Appunti di Statistica Non Parametrica

Alberto Cavalin × *Campigotto Federico*

2007/2008

Indice

Capitolo 1

Test parametrici

1.1 Test t di Student a 2 campioni

Assunzioni:

- per dati non appaiati:
  $X_{Ai} \sim N(\mu_A, \sigma^2)$, i.i.d.
  $X_{Bi} \sim N(\mu_B, \sigma^2)$, i.i.d.
- per dati appaiati:
  $X_{Ai} \sim N(\mu_A, \sigma^2)$, i.d.
  $X_{Bi} \sim N(\mu_B, \sigma^2)$, i.d.
  cioè indip. a coppie:
  $(X_{Ai},\, X_{Bi}) \perp\!\!\!\perp (X_{Aj},\, X_{Bj}) \quad \forall i, j$

Ipotesi:

$$\begin{cases} H_0: & X_A \overset{d}{=} X_B \Leftrightarrow \mu_A = \mu_B \\ H_1: & X_A \overset{d}{<\neq>} X_B \end{cases}$$

per $X_A \perp\!\!\!\perp X_B$ (dati non appaiati): $t^{oss} = \dfrac{\left(\overline{X}_B - \overline{X}_A\right) \Big/ \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}{\sqrt{\frac{\sum (X_{Ai} - \overline{X}_A)^2 + \sum (X_{Bi} - \overline{X}_B)^2}{n_A + n_B - 2}}} \sim t_{n_A + n_B - 2}$

per $X_A \not\perp\!\!\!\perp X_B$ (dati appaiati): $t^{oss} = \dfrac{\sum d_i}{\sqrt{\frac{\sum (d_i - \overline{d})^2}{n(n-1)}}} \sim t_{n-1}, \quad n = n_A = n_B,\ d_i = X_{Bi} - X_{Ai}$

NB: Se $\rho = \sigma_2/\sigma_1$ è sconosciuto, $\forall \eta = n_2/n_1$, sotto condizioni di normalità, non esiste una soluzione esatta tramite la t di Student.

R: `t.test(x1,x2,var.equal=T)`

R: `t.test(x1,x2,var.equal=T,paired=T) # per dati appaiati`

1.2 Test di Welch a 2 campioni

Assunzioni:

- $X_{Ai} \sim N(\mu_A, \sigma_A^2)$, i.i.d.
- $X_{Bi} \sim N(\mu_B, \sigma_B^2)$, i.i.d.
- $\sigma_A^2 \neq \sigma_B^2$

Ipotesi:

$$\begin{cases} H_0: & X_A \overset{d}{=} X_B \Rightarrow \mu_A = \mu_B \\ H_1: & X_A \overset{d}{<\neq>} X_B \end{cases}$$

$$t^{oss} = \frac{\overline{X}_B - \overline{X}_A}{\sqrt{\frac{\sum(X_{Ai}-\overline{X}_A)^2}{n_A-1} + \frac{\sum(X_{Bi}-\overline{X}_B)^2}{n_B-1}}} \dot{\sim} t_{g^*}$$

g.l. = vedi appunti

R: `t.test(x1,x2,var.equal=F)`

1.3 ANOVA (analisi della varianza per confronti multipli)

Assunzioni:

- $X_1, \cdots, X_m \quad , \quad X_i \sim N(\mu_i, \sigma^2)$
- $X_i \perp\!\!\!\perp X_j \quad , \quad \forall\, i \neq j$

Ipotesi:

$$\begin{cases} H_0: & X_1 \overset{d}{=} \ldots \overset{d}{=} X_m \Leftrightarrow \mu_1 = \ldots = \mu_m \\ H_1: & X_i \overset{d}{\neq} X_j \quad \text{per almeno un } i, j \end{cases}$$

$$F^{oss} = \frac{n * \sum_i (\overline{x}_i - \overline{x})^2/(m-1)}{\sum_i \sum_j (x_{ij} - \overline{x}_i)^2/(m(n-1))} \sim F_{(m-1),m(n-1)}$$

R: `aov(formula) ; summary(aov.obj)`

1.3.1 Test t di Bonferroni

Effettuo un test t di Student per ogni coppia di gruppi, e rifiuto H_0 se si rifiuta almeno un test.
Il p-value globale va diviso tra il numero dei test (dalla disuguaglianza di Bonferroni).

Test t di Bonferroni con un solo gruppo di controllo

Con le stesse ipotesi vengono però effettuati i confronti solo con il gruppo di controllo; perciò ho meno test da effettuare e quindi un p-value locale più alto.

1.3.2 Test di SNK (Student-Newman-Keuls)

- si ordinano per media crescente i gruppi

- effettuo un test q (SNK) per ogni coppia di gruppi
- rifiuto H_0 se si rifiuta almeno un test
- individuo le coppie discordi dai test rifiutati

NB: Se non esiste differenza significativa fra due medie, si deve concludere che non esiste differenza neppure fra le medie comprese fra esse.

$$q^{oss} = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{\frac{S^2_{entro}}{2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \qquad \rightarrow \text{valori tabulati}$$

Test di Dunnet (SNK con un solo gruppo di controllo)

Con le stesse ipotesi dell'SNK vengono però effettuati i confronti solo con il gruppo di controllo. Questo test è più sensibile del test t di Bonferroni (tendo a rifiutare di più).

