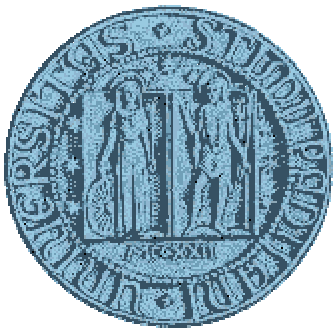


UNIVERSITÀ DEGLI STUDI DI PADOVA
FACOLTÀ DI SCIENZE STATISICHE

CORSO DI LAUREA TRIENNALE IN STATISTICA E GESTIONE DELLE IMPRESE

UNA ANALISI DELLE ESPERIENZE DI EDUCAZIONE FISICA NELLE SCUOLE: I DATI PACES

RELATRICE: CH. MA PROF. SSA LAURA VENTURA
CORRELATORE: PROF. ATTILIO CARRARO

LAUREANDA: BARBARA SARTORI

ANNO ACCADEMICO 2005-2006

Indice

Introduzione

Obiettivo di questo lavoro è analizzare, tramite strumenti statistici, un campione di 342 studenti (170 femmine e 172 maschi) di 5 scuole elementari di una provincia Veneta: Arcugnano, Torri, Monticello, Camisano e Scamozzi.

Lo scopo di questa analisi è valutare il gradimento del professore di educazione fisica da parte degli allievi attraverso un questionario redatto da alcuni psicologi del Dipartimento di Psicologia dell'Università degli Studi di Padova.

Il questionario, somministrato ai bambini, è strutturato in 16 affermazioni, sia di carattere emotivo (per esempio "Mi diverto") sia di carattere più "fisico" (per esempio "Mi da energia"), a cui i bambini possono rispondere, secondo la scala Likert, con un valore compreso da 1 a 5 (1 = per nulla, 5 = molto). Inoltre, sono stati rilevati la frequenza pratica (bassa, madia, alta), il sesso del bambino (M,F), la scuola di appartenenza e il sesso del docente di educazione fisica (M,F).

L'indagine è stata svolta due volte, in modo tale da verificare se i bambini avessero risposto accuratamente alle domande o se avessero dato le risposte con superficialità.

Per la natura dei dati e lo scopo dell'analisi, si è scelto di analizzare i dati attraverso l'analisi delle corrispondenze, una tecnica algebrica che si prefigge di rappresentare graficamente le modalità dei caratteri in un sottospazio di dimensionalità minima. Questo tipo di analisi è in grado non solo di supportare le ipotesi di lavoro di partenza, ma anche di orientare lo studio verso la formulazione di nuove ipotesi, di verificare qualitativamente l'esistenza di opportune assunzioni sulle variabili in esame e di suggerire eventuali modelli statistici parametrici. I supporti grafici facilitano la lettura dell'informazione statistica, rendendola maggiormente incisiva e rapida, soprattutto in presenza di grandi basi di dati, come in questo caso.

Tutte le analisi sono state svolte tramite il programma statistico R. R è un linguaggio e un ambiente statistico *open source* (scaricabile dal sito http://www.r-project.org/) per l'analisi dei dati, nato come estensione del linguaggio di

programmazione S. Inizialmente è stato scritto da Ross Ihaka e Robert Gentleman del Dipartimento di Statistica dell'Università di Auckland, Nuove Zelanda, ma dal 1997 lo sviluppo di R è affidato ad un gruppo internazionale, l'R Core Team.

Lo schema della tesi è il seguente: nel primo capitolo verrà descritta la tecnica utilizzata per analizzare i dati, cioè l'analisi delle corrispondenze, nel secondo verranno svolte le analisi preliminari sui dati, nel terzo, infine, sarà applicata l'analisi delle corrispondenze ai dati.

Capitolo 1

L'analisi delle corrispondenze

L'analisi delle corrispondenze è una tecnica algebrica utile per lo studio della struttura della dipendenza interna di una tabella di frequenza, basata su una rappresentazione grafica delle modalità dei caratteri in uno spazio di dimensionalità minima (cfr. Fabbris, 1983). Questa tecnica è utile quando si vuole analizzare una tabella di ampie dimensioni, contenente numeri non negativi, derivata dalla scomposizione di un fenomeno secondo due o più caratteristiche, per le quali ha interesse estrarre l'informazione utile in termini di similarità fra gli elementi appartenenti a ciascuno dei due insiemi di riga e di colonna. Tale similarità si osserva tramite la rappresentazione fattoriale della forma di nuvole di punti associate a tali insiemi.

Questo tipo di analisi è molto importante in quanto permette di studiare l'informazione proveniente sia da caratteri qualitativi che da caratteri quantitativi (se la variabile è di tipo continuo bisogna però suddividerla opportunamente in classi), al contrario dell'analisi fattoriale che è utilizzabile solo con variabili quantitative.

Oggi uno dei campi privilegiati d'applicazione di questa tecnica è l'analisi dei questionari, in quanto di solito coesistono informazioni sia numeriche (come età, reddito o varie misure di durata, distanza o intensità), sia ordinate (come scale di atteggiamento, preferenza o d'accordo), sia nominali (risposte a scelta multipla o dicotomiche).

L'analisi delle corrispondenze nacque contemporaneamente negli Stati Uniti e nel Regno Unito, e successivamente fu riscoperta indipendentemente da numerosi studiosi. All'inizio non ebbe un nome fisso; infatti Richardson la chiamò "method

of reciprocal averages", mentre Hill la definì "reciprocal averaging". Il nome definitivo lo diede Benzécri, uno studioso francese a cui si deve la diffusione di questo metodo in Francia negli anni sessanta, che la chiamò prima "analyse factorielle des correspondences", e poi più semplicemente "analyse des correspondence".

1.1 Scelte per l'analisi delle corrispondenze

Per effettuare una analisi delle corrispondenze è necessario prendere delle decisioni affinché i risultati ottenuti portino a delle conclusioni corrette e utili per lo scopo dell'analisi.

i. Individuazione delle variabili e delle modalità da considerare per l'analisi: poiché la matrice da analizzare è di grandi dimensioni, l'ispezione visiva delle tabelle non basta per rilevare le relazioni tra le variabili; bisogna quindi scegliere le variabili in modo che queste rispettino i criteri di omogeneità delle misure (cioè le modalità devono essere quantità o punteggi espressi in unità di misura omogenee, per poter calcolare distanze sensate), omogeneità del contenuto (vanno individuate le informazioni che si rapportano, anche in senso ampio, ad uno stesso fenomeno) ed esaustività dell'insieme (cioè tutti gli aspetti del fenomeno devono essere presi in considerazione).

   Anche la scelta delle modalità è molto importante in quanto più modalità di una variabile si inseriscono nell'analisi, tanto più elevata è la probabilità che quella variabile sia importante nel determinare la soluzione analitica.

   Queste due scelte sono molto importanti perché influiscono significativamente sull'esito dell'analisi.

ii. Ripartizione delle variabili osservate in attive e supplementari: come già detto precedentemente, non tutte le variabili vengono utilizzate per la determinazione della soluzione. E' quindi necessario suddividere le variabili in "attive", cioè quelle che si impiegano nella ricerca della soluzione fattoriale, e in "supplementari" o "illustrative", cioè le variabili che non si utilizzano per il ritrovamento della soluzione, ma che si proiettano alla fine sugli assi trovati analizzando le modalità attive.

   L'assegnazione delle variabili ad una di queste due categorie è arbitraria e influisce sull'esito della soluzione. Infatti, lo scambio di identità tra variabili

attive e supplementari è utilizzato alcune volte per verificare la stabilità della soluzione.

iii. Ripartizione delle unità osservate in attive e supplementari: alcune unità possono essere escluse nella fase di ricerca della soluzione fattoriale ed essere introdotte in seguito come ausilio nell'interpretazione dei risultati.

iv. Tipo di approccio analitico: l'approccio può essere "semplice", se vengono utilizzate due sole variabili, o "multiplo" se si utilizzano più di due gruppi di variabili.

v. Dimensionalità della soluzione: poiché non conosciamo il numero esatto di fattori, si può iniziare l'analisi presumendo una dimensionalità alta (al massimo quattro o cinque assi) e poi costringere la soluzione ad un numero inferiore di assi (solitamente due, così da poter rappresentare graficamente le modalità per una più facile interpretazione dei risultati).

vi. Ritorno all'indietro: è possibile valutare la bontà della soluzione trovata ricostruendo la tabella iniziale tramite gli autovalori e gli autovettori ricavati.

1.2 Metodo di calcolo delle corrispondenze

Si consideri la matrice originaria dei dati indicata con *R(n,s)* con *n* righe (cioè il numero di unità statistiche) e *s* colonne (il numero di variabili rilevate).

$$R(n,s) = \begin{bmatrix} r_{1A} & \dots & r_{1S} \\ \vdots & \ddots & \vdots \\ r_{nA} & \dots & r_{nS} \end{bmatrix}$$

Ora si considera la matrice disgiuntiva completa, indicata con *D(n,p)*, che ha le *n* unità statistiche nel senso delle righe e tutte le modalità di ogni variabile nel senso delle colonne. Ogni colonna di questa matrice rappresenta una nuova variabile, indicata con il nome della modalità, il cui valore sarà 1 se l'unità statistica della riga considerata assume quella modalità per la variabile originaria corrispondente, altrimenti sarà 0 (quindi queste nuove variabili sono dicotomiche). Questa matrice ha alcune caratteristiche:

- la somma degli elementi di ogni riga è uguale al numero delle variabili di partenza;
- ogni blocco di colonne corrispondente ad una variabile contiene uno e un solo 1 per ogni riga.

