Mètodes basats en rangs
tècniques concretes: test de Kruskal-Wallis

Mètodes no paramètrics i de remostratge
Grau interuniversitari en Estadística UB – UPC

Prof. Jordi Ocaña Rebull
Departament d'Estadística, Universitat de Barcelona

- Considerat l'alternativa basada en rangs a l'ANOVA d'un factor *A* amb *a* nivells. Adequat per a **comparar** paràmetres de **localització** d'***a* grups independents**
- $\mathbf{Y} = (Y_{11},\ldots,Y_{1n_1},\ldots, Y_{a1},\ldots,Y_{an_a})$ mostra aleatòria formada per *a* submostres de mides n_1, n_2, ..., n_a, obtingudes independentment, sota "*a*" condicions diferents associades a distribucions F_1, F_2, ..., F_a, respectivament
- F_i contínues univariants
- $N = n_1 + \ldots + n_a$ mida mostral total

Test de Kruskal-Wallis
planteig – condicions de validesa

- μ_i i σ_i paràmetres de localització i escala (o dispersió) per cada grup i = 1,2,…,a
  - (no necessàriament mitjana i desviació típica)
- La distribució és la mateixa entre nivells excepte possibles ≠ de localització i escala:

$$F_1\left(\frac{y-\mu_1}{\sigma_1}\right) = \ldots = F_a\left(\frac{y-\mu_a}{\sigma_a}\right)$$

- Però també suposarem que $\sigma_1 = \ldots = \sigma_a$ → hipòtesi d'igualtat de distribucions esdevé:

Test de Kruskal-Wallis

planteig – condicions de validesa

$$H_0: \; \mu_1 = \mu_2 = \ldots = \mu_a$$

$$H_1: \; \mu_i \neq \mu_j$$

per alguna parella de nivells *i*, *j*

Test de Kruskal-Wallis

hipòtesis nul·la i alternativa

- **R** = (R_{11},…,R_{1n_1},…, R_{a1},…,R_{an_a}) rangs de la mostra
- Suma de rangs dins cada grup:

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

- Com més heterogenis siguin els $R_{i.}$, o les seves mitjanes $\bar{R}_{i.}$ en cas no balancejat, més evidència a favor de H_1

Test de Kruskal-Wallis
procediment

- Kruskal i Wallis van tenir la idea de fer un ANOVA sobre els rangs

$$F=\frac{MS_A}{MS_R}=\frac{SS_A/(a-1)}{SS_R/(N-a)}=\frac{\sum_{i=1}^{a} n_i\left(\bar{R}_{i.}-\bar{R}_{..}\right)^2 \Big/ (a-1)}{\sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(R_{ij}-\bar{R}_{i.}\right)^2 \Big/ (N-a)}$$

$$SS_{Tot}=\sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(R_{ij}-\bar{R}_{..}\right)^2=SS_A+SS_R$$

Test de Kruskal-Wallis procediment: estadístic de test

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{a} \frac{R_{i.}^{2}}{n_i} - 3(N+1)$$

si H_0 és certa, $H \approx \chi^2(a-1)$

és a dir, "H té distribució asimptòticament khi-quadrat amb $a-1$ graus de llibertat"

Test de Kruskal-Wallis

procediment: estadístic de test

$$H_0: \ \mu_1 = \ldots = \mu_a$$

$H_1: \ \mu_i \neq \mu_j$ per algun $i \neq j$.

"Rebutjarem H_0 si $H \geq \chi^2_\alpha(a-1)$"

(o bé si $p-valor = \Pr\left\{\chi^2 \geq H \middle| H_0\right\} \leq \alpha$)

per $\chi^2 \sim \chi^2(a-1)$

Test de Kruskal-Wallis
procediment: criteri de test

- Si en realitat *Y* es mesura en escala ordinal o amb insuficient precisió, hi poden haver "empats" (ties)
- Estratègia habitual: assignar a cada sèrie de valors empatats els rangs que els tocarien (com si no estesin empatats) i posteriorment substituir-los per la seva mitjana → tota la sèrie de valors empatats queda amb el mateix rang mitjà

Test de Kruskal-Wallis
empats

- Alguns autors recomanen l'estadístic "corregit pels empats", $H' = H / C$, amb la mateixa distribució asimptòtica khi-quadrat, on:

$$C = 1 - \frac{\sum_{i=1}^{s}\left(t_i^3 - t_i\right)}{N^3 - N}$$

$s =$ nombre de sèries de valors empatats

$t_i =$ llargada de sèrie i de valors empatats

Test de Kruskal-Wallis
empats

- Alternativa vàlida a l’enfoc paramètric normal si possibles diferències als paràmetre de localització, però no als d’escala
- El test de Kruskal-Wallis quan $a = 2$ coincideix amb la versió asimptòtica del test de Mann-Whitney-Wilcoxon per una alternativa bilateral

Test de Kruskal-Wallis
comentaris finals