1.3.3 Test di Tukey

Stesse ipotesi e procedura dell'SNK,ma con un p-value diverso; ciò implica che questo test è più prudente dell'SNK (tendo a rifiutare di più).

R: `TukeyHSD(aov.obj)`

1.4 Test di Shapiro-Wilk

Assunzioni:

- $X_i \sim D(\cdots),$ i.i.d.

Ipotesi:

$$\begin{cases} H_0: & X \overset{d}{=} N(\mu, \sigma^2) \\ H_1: & X \overset{d}{\neq} N(\mu, \sigma^2) \end{cases}$$

$$T_{SW} = \frac{\sum\left(w_i X_{(i)}\right)}{\sum\left(X_i - \overline{X}\right)^2} \qquad \rightarrow \text{valori tabulati}$$

NB: I w_i sono dei pesi "ad-hoc" e gli $X_{(i)}$ sono i valori osservati ordinati.

R: `shapiro.test(x) # x=vettore numerico`

1.5 Test Chi-quadro

Assunzioni:

- X_{Ai}, X_{Bi} categoriali

Ipotesi:

$$\begin{cases} H_0: & X_A \perp\!\!\!\perp X_B \\ H_1: & X_A \not\perp\!\!\!\perp X_B \end{cases}$$

Procedura del test:

- si calcoli la tabella di contingenza (rxc) per X_A e X_B
- $T_{\chi^2} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij}-n^*_{ij})^2}{n^*_{ij}} \sim \chi^2_{(r-1)(c-1)}\,, \quad n^*_{ij} = (n_{i\cdot} + n_{\cdot j})/n$

R: `chisq.test(x1,x2)` oppure `fisher.test(x1,x2)`

R: `chisq.test(x1,x2,correct=T) # con correzione per freq basse` ≤ 5

1.6 Estensione multivariata del test t di Student

1.6.1 Test T-square di Hotelling

Assunzioni:

- $X_{jAi} \sim N(\mu_A, \sigma^2)$, i.i.d.
  $X_{jBi} \sim N(\mu_B, \sigma^2)$, i.i.d.
  $j = 1, \ldots, q$, q = numero v.c.

Ipotesi:

$$\begin{cases} H_0: & X_A \overset{d}{=} X_B \Leftrightarrow \mu_A = \mu_B \\ H_1: & X_A \overset{d}{\neq} X_B \end{cases}$$

$$T^2_H = \frac{n_A n_B}{n_A+n_B}(\overline{\mathbf{x}}_A - \overline{\mathbf{x}}_B)^T \mathbf{W}^{-1} (\overline{\mathbf{x}}_A - \overline{\mathbf{x}}_B)$$

dove

$$\mathbf{W} = \frac{\sum_{i=1}^{n_A}(\mathbf{x}_{Ai}-\overline{\mathbf{x}}_A)(\mathbf{x}_{Ai}-\overline{\mathbf{x}}_A)' + \sum_{i=1}^{n_B}(\mathbf{x}_{Bi}-\overline{\mathbf{x}}_B)(\mathbf{x}_{Bi}-\overline{\mathbf{x}}_B)'}{n_A+n_B-2} \quad \text{e} \quad \frac{n_A+n_B-q-1}{(n_A+n_B-2)q} T^2_H \sim F_{(q,n_A+n_B-1-q)}$$

1.6.2 Test OLS di O'Brien

Assunzioni:

- $X_{jAi} \sim N(\mu_A, \sigma^2)$, i.i.d.
  $X_{jBi} \sim N(\mu_B, \sigma^2)$, i.i.d.
  $j = 1, \ldots, q$, q = numero v.c.

Ipotesi:

$$\begin{cases} H_0: & X_A \overset{d}{=} X_B \Leftrightarrow \mu_A = \mu_B \\ H_1: & X_A \overset{d}{<>} X_B \end{cases}$$

$$T_O = (T_1 + T_2) \Big/ \sqrt{\begin{pmatrix}1\\1\end{pmatrix}^T \begin{pmatrix}1 & \hat{\rho}_{12}\\ \hat{\rho}_{12} & 1\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix}} \dot{\sim} N(0,1) \qquad \hat{\rho}_{12} = \hat{\text{Cov}}(X_i, X_j)$$

dove T_h è la statistica t di Student applicata all'h-sima dimensione.