Adesso possiamo costruire la matrice *B(p,p)*, di tante righe e tante colonne quante sono le colonne di *D*, detta matrice di Burt (o delle corrispondenze multiple). Questa matrice quadrata è data dalla moltiplicazione della trasposta di *D* per *D*, ossia

$$B = D' \cdot D \ .$$

In ogni casella ci sono dei numeri che rappresentano quante volte si è registrata la compresenza delle due modalità, quella sulla riga e quella sulla colonna. Gli elementi sulla diagonale contano semplicemente le frequenze della modalità corrispondente. I valori possibili per gli elementi della matrice di Burt sono tutti i valori compresi tra 0 e *n*.
Si può dividere la matrice *B* in p^2 sottomatrici:

- quelle sulla diagonale sono le matrici dove si incrociano le modalità della stessa variabile e sono matrici diagonali in quanto non è possibile per una variabile assumere due modalità distinte in un singolo caso; la traccia di ogni singola sottomatrice diagonale è pari a *n*;
- le sottomatrici extra-diagonali sono le tabelle di contingenza delle diverse variabili.

La matrice su cui si effettua l'analisi è la matrice ottenuta dalla matrice di Burt "normalizzata", ossia dividendo gli elementi di ogni colonna per l'elemento appartenente alla diagonale della colonna stessa, ovvero la matrice i cui elementi sono:

$$b^*_{ij} = \frac{b_{ij}}{b_{ii}} \qquad \begin{matrix} i = 1,..., \ p \\ j = 1,..., \ p \end{matrix} \ .$$

Questa matrice è ancora una matrice simmetrica ed è quindi possibile "diagonalizzarla", cioè trovarne gli autovalori e gli autovettori.
A questo punto si procede prendendo i primi autovettori come "variabili latenti" (fattori) e calcolando le coordinate fattoriali dei punti-unità sull'asse fattoriale (cfr. ad esempio Fabbris, 1983).

1.3 Criteri per determinare il numero ottimo di fattori

Nell'analisi delle corrispondenze un fattore è una combinazione lineare delle modalità individuate, ognuna considerata come una variabile a se stante.
Come già detto il numero di assi su cui rappresentare la nuvola di punti non è predeterminabile; ci sono però alcuni criteri che possono aiutare a farsi un'idea della dimensionalità dei dati.

- ***Numero di fattori prefissato***

  Questo criterio viene utilizzato per far si che non si trovino troppi assi fattoriali che produrrebbero una soluzione troppo complicata. Poiché questo criterio è molto arbitrario spesso viene accompagnato con un altro dei seguenti criteri.

- ***Soglia di inerzia globale***

  Nell'analisi delle corrispondenze semplici, la somma degli autovalori non banali è uguale a χ^2/n, dove χ^2 è il coefficiente di Pearson per la misura della dipendenza tra le variabili. L'inerzia (variabilità) tra osservazioni spiegata dalla soluzione va quindi valutata in base al rapporto tra gli autovalori degli assi e la somma degli autovalori della tabella esaminata. La frazione di inerzia spiegata dai primi fattori è una misura della loro idoneità a rappresentare la variabilità delle modalità analizzate (poiché la frazione di inerzia dipende dal numero di modalità attive non si può stabilire a priori una soglia di inerzia, ma solitamente percentuali di inerzia spiegata superiori al 60% possono considerarsi buone).

  Nell'analisi delle corrispondenze multiple la frazione di variabilità spiegata dai fattori non può essere confrontata con quella di soluzioni ricavate seguendo altri approcci né con quella di eventuali analisi delle componenti principali su dati resi confrontabili, per questo per determinare la soglia di inerzia si utilizzano delle simulazioni.

- ***Significatività statistica della soluzione***

  Sia la somma degli autovalori non banali che la somma degli autovalori della soluzione moltiplicate per *n* possono essere confrontate con i valori critici della distribuzione del χ^2 di Pearson con (*H*-1)x(*M*-1) gradi di libertà per valutare la significatività statistica della dipendenza tra le variabili della tabella originaria e della tabella ricostruita in base agli autovalori e agli autovettori della soluzione.

- ***Effetto Guttman***

  Spesso quando si riportano i punti sui piani definiti da copie di assi si possono notare delle configurazioni per cui le coordinate del secondo fattore non sono una combinazione lineare di quelle del primo; la presenza di questo effetto indica che il fenomeno è unidimensionale (cfr. Fabbris, 1983). Le forme tipiche dell'effetto Guttman sono quelle "a ferro di cavallo", in cui il secondo fattore è funzione quadratica del primo, e quella "a onde" (polinomi di grado superiore al secondo).

1.4 Criteri per l'interpretazione della soluzione

L'interpretazione del risultato è una delle fasi più importanti dell'analisi delle corrispondenze ed esistono vari criteri formali per valutare la bontà della soluzione e per interpretarla. Di seguito saranno passati in rassegna i vari metodi.

- ***Contributo assoluto***

  Il contributo assoluto indica l'importanza che la *i*-esima modalità riga della tabella di contingenza fa assumere al fattore o asse principale sul quale è rappresentata:

$$C_{h|k} = \frac{p_{h.} \cdot f_{kh}^2}{\Omega_k} \quad ,$$

  dove $p_{h.}$ è la frequenza relativa marginale di riga, f_{kh} è l'autovettore *h*-esimo sull'asse *k*-esimo e Ω_k è l'autovalore *k*-esimo.

La stessa cosa può essere fatta in riferimento alle modalità colonna:

$$C_{m|k} = \frac{p_{.m} \cdot f_{km}^2}{\Omega_k} \quad ,$$

dove $p_{.m}$ è la frequenza relativa marginale di colonna, f_{km} è l'autovettore m-esimo sull'asse k-esimo e Ω_k è l'autovalore k-esimo.

Al fine di interpretare gli assi saranno maggiormente coinvolte le modalità per le quali le contribuzioni assolute sono prevalenti sulle altre.

- ***Contributo relativo***

  Il contributo relativo fornisce una valutazione numerica sulla bontà di descrizione di un particolare asse da parte di una modalità, ovvero dà un'idea di quanto ben rappresentati siano i punti vettore relativamente alle corrispondenti ordinate. Una modalità non è ben rappresentata sul piano fattoriale delle corrispondenze quando il suo contributo relativo è basso (cfr. Bolasco, 1999).

  Se si considerano le modalità poste sulle righe, il contributo relativo è

$$C_{k|h} = \frac{f_{kh}^2}{\sum_k f_{kh}^2} \quad ,$$

  dove f_{kh} è l'autovettore h-esimo sull'asse k-esimo e la sommatoria si estende a tutti gli autovalori non banali e non nulli.

  Se invece si considerano le modalità colonna il contributo relativo è

$$C_{k|m} = \frac{f_{km}^2}{\sum_k f_{km}^2} \quad ,$$

  dove f_{km} è l'autovettore m-esimo sull'asse k-esimo e la sommatoria si estende a tutti gli autovalori non banali e non nulli.

- ***Ispezione della configurazione***

  Dall'osservazione del grafico ottenuto proiettando i punti delle modalità analizzate su un sistema di assi si possono trarre delle conclusioni riguardo l'interpretazione degli assi e sulla correlazione delle variabili.

L'origine degli assi è il baricentro della distribuzione di punti; quindi i punti più lontani dall'origine sono quelli correlati con il fattore e che concorrono a denominarlo. Se due modalità hanno coordinate con valori notevoli (sono cioè lontani dal baricentro) e stanno dalla stessa parte, significa che tendono ad essere direttamente associate; viceversa se due modalità hanno coordinate con valori elevati ma segno opposto, tendono ad essere inversamente associate (cfr. Fabbris, 1983).

In genere le rappresentazioni grafiche di insiemi di modalità si presentano come nuvole di punti concentrati nella parte centrale (origine del sistema di assi) e gradualmente meno dense man mano che ci si allontana dal centro. Esistono però delle configurazioni di particolare significato per l'interpretazione.

- **Ellissoide**

  La forma più comune è quella dell'ellissoide con l'asse maggiore nella direzione del fattore più importante e con quello minore nella direzione del fattore meno importante tra i due; la forma dell'ellissoide si affusola con l'aumentare del rapporto tra gli autovalori dei due assi sui quali i punti sono rappresentati.

- **Nuvole separate**

  Se si riscrive la matrice di frequenze ponendo vicine tra loro le M_1 entità della prima nuvola di punti e facendo seguire le M_2 della seconda ($M=M_1+M_2$) si ottiene una matrice ripartita in blocchi diagonali, uno $H_1 \times M_1$ di frequenze non nulle riguardanti la prima nuvola di punti, un altro $H_2 \times M_2$ riguardanti la seconda nuvola. Poiché i due sistemi sono indipendenti tra loro, si possono fare due analisi distinte, una per ciascun blocco.

- **Ferro di cavallo**

  Un andamento di questo tipo indica una sostanziale unidimensionalità delle frequenze osservate; il secondo fattore, cioè, è una riproduzione del primo e aggiunge solo alcune sfumature per l'interpretazione (effetto Guttman).

- **Triangolo o tetraedro**

  Una configurazione a triangolo si ha quando le modalità rappresentate rispetto al secondo fattore variano considerevolmente in corrispondenza

dei valori alti e poco sui valori negativi del primo fattore (o viceversa). Una configurazione a tetraedro si presenta quando lo stesso andamento triangolare si trova anche su un terzo fattore.

