sum_{j=1}^{n_i} a(R_{ij}) - n_j \overline{a})^2}{n_j},$$

kus $\overline{a} = \frac{1}{N} \sum_{i=1}^{N} a(i)$ ja $s_a^2 = \frac{1}{N-1} \sum_{i=1}^{N} (a(i) - \overline{a})^2$.

Juhul kui H_0 kehtib, siis $K_a \to \chi^2_{df=k-1}$.

Võrdsete vaatluste olemasolu korral võib kasutada keskmiseid astakuid (midrank), kuid lisaks tuleb jagada statistiku väärtus hajuvuse vähenemist kirjeldava kordajaga:

$$K^* = K/const,$$

kus

$$const = 1 - \sum (d_i^3 - d_i)/(N^3 - N).$$

Toodud valemis võetakse summa üle kõigi unikaalsete väärtuste ja d_i näitab, mitu korda esines i. unikaalne vaatlus valimis.