UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA TRIENNALE IN SCIENZE E TECNOLOGIE INFORMATICHE

TESI DI LAUREA

IDENTIFICAZIONE E VALUTAZIONE DELLE OSSERVAZIONI INFLUENTI NEI TEST DI PERMUTAZIONE

RELATORE : DOTT. FINOS LIVIO

CORRELATORE : DOTT. CALLEGARO ANDREA

LAUREANDA : ZENERE ANGELICA

MATRICOLA N° 599718

Anno accademico 2010/2011

Sommario

Capitolo 1

Introduzione al problema

Quando si utilizzano i test di permutazione ci si avvale soltanto dell'informazione data dal singolo p-value dei dati osservati. Lo scopo di questa ricerca è di affrontare ed esaminare i test di suddetti, per capire come si possono estrapolare ulteriori approfondimenti che possiamo ottenere dal procedimento. Ciò deriva dal fatto che, dentro i test di permutazione c'è dell'informazione ulteriore che non vogliamo lasciare inutilizzata, in modo da estrarre tutto ciò che può essere utile nel capire e identificare i valori che portano una significativa influenza nella statistica test.

Una possibilità, per affrontare il problema, è valutare l'importanza marginale delle osservazioni come una scomposizione in singoli contributi al p-value, così da identificare quali siano le componenti che detengono un'influenza maggiore delle altre.

Studi precedenti

Questo approccio al problema ha analogie con altre tecniche d'indagine, infatti, possiamo trovare l'influenza delle singole componenti anche attraverso la distanza di cook oppure la cross validation (leave-one-out). Questi due metodi, però non concorrono nel risolvere il problema introdotto perché sono due metodi basati su modelli lineari e sono valutati sotto H_1.

In letteratura, lo studio dei singoli contributi, generalmente in ambito non parametrico, non è facilmente riscontrabile e viene introdotto soltanto attraverso procedure differenti da quelle di permutazione. Di solito, l'analisi dei singoli contributi è effettuata tramite jackknife. Questa procedura, utilizzata per la stima non distorta della varianza, introduce una valutazione utilizzando tutte le osservazioni meno una. Per cui, combinando un procedimento di ricampionamento e il jackknife, possiamo riscontrare un risultato similare a quello introdotto come obbiettivo. Un esempio che pare avere riscontri simili è introdotto da John Fox in Bootstrapping Regression Models in cui avvia un'analisi sull'influenza di un elemento sulla stima di un parametro tale $\theta = t(P)$ attraverso un'applicazione combinata tra bootstrap e jackknife, la cui stima non è altro che il vettore delle statistiche $T=t(S)$, dove P rappresenta la popolazione e S rappresenta un campionamento via bootstrap della popolazione, ovvero di P.

Per capire meglio come può essere utilizzata questo tipo di analisi, si può fare riferimento ai casi di studio su test clinici, in cui a causa dello sforzo compiuto per ottenere le rilevazioni e il costo di tali ricerche, si cerca di estrarre tutte le informazioni possibili dai dati in possesso. Per cui capire quali sono i soggetti che contribuiscono maggiormente a rifiutare un'ipotesi, può essere utile in quanto si possono identificare e analizzare specificatamente, così da non accontentarsi soltanto di un responso univoco.

Esempio introduttivo

Un esempio a riguardo può essere esplicato portando come dimostrazione il caso di studio sul metabolismo alcolico differenziato tra maschi e femmine che usufruiscono dell'alcol o meno. La variabile risposta è metabol, ovvero la differenza nei metodi di assunzione dell'alcol: o quando viene iniettato direttamente nel sangue o quando invece viene consumato per via orale e quindi passa prima attraverso lo stomaco.
L'unica variabile esplicativa quantitativa è gastric, una misura dell'attività di enzimi nello stomaco che in parte metabolizza l'alcol.
Ci sono due variabili categoriali, sesso e alcol, che indicano come categorizzare l'individuo.

Detto ciò, si vuole ricercare un'influenza nella determinazione di un test d'ipotesi introdotto come :

$$\begin{cases} H_0 : X_1 =^d X_2 \\ H_0 : X_1 >^d X_2 \end{cases}$$

Dove X_1 e X_2 identificano i due gruppi influenti come i due sessi, oppure uso di alcol o meno. Questa specificazione verrà introdotta dal momento in cui si saprà se vi è l'influenza del tipo di classificazione.

In questo esempio si vedrà quindi la differenziazione tra maschi e femmine e l'utilizzo di alcol o meno. Inoltre si noterà un'importante dipendenza tra metabol e gastric, infatti al crescere della seconda componente, aumenteranno i valori della prima. In relazione a questo si valuterà l'effettiva influenza di gastric e si ricercherà se si può reputare come una componente di confusione (ovvero se comporta un'influenza casuale).

Capitolo 2

Test di permutazione

Introduzione

Solitamente utilizzati per saggiare l'ipotesi di uguaglianza delle distribuzioni (ipotesi nulla), i test di permutazione nascono con il principio che se diversi esperimenti prendono valori generati da un medesimo spazio X^n con distribuzioni P1 ,P2,..,Pn , anche non note, tutti i membri generati da una stessa famiglia di distribuzione ,danno uno stesso campione X, allora nelle diverse inferenze, condizionate a X e ottenute usando uno stesso test statistico, dovrebbero portare sotto l'ipotesi nulla alla possibilità di permutare i dati osservati rispetto i gruppi.

2.1 Test su campioni appaiati

Considerando un esempio di un problema, andiamo a osservare un campione di dati appaiati che potrebbero riferirsi a un esame pre (Y) e post (W) trattamento. I dati bivariati sono dipendenti perché vengono considerate le stesse persone, però nel contesto delle coppie, si possono considerare indipendenti, ovvero il data set può essere visto come un campione casuale di n coppie i.i.d. di un bi-variato campione casuale (Y,W).

L'ipotesi d'interesse è il cambiamento o meno della distribuzione dei due campioni dopo un trattamento, per cui l'ipotetico problema viene introdotto come:

$$H_0 = \{Y =^d W\} = \{P_Y(z) = P_W(z), \forall z \in R^1\}$$

Avendo l'ipotesi alternativa

$$H_1 = \{Y <^d (>^d) W\}$$

Dove P_Y e P_W rappresentano le distribuzioni marginali di Y e W.

Ulteriori condizioni di simmetria attorno allo zero che devono essere soddisfatte sotto H_0 sono:

$$H_0 = \begin{cases} Pr(Y - W > z) = \Pr(Y - W < -z), \forall\, z \in R^1 \\ \Pr\{(Y - W) > 0\} = 1/2 \end{cases}$$

Ovvero, che le due code abbiano la stessa probabilità, per cui H_0 è valida soltanto nel caso in cui X =Y-W sia distribuita simmetricamente attorno all'origine δ (presupponendo δ tale che sia E(X), oppure la M(X)).

Alcuni risultati sotto l'ipotesi nulla (permutabilità delle osservazioni) possono essere riassunti in:

$$F_Y(z) = F_W(z) \qquad \forall\, z \in R$$

$$F_{Y|t}(z|W = t) = F_{w|t}(z|Y = t) \qquad \forall\ (z, t) \in R^2$$

Quando questi risultati non sono presenti allora si afferma che $Y >^d (<^d) W$.

Possibili approcci :

- Test di student

Quando vi si è in un contesto parametrico si assume l'esistenza E(X), però non appena si affronta un problema che comporta soluzioni non parametriche spesso perdiamo questo assunto per cui, anche per introdurre una misura robusta, si può utilizzare la mediana M(X).

Per un'analisi più intuitiva si può considerare una trasformazione φ(.), rispetto le variabili e la distribuzione, in modo da migliorare l'interpretazione e la potenza del test statistico. Tale test viene ottenuto, per un campione a grandezza limitata da $\sum_i \varphi_i$.
A questo scopo si può ricorrere ad una struttura parametrica che comporta l'assunzione di normalità della variabile X con varianza ignota positiva.

$$X_i \sim N(\delta, \sigma_x^2)$$

proponendo una trasformazione $T = \overline{X}\,\frac{\sqrt{n}}{\hat{\sigma}}$ dove $\hat{\sigma}^2 = \frac{\sum_i (X_i - \overline{X})^2}{n-1}$ e $\overline{X} = \frac{\sum_i X_i}{n}$, sotto H_0 , la trasformazione T si distribuisce come una t di student centrale con n-1 gradi di libertà.
Affrontando questo metodo si deve fare la premessa che la varianza sia costante e indipendente dai valori, quando ciò non è verificato e vi è una dipendenza della varianza dalle componenti, non si può affrontare una soluzione parametrica.

- Test sui ranghi

Nel momento che non si può affrontare una via parametrica si può ricorrere al test sui ranghi di Wilcoxon.

Tale soluzione supporta i presupposti che la variabile X sia continua , e la distribuzione in probabilità può essere completamente sconosciuta, inoltre non si fanno presupposti su E(X).

Il test sui ranghi di wilcoxon è tale che:

$$W = \sum_i R_i * w_i$$

Dove

$R_i = \sum_{1<j<n} I(|X_j| \leq |X_i|)$ e $w_i = \begin{cases} 1\ se\ x_i > 0 \\ 0\ altrimenti \end{cases}$, ∀ i=1…n

La cui distribuzione, sotto H_0 , si distribuisce come una normale N(0,1)

$$\frac{W - E(W|H_0)}{\sqrt{V(W|H_0)}} \sim N(0,1)$$

Dove $E(W \mid H0) = n(n+1)/4$ e $V(W \mid H0) = n(n+1)(2n+1)/24$.

Un aspetto importante sul test di wilcoxon è che non è influenzato dalle distribuzioni, infatti è invariante rispetto ad esse.

- Soluzione binomiale

Tramite questo approccio è possibile evitare di, condurre assunzioni sulla continuità, mentre l'influenza della distribuzione è nulla; infatti il test adottato per questa soluzione è composto da :

$U = \#(x_i > 0)$ e $V = \#(x_i \neq 0)$

Per cui sotto H_0 si può approssimare $U \sim Bin(V, 1/2)$.

La particolarità di questo test può essere applicato anche in situazioni di non omogeneità, infatti si può utilizzare quando le componenti sono indipendenti, però non identicamente distribuite, oppure quando σ_i^2 non sono costanti.

- Soluzione via permutazione

La permutazione è una via che non comporta assunzioni specifiche, infatti va a sfruttare totalmente e soltanto l'informazione fornita dai dati. La struttura del procedimento non necessita della distribuzione della popolazione.

Quando si utilizza un approccio permutativo, spesso si vuole verificare l'ipotesi nulla $H_0 = \{X_A =^d X_B\}$, ovvero si cerca si verificare l'uguaglianza in distribuzione di due o più gruppi attraverso il ricampionamento condizionato delle unità tra i gruppi. In questo modo, sotto H_0 , si afferma che i valori appartenenti ad ogni gruppo sono generati da una medesima distribuzione.

Considerando il caso di due campioni l'analisi si può ricondurre ad un approccio basato sulle differenze dei due gruppi (questo ritorna utile quando si hanno dati appaiati).

Considerando ciò, quel che si vuole attuare, è niente meno che la randomizzazione dei segni delle differenze dei due gruppi, dove un segno (+ o -) viene assegnato al valore $X = X_A - X_B$ con probabilità ½ , e il test che si va a considerare è $T = \frac{1}{n}\sum_i X_i$.

Denominato T_{ob} il test sviluppato sui valori originari, si considera la distribuzione di $T^* = \frac{1}{n}\sum_i X_i^*$ dove $X_i^* \in \mathbf{X}^*$ è il campione modificato con l'assegnazione casuale dei segni + o - . Si ottiene, attraverso la permutazione, che la probabilità dell'evento $\mathbf{X}^*$ condizionato a $\mathbf{X}$ è uniforme all'interno dello spazio campionario ($\mathcal{X}$ |$\mathbf{X}$) dove vi sono M=#[X* ∈ ($\mathcal{X}$ |$\mathbf{X}$)] punti. Considerando ciò, per ottenere un p-value dei valori osservati si calcola :

$$\alpha = \Pr\{T^* \geq T_{ob} \,|\boldsymbol{X}\} = \frac{1}{M}\sum_i I(T_i^* \geq T_{ob})$$

Considerando la randomizzazione dei segni come trasformazione $X_i^* = X_i \cdot S_i^*$, i =1..n, dove S_i^* è una assegnazione casuale dei valori -1 e +1 si ottengono

$$E\left\{\left(\frac{1}{n}\sum_i X_i \cdot S_i^*\right)\middle| X\right\} = 0$$

$$V\left\{\left(\frac{1}{n}\sum_i X_i \cdot S_i^*\right)\middle| X\right\} = \left(\sum_i \frac{X_i^2}{n^2}\right)^{\frac{1}{2}}$$

Così, quando si vanno a considerare campioni con n grande la distribuzione di T^*, si può approssimare il test ad una normale standard

$$K^* = \frac{\left\{\left(\frac{1}{n}\sum_i X_i \cdot S_i^*\right)\middle| X\right\}}{\left(\sum_i \frac{X_i^2}{n^2}\right)^{\frac{1}{2}}} \sim N(0,1)$$

2.2 I test di simmetria (test a due campioni appaiati)

Per questo tipo di verifica vengono considerati dati del tipo $X=Y_a-Y_b$ dove le realizzazioni X_i sono continue, indipendenti e identicamente distribuite. Anche se non siamo a conoscenza della distribuzione di X, sappiamo che per saggiare l'ipotesi H_0 = (simmetria attorno a 0 di X oppure $Y_a =^d Y_b$), si può eseguire un test sul campione attribuendo casualmente i segni ai valori di X.