- ***Variabili supplementari***

  Alla fine dell'analisi si possono proiettare delle variabili (se quantitative) o di singole modalità (se qualitative) supplementari sugli assi ortogonali trovati per aggiungere delle sfumature di significato alla denominazione degli assi. Le modalità supplementari si rappresentano sugli assi allo stesso modo di quelle attive ma, allo scopo di agevolare la lettura dei risultati, conviene distinguerle (ad esempio riquadrandole); se la variabile supplementare è su scala ordinale è conveniente, invece, collegare le modalità per evidenziare eventuali relazioni tra la sequenza di modalità e gli assi trovati.

  Quando una variabile dal significato chiaro si colloca in una posizione lontana dal baricentro si ricava l'indicazione di una forte correlazione tra la variabile stessa e l'asse e questo contribuisce a far intuire il significato del fattore che l'asse rappresenta. La proiezione di modalità supplementari sugli assi trovati, invece, è un modo per scoprire interazioni di ordine superiore a quelle analizzate.

- ***Unità supplementari***

  Sulla configurazione possono essere rappresentati anche punti inerenti ad unità statistiche ignorate nella fase di ricerca della soluzione; questi punti si ottengono combinando i valori osservati presso queste unità con i coefficienti della soluzione ottenuta in base alle unità attive. Gli scopi di proiettare le unità supplementari sul sistema di assi sono due:

  - assegnare significati agli assi;
  - confrontare la posizione delle unità attive nell'analisi con quelle di altre unità.

- ***Impiego in sequenza dell'analisi delle corrispondenze e dell'analisi dei gruppi***

  L'analisi delle corrispondenze è una tecnica adatta a trattare insiemi di dati di numerosità notevole per quanto riguarda sia le unità statistiche, sia le variabili esaminate. L'analisi dei gruppi, invece, è un metodo proposto per compattare gli insiemi di unità o di variabili in pochi gruppi. Spesso è utile

applicare una procedura di calcolo dei gruppi dopo aver fatto una analisi delle corrispondenze al fine di aggiungere alla configurazione solo pochi punti ed ottenere una rappresentazione più essenziale.

- ***Ritorno all'indietro***

  Ricostruendo la matrice iniziale in base agli autovalori e agli autovettori della soluzione rappresentata è possibile eliminare le unità o le modalità che non hanno partecipato alla determinazione della soluzione.

Nel prossimo capitolo verranno presentati i dati e svolte le analisi preliminari sui dati, mentre nel capitolo 3 verrà applicata ai dati la tecnica appena descritta.

Capitolo 2

I dati e le analisi preliminari

Obiettivo di questo capitolo è descrivere i dati utilizzati nell'analisi e presentare le analisi esplorative preliminari operate. Gli scopi di questa analisi sono di verificare la stabilità nel tempo delle risposte, analizzare se il sesso dei bambini influenza il gradimento del professore e studiare la correlazione tra le variabili presenti nel questionario.

2.1 I dati

Come già detto nel primo capitolo, i due dataset (corrispondenti al primo e al secondo questionario) sono formati da 20 variabili, le prime 16 corrispondenti alle affermazioni dei questionari a cui sono stati sottoposti i bambini delle 5 scuole prese in esame, le altre quattro a caratteristiche degli intervistati e dei relativi professori di educazione fisica.

Le domande dei questionari prevedono come risposta un valore da 1 a 5 a seconda del grado di accordo/disaccordo con l'affermazione (scala Lickert): 1 = per nulla d'accordo, 2 = poco d'accordo, 3 = abbastanza d'accordo, 4 = molto d'accordo, 5 = pienamente d'accordo. Poiché la scala utilizzata è formata da un numero dispari di possibili risposte, è prevista l'opportunità di una risposta "neutrale" rispondendo con il valore centrale della scala; coloro che hanno risposto 3, quindi, sono i cosiddetti "neutri".

Si ricorda una sintetica descrizione delle variabili:

V1: Mi diverto;

V2: Mi annoio;

V3: Non mi piace;
V4: Lo trovo piacevole;
V5: Non mi diverto per niente;
V6: Mi da' energia;
V7: Mi fa sentire depresso;
V8: È molto piacevole;
V9: Il mio corpo si sente bene;
V10: Ottengo qualcosa;
V11: È molto eccitante;
V12: Mi da' frustrazione;
V13: Non è per niente interessante;
V14: Mi da una forte sensazione di successo;
V15: Mi fa sentire bene;
V16: Mi sento come preferissi fare qualcos'altro;
V17: frequenza pratica dell'attività (A = bassa, B = media, C = alta);
V18: sesso dello studente (M, F);
V19: scuola (in questa variabile sono sintetizzati la classe, la sezione e la scuola di appartenenza del bambino);
V20: docente (comprendente il sesso e l'anno di nascita del professore di educazione fisica).

2.2 Analisi preliminari

In questo paragrafo vengono presentati alcuni grafici e test statistici relativa alle prime analisi sui dati.

2.2.1 Stabilità nel tempo

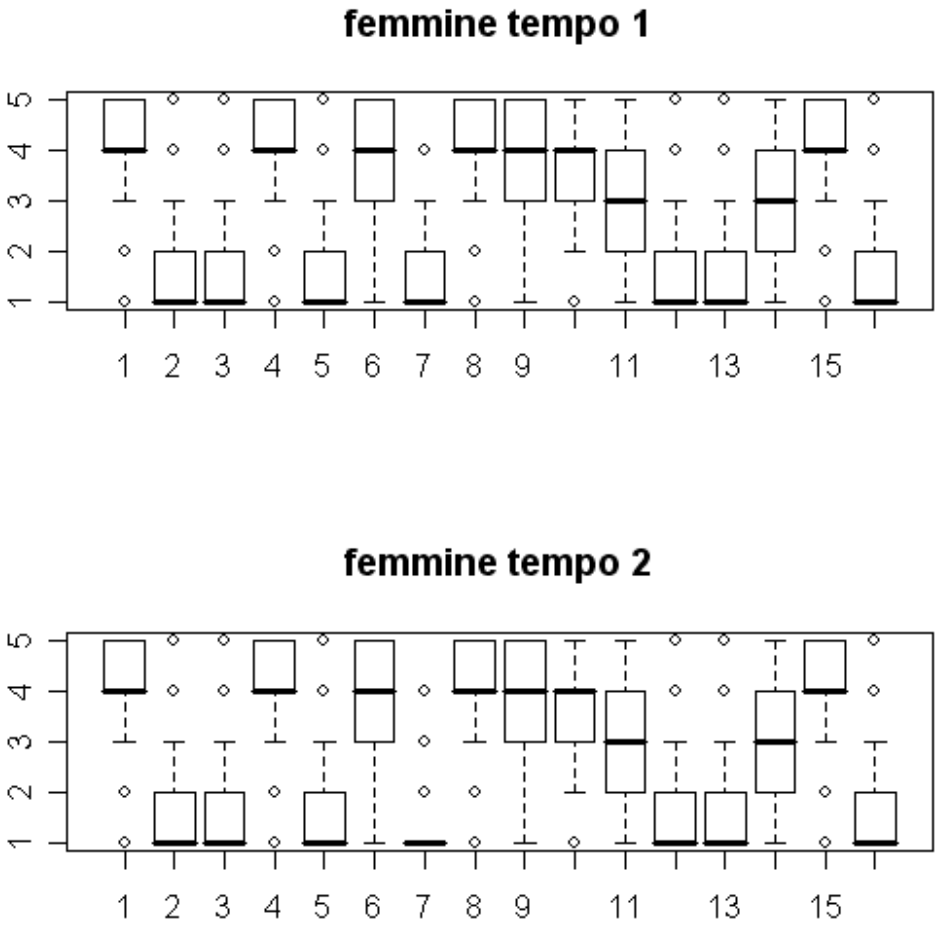
Come prima cosa è utile verificare la stabilità nel tempo delle risposte, cioè se le risposte date dai bambini nei due questionari sono rimaste pressoché identiche o se sono cambiate perché dettate dall'umore. Per far ciò si possono confrontare le risposte del primo questionario con quelle del secondo tramite dei box-plot (cioè dei grafici in cui vengono rappresentati la forma della distribuzione, il valore centrale e la variabilità) e il test di Wilcoxon, uno dei più potenti test non parametrici per verificare, in presenza di valori ordinali, se due campioni statistici provengono dalla

stessa popolazione (cfr. Ercolani, 2002). Queste analisi vengono svolte separatamente per sessi: prima si è svolta l'analisi sulle femmine e poi quella sui maschi.

2.2.1.1 Femmine

Per avere una prima idea sulla stabilità dei dati si sono considerati i box-plot di ciascuna variabile del primo e del secondo questionario in modo tale da poter confrontare le risposte delle corrispondenti domande (Grafico 1).

Grafico 1



Come si può notare, i box-plot delle varie risposte sono praticamente identici per tutte le domande tranne che per la domanda numero 7 in cui nel secondo questionario (tempo 2) le risposte sono molto più concentrate attorno al valore 1. Questi grafici portano a supporre che la stabilità nel tempo per le femmine sia verificata, ma per averne la certezza si è svolto un test di Wilcoxon per dati appaiati

sulle variabili (nella tabella seguente vengono riportati solo i valori dei test e i relativi *p-value*).