Capitolo 2

Test non parametrici

Sia $\mathscr{X}^{(n)}$ lo spazio campionario, X il campione osservato, e $\mathscr{X}^{(n)}|_X$ lo spazio campionario di permutazione:

$$\begin{aligned}\mathscr{X}^{(n)}|_X &= \{X^* : \text{rapporto di verosim. } \rho(X,\, X^*) = 1\} \\ &= \left\{(X_1^*,\, \ldots,\, X_n^*) = (X_{u_1^*},\, \ldots,\, X_{u_n^*})\,,\ \forall (u_1^*,\, \ldots,\, u_n^*) \text{ permutazione di } (1,\, \ldots,\, n)\right\} \\ &= \{\text{tutte le possibili permutazioni di } X\}\end{aligned}$$

Se vale il principio di scambiabilità dei dati, è cioè possibile applicare un processo di randomizzazione, allora:

$$\mathbb{P}(X^* \in A^{(n)} | X^* \in \mathscr{X}^{(n)}|_X) = \frac{\mathbb{P}(X^* \in A^{(n)} \cap X^* \in \mathscr{X}^{(n)}|_X)}{\mathbb{P}(X^* \in \mathscr{X}^{(n)}|_X)} = \frac{\#(X^* \in A^{(n)})}{\#(X^* \in \mathscr{X}^{(n)}|_X)}, \quad A^{(n)} \subseteq \mathscr{X}^{(n)}|_X$$

2.1 Statistiche associative

È detta statistica associativa una funzione del tipo:

$$T(X) = \text{mean}\,(\varphi(X)) = \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i), \quad \varphi(\cdot) = \text{una funz. qualunque di } (\cdot)$$

NB: Test associati a statistiche associative non costanti sono non distorti uniformemente.

Oss: La funzione di ripartizione empirica $\hat{F}(x)$ è per costruzione una statistica associativa, ed è invariante rispetto a permutazioni dei dati:

$$\hat{F}(X, x) = \frac{1}{n}\#(X \le x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \le x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i^* \le x) = \hat{F}(X^*, x)$$

2.2 Come costruire un test di permutazione

Solitamente un test parametrico è una funzione $T : \mathscr{X}^{(n)} \to \mathbb{R}$, mentre per un test (non parametrico) di permutazione si ha:

$$T^* : \mathscr{X}^{(n)}|_X \to \tau_X = \{T_1, \ldots, T_\omega\}, \quad T_i \in \mathbb{R}, \quad \omega = \text{\# permutazioni per X}$$

dove $\tau_X \in \mathbb{R}^\omega$ è la distribuzione (condizionata ai dati) di permutazione del test; allora considerando i valori grandi come significativi e fissato un α, il valore di soglia viene definito come:

$$T_\alpha := \left\{ \frac{\#(T_i \geq T_\alpha)}{\omega} = \alpha \right\}$$

2.2.1 Algoritmo del test

Dati due campioni X_A e X_B aventi rispettivamente n_1 e n_2 osservazioni:

1. si calcoli una statistica test T_0 sui dati di partenza
2. calcolare T_i per ogni possibile permutazione dei dati:
   - se $X_A \perp\!\!\!\perp X_B \rightsquigarrow$ permutazione di X_{Ai} e X_{Bj}, $\forall i = 1, \ldots, n_1$ e $\forall j = 1, \ldots, n_2$ $\hookrightarrow \omega = \binom{n_1+n_2}{n_1}$
   - se $X_A \not\perp\!\!\!\perp X_B$ (dati appaiati) $\rightsquigarrow$ permutazione di X_{Ai} e X_{Bi}, $\forall i = 1, \ldots, n$ $\hookrightarrow \omega = 2^n\,, \qquad n = n_1 = n_2$
3. si calcoli il p-value $\lambda = \#(T_i \geq T_0)/\#(T_i) = \#(T_i \geq T_0)/\omega$
4. se risulta essere $\lambda < \alpha \Rightarrow$ rifiuto H_0

NB: Se ω è molto grande scelgo in modo casuale un sottoinsieme delle permutazioni possibili.

2.2.2 Test equivalenti

Due test T_1^* e T_2^* sono equivalenti se uno è funzione biunivoca dell'altro, o equivalentemente, essi devono generare il medesimo p-value per ogni possibile campione: $\lambda_{T_1} = \lambda_{T_2}, \ \forall X \in \mathscr{X}|_X$. È quindi conveniente utilizzare la statistca test computazionalmente meno onerosa tra le equivalenti.

2.3 Elenco dei test

2.3.1 Test di permutazione standard

NB: È un test non parametrico analogo alla t di Student a due campioni.

Assunzioni:

- $X_{Ai} \sim D_A(\cdots),$ i.d.
  $X_{Bi} \sim D_B(\cdots),$ i.d.
- $X_{\cdot i} \in \mathbb{R}$

Ipotesi:

$$\begin{cases} H_0: & X_A \overset{d}{=} X_B \\ H_1: & X_A \overset{d}{<\neq>} X_B \end{cases}$$

Procedura del test:

- per $X_A \perp\!\!\!\perp X_B$ (dati non appaiati): $T^* = \frac{\overline{X}^*_A - \overline{X}^*_B}{\sigma} \equiv \overline{X}^*_A$
- per $X_A \not\perp\!\!\!\perp X_B$ (dati appaiati): $T^* = \frac{\sum(X_{Ai} - X_{Bi})}{\sigma} \equiv \sum (X_{Ai} - X_{Bi}) \equiv \sum d_i s_i^* \dot{\sim} \mathrm{N}(0,1)$
  $d_i = |X_{Ai} - X_{Bi}|, \quad s_i = \mathrm{sgn}(X_{Ai} - X_{Bi})$

2.3.2 Test di Mann-Whitney

NB: È un test non parametrico analogo alla t di Student a due campioni.