Per cui considerando un finito gruppo di osservazioni (2^n) abbiamo che :

$$X^* = g^*(X) = \{x_1 \cdot S_1^* \dots \dots \dots X_n \cdot S_n^*\}$$

Dove S* è considerato come assegnazione casuale dei segni + o -.

Un test ottimale che è possibili eseguire sotto H_0 e che può essere facilmente affrontato anche sotto H_1 è $\mathrm{T} = \sum_{\mathrm{i}} \mathrm{X_i}$ dove la distribuzione del test è definita come

$$\mathrm{T} = \{T^* = T(X^*) = \sum_i X_i \cdot S_n^*\}$$

Si nota che, solo sotto ipotesi nulla, il data set fornito X può essere considerato come una selezione casuale dallo spazio campionario ($\mathcal{X}|\mathbf{X}$) con probabilità $1/(2^n)$, però sotto H_1 ciò non si può considerare perché la probabilità è condizionata da δ (parametro di centralità).

2.3 I test a 2 campioni indipendenti

Finora si è parlato di test riguardanti gruppi con dati appaiati, però nel caso volessimo confrontare insiemi di dati indipendenti possiamo affrontare la questione sotto diversi punti di vista. Considerando il data set formato dall'insieme di dati $X = X_1 \oplus X_2 = \{X(i), i = 1, \dots, n;\ n_1, n_2\}$. Nel verificare l'ipotesi $H_0 = \left\{X_1 =^d X_2\right\}$ e assumendo

l'ipotesi di omoschedasticità[1] e di scambiabilità sotto H_0, qualora si presumono effetti gruppo fissi[2], si può verificare direttamente $H_0 = \{\delta = 0\}$ dove con δ viene indicata la tendenza dell'effetto. In questo modo avendo la stima della media (oppure della mediana, volendo restare in un contesto più robusto), una statistica utilizzabile per la verifica è:

$$S = \overline{X}_1 - \overline{X}_2$$

dove la versione di permutazione è

$$S^* = \overline{X}_1^* - \overline{X}_2^*$$

Tuttavia, S* è permutazionalmente equivalente a $T^* = \sum_i X_{1i}^*$, poichè c'è una relazione biunivoca e crescente tra le due statistiche dal momento che, nell'analisi basata sulle permutazioni, il data set X resta fissato. Infatti, le quantità $\sum_{ij} X_{ij}^* = \sum_{ij} X_{ij}$, essendo uguali, sono permutazionalmente invarianti, tale concetto viene introdotto nel successivo capitolo .

Come nei casi precedenti, anche attraverso quest'analisi il test si può ricondurre, quando vi è un campione abbastanza grande, ad una distribuzione normale considerando

$$\mu_T = n_1 \cdot \mu_X = n\_1 \sum_{ij} \frac{X_{ij}}{n}$$

$$\sigma_T^2 = n_1 \cdot n_2 \cdot \sigma_X^2/(n-1)$$

Dove $\sigma_X^2 = \frac{\sum_{ij} X_{ij}^2 - n \cdot \mu_X^2}{n}$.

- Test d'uguaglianza con C>2 distribuzioni (ANOVA ad una via)

Qualora si cerchi di risolvere problemi riguardanti l'uguaglianza di C ≥ 2 gruppi di campioni, possiamo ricorrere all'utilizzo dell'ANOVA ad una via.

Andando a considerare il data set $X = \{X_{ij} = \mu + \delta_j + \sigma \cdot Z_{ij} \; ; i = 1 \dots n_j, j = 1, \dots, C\}$ si vuole verificare l'ipotesi $H_0 = \{X_1 =^d X_2 =^d \dots =^d X_C\}$, dove l'ipotesi alternativa risulta

[1] Qualora non si verificasse l'omoschedasticità ($\sigma_1 \neq \sigma_2$) , il data set risulta diviso in due sotto gruppi X_1 e X_2 e l'assunto di scambiabilità non è più verificabile. Per cui non si possono scambiare i dati di un gruppo in un altro.

[2] Il data set è fissato

$H_1 = \{\text{almeno una media è differente}\}$. δ_j , che indicano l'effetto del trattamento j , soddisfano il vincolo $\sum_j n_j \cdot \delta_j = 0$.

Nel verificare H_0 si può ricorrere alla definizione :

$$S = \sum_j \left(\overline{Y}_j. - \overline{Y}\right)^2 \cdot n_j$$

In cui $\overline{Y}_j. = \sum_i \varphi(X_{ij})/n_j$ e $\overline{Y} = \sum_j \frac{\overline{Y}_j \cdot n_j}{n}$ che rappresenta una costante nella statistica di permutazione, per cui possiamo utilizzare la statistica test

$$T = \sum_j n_j \cdot \overline{Y}_j.$$

-Variabili categoriali

Nel momento in cui il campione si compone da osservazioni raccolte in categorie, tale che X è diviso in j =1,…,C gruppi i quali sono suddivisi a loro volta in A_i[3] classi ordinate i=i,..,k.
in questo caso il test d'ipotesi risulta

$$H_0 = \{X_1 =^d X_2\} = \{F_1(A_i) = F_2(A_i)\ , \forall\ i = 1, \dots, k\}$$

Un modo per affrontare questo problema è attribuire un peso alle classi ω.
Per cui il test permutato risulta

$$S^* = \sum_i \omega_i (f^*_{1i} - f^*_{2i}) = n_1 \overline{\omega}^*_1 - n_2 \overline{\omega}^*_2$$

Dato che $f^*_{1i} + f^*_{2i} = f_{.i}$ sono valori fissati allora si può ricondurre il test a

$$T^*_\omega = \sum_i \omega_i \cdot f^*_{1i} = n_1 \cdot \overline{\omega}^*_1$$

[3] Possono essere sia qualitative che quantitative

2.4 Test di permutazione e valore critico

È importante notare che date due statistiche T_1 eT_2 , entrambe formulate su **X**, si dicono permutazionalmente equivalenti quando, per ogni punto in **X** e di X* ϵ ($\mathcal{X}$|**X**), la relazione $\{T_1(X^*) \leq T_1(X)\}$ è vera soltanto se $\{T_2(X^*) \leq T_2(X)\}$ è vera, definendo X* come la permutazione di **X** e ($\mathcal{X}$|**X**) come lo spazio campionario condizionato. Quando la relazione è permutazionalmente equivalente questa viene indicata come $T_1 \approx T_2$.

Si definisce il test di permutazione una via formale per raggiungere risultati di determinazione, a questo scopo la procedura utilizzata per arrivare ai risultati desiderati è:

1) Si elabora un insieme di dati X ed attraverso un test statistico T : Rn -> R1 ;
2) Si determinano i M test permutazioni T(X) = { T(X*) : X* ϵ (X|X) } e ridistribuiti in ordine non-decrescente $T^*_{(1)} \leq T^*_{(2)} \leq \cdots \leq T^*_{(M)}$
3) Ad ogni α ϵ (0,1), si determina $T_\alpha = T^*_{M\alpha}$ ovvero il valore critico α associato a (T|X)
4) Con Mα = int[(1-α)*M] si definisce quante permutazioni sono al disotto del valore critico T^*_α

È importante notare che il valore critico dipende da ($\mathcal{X}$|**X**) , ovvero è invariante rispetto allo spazio campionario condizionato,

$$T_\alpha = T_\alpha(X) = T_\alpha\left(X^{'}\right) \ , \forall\, X' \in (\chi|\mathbf{X})$$

Ciò è una conseguenza del fatto che $T_\alpha(X)$ è un valore fisso all'interno del supporto (T|**X**), però tale supporto varia generalmente al variare di X.

La versione randomizzata del test di permutazione ϕ_R associata a (T,**X**) viene definita come :

$$\phi_R = \begin{cases} 1 & se \quad T_{ob} > T_\alpha \\ \gamma & se \quad T_{ob} = T_\alpha \\ 0 & se \quad T_{2b} < T_\alpha \end{cases}$$

Dove $\gamma = [\alpha - PR\{\, T_{ob} > T_\alpha | X\}\,] / \Pr\{\, T_{ob} = T_\alpha \,\}$.

Con questo procedimento si considera ϕ_R come l'effettiva posizione del test del campione osservato all'interno del supporto permutato di T(X). Essendo sotto ipotesi nulla si arriva a verificare che l'esatta misura di α viene ottenuta da :

$$\int_{(\chi|\mathbf{X})} \phi_R \cdot dP|_x = \mathrm{E}_{(\chi|\mathbf{X})}(\phi_R) = \alpha$$

-Anche se tale metodo è usato per avere un valore di α esatto ci si può limitare, per semplicità, a calcolare α utilizzando il valore di ϕ_R calcolato come

$$\phi = \begin{cases} 1 & se \quad T_{ob} \geq T_\alpha \\ 0 & se \quad T_{ob} < T_\alpha \end{cases}$$

Di conseguenza sotto ipotesi nulla si ottiene un α' che generalmente è maggiore uguale al valore esatto dello stesso.

$$\int_{(\chi|\mathbf{X})} \phi \cdot dF(t|Z) = E_X(\phi) = \alpha' \geq \alpha$$

Un interesse maggiore viene dato dal p-value, sotto H_0 e associato a (T,**X**), tale viene definito come :

$$\lambda(X) = \Pr\{T^* \geq T_{ob}|X\}$$

mentre sotto H_1, il p-value, è condizionato al parametro di centralità δ e viene definito come :

$$\lambda(X(\delta)) = \Pr\{T^*(\delta) \geq T_{ob}(\delta)|\, X(\delta)\}$$

Tale valore è strettamente legato al valore di α infatti dal momento che λ(X)>X si ha che $T_{ob}>T_\alpha$.

È da considerare il fatto che anche se viene considerato il test di permutazione esatto come un'importante proprietà di T, questo non implica che T sia un buon test. Ovvero per far si che un test sia ottimale, questo non deve essere distorto e deve avere la proprietà di consistenza. A differenza di ciò il test di permutazione esatto non garantisce la consistenza.
Per verificare che un test abbia queste proprietà ci si avvale di due semplici

dimostrazioni. La non distorsione di un test viene riscontrata dal momento in cui si verifica che

$$E_{(\chi|\mathbf{X})}\left[\mathrm{I}\left\{\lambda\left(\mathrm{X}^{*}(\delta)\right) \leq \alpha|\mathrm{X}\right\}\right] \geq E_{(\chi|\mathbf{X})}\left[\mathrm{I}\{\lambda(\mathrm{X}^{*}) \leq \alpha|\mathrm{X}\}\right] = \alpha$$

sia vera.

Per quanto riguarda la consistenza, la si può verificare nel momento in cui si ha il valore critico T_α . Essa si dimostra verificando che sia sotto H_0, sia per n considerevolmente grande (n ->∞) si abbia :

$$\Pr\{T(X_n^*) \geq T_\alpha(X_n)|X_n\} = \alpha$$

Mentre quando si è sotto ipotesi H_1 la proprietà viene dimostrata quando, al tendere di n a numerosità molto grandi, risulta che :

$$\Pr\left\{T\left(X_n^*(\delta)\right) \geq T_\alpha(X_n)\middle|X_n\right\} = 1$$

2.5 Proprietà dei test di permutazione

La lettura di questo paragrafo può essere omessa senza perdita di compressione del lavoro presentato.