Tabella1: Stabilità nel tempo per le femmine

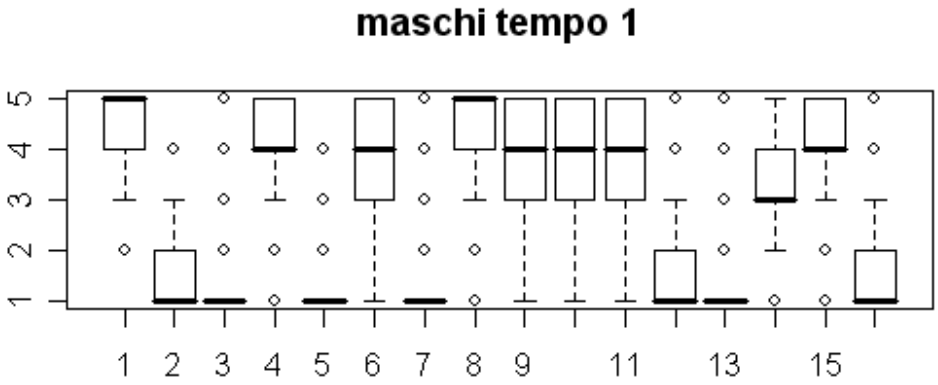
VARIABILE	VALORE TEST	*P-VALUE*
V1	V = 434.5	*0.03236*
V2	V = 335.5	0.7991
V3	V = 664.5	0.4030
V4	V = 619	*0.03299*
V5	V = 384.5	0.05356
V6	V = 1162	0.7767
V7	V = 474	0.1158
V8	V = 841.5	0.2208
V9	V = 1422	0.9886
V10	V = 822.5	0.1827
V11	V = 779	0.1554
V12	V = 972	0.977
V13	V = 453	0.205
V14	V = 839	0.1136
V15	V = 703	*0.04265*
V16	V = 956.5	0.888

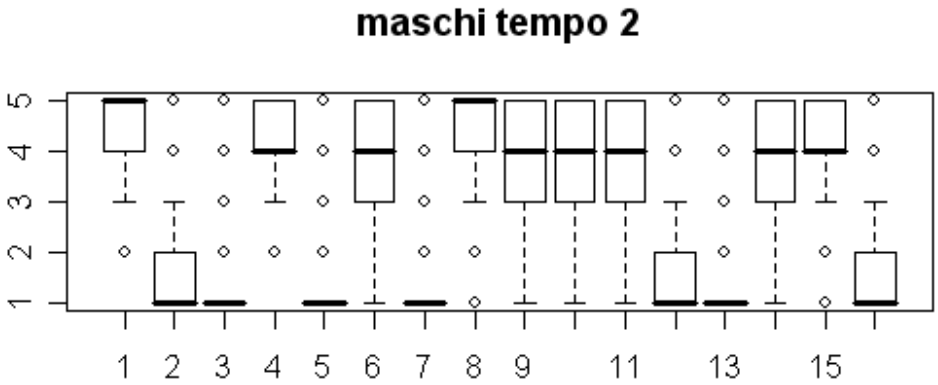
Prendendo come livello di significatività α=0.01, i *p-value* sono tutti maggiori di α, quindi si accetta l'ipotesi di stabilità nel tempo delle risposte per le femmine.

2.2.1.2 Maschi

Per i maschi è stata svolta la stessa analisi del sottoparagrafo precedente, cioè sono stati fatti i box-plot (Grafico 2) ed è stato calcolato il test di Wilcoxon per dati appaiati (Tabella 2).

Grafico 2:





Anche per i maschi è ipotizzabile che la stabilità nel tempo sia verificata dato che i box-plot sono molto simili tra loro, tranne che per la variabili V14 in cui le risposte nel secondo questionario hanno valori più alti (la mediana è 4 mentre nel primo questionario la mediana è 3).

La Tabella 2 riporta i valori e i *p-value* del test di Wilcoxon.

Tabella 2: Stabilità nel tempo per i maschi

VARIABILE	VALORE TEST	*P-VALUE*
V1	V = 696.5	0.3668
V2	V = 450	0.2868
V3	V = 417.5	0.2991
V4	V = 1092.5	0.9307
V5	V = 179	0.1544
V6	V = 1483	0.9166
V7	V = 258.5	0.4984
V8	V = 862	0.6735
V9	V = 1397.5	0.2277
V10	V = 1752.5	0.8818
V11	V = 2249	0.5218
V12	V = 867	0.2334
V13	V = 269.5	0.924
V14	V = 1163.5	*0.03803*
V15	V = 1393	0.803
V16	V = 1021	0.2573

Il test conferma che la variabile V14 non è molto stabile, ma prendendo come livello di significatività α=0.01, è possibile accettare l'ipotesi di stabilità nel tempo delle risposte dei maschi.

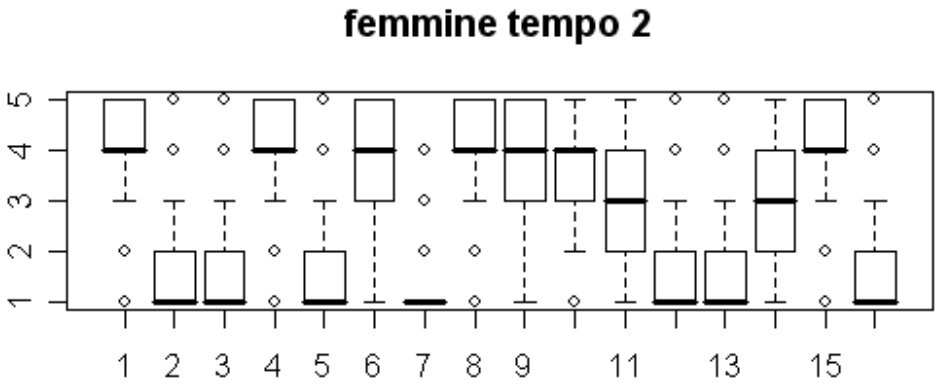
2.2.2 Confronto tra sessi

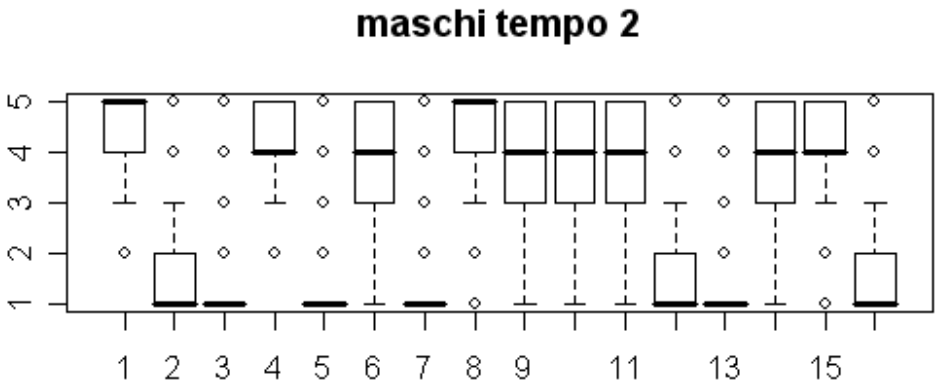
Nel paragrafo precedente è stato dimostrato che i dati sono stabili nel tempo, quindi è possibile proseguire le analisi scegliendo i dati di uno dei due questionari; qui è stato scelto il secondo poiché in questo modo i dati sono più recenti.

Un'altra analisi molto importante da considerare riguarda il confronto tra i maschi e le femmine per verificare se il sesso del bambino influenza le risposte e quindi il gradimento del professore.

Graficamente è possibile vedere se la distribuzione delle risposte è uguale per entrambi i sessi sia con i box-plot (Grafico 3), sia con i diagrammi a barre (Grafici 4 e 5).

Grafico 3:





Come si può notare dal grafico, per alcune variabili la distribuzione delle risposte è abbastanza diversa tra i due sessi.

Per maggiore chiarezza sono state riportate le misure di posizione nella Tabella 3.

Tabella 3:

		min	1° quartile	mediana	3° quartile	max
V1	M	2	4	5	5	5
	F	1	4	4	5	5
V2	M	1	1	1	2	5
	F	1	1	1	2	5
V3	M	1	1	1	1	5
	F	1	1	1	2	5
V4	M	2	4	4	5	5
	F	1	4	4	5	5
V5	M	1	1	1	1	5
	F	1	1	1	2	5
V6	M	1	3	4	5	5
	F	1	3	4	5	5
V7	M	1	1	1	1	5
	F	1	1	1	1	4
V8	M	1	4	5	5	5
	F	1	4	4	5	5
V9	M	1	3	4	5	5
	F	1	3	4	5	5
V10	M	1	3	4	5	5
	F	1	3	4	5	5
V11	M	1	3	4	5	5
	F	1	2	3	4	5
V12	M	1	1	1	2	5
	F	1	1	1	2	5
V13	M	1	1	1	1	5
	F	1	1	1	2	5
V14	M	1	3	4	5	5
	F	1	2	3	4	5
V15	M	1	4	4	5	5
	F	1	4	4	5	5
V16	M	1	1	1	2	5
	F	1	1	1	2	5

Questo può essere visto meglio con i grafici a barre, i quali mostrano in modo più chiaro la distribuzione di ciascuna variabile.

