Mètodes basats en rangs
idees generals

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

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

- $\mathbf{Y} = (Y_1,\dots,Y_n)$ mostra aleatòria de distribució **contínua** univariant F
- Idea intuïtiva de rang d'Y_i: posició que ocuparia aquell valor dins la mosta ordenada:
  - P.e. si $n = 3$, $Y_1 = 17.1$, $Y_2 = 82.7$, $Y_3 = 6.9$, els rangs serien $R_1 = 2$, $R_2 = 3$, $R_3 = 1$ ja que en la mostra ordenada, el primer valor seria el segon, el segon el tercer i el tercer el primer: $Y_{(1)} = Y_3 = 6.9$, $Y_{(2)} = Y_1 = 17.1$, $Y_{(3)} = Y_2 = 82.7$

Concepte de rang

- $\mathbf{Y} = (Y_1,\ldots,Y_n)$ mostra aleatòria de distribució contínua univariant F
- $\mathbf{Y}_{()} = \{\mathbf{Y}\} = (Y_{(1)},\ldots,Y_{(n)})$ estadístic ordinal corresponent
- Direm que Y_i té rang R_i si $Y_i = Y_{(R_i)}$
  - (Comproveu-ho en l'exemple anterior...)
- (Idea operativa darrere aquesta definició: substituint els índexs de la mostra ordenada pels rangs reconstruïm la mostra original)

Definició més formal de rang

Estadístic basat en rangs

- "Estadístic basat en rangs": Estadístic, diguem-ne "*u*", que només depèn de la mostra **Y** a través dels seus rangs.
- Si **R** = rangs(**Y**) = $(R_1,\ldots,R_n)$, llavors $u(\mathbf{Y}) = g(\mathbf{R})$
- Si tots els elements de la mostra provenen de la mateixa *F* contínua, fàcil determinació de la distribució mostral d'*u*

- En general s'hi perd "relativament poc"
- Correlacions altes entre dades originals i rangs, normalment d'ordre 0,8 o superiors
  - tant a nivell mostral (càlcul de r)
  - com en estudis teòrics per determinades distribucions (càlcul de ρ)
- Potència relativa test rangs vs test paramètric: normalment < 1 (però no molt allunyada) sota condicions validesa test paramètric

Què s'hi perd treballant amb rangs i no amb la mostra original?

- **Y** = ($Y_1,\ldots,Y_n$) mostra aleatòria de distribució contínua univariant F
  - (recordeu, totes les Y_i provenen d'idèntica F)
- **R** = ($R_1,\ldots,R_n$) els seus rangs
- "Qualsevol permutació **r** = ($r_1,\ldots, r_n$) dels valors 1, 2, … n, té la mateixa probabilitat: $P\left\{\mathbf{R}=\mathbf{r}\right\}=1/n!$ "
- Conseqüència: la distribució de qualsevol estadístic u basat en **R** es podrà obtenir, a priori enumerant els possibles valors d'u

Teorema fonamental sobre rangs

- Sota hipòtesi que tots els elements de la mostra provenen d'idèntica distribució, qualsevol permutació de rangs (i de valors) té la mateixa probabilitat: 1 / (n!)
- Avantatge de treballar amb rangs: sempre són els mateixos: 1, 2, ..., n
- Distribució d'estadístic basat en rangs **es pot tabular a priori**, abans de tenir la mostra

Fonament de tests de rangs i de permutacions: el mateix

$$P\{R_i = r\} = \frac{1}{n} \text{ per } r = 1,\dots,n$$

$$P\{R_i = r, R_j = s\} = \frac{1}{n(n-1)} \text{ per } r \neq s = 1,\dots,n$$

$$E(R_i) = \frac{n+1}{2}$$

$$\mathrm{var}(R_i) = \frac{(n+1)(n-1)}{12}$$

$$\mathrm{cov}(R_i, R_j) = -\frac{n+1}{12} \text{ per } i \neq j$$

Conseqüències directes dels resultats anteriors

El problema dels empats

- Si "empats" a ***Y***: rangs no ben definits
- En principi si *F* contínua, Pr{empats} = 0
- Però empats freqüents si ***Y*** en escala ordinal o poca precisió de les mesures
- Normalment començarem suposant que no hi ha empats
- Però a la pràctica n'hi acostuma a haver: caldrà considerar com afecten els mètodes, si es pot corregir per empats, etc.