2.5.1 In caso di eterogeneità dei campioni

Precedentemente si è parlato solamente di effetti fissi additivi, però nel caso in cui σ è dipendente da δ, ovvero i dati sono generati da

$$X_{ij} = \mu + \delta_j + \sigma(\delta_i) \cdot Z_{ij}$$

L'ipotersi di permutabilità è validata soltanto sotto H_0, ovvero qualora è vero $(\delta_1 = \delta_2)$ gli errori $\sigma(\delta_j)Z_{ij}$ sono permutabili. Infatti formulando il test statistico come

$$T = \sum_i X_{1i} - \sum_i X_{2i}$$

Supponendo ν* siano gli scambi casuali tra i due gruppi vediamo che

$$T^* = \sum_i X_{1i} - \sum_i X_{2i} = (n_1 - 2v*)\delta_1 - (n_2 - 2v*)\delta_2 + \sum_i \sigma(\delta_1^*)\, Z_{1i}^* - \sum_i \sigma(\delta_2^*)\, Z_{2i}^*$$

La struttura del test denota che solamente nel caso in cui H_0 sia vera T_i^* si basa solo sulla permutazione degli errori.

- Anova a una via

Considerando un campione composto da

$$X = \{X_{ij}\ ,\qquad i = 1..n_j\ ;\ j = 1 \dots C\ \}$$

Dove le $X_{ij} = \mu + \delta_j + \sigma(\delta_j)Z_{ij}$ e l'effetto del trattamento soddisfa $\sum_j n_j \cdot \delta_j = 0$.
Andando a testare l'ipotesi nulla $H_0 = \{\delta_1 = \cdots = \delta_C\}$ contro l'ipotesi $H_1 = \{H_0 \text{ è } falsa\}$ si prende in considerazione il test :

$$T = \sum_j n_j \overline{X}_1^{\,2}$$

Come precedentemente gli errori sono permutabili soltanto qualora la relazione $\delta_1 = \delta_2$ risulta vera, per cui considerando la permutazione dell'intero data set $X = X_1 \oplus \dots \oplus X_C$, il test risulta :

$$T^* = \sum_{ij} [\delta_j^* + \sigma(\delta_j^*) \cdot Z_{ij}]$$

Anche in questo caso si ottiene che sotto H_0 il test dipende solo dagli errori.

2.5.2 La non distorsione dei test di permutazione

Ora andiamo a considerare un insieme di dati così composti

$X \sim > H_0 \qquad X = X_1 \oplus X_2$

$X_\Delta \sim > H_1 \qquad X = X_1 \oplus Y \qquad Y = \{Y_i \leq X_{2i} \ , \ i = 1 \dots n_2\}$

Ovvero

$X_\Delta = X - \Delta \qquad dove \quad \Delta = (0_i \ , \qquad i = 1..n_1) \oplus (\ X_{2i} - Y_i < 0 \ , i = 1..n_2\ )$

Ovvero i Δ rappresentano gli effetti fissi stocastici.

- comparazione di due popolazioni

Supponiamo di avere un problema di confronto tra due campioni per cui verificare $H_0=\{X_1 =^d X_2\}$ contro $H_1=\{X_1 >^d X2\}$, e i dati sono generati da un modello con effetti gruppo fissi

$$X_{ij} = \mu + \delta_j + \sigma Z_{ij}$$

così formulato, possiamo introdurre la statistica test $T = \sum_i^n X_{1i}$, la cui distribuzione è stocasticamente ordinata rispetto all'effetto generale del trattamento, per cui i valori associati a $T_{ob}\ sono\ \sum_i\ X_{1i} = T_{ob}(X_\Delta)$. Riassumendo, il test rivolto a X_Δ può essere visto come $T_\Delta^* = T^* - \sum_i\ \Delta_i^*$ si ha che

$$\lambda(X_\Delta) = \Pr\{T_\Delta^* \geq T_{ob}(X_\Delta)|X_\Delta\} = \Pr\left\{T^* - \sum\nolimits_{\mathrm{i}}^{\mathrm{n}} \Delta_i \geq T_{ob}\middle| X\right\} \leq \Pr\{T^* \geq T_{ob}|X\} = \lambda(X)$$

In quanto sotto $H_0 : \sum_i\ \Delta_i = 0$

Generalmente, quando si applica una struttura di permutazione, ogni statistica si può riportare con la formulazione $\{T^* - \sum_{\mathrm{i}}^{\mathrm{n}} \Delta_i \geq T_{ob}|X\}$, allora la statistica si dice permutazionalmente esatta, o non distorta, e consistente.
Le proprietà di non distorsione e di consistenza sono valide a condizione che $\sum_i\ \Delta_i^*$ diverga con probabilità 1 quando la dimensione del campione tende ad infinito, ovvero che qualora si veramente sotto H_1 si rifiuta H_0 con certezza[4].

[4] La potenza del test è pari a 1

- La non distorsione dei test a C campioni (ANOVA a una via)

Consideriamo i punti che andiamo ad osservare sono rappresentati da X quando sono sotto H_0, mentre sotto H_1 sono rappresentati da $X(\delta)$[5].
Si constata così che i test osservati sono :

$T(X) = t_{ob}$

$t_{ob}(\delta) = t_{ob} + \sum_j \ n_j \cdot \delta_j^2 + \ 2 \sum_{ij} X_{ij} \cdot \delta_{ij}$

mentre i test permutati sono :

$T^* = \sum_j \ n_j \cdot \left(\overline{X}_j^*\right)^2$

$T^*(\delta) = T^* + \sum_j \ n_j \cdot \left(\overline{\delta}_j^*\right)^2 + 2 \sum_{ij} X_{ij}^* \cdot \delta_{ij}^*$

$\overline{\delta}_j^* = \sum_i \ \delta_{ij}^* / n_j$

Si verifica che :

1) $\sum_{ij} \delta_{ij} = \sum_{ij} \delta_{ij}^* = 0$

2) $\sum_{ij} X_{ij} \cdot \delta_{ij} = \sum_{ij} X_{ij}^* \cdot \delta_{ij}^*$

3) $\sum_j \ \left(\overline{\delta}_j^*\right)^2 = \sum_j \ \sum_i \ \left(\delta_{ij}^* - \overline{\delta}_j^*\right)^2 + \sum_j \ n_j \cdot \left(\overline{\delta}_j^*\right)^2 = \sum_j \ n_j \cdot \delta_j^2$

Allora si ha :

$$\Pr\{T^*(\delta) \geq T_{ob}(\delta) | X(\delta)\} = \Pr\left\{T^* - \sum_{ij} \left(\delta_{ij}^* - \overline{\delta}_j^*\right)^2 \geq t_{ob} \middle| X\right\} \leq \Pr\{T^* \geq t_{ob} | X\}$$

[5] X(δ) = Xij + δj , i=1..nj , j=1..C

2.6 Combinazione non parametrica multidimensionale

L'utilizzo delle permutazioni, precedentemente illustrato in un ambito unidimensionale, viene ad assumere un'importanza e utilità maggiore qualora si presentino casi con data set multidimensionali.
Infatti grazie a questa possibilità di condurre un'analisi attraverso un ambito non parametrico, possiamo affrontare la gran parte dei problemi che sono presentati senza avere informazioni sulle eventuali distribuzioni, oppure delle dipendenze che si potrebbero rilevare all'interno del campione tra le differenti dimensioni.
La metodologia delle permutazioni non parametriche viene attuata relativamente ad un finito numero di test di permutazioni dipendenti o meno.

$$X = \{X_{ijh} \quad i = 1..n_j \ ; \ j = 1..C \ ; \ h = 1..q\}$$

In quanto X viene generato da uno spazio $\mathcal{X}$ multidimensionale distribuito su C≥2 sono i livelli, per cui X_j sono campioni multidimensionali indipendenti e identicamente distribuiti con distribuzioni $P_j \in \mathcal{P}$ j=1..C.

Nel caso in cui il problema sia di natura multidimensionale, si ha che l'ipotesi nulla adottata è introdotta come H_0={P_1 =….= P_{C-1} = P_C } contro H_1={almeno una distribuzione è falsa}. In questo caso è possibile decomporre H_0 in k sotto ipotesi, per cui H_0 è vera qualora congiuntamente $H_{0\,i}$, per i=1..K, sono vere e H_1 è vera quando almeno un'ipotesi $H_{0\,i}$ non è verificata.
Relativamente all'ipotesi nulla, decomponibile in k sottoipotesi, vi è il vettore dei test statistici T=T(X) ∈ $\Re^k$ e il test parziale univariato è rappresentato da T_i = T_i(X) per i=1..k.

La procedura con cui si affronta il ricampionamento in ambito multidimensionale è:

1) Calcolare il vettore dei test osservati t_{ob}=t_{ob1},….,t_{obK}

2) Considerare un'applicazione casuale g* derivante dall'insieme di trasformazioni G e considerare il vettore dei test di permutazioni T*=T(X*) X*=g*(X) T*∈$\mathcal{R}^B$.

In cui g* sia una permutazione casuale $(u_1^*, \dots, u_n^*)$ con cui si va a permutare i vettori dei vettori n dimensionali, mantenendo così intatte le relazioni tra le osservazioni date dai vettori q dimensionali. E B è il numero di permutazioni effettuate sul campione.

3) Considerando le k variate funzioni di distribuzione empiriche : [6]

$$\widehat{F}_B(z|X) = \left[\frac{1}{2} + \sum_r^B I\left(T_r^* \leq z\right)\right] \backslash (B+1) \quad \forall z \in \mathcal{R}^k$$

la significatività marginale di permutazione è

$$L_i(z|X) = \Pr\{\mathrm{T}_\mathrm{i}^* \geq \mathrm{z} \,|\mathrm{X}\} = \Pr\{\mathrm{T}_\mathrm{i}^* \geq \mathrm{T}_{\mathrm{os\,i}} \,|\mathrm{X}\} = \lambda_i$$

I test T_i calcolati per i =1…k risultano essere marginalmente non distorti e significativi per valori risultanti alti e, inoltre, sono consistenti qualora $Pr\{T_i \geq t_{i\alpha}|H_{1i}\} \to 1$ per n che tende all'infinito. Si deve considerare che queste affermazioni non necessariamente sono verificate contemporaneamente in tutte le ipotesi, tuttavia può risultare che queste condizioni si presentino per H_{0i} contro H_{1i} e non per H_{0j} contro H_{1j} per i≠j .

Per dare un'interpretazione univoca del problema e cercare un approccio più semplice nella lettura del quesito, si può apportare una combinazione dei test non parametrici in modo da ottenere un test di second'ordine del tipo

$$T'' = \varphi(\lambda_1 \dots \lambda_k)$$

in cui φ è definita come funzione di combinazione univariata, continua e che non altera i valori considerati[7] , $\varphi = (0,1) \to \mathcal{R}^1$, inoltre risulta

$$\varphi(\dots \lambda_i \dots) > \varphi(\dots \lambda_i^* \dots) \;\; si\ ha\ che\ \; \lambda_i < \lambda_i^* \;\; per\ i = 1..k\,.$$

Nella stima del test di second'ordine vi è il valore massimo φ^{--} qualora esista almeno un $\lambda_i \to 0$, tale che $\varphi(\dots \lambda_i \dots) \to \varphi^-$. Grazie a ciò , il valore critico T''_α esiste, finito, quando è minore di φ^- .

Questo insieme di proprietà definisce la classe delle funzioni combinatorie C. Considerate queste caratteristiche, viene da chiedersi quale sia la migliore funzione in C che porti a $T'' = \varphi(\lambda_1 \dots \lambda_k)$ per cui T'' risulti un test di second'ordine esatto.

[6] ½ viene inserito nella funzione di distribuzione empirica per ottenere una stima della funzione di distribuzione cumulativa F(z|X) e del p-value in un intervallo aperto (0,1)

[7] Applica un cambiamento del tipo di scala

A questo riguardo, in letteratura, sono stati introdotte diverse combinazioni non parametriche che soddisfano i requisiti necessari. Alcune di queste trasformazioni sono riportate a seguito :

a) La combinazione Fisher omnibus viene proposta come un modellamento dei k p-value, valutati come indipendenti e continui. Sotto ipotesi nulla la distribuzione di T'' viene delineata come un chi-quadro con 2·k gradi di libertà:

$$T_F^{''} = -2 \cdot \sum_i^k \log(\lambda_i)$$

b) La combinazione Liptak viene basata sulla statistica

$$T_L^{''} = \sum_i^k \phi^{-1}(1-\lambda_i)$$

in questo caso la combinazione dei k p-value, indipendenti e continui, che compongono il test di second'ordine T'' si può approssimare con una distribuzione normale con media 0 e varianza k.