Grafico 4 (Femmine):

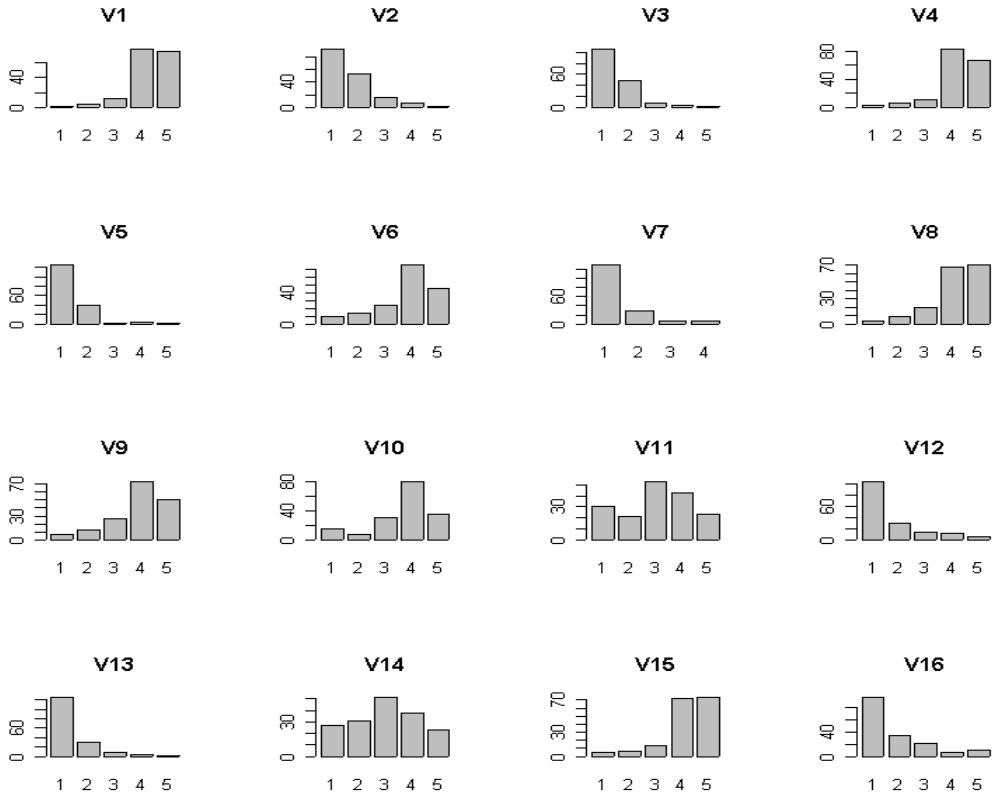
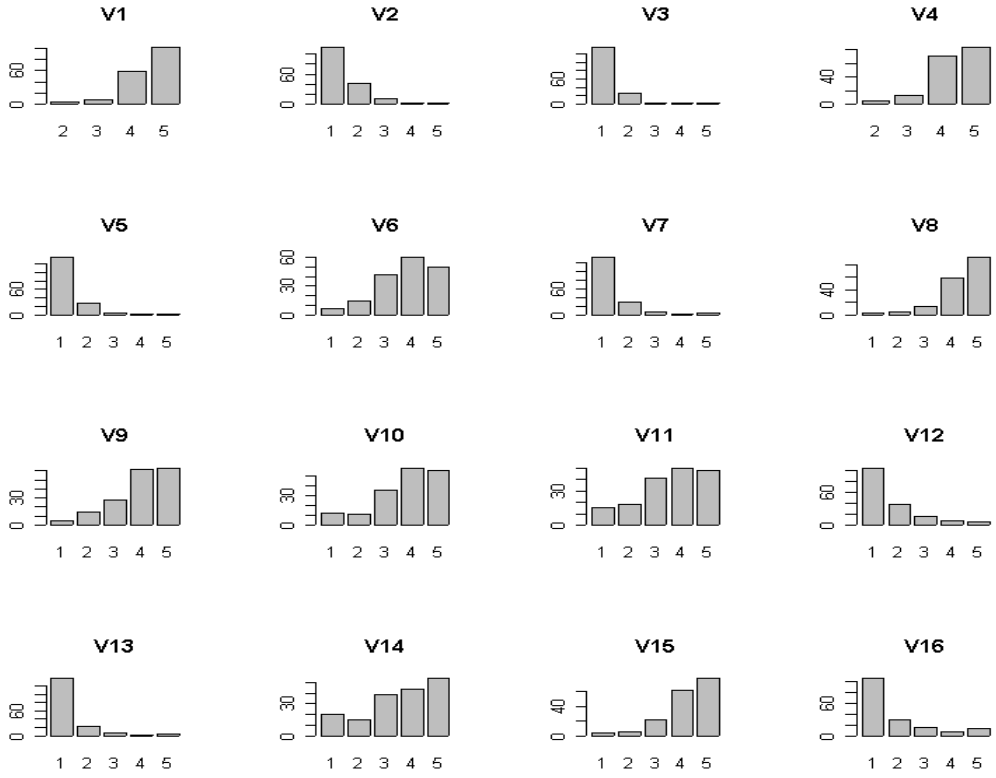


Grafico 5 (Maschi):



Sia dalla Tabella 3 che dai Grafici 4 e 5, si nota che le variabili che mostrano una distribuzione notevolmente diversa sono la V1, la V3 e la V14.
Analiticamente, invece, per verificare l'effetto del sesso del bambino, si usa il test di Wilcoxon, come nel paragrafo precedente, ma in questo caso in modo leggermente differente poiché si utilizzano dati provenienti da popolazioni diverse.
I risultati ottenuti sono sintetizzati nella Tabella 4.

Tabella 4:

VARIABILE	VALORE TEST	*P-VALUE*
V1	W = 12211	*0.00321*
V2	W = 16502	*0.0182*
V3	W = 17018	*0.0009955*
V4	W = 13120.5	0.07116
V5	W = 15257.5	0.3488
V6	W = 14950	0.7051
V7	W = 14657	0.957
V8	W = 12577.5	*0.01489*
V9	W = 13960	0.4466
V10	W = 13604.5	0.2439
V11	W = 11225.5	*0.0001356*
V12	W = 14620	1
V13	W = 15552	0.1661
V14	W = 10885	2.819^{-5}
V15	W = 14577	0.9598
V16	W = 15118	0.5384

Il test conferma le ipotesi sulla diversità di distribuzione delle variabili V1, V2, V3, V8, V11 e V14: infatti il *p-value* sono inferiori al livello di significatività α=0.05.

2.2.3 Analisi della correlazione tra variabili

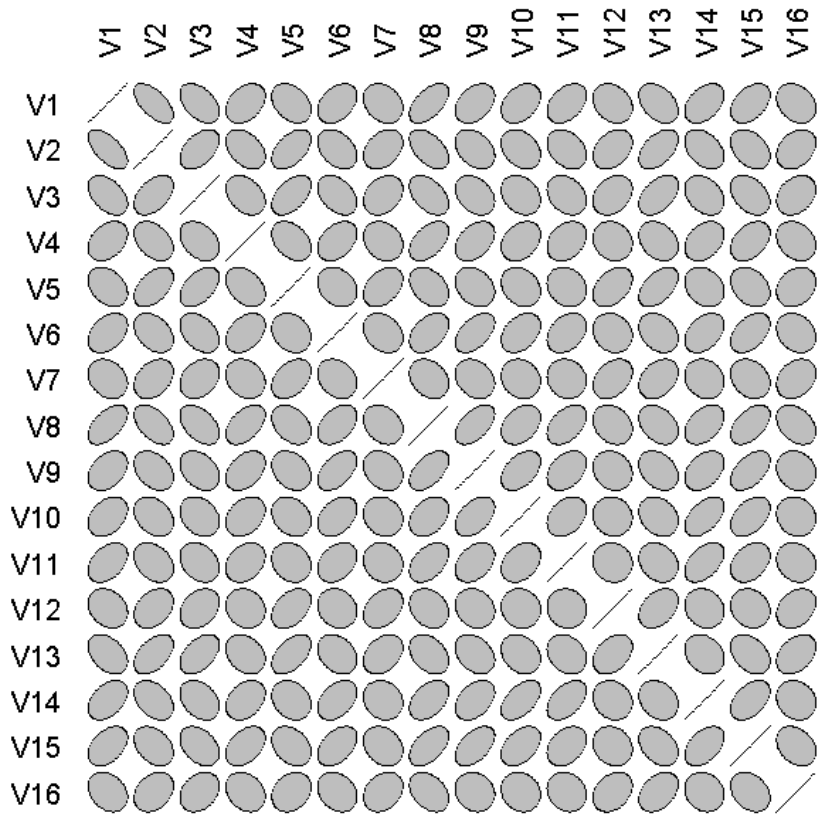
Per analizzare se le variabili prese in esame sono correlate fra loro si è utilizzato il test di correlazione di Spearman. L'indice di correlazione per ranghi di Spearman è una misura statistica non parametrica della correlazione e misura pertanto il grado

di relazione tra due variabili per le quali non si fa altra ipotesi che non la misura ordinale ma possibilmente continua (cfr. Ercolani, 2001).

I risultati di questa analisi (riportati nell'Appendice 1) fanno concludere che le variabili sono tutte correlate tra loro in quanto i *p-value* dei test sono tutti uguali o molto vicini a zero e quindi l'ipotesi nulla di incorrelazione viene rifiutata.

Graficamente le correlazioni tra le variabili possono essere rappresentate con un *plotcorr*, un grafico in cui la correlazione tra le variabili viene descritta con delle ellissi. Se l'ellisse è inclinata verso destra significa che c'è una correlazione positiva tra le variabili, mentre se è inclinata verso sinistra significa che sono inversamente correlate; le dimensioni dell'ellisse, invece, mostrano l'intensità della correlazione.