Assunzioni:

- $X_{Ai} \sim D_A(\cdots),$ i.i.d.
  $X_{Bi} \sim D_B(\cdots),$ i.i.d.
- $X_A \perp\!\!\!\perp X_B$
- $X_{\cdot i} \in \mathbb{R} \Rightarrow$ no doppioni
- sotto H_1 si ha che $F_A <> F_B$, cioè le f.d.r. $F_{\cdot}$ non si incrociano mai

Ipotesi:

$$\begin{cases} H_0: & X_A \overset{d}{=} X_B \\ H_1: & X_A \overset{d}{<\neq>} X_B \end{cases}$$

Procedura del test:

- si costruisca il vettore unico $X = X_A \biguplus X_B$
- si calcoli $R_i = \mathrm{Rank}(X_i) = \#(X_j \leq X_i)$
- si separi R^* in 2 vettori $R^*_A,\ R^*_B$ in base alla numerosità dei campioni di origine
- $T^*_{MW} = \frac{\overline{R}^*_A - \overline{R}^*_B}{\sqrt{\mathrm{Var}(R^*)}} \equiv \frac{\sum\left(R^*_{Ai} - \mathrm{M}'\right)}{\sqrt{\mathrm{V}'}} \equiv \sum R^*_{Ai}$
  - se $n_A \leq 7$ o $n_B \leq 7,\ T^*_{MW} \rightarrow$ valori tabulati
  - se $n_A \geq 7$ e $n_B \geq 7,\ T^*_{MW} \dot{\sim} \mathrm{N}(0,1)$

Dove $\mathrm{M}' = n_A(n+1)/2$ e $\mathrm{V}' = n_A n_B (n+1)/12$.

Prop: Grazie alla trasformazione in ranghi, il test non è più condizionato ai dati osservati.

NB: È possibile inoltre utilizzare la correzione di Fisher-Yates per migliorare la normalità:

$$T^*_{FY} = \overline{\psi}^*_A - \overline{\psi}^*_B \equiv \overline{\psi}^*_A \dot{\sim} N\left(0,\ \frac{n-n_A}{n(n-1)}\right), \quad \psi_{ji} = \psi(R_{ji}) = \Phi^{-1}(R_{ji}/(n+1))$$

R: `wilcox.test(x1,x2,exact=F) # p-value asintotico`
R: `wilcox.test(x1,x2,exact=T) # p-value esatto`

2.3.3 Test di Wilcoxon

NB: È un test non parametrico analogo alla t di Student a due campioni per dati appaiati.

Assunzioni:

- dati appaiati X_{Ai}, X_{Bi} continui
  cioè indip. a coppie:
  $(X_{Ai},\ X_{Bi}) \perp\!\!\!\perp (X_{Aj},\ X_{Bj}) \quad \forall i, j$
- $d_i = (X_{Bi} - X_{Ai})$ i.i.d.
- $|d_i|$ non ripetuti

Ipotesi:

$$\begin{cases} H_0: & X_A \overset{d}{=} X_B \\ H_1: & X_A \overset{d}{<\neq>} X_B \end{cases}$$

Procedura del test:

- si calcoli $|d_i|$
- si calcoli $R_i = \text{Rank}(|d_i|) = \#(|d_j| \leq |d_i|)$
- $T^*_W = \frac{\sum(R_i \text{sgn}(d_i) - \text{Mean}(R))}{\sqrt{\text{Var}(R)}} \equiv T^*_W{}' = \frac{\sum\left(R_i \mathbb{I}_{[0,+\inf)}(d_i) - \text{M}'\right)}{\sqrt{\text{V}'}} \dot{\sim} \text{N}(0,1)$

Dove $\text{M}' = n(n+1)/4$ e $\text{V}' = n(n+1)(2n+1)/24$.

Prop: Grazie alla trasformazione in ranghi, il test non è più condizionato ai dati osservati.
NB: In modo analogo al test di Mann-Whitney, è possibile inoltre utilizzare la correzione di Fisher-Yates, o quella di Van Der Waerden per migliorare la normalità.

R: `wilcox.test(x1,x2,paired=T,exact=F) # p-value asintotico`
R: `wilcox.test(x1,x2,paired=T,exact=T) # p-value esatto`

2.3.4 Test di McNemar

NB: È un test non parametrico analogo alla t di Student a due campioni per dati binari.

Assunzioni:

- dati appaiati X_{Ai}, X_{Bi}
- $X_{Ai} \sim D_A(\cdots)$, i.d.
  $X_{Bi} \sim D_B(\cdots)$, i.d.