Viene proposta anche un'ulteriore combinazione di Liptak che considera la trasformazione logistica dei p-values :

$$T_P^{''} = -\sum_i^k log\left[\frac{\lambda_i}{(1-\lambda_i)}\right]$$

c) La combinazione di Tippet è data da:

$$T'' = \max_{1 \le i \le k}(1-\lambda_i)$$

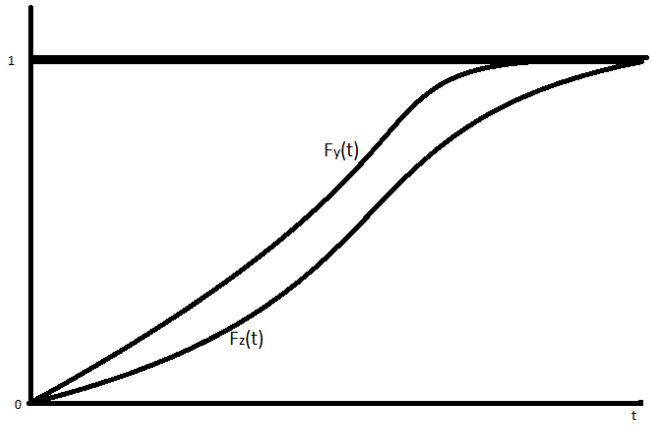
la cui distribuzione sotto H_0 , qualora risultano i k test indipendenti e continui, si assume asintoticamente la distribuzione di una distribuzione uniforme (0,1).

d) La funzione di combinazione non parametrica viene considerata qualora vi siano i k test statistici parziali omogenei, ovvero che abbiano asintoticamente la stessa distribuzione nella permutazione. La formulazione di questa combinazione risulta :

$$T_D^{''} = \sum_i T_i$$

Considerando B il numero di permutazioni effettuate, n $\rightarrow\infty$, si dice che $\forall\ \alpha$ e k fissate, T={T_i i=1..k} sono fortemente consistenti nei test di permutazione quando $\forall\ \varphi \in C$, T''=$\varphi(\lambda_1 \ldots \lambda_k)$ è consistente nella combinazione di tutti i test d'ipotesi $H_0 = \cap_i^k H_{0i}$ contro $H_1 = \cup_i^k H_{1i}$.

Considerando y e z due variabili casuali definite in uno stesso campione unidimensionale, y è stocasticamente più grande di z allora le loro funzioni di distribuzione cumulativa soddisfano la relazione $F_y(t) \leq F_z(t)\ \ \forall t \in \mathcal{R}^1$



Se T={T_i, i=1..k} sono test marginalmente non distorti rispetto a H_{0i} contro H_{1i} , allora $\forall\ \varphi \in C$, T''=$\varphi(\lambda_1 \ldots \lambda_k)$ è non distorto nella combinazione di tutti i test d'ipotesi $H_0 = \cap_i^k H_{0i}$ contro $H_1 = \cup_i^k H_{1i}$

Capitolo 3

Indici di influenza individuale

3.1 Descrizione dei possibili approcci

Per la descrizione del metodo presentato prendiamo in considerazione il seguente sistema di ipotesi

$$\begin{cases} \mathrm{H}_0 : \ (\mathrm{X}_1 =^{\mathrm{d}} \mathrm{X}_2) \\ \mathrm{H}_1 : \ (\mathrm{H}_0 \text{ non è vera}) \end{cases}$$

Come precedentemente introdotto si vuole analizzare la procedura di permutazione per individuare delle informazioni che non sono riscontrabili soltanto con un valore unico come il p-value, vogliamo infatti capire come cambia la statistica test quando la i-esima unità viene scambiata.

A questo scopo si sono affrontati diversi approcci nella scomposizione in contributi individuali.

Rappresentando il complesso dei valori osservati in:

$X_{11} \ , \dots \dots \dots , X_{1n_1}$	$X_{21} \ , \dots \dots \dots , X_{2n_2}$

Si individua che alla struttura introdotta, basata sul test di permutazione per i valori osservati, viene affiancato nell'uso lo spazio di permutazionale, ovvero una matrice di dimensioni B x n contenente i ranghi del vettore originario permutati.

$X_{11}^{*1}, \dots\dots\dots, X_{1n_1}^{*1}$	$X_{21}^{*1}, \dots\dots\dots, X_{2n_2}^{*1}$
…… …… ……	…… …… ……
$X_{11}^{*B}, \dots\dots\dots, X_{1n_1}^{*B}$	$X_{21}^{*B}, \dots\dots\dots, X_{2n_2}^{*B}$

Le statistiche test sono considerate come

T_{ob}	$T_1^*, \dots\dots\dots\dots, T_B^*$

Considerando $n = (n_1 + n_2)$, lo spazio per mutazionale si riconduce a

$r_1^{*1}, \dots\dots\dots, r_{n_1}^{*1}$	$r_{n_1+1}^{*1}, \dots\dots\dots, r_n^{*1}$
…… …… ……	…… …… ……
$r_1^{*B}, \dots\dots\dots, r_{n_1}^{*B}$	$r_{n_1+1}^{*B}, \dots\dots\dots, r_n^{*B}$

Dove

$r_1, \dots\dots\dots, r_{n_1}$	$r_{n_1+1}, \dots\dots\dots, r_n$

$$r_i = i \quad per\ i = 1, .., n$$

Rappresentano le posizioni (o ranghi) dei dati osservati.

Grazie allo spazio permutazionale, possiamo identificare una matrice di valori binari che assumono i valori 1 o 0 in base alla localizzazione dell'elemento i-esimo nei gruppi (1 viene spostato di gruppo, 0 rimane nello stesso gruppo).

$S_{11}, \dots\dots\dots, S_{1n_1}$	$S_{1(n_1+1)}\ , \dots\dots\dots, S_{1n}$
$S_{21}, \dots\dots\dots, S_{2n_1}$	$S_{2(n_1+1)}\ , \dots\dots\dots, S_{2n}$
$\dots\dots\dots\dots$	$\dots\dots\dots\dots$
$S_{B1}, \dots\dots\dots, S_{Bn_1}$	$S_{B(n_1+1)}\ , \dots\dots\dots, S_{Bn}$

Dove :

$$S_{ij} = \begin{cases} 0\ se\ \ j \in r_1^{*i}\ , \dots, r_{n_1}^{*i} \\ 1\ \ se\ \ j \notin r_1^{*i}\ , \dots, r_{n_1}^{*i} \end{cases} \text{ per } j\ = 1, \dots, n_1\ \ e\ i = 1, \dots, B$$

e

$$S_{ij} = \begin{cases} 0\ se\ \ j \in r_{n_1+1}^{*i}\ , \dots, r_n^{*i} \\ 1\ \ se\ \ j \notin r_{n_1+1}^{*i}\ , \dots, r_n^{*i} \end{cases} \quad \text{per } j\ = n_1 + 1, \dots, n\ \ e\ i = 1, \dots, B$$

In questo modo è possibile ricercare una relazione data dall'influenza di un elemento , nel calcolo dei test, e la possibilità che questo abbia cambiato gruppo di appartenenza o meno, così da scomporre l'importanza di ogni osservazione nella significatività del test.

Inoltre vengono calcolati i test statistici, che vengono riassunti come vettori di valori legati alla composizione dei dati, e un vettore di valori binari che indicano se la statistica nella rispettiva posizione, sia oltre il valore osservato. L'avere sempre disponibile soltanto un vettore di statistiche test, è dato dal fatto che, qualora vi siano data set multidimensionali e si affrontano test unidimensionali per ogni dimensione, si combinano i diversi vettori di test statistici tramite una funzione combinatoria di second'ordine e si attua l'analisi utilizzando quest'ultima componente.

Allo scopo si sono adottate due diverse alternative nella scomposizione.

3.1.1 Indici a contributi correlati

Viene considerata la correlazione che si presenta tra la condizione di permanenza nel gruppo originario, durante il processo di permutazione, e il vettore dei valori ottenuti in ogni test statistico rilevato nei B processi. In questo modo si identificano i valori che, quando appartengono al gruppo primario, alzano la statistica test qualora si allontanino dalla media generale[8].
Agli indici ottenuti, si applica la correzione di $\sqrt{n}$, per ottenere una standardizzazione dei valori.

Questa modifica si avvale del risultato che nella stima della varianza e della correlazione :

$$var(X_i) = \sigma^2 \qquad var(\sum_{i \in n} x_1) = n \cdot \sigma^2$$

e

$$\rho^2(x_i, \sum_{i \in n} x_i) = \frac{1}{n}$$ così si ottiene la correzione $$\sqrt{n} \cdot \rho(x_i, \sum_{i \in n} x_i) = 1$$

Questo genere di test diventa rilevante qualora vi sia una significatività particolarmente alta per il rifiuto dell'ipotesi nulla. Infatti, in questo modo risulta sempre la covarianza tra statistiche test e la matrice dell'informazione di scambio. Per questo si può dire che è sempre esplicativa in quanto riporta risultati che in ogni situazione sono interpretabili in nodo corretto, in più è intuitiva perché si comprende subito che all'allontanarsi dal'indice nullo vi è un'influenza della componente sempre maggiore.
Nell'andare ad analizzare il grafico di quest'indice, non vi è un limite prefissato per cui applicare una soglia per determinare se un dato è influente o meno, infatti la

[8] Si precisa ciò perché quando il gruppo(G1) è stocasticamente al di sotto della media generale la correlazione è invertita rispetto ad un gruppo(G2) stocasticamente al di sopra della media.
X_G1 < mean → cor > 0 ;
X_G2 < mean → cor < 0 ;

considerazione di tenere come limite ± 1 (come vedremo poi nei grafici), nasce dal fatto che la maggior parte dei dati sono all'interno di quell'intervallo e quindi si ricercano soltanto quelle osservazioni che risaltano da un'analisi visiva, inoltre, se si volesse essere più scrupolosi nel catalogare l'influenza, si potrebbe inserire una soglia a intervalli maggiori (es. ± 1.5 oppure ± 2). Nonostante ciò bisogna considerare che questo tipo di approccio critico può essere affrontato senza problemi in ambito unidimensionale, mentre andando a considerare un ambiente multidimensionale, anche attuando la standardizzazione imposta precedentemente, non si riscontrano i medesimi intervalli dato che la correlazione, tra test e la matrice condizionata ai valori scambiati, si abbassa man mano che si ha un numero di permutazioni che si avvicina al reale valore di scampi possibili in questo caso vi sono indici inferiori e quindi non si potrebbero applicare le precedenti soglie di determinazione per l'importanza, per cui vengono fissate osservando dove si collocano la maggior parte delle osservazioni, così si considerano influenti i valori che appaiono scostarsi dal gruppo.

3.1.2 Indici a Contributi Individuali

In un primo momento si è scomposto il p-value, visto come $p - value = \frac{(\sum_{b=1}^{B} y_b)}{B}$, in contributi marginali di ogni osservazione (elemento nel caso di dati unidimensionali, vettore nel caso in cui i dati siano multidimensionali) in cui non vengono coinvolti valori legati strettamente a statistiche, infatti vengono utilizzate soltanto componenti composte da valori binari e che combinati risultano restituire le influenze marginali. Grazie a ciò si aggirano problemi che insorgono quando una componente è particolarmente influente, e maschera tutte le altre osservazioni che marginalmente hanno un'importanza significativa.
La decomposizione prevede che qualora la statistica di permutazione sia superiore alla statistica Tα , si somma l'osservazione i-esima tutte le volte che risulta rimanere nel gruppo originario durante le B permutazioni, per poi pesarla sulla base del numero di scambi effettuati rispetto al vettore primario.

$$p - value = \sum_{i \in n} \frac{1}{B} \sum_{b}^{B} \frac{scambiati_{bi} * I(T_{bi} > T_{ob}) * n}{\#scambi\ del\ campione\ nella\ permutazione\ b}$$

L'interpretazione del grafico, ottenuto dai contributi individuali, è simile a quella precedentemente introdotta per l'indice a contributi correlati, infatti, i valori influenti sono i punti che si identificano alle estremità superiori ed inferiori del grafico .

La differenza sostanziale è che in questo caso viene riportato come scala un valore che appare come una successione di p-value per le componenti. Infatti, qualora un'osservazione porti ad avere un'alta significatività[9], si avrà un valore basso per l'indice e, viceversa, quando una componete porterà fortemente ad accettare l'ipotesi si avrà un valore alto di indice.
Questo non è interpretabile come il p-value introdotto dal valore, perché anche osservazioni con influenze al di sopra di 0.05 possono portare al rifiuto del test, inoltre vi è anche la possibilità di ottenere indici maggiori di 1 quando vi è una significatività del test ≅ 1.