Grafico 6:



Dal grafico risaltano particolarmente alcune coppie di variabili, in particolare la V8/V9 e la V11/V14. Guardando la tabella con i *p-value* del test di correlazione di Spearman (Appendice 1), si ha la conferma della forte correlazione tra queste coppie di variabili, in quanto i *p-value* sono uguali a zero

Capitolo 3

Analisi dei dati

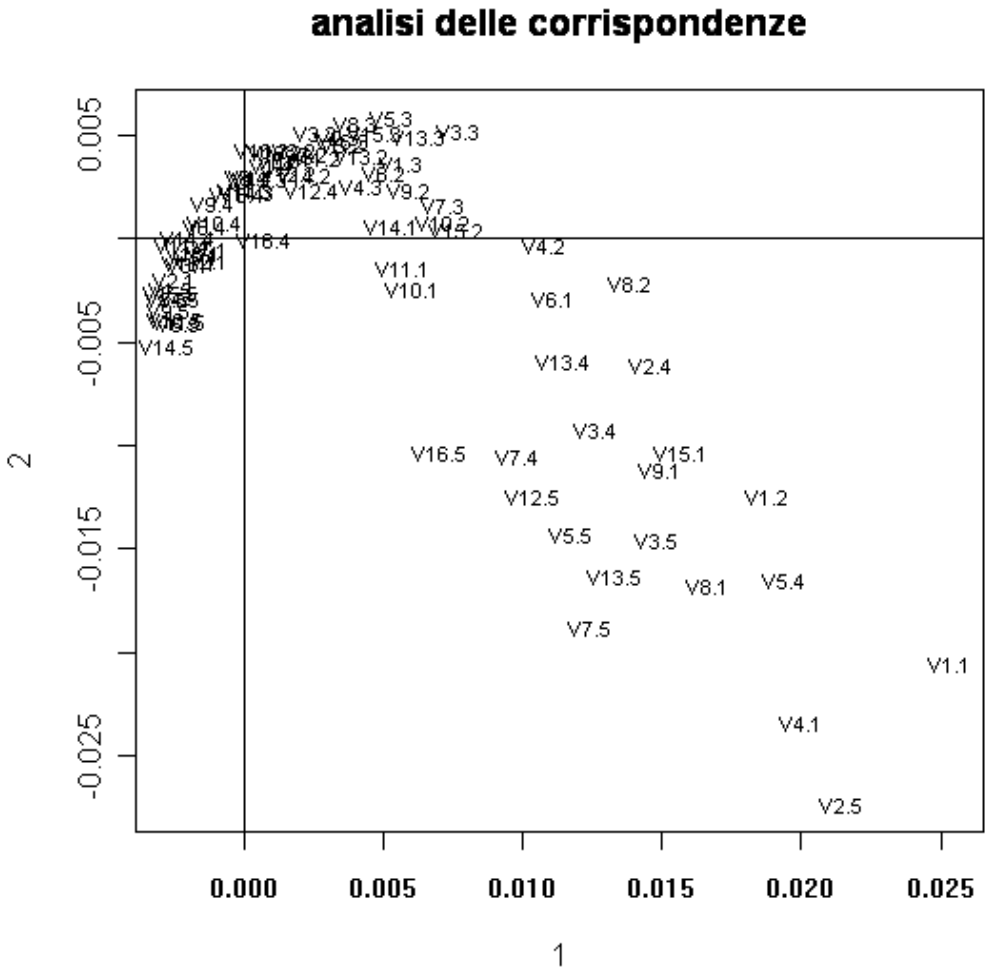
In questo capitolo verrà svolta l'analisi vera e propria attraverso l'analisi delle corrispondenze. Alla fine è stata svolta anche un'analisi cluster in modo tale da compattare l'insieme di unità in pochi gruppi ed ottenere una rappresentazione più essenziale.

3.1 Analisi delle corrispondenze

La prima scelta da fare per applicare l'analisi delle corrispondenze ai dati è la distinzione tra variabili attive e variabili supplementari. In questo caso sono state scelte le prime sedici variabili (V1-V16), corrispondenti alle domande del questionario, come variabili attive e le restanti 4 variabili, corrispondenti alle caratteristiche degli intervistati e dei professori, come variabili illustrative.
In seguito è stato scelto il numero di fattori da utilizzare nell'analisi; poiché le variabili non sono molto numerose è stato deciso di utilizzare solo due fattori cosi da poter rappresentare graficamente i risultati delle analisi.

Il grafico che si ottiene applicando l'analisi delle corrispondenze al dataset è il seguente:

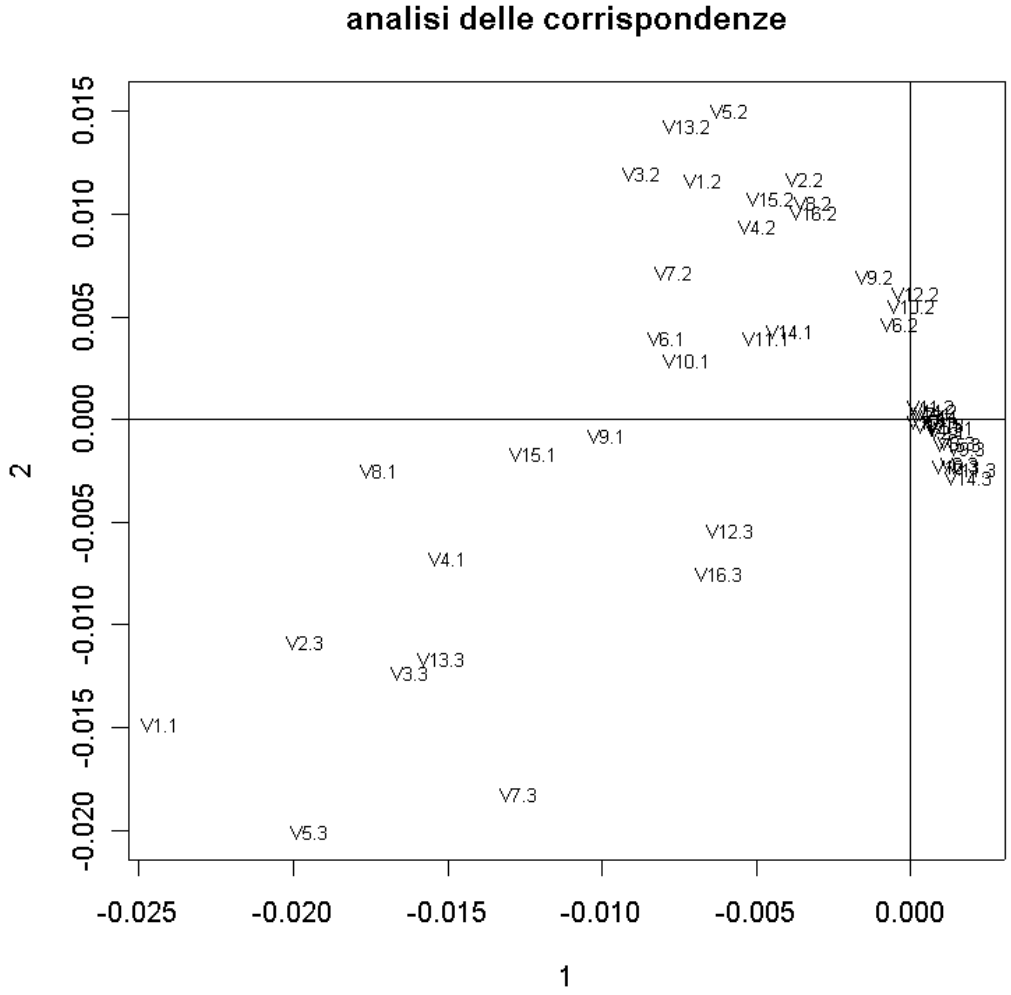
Grafico 7:



Come si può notare, i punti si distribuiscono nel piano secondo la cosiddetta forma a "ferro di cavallo"; questo implica che è presente l'effetto Guttman. Questo effetto è tipico delle Scale Lickert e per cercare di eliminarlo sono state ricodificate le variabili in tre classi per diminuire il numero di modalità: per nulla d'accordo e poco d'accordo sono state unite in un'unica classe, abbastanza d'accordo è rimasta uguale, molto d'accordo e pienamente d'accordo sono state unite in un'altra classe.

Il risultato ottenuto con questa ricodifica è il seguente:

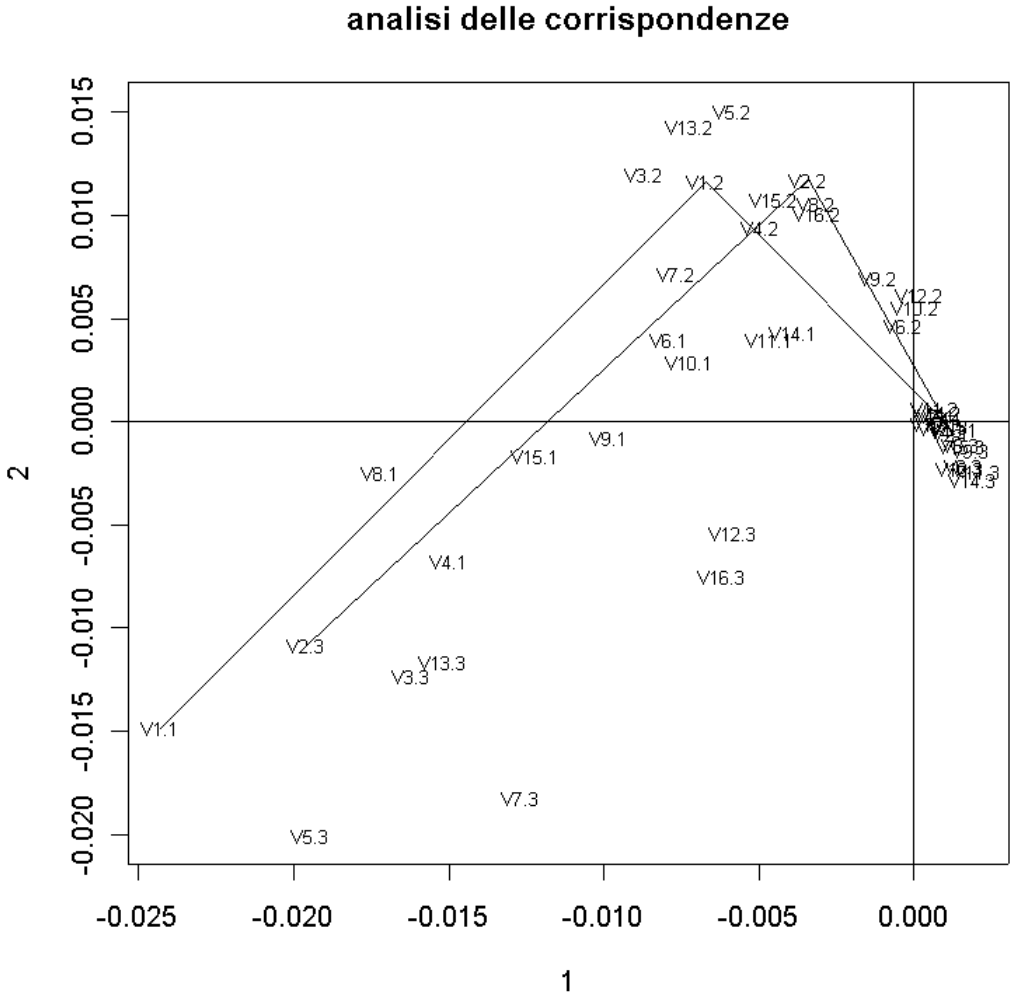
Grafico 8:



Adesso l'effetto Guttman è diminuito sensibilmente ed è ora visibile una distribuzione a triangolo dei punti modalità sugli assi fattoriali. Una configurazione di questo tipo indica che le modalità rappresentate rispetto al secondo fattore variano considerevolmente in corrispondenza dei valori negativi e poco sui valori positivi vicini all'origine.
Nell'Appendice 3 sono riportate le coordinate dei punti modalità.

Poiché in questo caso le variabili sono ordinabili, è possibile unire i punti modalità delle variabili con una spezzata per interpretare meglio gli assi fattoriali (Grafico 9). Per non appesantire troppo il grafico sono state unite solo le modalità delle prime due variabili (V1 e V2), in ogni modo l'andamento delle spezzate delle altre variabili è molto simile a quello delle prime due.

Grafico 9:



Come si può notare dal grafico, il primo fattore (l'asse orizzontale) rappresenta la scala di misura delle variabili. Infatti a destra sono concentrate le modalità che corrispondono alle risposte positive sulle esperienze di educazione fisica (corrispondenti alle modalità 1 se l'affermazione del questionario era negativa, per esempio "Mi annoio", alla modalità 3 se invece l'affermazione era positiva, per esempio "Mi diverto"), in mezzo sono riuniti i cosiddetti "neutri" (corrispondenti alla modalità 2), mentre a sinistra sono raggruppate quelle negative.

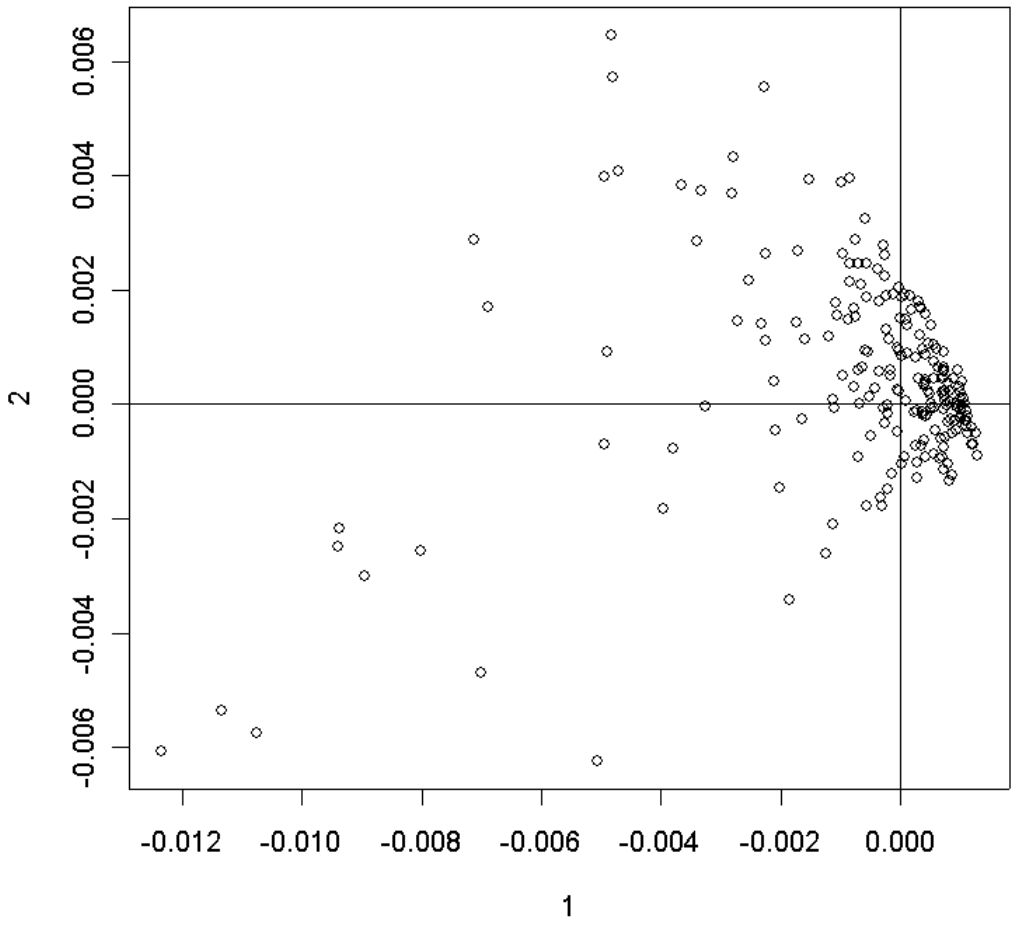
Il secondo fattore (asse verticale) è di più difficile interpretazione: in alto dominano le variabili di tipo fisico (per esempio la V6: mi da energia), mentre in basso prevalgono le variabili di tipo emotivo (per esempio la V7: mi fa sentire depresso).

Il grafico evidenzia inoltre tre gruppi di modalità: quelle in basso a sinistra che sono quelle "negative" e che rappresentano gli scontenti, quelle al centro verso destra che

sono quelle positive e che rappresentano gli individui contenti dell'attività fisica e quelle in alto che rappresentano gli "indecisi", cioè quelli che hanno risposto con il valore centrale della scala Likert.

Per vedere la numerosità di ciascuno di questi gruppi è stato fatto un grafico proiettando le unità statistiche (di numerosità 342) nel piano fattoriale (Grafico 10).

Grafico 10:



Il grafico mostra che le unità statistiche sono concentrare leggermente a destra dell'origine degli assi, cioè dove sono raggruppate le modalità positive delle variabili.

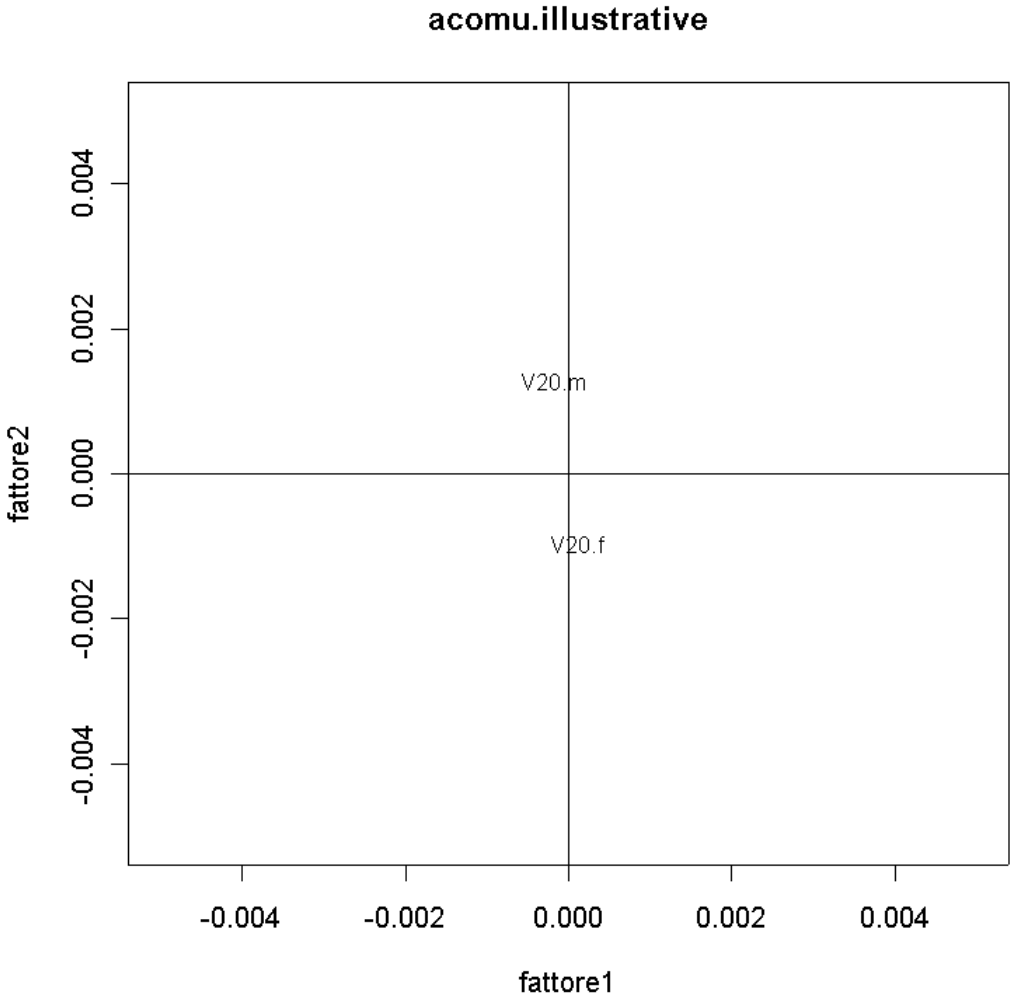
Ciò dimostra che i bambini che hanno dato valori alti alle domande, e che quindi sono soddisfatti dell'attività fisica svolta, sono molto numerosi, mentre quelli che hanno dato valori bassi sono molto pochi.