Ipotesi:

$$\begin{cases} H_0: & X_A \overset{d}{=} X_B \\ H_1: & X_A \overset{d}{<>} X_B \end{cases}$$

Procedura del test:

- $X_{\cdot i} \in \mathbb{R}$
- $T^*_{Mc} = \sum \frac{d_i s^*_i}{|d_i|} \equiv \frac{d_i s^*_i}{\sqrt{\sum d_i^2}}$

Procedura alternativa del test:

- $X_{\cdot i} \in \mathbb{R}$, o $X_{\cdot i}$ qualitativi ordinali, o binari
- si considerano i segni delle d_i
- si calcoli la tabella di contingenza (freq\differenza dei valori):

A\B	0	1
0	$f_{00}\backslash 0$	$f_{01}\backslash 1$
1	$f_{10}\backslash -1$	$f_{11}\backslash 0$

$f_{01}, f_{10} \sim \text{Bin}(f_{01} + f_{10}, \frac{1}{2})$

- $T_{Mc} = f_{01} \equiv f_{10} \sim \text{Bin}(f_{01} + f_{10}, \frac{1}{2})$

R: `mcnemar.test(x1,x2) # xi=vettore categoriale`

R: `wilcox.test(tbc) # tbc=tabella di contingenza`

2.3.5 Test Chi-quadro di permutazione e Test esatto di Fisher

Assunzioni:

- X_{Ai} i.d., X_{Bi} i.d. ambedue qualitative con $k \geq 2$ livelli ciascuna

Ipotesi:

$$\begin{cases} H_0: & X_A \overset{d}{=} X_B \\ H_1: & X_A \overset{d}{\neq} X_B \end{cases}$$

Solitamente per verificare tali ipotesi viene utilizzato il test parametrico Chi-quadro, ma quando nella tabella di contingenza si è in presenza di frequenze di cella basse (< 5), è possibile far affidamento alla sua distribuzione di permutazione, svincolandosi così da questa debolezza.

Procedura del test:

- si permuta come nel caso $X_A \perp\!\!\!\perp X_B$ (tra i gruppi)
- $T^*_{\chi^2} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - f^*_{ij})^2}{f^*_{ij}}$, $f_{ij} =$ freq. oss, $f^*_{ij} =$ freq. attese

NB: Nel caso $k = 2$ viene utilizzato il test (delle probabilità esatte) di Fisher, il quale si basa sulla distribuzione ipergeometrica per calcolare il valore esatto da ottenere nelle celle.

→ per numerosità basse viene utilizzato il test esatto di Fisher

→ per numerosità alte viene usato il test Chi-quadro di permutazione

2.3.6 Test per v.c. qualitative ordinali

Assunzioni:

- X_{Ai} i.d., X_{Bi} i.d. ambeue qualitative ordinali con $k \geq 2$ livelli ciascuna

Ipotesi:

$$\begin{cases} H_0: & X_A \stackrel{d}{=} X_B \\ H_1: & X_A <> X_B \end{cases}$$

Per questo problema sono state proposte diverse soluzioni, come: il $\overline{\chi}^2$, il χ^2 cumulato, la trasformazione in ranghi delle osservazioni o delle classi; ma queste presentano alcuni inconvenienti anche severi. La soluzione viene dalla famiglia di test detta goodness of fit (o di bontà di adattamento, o di divergenza tra distribuzioni):

- Cramer-Von Mises
  $T^*_{CM} = \sum_{ji} \hat{F}^*_A(X_{ji}) - \hat{F}^*_B(X_{ji}) \equiv \sum_{j=1}^{k} \hat{F}^*_A(L_j) - \hat{F}^*_B(L_j)$
- Anderson-Darling
  $T^*_{AD} = \sum_{j=1}^{k} \left[\hat{F}^*_A(L_j) - \hat{F}^*_B(L_j)\right] \Big/ \sqrt{\hat{F}_C(L_j)\left(1-\hat{F}_C(L_j)\right)\frac{4n_A}{n_n(n_a+n_b-1)}}$
- Kolmogorov
  $T_K = \text{sup}_t \left|\hat{F}_A(t) - \hat{F}_B(t)\right|, \qquad$ NB: $\text{sup}_t \left|\hat{F}_n(t) - F^0(t)\right| \stackrel{q.c.}{\to} 0$
- Kolmogorov-Smirnov
  $T^*_{KS} = \text{max}_i \left(\hat{F}^*_{Ai} - \hat{F}^*_{Bi}\right)$
- Kolmogorov-Smirnov-Anderson-Darling
  $T^*_{KSAD} = \text{max}_i \left(\hat{F}^*_{Ai} - \hat{F}^*_{Bi}\right) \Big/ \sqrt{\hat{F}_C\left(1-\hat{F}_C\right)}$
- test della mediana di Brown-Mood
  si dicotomizza X_A e X_B rispettivamente alla mediana $\tilde{X}$ di $X_A \biguplus X_B$
  si crea la tabella di contingenza:

	A	B
$\leq \tilde{X}$	f_{11}	f_{12}
$> \tilde{X}$	f_{21}	f_{22}

  e si considera il test $T^*_{BM} = f^{dx}_{11}$ oppure f^{sx}_{12}

Notazioni e proprietà:
$\hat{F}_C =$ f.d.r. per $X_A \biguplus X_B$ $\qquad L_j =$ livello j-mo
$\mathbb{E}(\hat{F}_A(L_j)) = \hat{F}_A(L_j) \qquad \mathbb{V}(\hat{F}_A(L_j)) = F_A(L_j)(1-F_A(L_j))/n_A$

NB: T^*_{KSAD} e T^*_{KS} sono spesso i migliori ma a volte T^*_{MW} prevale.