Le precedenti scomposizioni operano anche qualora vi si presentino ipotesi alternative unilaterali e quindi saranno considerati i seguenti sistemi d'ipotesi:

$$\begin{cases} H_0\colon\ X_1 =^d X_2 \\ H_1\colon\ X_1 <^d X_2 \end{cases}$$

$$\begin{cases} H_0\colon\ X_1 =^d X_2 \\ H_1\colon\ X_1 >^d X_2 \end{cases}$$

$$\begin{cases} H_0\colon\ X_1 =^d X_2 \\ H_1\colon\ X_1 \neq X_2 \end{cases}$$

[9] porta un p-value basso

Capitolo 4

Applicazione a dati reali

Per valutare la validità del progetto si è effettuato uno studio utilizzando il software statistico R, generando un numero di permutazioni pari a B=10000 su un campione, per il quale si identificano lo spazio di permutazione, il vettore dei test statistici e il vettore dei p-value sopra descritti.

4.1 studio su 2 campioni indipendenti

Continuando lo studio utilizzato precedentemente, andiamo ad osservare un caso di test per due campioni indipendenti, i quali erano classificati da maschi e femmine, uso di alcol o meno. Andando a cercare una distinzione significativa tra due tipi di classificazione.

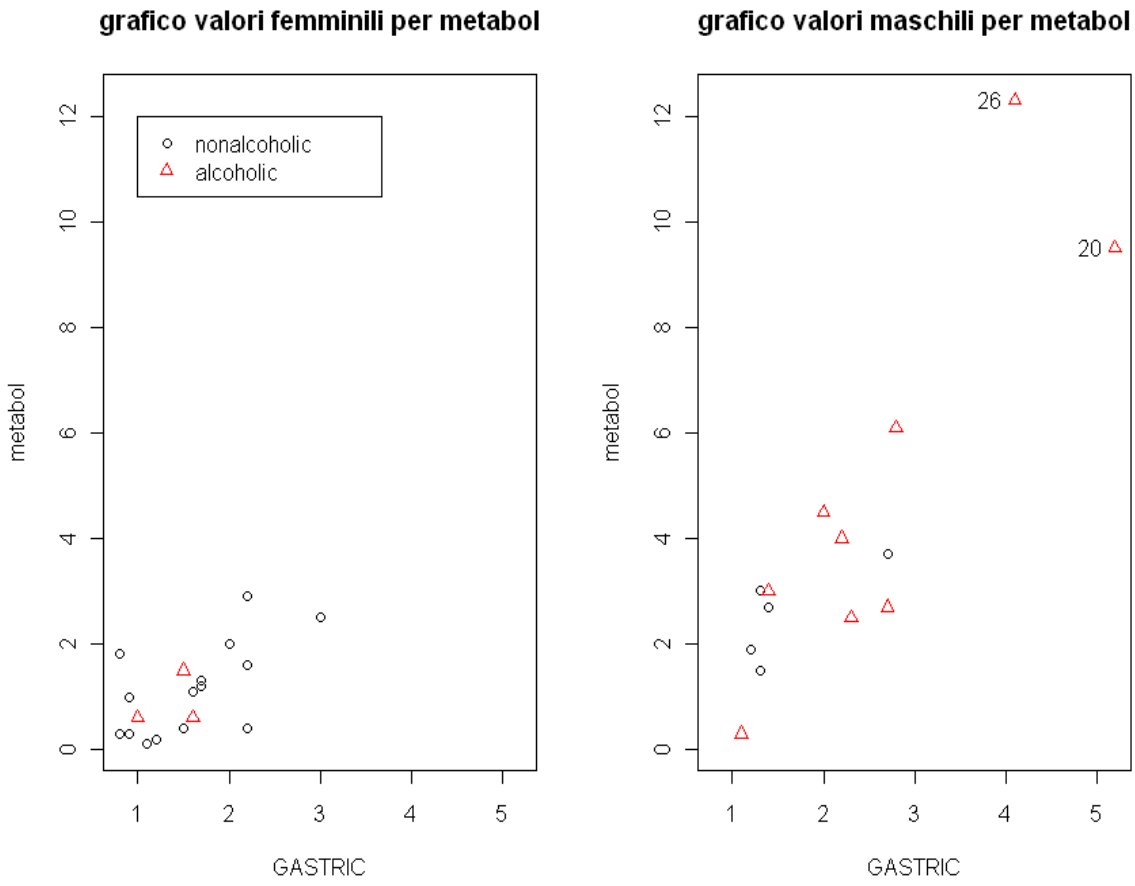


Figura 1

Dai grafici di figura 2 si nota che solo nella distinzione di metabol in maschi e femmine comporta una sostanziale differenza, mentre tra alcolic e non vi è rilevanza nella distinzione dei due gruppi. A conferma di ciò confrontiamo la differenza tra i gruppi nelle due circostanze. Per saggiare i confronti indotti utilizziamo un sistema di ipotesi:

$$\begin{cases} H_0\colon\ X_1 =^d X_2 \\ H_1\colon\ X_1 \neq X_2 \end{cases}$$

Dove X_1 e X_2 identificano i due gruppi di rilevazioni che si vanno a identificarsi nel momento in cui stratifichiamo per un criterio.

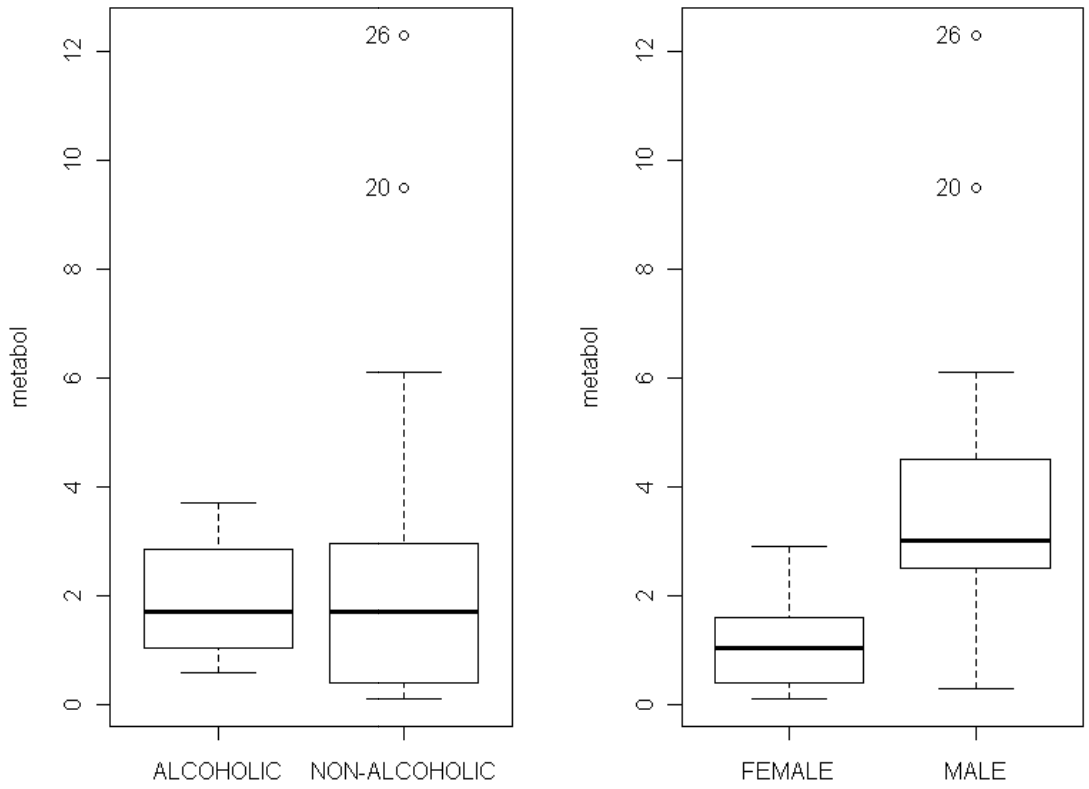


Figura 2

Nel caso di uso di alcol, vi è una significatività oltre al 90% per cui possiamo ribadire che non vi è una differenza in metabol. Nel confronto tra maschi e femmine troviamo, differentemente da prima, che risulta molto significativo, infatti il p-value risultante è ≅0, per cui possiamo stratificare le rilevazioni attraverso il sesso.

Quello che ci risulta ora è una situazione in cui, oltre ad avere una componete risposta y (metabol) e una differenziazione di genere (sex), ci troviamo ad osservare una covariata che comporta un'influenza su metabol (figura 3). Vediamo, infatti, che all'aumentare di gastric aumenta anche il valore di metabol.

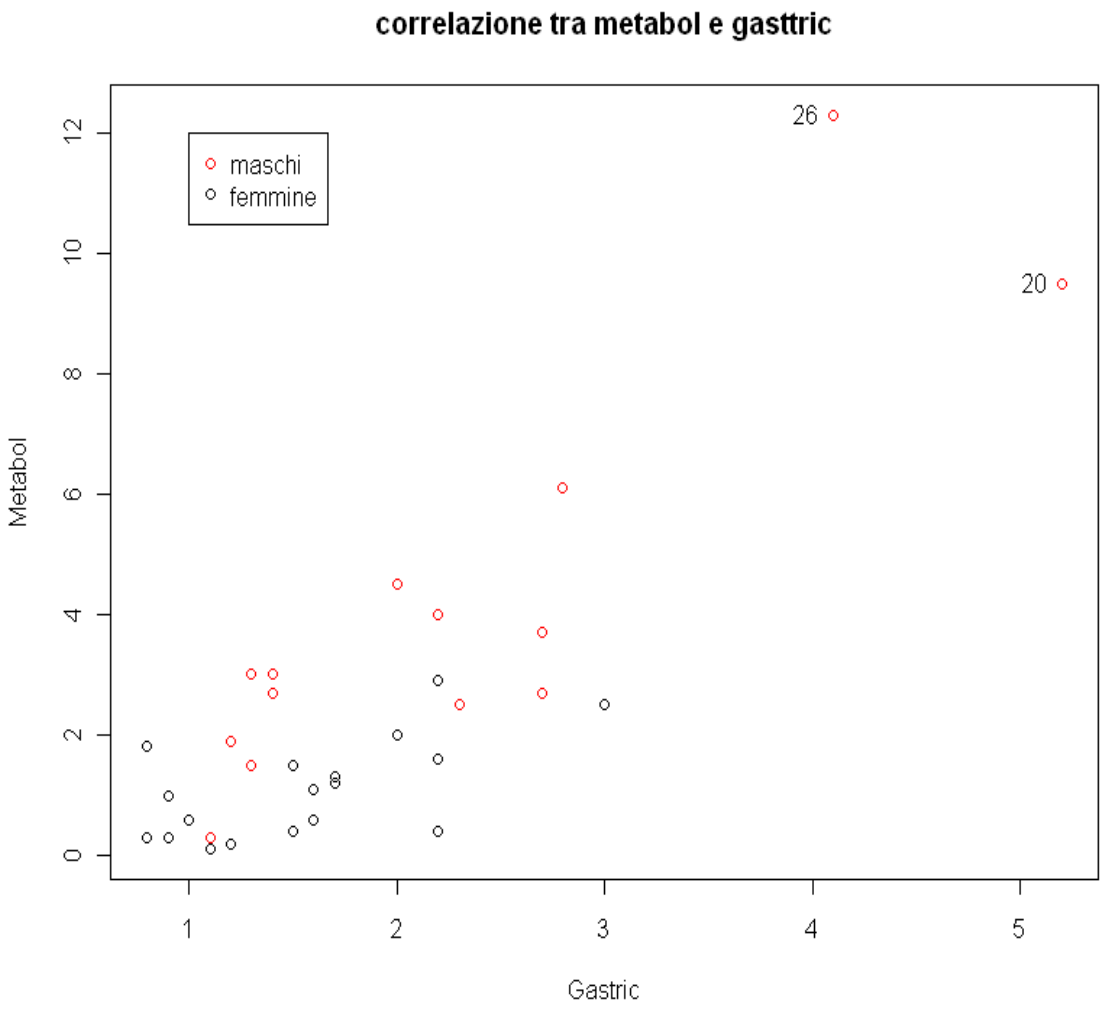


Figura 3

Per individuare la reale influenza di gastric, si andranno poi ad analizzare come si comportano gli indici calcolati in relazione alla variabile. In tal modo si vedrà se, al crescere della seconda componente, gli indici delle influenze saranno costanti, e quindi non vi sarà una reale influenza di gastric, oppure si avrà che al crescere dell'importanza corrisponderà una variazione di gastric. Se questo accadrà vorrà dire che in metabol vi è presente un'influenza derivante non solo da sex, ma anche da gastric.

Gli indici di influenza tra metabol e il sesso risultano:

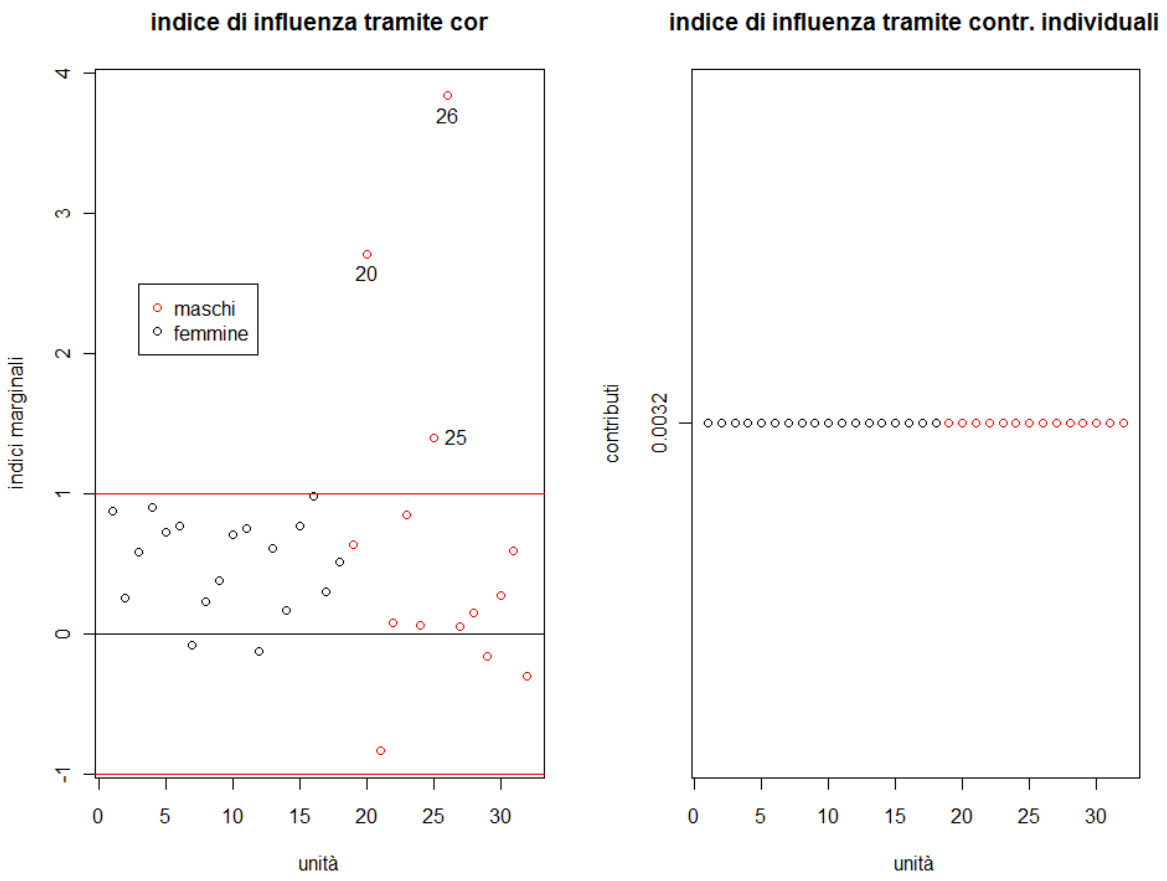


Figura 4

In questo caso, dato che la significatività del test è molto significativa (≅0), il grafico dei contributi individuali risulta non informativo ed i dati risultano come insieme di punti allineati a causa della grande significatività del p-value.

Guardando i grafici finora inseriti, si notano due punti che si assumono tutta l'influenza data, per cui saranno tolti dal complesso dei dati per ricavare ulteriori informazioni nascoste da questi due elementi.

Nei successivi due grafici si vedono ora tutte le reali influenze sul test da parte di metabol.

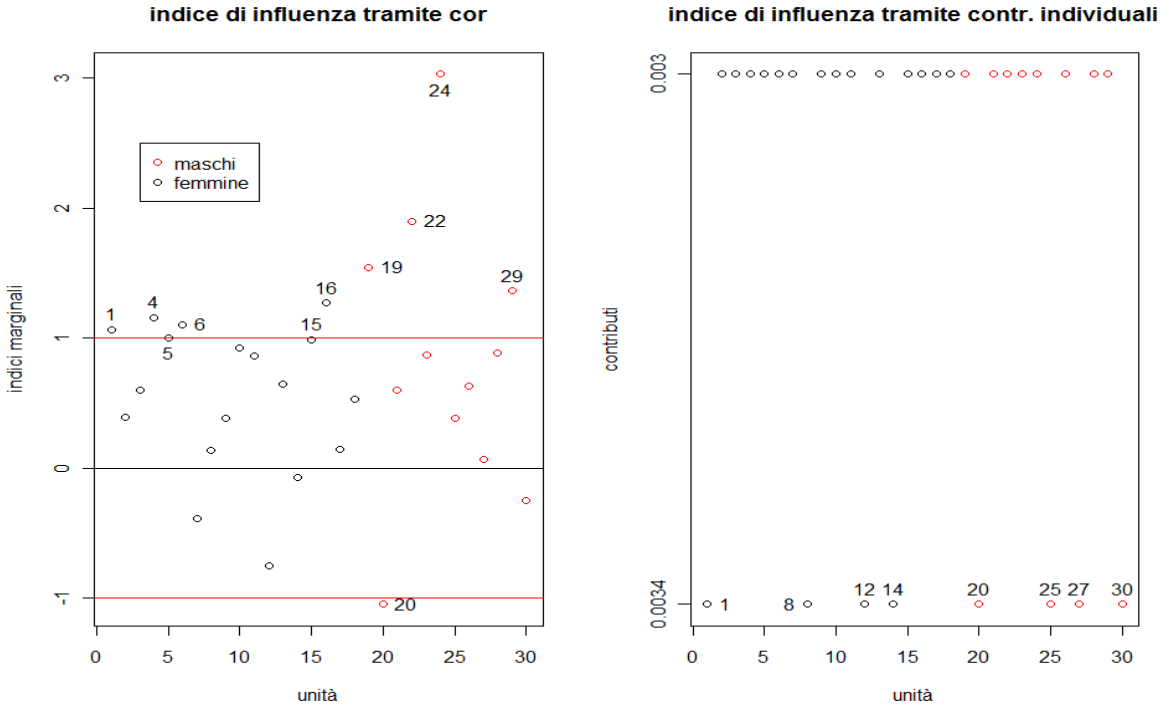


Figura 5

A questo punto andiamo a verificare il problema, avanzato precedentemente, dell'informazione nascosta in metabol di gastric. Per affrontare ciò andiamo ad analizzare la covarianza che si ha tra gli indici di influenza, in questo caso estratti con la correlazione, con i valori di gasrtic.

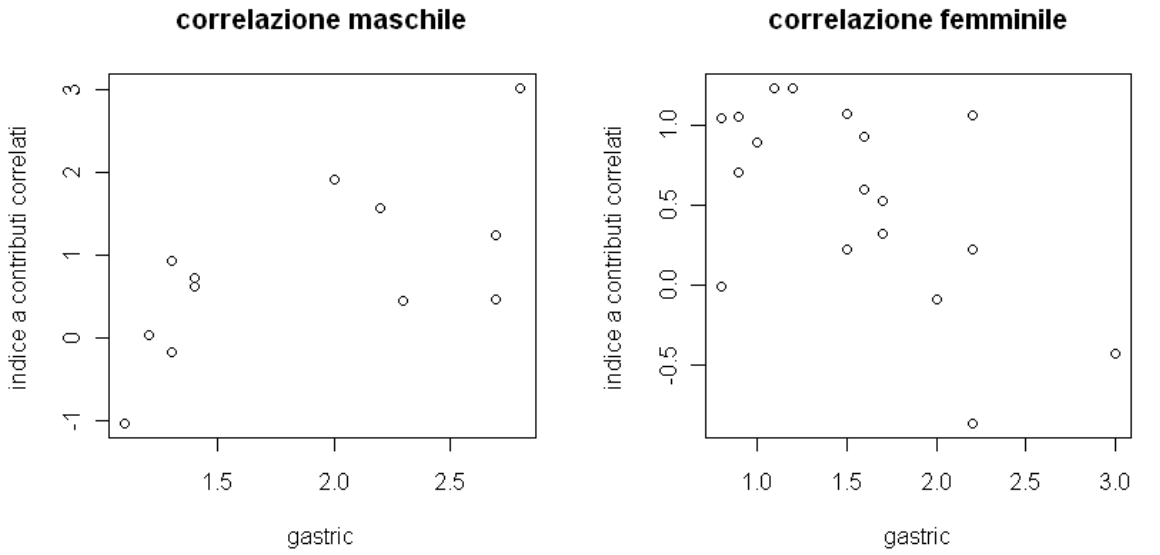


Figura 6

Da questi grafici si può vedere che a maggiore influenza della componente i-esima, varia il valore di gastric. Si vede, infatti, che nel secondo grafico si ha una correlazione per così dire inversa, questo è dovuto al fatto che a bassi valori di metabol, e quindi alti d'influenza, si hanno valori minori di gastric, questo perché vi è una relazione univoca crescente tra metabol e gastric. Per cui si può affermare che la prima variabile è molto legata al fattore gastrico che porta un'influenza maggiore del sesso a cui appartengono gli individui.

Per capire quanto è veramente influente il sesso, apporto un approccio basato sulla stratificazione di gastric in valori maggiori o minori della media, così facendo si può vedere come il sesso va ad influire sugli indici calcolati per la classificazione appena introdotta.

Come è stato riscontrato precedentemente, vi è una differenza tra gastric superiore e inferiore alla media, infatti, si ottiene un p-value del 0.01. Dall'analisi fatta si ottiene:

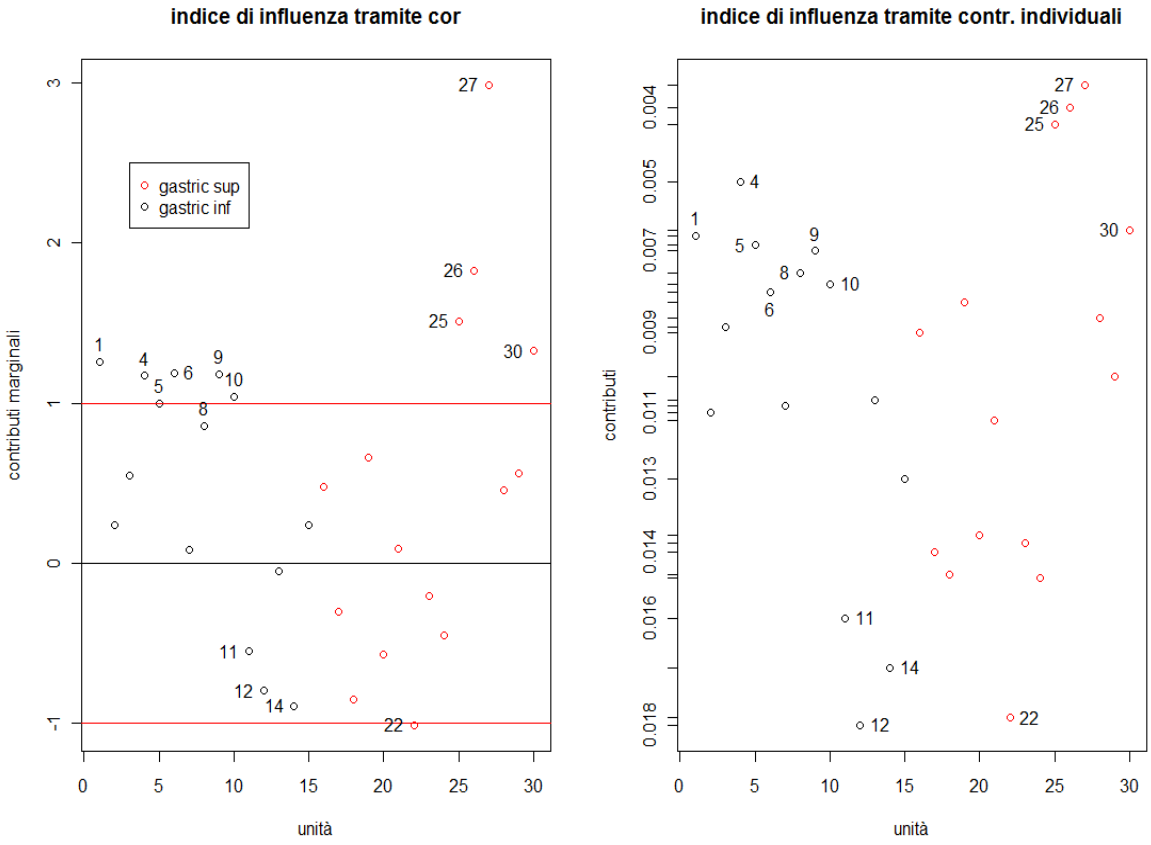


Figura 7

Per andare a ricercare questi indici si deve tener presente che le posizioni sono variate, dato che i due gruppi, in questo caso, sono stati ordinati in base alla classificazione[10] (ad esempio il valore 18 rappresenta il 3° valore che si ottiene selezionando soltanto i valori del secondo gruppo).