2.3.7 Analogo dell'Anova ad una via

Sia $X = \{X_{hji} \ : \ h = 1, \ldots, q; \ j = 1, \ldots, c; \ i = 1, \ldots, n_j\}$ v.c. q-dimensionale con c campioni di numerosità n_j ciascuno. È possibile trasformare X da q aspetti d'interesse a $k <=> q$.

$$\text{Ipotesi:} \begin{cases} H_0: & X_1 \overset{d}{=} \cdots \overset{d}{=} X_c \equiv \left\{\bigcap_{i=1}^{k} H_{0i}\right\} \\ H_1: & \left\{\bigcup_{i=1}^{k} H_{1i}\right\} \end{cases}$$

Per ogni componente dell'ipotesi nulla globale deve esistere un test T_i. Ogni test dev'essere non distorto ed almeno uno dev'essere consistente $\Rightarrow$ i λ_i sono positivamente dipendenti, cioè sono stocasticamente ordinati.

Per ottenere il p-value globale si utilizza una combinazione non parametrica $\psi(\lambda_1, \ldots, \lambda_k)$ di quelli parziali, la quale deve soddisfare le seguenti proprietà:

- ψ continua e non crescente in ogni argomento
- $\lim_{\lambda_i \to 0} \psi(\cdot) = \overline{\psi} = \text{sup}(\psi(\cdot))$ possibilmente infinito
- $\forall \alpha$ si ha che il valore critico della funzione di combinazione $\psi_\alpha < \overline{\psi}$

Funzioni di combinazione più usate:

- Fisher: $T_F'' = -2 \sum \log(\lambda_i)$
- Tippet: $T_T'' = \min(\lambda_i)$ o $\max(1 - \lambda_i)$
- Liptak-normal: $T_{LN}'' = \sum \Phi^{-1}(1 - \lambda_i), \quad \Phi = \text{f.d.r di una } N(0,1)$
- Liptak-gamma: $T_{LG}'' = \sum \Gamma_{a,b}^{-1}(1 - \lambda_i)$
- Liptak-logistic: $T_{LL}'' = \sum \log\left[(1 - \lambda_i)/\lambda_i\right]$
- Quadratica: $T_Q'' = U^T R_U^{-1} U$
  con $U = [\Phi^{-1}(1 - \lambda_1), \ldots, \Phi^{-1}(1 - \lambda_k)]\,,\ R_U = \{\mathbb{C}\text{ov}(U_i, U_j),\ i, j = 1, \ldots, k\}$
- Diretta: $T_D'' = \sum F_T^{-1}(1 - \lambda_i) = \sum T_i$, tutti i T_i devono avere la stessa distribuzione

Procedura del test:

1. nel caso $c = 2$, si consideri la matrice $X = X_{\cdot 1 \cdot} \biguplus X_{\cdot 2 \cdot}$:

$$\left[\begin{array}{ccc|ccc} X_{111} & \cdots & X_{11n_1} & X_{121} & \cdots & X_{12n_2} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ X_{q11} & \cdots & X_{q1n_1} & X_{q21} & \cdots & X_{q2n_2} \end{array}\right]$$

2. si calcoli il T^0 per ogni variabile, e la relativa distribuzione di permutazione:

$$\left[\begin{array}{c|ccc} T_1^0 & T_{11}^* & \cdots & T_{1B}^* \\ \vdots & \vdots & \ddots & \vdots \\ T_q^0 & T_{q1}^* & \cdots & T_{qB}^* \end{array}\right]$$

3. per ogni variabile i, si calcoli il λ_i^0 e la funzione p-value su ogni T_{ij}^*:

$$\left[\begin{array}{c|ccc} \lambda_1^0 & \lambda_{11}^* & \cdots & \lambda_{1B}^* \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_q^0 & \lambda_{q1}^* & \cdots & \lambda_{qB}^* \end{array}\right], \quad \lambda_{ib} = \frac{\frac{1}{2} + \#(T_{ih}^* \geq T_{ib}^*)}{B+1}, \quad b = 1, \ldots, B$$

4. scelta una funzione di combinazione, si effettuano le $B+1$ combinazioni dei p-value parziali:

$$\left[\begin{array}{c|ccc} \psi^0 & \psi_1^* & \cdots & \psi_B^* \end{array}\right]$$

5. ottengo il p-value globale per H_0: $\lambda_G^0 = \frac{\#(\psi_b^* \geq \psi^0)}{B}, \quad b = 1, \ldots, B$

Procedura di combinazione iterata:

La conclusione finale dipende però dalla scelta più o meno arbitraria di ψ. Per svincolarsi da tale scelta si procede nel seguente modo:

- arrivati al passo (4), si calcolano $k > 1$ funzioni di combinazione ψ:

$$\left[\begin{array}{c|ccc} \psi_1^0 & \psi_{11}^* & \cdots & \psi_{B1}^* \\ \vdots & \vdots & \ddots & \vdots \\ \psi_k^0 & \psi_{1k}^* & \cdots & \psi_{Bk}^* \end{array}\right]$$

- si calcolano i k p-value globali come al punto (5), e:
  – se i $\lambda_{Gi}^0,\ i = 1, \ldots, k$ coincidono tutti tra loro $\Rightarrow$ STOP, $\lambda_G^0 = \lambda_{G0}^0$
  – si riparte dal punto (2) utilizzando la matrice corrente

2.3.8 Test di Kruskal-Wallis

NB: È un test non parametrico analogo all'analisi della varianza.

Assunzioni:

- $X_{ij} \sim D_i(\cdots),$ i.d.
  $i = 1, \cdots, g,$ g=# gruppi
- $X_i \perp\!\!\!\perp X_k \quad \forall i, k$
- $X_{\cdot i} \in \mathbb{R} \Rightarrow$ no doppioni
- sotto H_1 si ha che $F_i <> F_k \quad \forall i, k$, cioè le f.d.r. $F_{\cdot}$ non si incrociano mai

Ipotesi:

$$\begin{cases} H_0: & X_i \overset{d}{=} X_k, \ \forall i, k \\ H_1: & X_i \overset{d}{\neq} X_k \quad \text{per almeno un } i, k \end{cases}$$

Procedura del test:

- si costruisca il vettore unico $X = \biguplus_{i=1}^{g} X_j$

- $T^*_{KW} = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot}-\bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij}-\bar{r})^2} = \frac{12}{N(N+1)}\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot}-\bar{r})^2 \dot{\sim} \chi^2_{g-1}$
  - $n_g =$ numero di osservazioni nel gruppo g
  - $r_{ij} =$ rango globale dell'oss. j dal gruppo i
  - $N = \#(X) =$ numero totale di osservazioni
  - $\bar{r}_{i\cdot} = \sum_{j=1}^{n_i} r_{ij}/n_i$
  - $\bar{r} = (N+1)/2 =$ media di tutti gli r_{ij}

R: `kruskal.test(x~factor)`

R: `kruskal.test(list_of_vectors)`

2.3.9 Analogo dell'Anova a due vie

Fissato uno dei due fattori avente L livelli:

- per ogni livello L_i:
  - viene calcolata un'Anova a una via
  - si ottiene così la distribuzione di permutazione del test: $[\psi_i^0|\psi_{i1}^*\cdots\psi_{iB}^*]$
- si concatenano in sequenza le distribuzioni ottenute: $\left[\begin{array}{c|c} \psi_i^0 & \psi_{i1}^*\cdots\psi_{iB}^* \\ \vdots & \vdots \\ \psi_L^0 & \psi_{L1}^*\cdots\psi_{LB}^* \end{array}\right]$
- eseguo nuovamente un'Anova a una via partendo dal punto (2) sulla matrice ottenuta
- il p-value ottenuto al punto (5) è quello globale finale

2.3.10 Test di Kolmogorov-Smirnov

Assunzioni:

- $X_i \sim D_A(\cdots),$ i.d.
- D distribuzione continua

Ipotesi:

$$\begin{cases} H_0: & X \overset{d}{=} D', D' = \text{ditrib. ipotizzata} \\ H_1: & X \overset{d}{<\neq>} D' \end{cases}$$

Procedura del test:

- si calcoli la F_X f.d.r. empirica, e la F_T f.d.r. teorica
- $T_{KS} = \sup_x |F_X(x) - F_T(x)|$ $\rightarrow$ valori tabulati

R: `ks.test(x1,x2) # confronta le 2 distrib.`

R: `ks.test(x,'pdistrib') # confronta con la distrib. teorica`

2.4 Dati mancanti

Se i dati mancano a caso, cioè sono equamente distribuiti tra i gruppi, allora si possono ignorare e procedere con l'analisi del dataset ridotto ($X|O=1$).

A volte però i dati possono mancare non completamente a caso, ad esempio quando la loro percentuale dipende dal tipo di trattamento utilizzato. Ad ogni variabile X_h viene quindi associata una variabile binaria O_h:

$$\begin{aligned}[O] &= O_{hrj}; \quad h=1,\dots,q; \quad r=1,\dots,c; \quad j=1,\dots,n_r \\ &= \begin{cases} 0 & \text{, se } X_{hrj} \text{ è mancante} \\ 1 & \text{, se } X_{hrj} \text{ è osservato} \end{cases}\end{aligned}$$

Per ovviare a questo inconveniente si considerano statistiche associative del tipo $W=\sum \varphi(X_{hrj})O_{hrj}$ da applicare in $(\mathscr{X}^{(n)}|_X; O)$, cioè una fetta dello spazio di permutazione nel quale rimane costante il numero di dati mancanti nei vari campioni X_{hr}.