Così, andando a osservare l'influenza rispetto ai sessi, si osserva che:

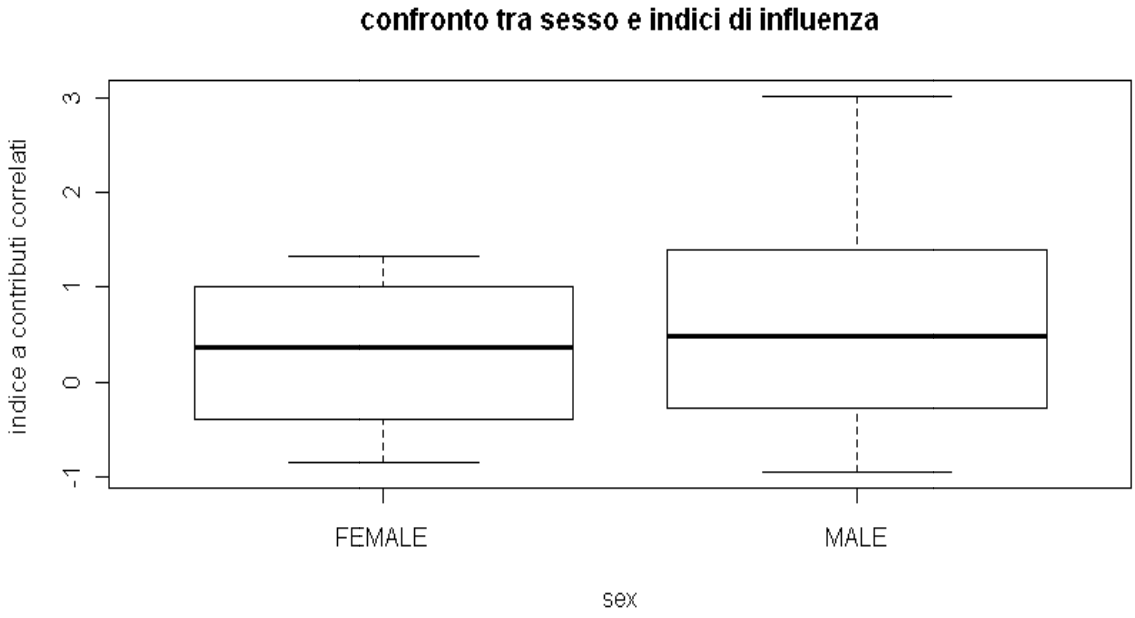


Figura 8

Il p-value che si ottiene dal test per il confronto di due campioni indipendenti è >0.5, per cui non ora non si rileva una differenziazione tra maschi e femmine.

Si può capire ora che l'influenza precedentemente riscontrata era dovuta dalla presenza di gastric, infatti, ora confrontando gli indici in base al sesso, non si riscontrano differenze di classificazione. L'unica affermazione che si potrebbe portare a riguardo è che per il sesso maschile si ha una maggiore variabilità d'importanza.

[10] Precedentemente questo problema non veniva appariva perché i valori erano ordinati distintamente in base al sesso

4.2 Studio su campioni a dati appaiati

Prendendo in considerazione i tassi d'occupazione delle donne nel 1972 e nel 1968[11] in 19 città degli Stati Uniti, si vuole testare $H_0: \{Y_{1972} =^{d} Y_{1968}\}$ contro $H_0: \{Y_{1972} \neq^{d} Y_{1968}\}$, ovvero si vuole verificare se negli anni il tasso di occupazione aumenta.

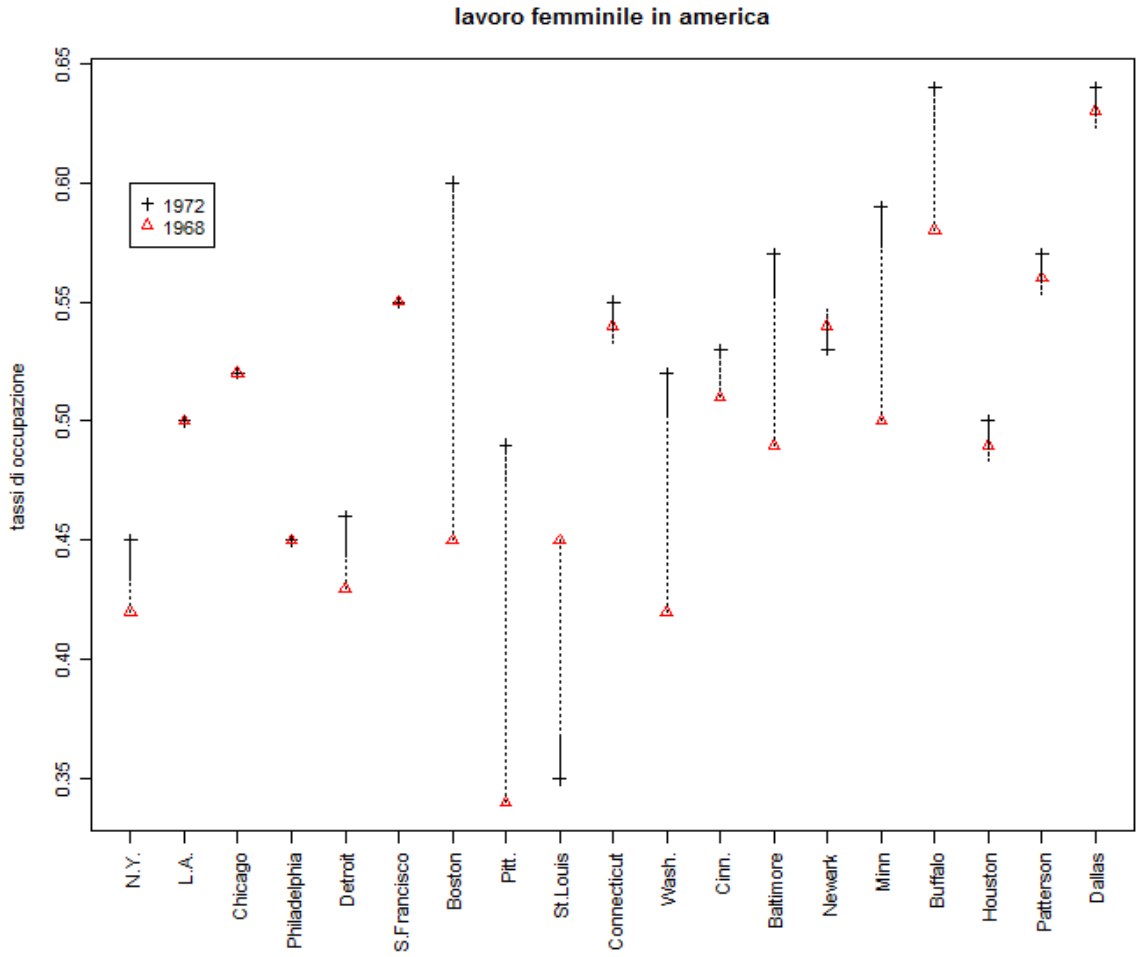


Figura 9

[11] "City: City United States, 1972: Labor Force Participation rate of women in 1972, 1968: Labor Force Participation rate of women in 1968"

Dal grafico notiamo che nel trascorrere degli anni, il tasso d'occupazione, sembra aumentare, anche se vi sono osservazioni talvolta uguali o minori. Infatti, a parità di osservazioni, troviamo spesso che le rilevazioni fatte nel 1972 hanno un tasso maggiore delle rispettive del 1968. Per affrontare questo tipo di problema analizzo le differenze dei logaritmi rilevate, per ogni città, nei due anni presi in considerazione. Utilizzando un test di permutazione ad un campione verifico che l'ipotesi nulla delle differenze viene rifiutata con un p-value <0.05 (0.0201). Nell'affrontare questo tipo di problema ci si avvale del metodo che impone la correlazione corretta tra statistiche test e l'evento scambiato.

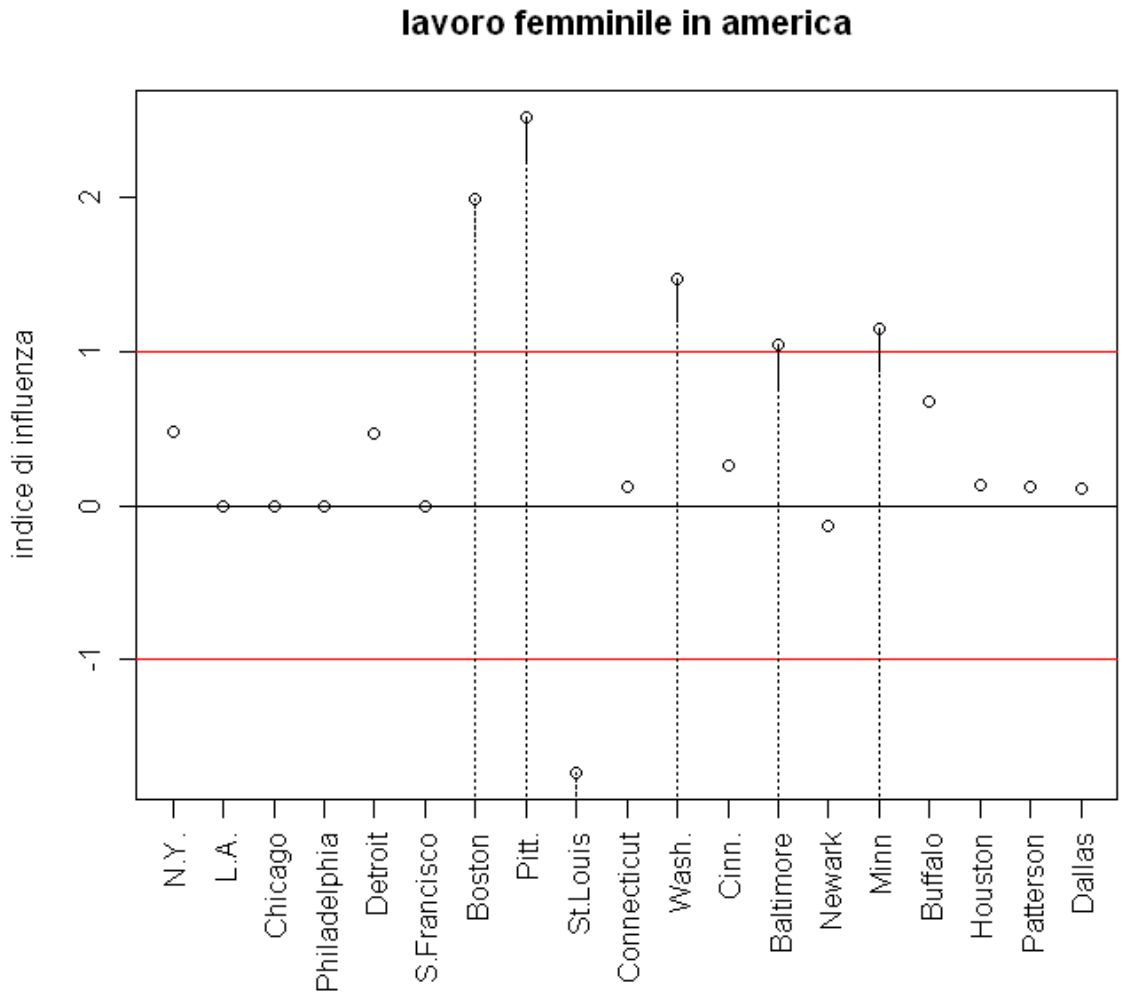


Figura 10

Nel valutare quali siano le città che portano al rifiuto dell'ipotesi scompongo il p-value in contributi individuali tramite la correlazione corretta tra la condizione di aver subito un cambio di segno e la statistica test osservata in tale permutazione.