È però difficile rendere tutte uguali le distribuzioni di T^* all'interno di una fetta, si cerca quindi di renderle uguali almeno in media e varianza; ciò è possibile considerando un campionamento da popolazione finita e utilizzando la seguente statistica test per due campioni:

$$T^* = W_1^*\sqrt{f_2^*/f_1^*} - W_2^*\sqrt{f_1^*/f_2^*}, \qquad f_r^* = \sum_j O_{rj}^*$$

2.5 Conclusioni sull'ipotesi globale date k parziali

Dato il seguente sistema di ipotesi:

$$\begin{cases} H_0 &= \left\{\bigcap_{i=1}^k H_{0i}\right\} \\ H_1 &= \left\{\bigcup_{i=1}^k H_{1i}\right\} \end{cases} \qquad H_{\cdot i} = \text{ipotesi parziale i-esima}$$

e fissato un α globale, per testare singolarmente le k ipotesi parziali è possibile utilizzare una delle seguenti procedure:

- Procedura Single-step di Bonferroni:
  $\hookrightarrow$ si aggiusta ogni p-value: $\lambda_i^{adj} = \lambda_i \cdot k, \quad i=1,\dots,k$
- Procedura Step-down di Bonferroni-Holm:
  $\hookrightarrow$ si aggiusta ogni p-value: $\lambda_{(i)}^{adj} = \lambda_{(i)} \cdot (k-i+1), \quad i=1,\dots,k; \quad \lambda_{(1)} \le \dots \le \lambda_{(k)}$
- Procedura Closed Testing di Marcus:
  $\hookrightarrow$ si considerano tutte le possibili intersezioni delle singole ipotesi:

  $$H_1 = H_1, \quad H_{12} = H_1 \cap H_2, \quad H_{123} = H_1 \cap H_2 \cap H_3, \quad \dots$$

  e si aggiusta il p-value per H_i col max delle ipotesi che la includono: $\lambda_i^{adj} = \max(\lambda_i, \dots, \lambda_{i\cdots k})$

infine si rifiuta l'ipotesi globale a livello ($\le$) α se si rifiuta almeno un'ipotesi parziale.

Appendice A

Annotazioni generali

A.1 Notazioni utilizzate

- $\overset{+}{\cup}$ = concatenamento
- $\perp\!\!\!\perp$ = indipendenza
- $\not\perp\!\!\!\perp$ = dipendenza

A.2 Ranghi

Dato il campione $X = (X_1, \ \ldots, \ X_n)$, il vettore dei ranghi $R = (R_1, \ \ldots, \ R_n), \ R_i \in \mathbb{N}^+$ associato ad X avrà generica componente $R_i = \text{Rank}(X_i) = \sum_{h=1}^{n} \mathbb{1}(X_h \leq X_i)$.

Equivalentemente il $R_i = \text{Rank}(X_i) = j \ \Leftrightarrow \ X_i = X_{(j)}$, dove $X_{(j)}$ è un elemento del vettore X ordinato.

A.3 Tabella riassuntiva dei test per l'analisi univariata

Tipo variabili	P	B	**2 camp. indip.**	P	B	**2 camp. dip.**	P	B	**C camp. indip.**	P	B	**C camp. dip.**
quantitative	✓	–	t-Student	✓	–	paired t-Student	✓	✓	1 way ANOVA			
	✓	–	Welch			McNemar		✓	1 way ANOVA*		✓	1 way ANOVA*
		–	Test di perm		–	Test di perm		✓	Kruskal-Wallis			
		–	Mann-Whitney		–	Wilcoxon						
qualitative ordinali	✓	✓	χ^2			McNemar	✓	✓	1 way ANOVA			
		✓	χ^2 di perm					✓	1 way ANOVA*		✓	1 way ANOVA*
			$T^*_{CM}, T^*_{AD}, T^*_{KS}, T^*_{KSAD}, T^*_{BM}$									
qualitative nominali	✓	✓	χ^2				✓	✓	1 way ANOVA			
		✓	χ^2 di perm					✓	1 way ANOVA*		✓	1 way ANOVA*
dicotomiche	✓	✓	χ^2		–	McNemar	✓	✓	1 way ANOVA			
		✓	Fisher Exact					✓	1 way ANOVA*		✓	1 way ANOVA*

Legenda

P	=	test parametrico	B	=	test bilaterale			
"✓"	=	condizione verificata	"␣"	=	condizione non verificata	"–"	=	condizione ininfluente

A.4 Tabella riassuntiva dei test per l'analisi multivariata

Tipo variabili	P	B	**2 camp. indip.**	P	B	**2 camp. dip.**	P	B	**C camp. indip.**	P	B	**C camp. dip.**
quantitative	✓	✓	T^2 Hotelling				✓	✓	2 way ANOVA			
	✓		OLS di O'Brien					✓	2 way ANOVA*		✓	2 way ANOVA*
qualitative ordinali							✓	✓	2 way ANOVA			
								✓	2 way ANOVA*		✓	2 way ANOVA*
qualitative nominali							✓	✓	2 way ANOVA			
								✓	2 way ANOVA*		✓	2 way ANOVA*
dicotomiche							✓	✓	2 way ANOVA			
								✓	2 way ANOVA*		✓	2 way ANOVA*

Legenda

P	=	test parametrico	B	=	test bilaterale			
"✓"	=	condizione verificata	"␣"	=	condizione non verificata	"–"	=	condizione ininfluente

Finito di stampare il 22 novembre 2007 con LaTeX