Osservando il grafico notiamo che le città che comportano delle notevoli differenze sono, in ordine: Boston, Pittsburgh, St.Louis, Washinton, Baltimore, Minnesota.

Dal grafico degli indici (figura 10) si notano come le componenti si pongono, si nota, infatti, che l'elemento 9 (St.Louis) da un'influenza negativa, in quanto in relazione al suo valore comporta un peso sul p-value inversamente proporzionale. Questa discrepanza la si può giustificare andando a verificare che attorno al 1971 a St.Louis vi sono state proteste nei settori meccanici e, di conseguenza, sono stati attuati licenziamenti di massa. Essendo questa una città operaia, il tasso occupazionale ne ha risentito in modo molto evidente.

4.3 Studio su due campioni indipendenti multidimensionali

Per affrontare il problema di campioni multidimensionali ci si può riportare ad uno studio condotto da Potthoff e Roy (1964) in cui viene calcolata la distanza in millimetri dal centro della ghiandola pituitaria alla fessura pterigo-mascellaredi 11 ragazze e 16 ragazzi. Le misurazioni sono state effettuate sugli stessi soggetti in età differenti (8, 10, 12 e 14 anni). Uno dei problemi statistici che si vogliono verificare su questi dati è quello di vedere se la distribuzione dei singoli profili delle ragazze è diverso da quello dei ragazzi e, in particolare, se gli incrementi dei profili delle ragazze sono stocasticamente dominanti su quelle dei maschi, nel senso che le ragazze si avvicinano alle loro dimensioni adulte rispettivi antropometriche prima dei ragazzi.

A questo scopo si andrà ad analizzare le differenze delle rilevazioni riscontrate dall'anno precedentemente, per cui si formulerà il problema come $\mathrm{Y_{ij}(t) = Xji(t) - Xji(t-1),\ t = 2,3,4}$. Per cui il sistema d'ipotesi risulta

$$\mathrm{H_0 : \{\bigcap_2 Y_G(t) =^d Y_B(t)\}}$$

Contro

$$\mathrm{H_1 : \{\bigcup_2 Y_G(t) <^d Y_B(t)\}}$$

In quanto, se si vuole confutare l'ipotesi che le ragazze si sviluppano prima, allora queste dovrebbero avere delle differenze minori negli anni.

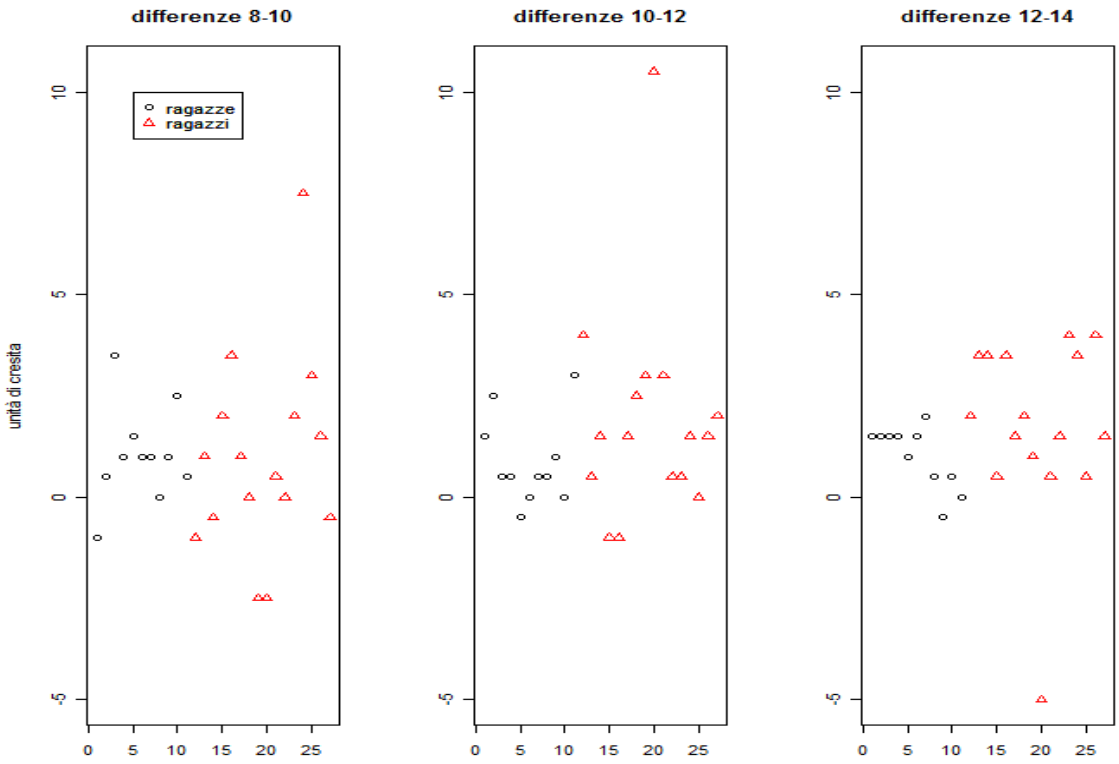


Figura 11

Guardando i grafici di figura 11 sembrano constatare l'ipotesi fatta, infatti, le differenze sono minori. Ciò indica che le ragazze raggiungono prima dei ragazzi le misure adulte. Nel verificare l'ipotesi viene introdotto il test per ipotesi unilaterali:

$$T_j = \sum_{i \leq n_G} Y_{ij} \ \ per\ j = 1{,}2{,}3$$

In questo modo si compiono tre test d'ipotesi che risultano avere p-value rispettivamente : 0.8187181, 0.1286871, 0.2224778.

Anche se singolarmente i test non risultano significativi, utilizzando il test di second'ordine che somma diretta dei test statistici, si ottiene che H_0 viene rifiutata con p-value = 0.03989601.
Grazie a ciò si deduce che vi è una reale differenza tra i due gruppi, infatti, scomponendo in contributi marginali degli elementi si nota che vi sono osservazioni che risultano avere un'influenza maggiore delle altre in particolare appare da entrambe gli approcci che l'osservazione 25 risulta la più influente nella differenziazione dei gruppi.

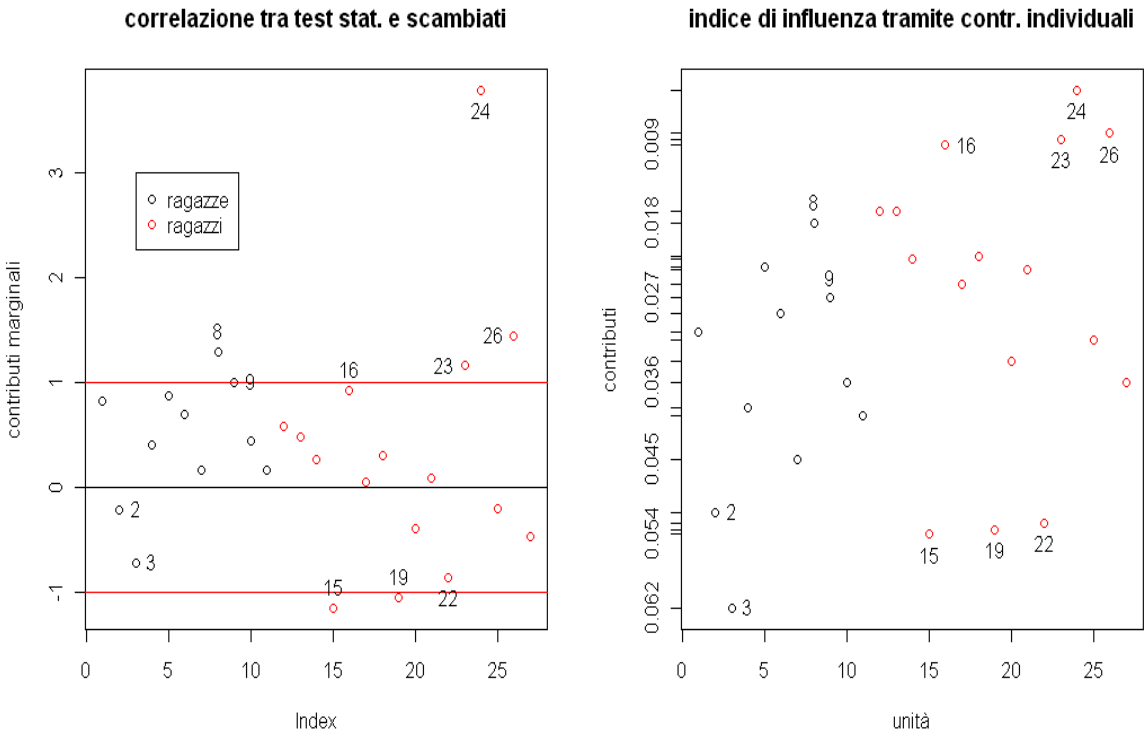


Figura 12

Guardando il primo grafico nella figura 12 si vedono totalmente i punti che influiscono sulla nullità della distribuzione.

Osservando la disposizione dei punti si deve introdurre una considerazione sull'influenza, infatti, si deve valutare che nei due gruppi l'indice a contributi correlati deve essere letta in due modi opposti. Per questo se andiamo a leggere le rilevazioni delle ragazze si deve dire che le osservazioni 8 e 9 sono molto piccole rispetto ad un valore della media (o della mediana), mentre le osservazioni 24 e 26 saranno molto grandi rispetto alla medesima misura. In questo modo si può dedurre ciò che fin dal principio si vuole verificare, infatti, si nota che i valori indicati in nero (le ragazze) hanno in maggior parte indici positivi, quindi si può affermare che le differenze nelle età delle ragazze sono stocasticamente minori di quelle dei ragazzi.

Tale affermazione significa allora che, essendoci minori differenze nelle ragazze, queste raggiungono prima le misure che si andranno a riscontrare in età adulta (in questo caso considerata come la misurazione finale).

Nel secondo grafico di figura 12 si vanno ad analizzare i contributi individuali, si riscontrano la maggior parte dei valori influenti evidenziati precedentemente : 26, 24, 23, 19, 16, 15.
Questi valori hanno molta rilevanza nel confutare l'ipotesi fatta, però vi sono altre componenti che si identificano e risultano essere significative anch'esse, ad esempio la terza osservazione risulta distinguersi.
In questo caso si potrebbe affermare che, la rivelazione individuata come terzo elemento, comporta un'influenza quasi pari, anche se opposta, all'osservazione 8. Infatti, prendendo come punto di riferimento l'indice centrale dei test riguardanti le ragazze, queste due osservazioni evidenziate hanno una distanza approssimativamente simile.
La differenza riscontrata con il primo grafico potrebbe essere dovuta da due fattori :

1) Qualora vi siano componenti molto rilevanti e quindi con valori di test statistici alti le altre, influenti anch'esse, sono in qualche modo eclissate dalla prima perché la correlazione si concentra su quest'ultima.

2) Il grafico dei contributi correlati, essendo basato sui valori dei test statistici, viene centrato sulla media generale. Per cui se realmente la media di un primo gruppo si discosta da quella del secondo, però vi sono valori che influiscono talmente tanto che riportano il valore della media del gruppo sul valore generale, allora le influenze del gruppo ora considerato sono tutte traslate verso la correlazione corretta assunta dalla tendenza del gruppo.

Questo risultato non si riscontra nel grafico a contributi individuali perché considera una scomposizione che non comprende i valori dei test, ma soltanto la condizione se è oltre il valore critico o meno, per cui anche se vi sono test statistici molto grandi, questi non influiscono sulla stima degli indici delle influenze.

Bibliografia

Fortunato Pesarin, “Multivariate Permutation Tests With Applications in Biostatistics “(2001), Wiley, Chichester,

Fortunato Pesarin, “Permutation testing of multidimensional hypotheses, by non parametric combination of dependent tests” (1999), , cleup editrice – padova

John Fox, “An R and S-Plus Companion to Applied Regression” (2002)

John Fox, “Bootstrapping Regression Models, Appendix to An R and S-PLUS Companion to Applied Regression” January 2002,

Pace, L. e Salvan, A., “Introduzione alla Statistica – II Inferenza, Verosimiglianza, Modelli “(2001). Cedam, Padova.

Richard G. Pearson, “Journal of Biogeography (J. Biogeogr.) “(2007) 34, 102–117, Christopher J. Raxworthy

Sidney Siegel & n. John Castellan jr., “Statistica non parametrica, seconda edizione” (1992), mcgraw-hill