3.2 Variabili illustrative

Dopo aver svolto l'analisi delle corrispondenze è utile proiettare sugli assi le variabili non utilizzate nell'analisi, cioè le cosidette variabili illustrative o supplementari, al fine di una migliore interpretazione degli assi.

Come prima cosa è stata proiettata la variabile V20, cioè quella riguardante il sesso del professore di educazione fisica (Grafico 11).

Grafico 11:

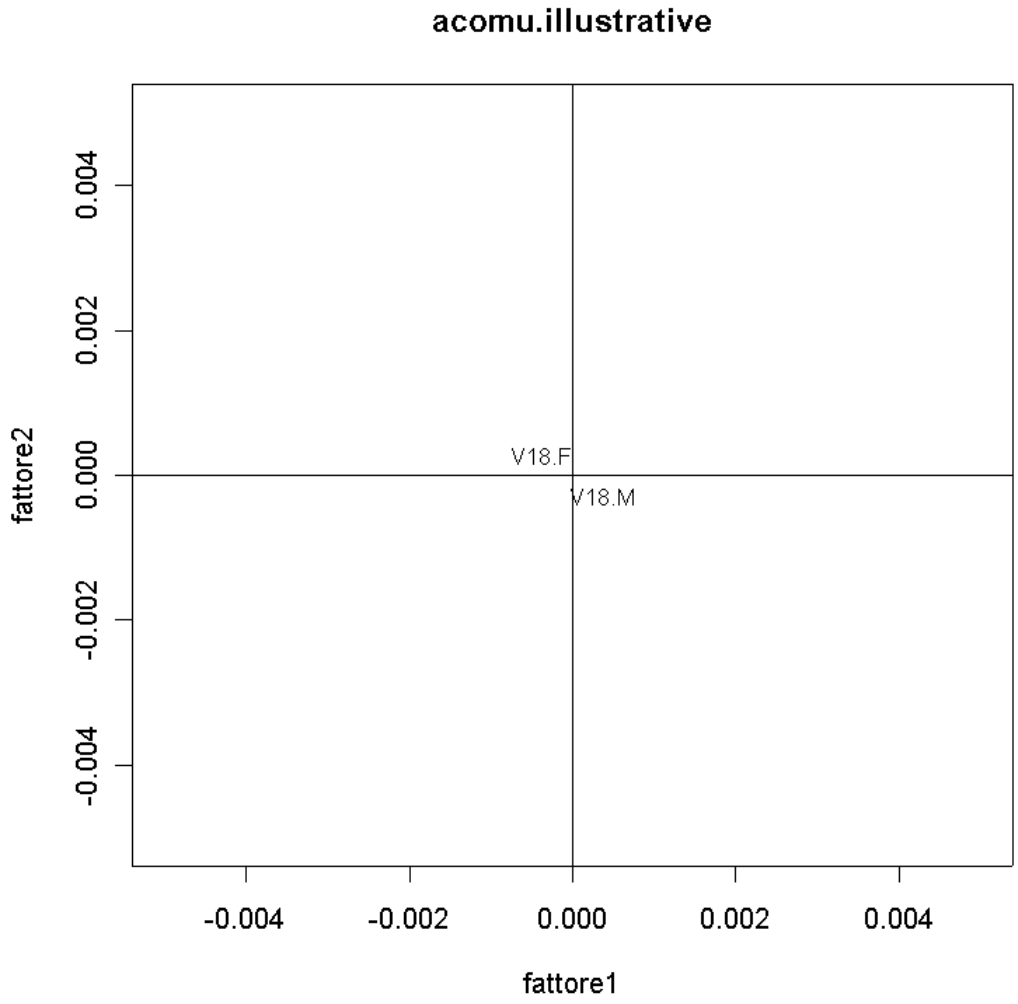


Il grafico ci mostra che il sesso del professore non influenza particolarmente la soddisfazione dei bambini in quanto i punti modalità sono molto vicini all'origine degli assi fattoriali. Si può notare, però, che la modalità m (corrispondente al sesso maschile del professore) è leggermente spostata verso il punto in cui sono addensate le variabili di tipo "fisico", mentre la modalità f (corrispondente al sesso femminile dell'insegnante) è leggermente spostata verso il gruppo di variabili di tipo

"emotivo". Questo fa supporre che gli insegnanti di sesso femminile curino di più gli aspetti emozionali della materia mentre quelli di sesso maschile curino di più gli aspetti fisici della materia.

In seguito è stata proiettata la variabile V18, cioè il sesso dello studente (Grafico 12).

Grafico 12:

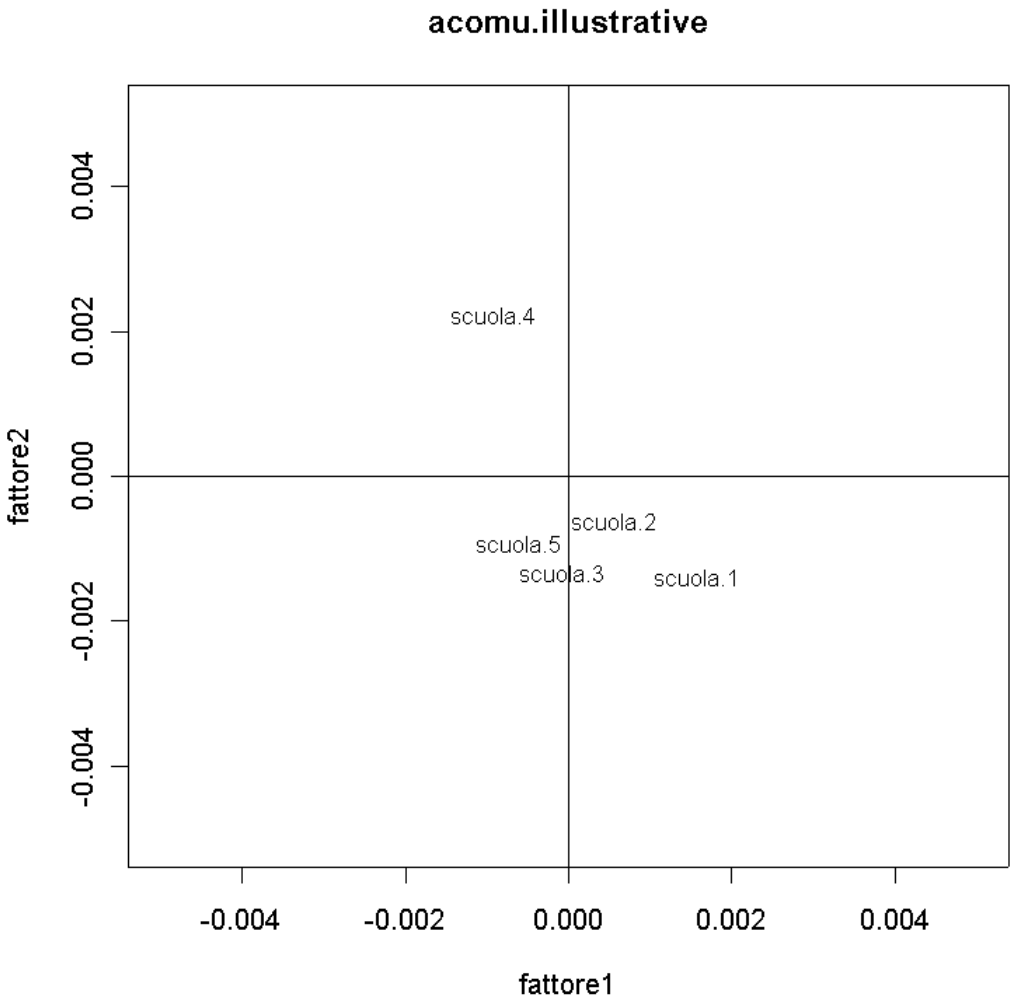


Anche in questo caso il sesso non influenza il gradimento del professore. Questo contrasta con le analisi fatte in precedenza in quanto nel confronto tra i sessi dei bambini (paragrafo 2.2.2) si era notata una differenza tra la distribuzione delle risposte dei maschi e la distribuzione delle risposte delle femmine.

È possibile notare, tuttavia, che la modalità f (corrispondente al sesso femminile degli studenti) è leggermente spostata verso il gruppo di bambini definiti "neutri", mentre la modalità m (corrispondente al sesso maschile degli studenti) si trova sotto l'asse orizzontale, cioè in direzione del gruppo di bambini contenti dell'attività fisica svolta.

Poi sono stati proiettati i punti modalità relativi alla variabile V19, cioè la scuola frequentata (Grafico 13). Prima di far ciò, però, questa variabile è stata riclassificata in modo tale da eliminare la classe e la sezione, in modo tale da discriminare i bambini solo tramite la scuola di appartenenza.

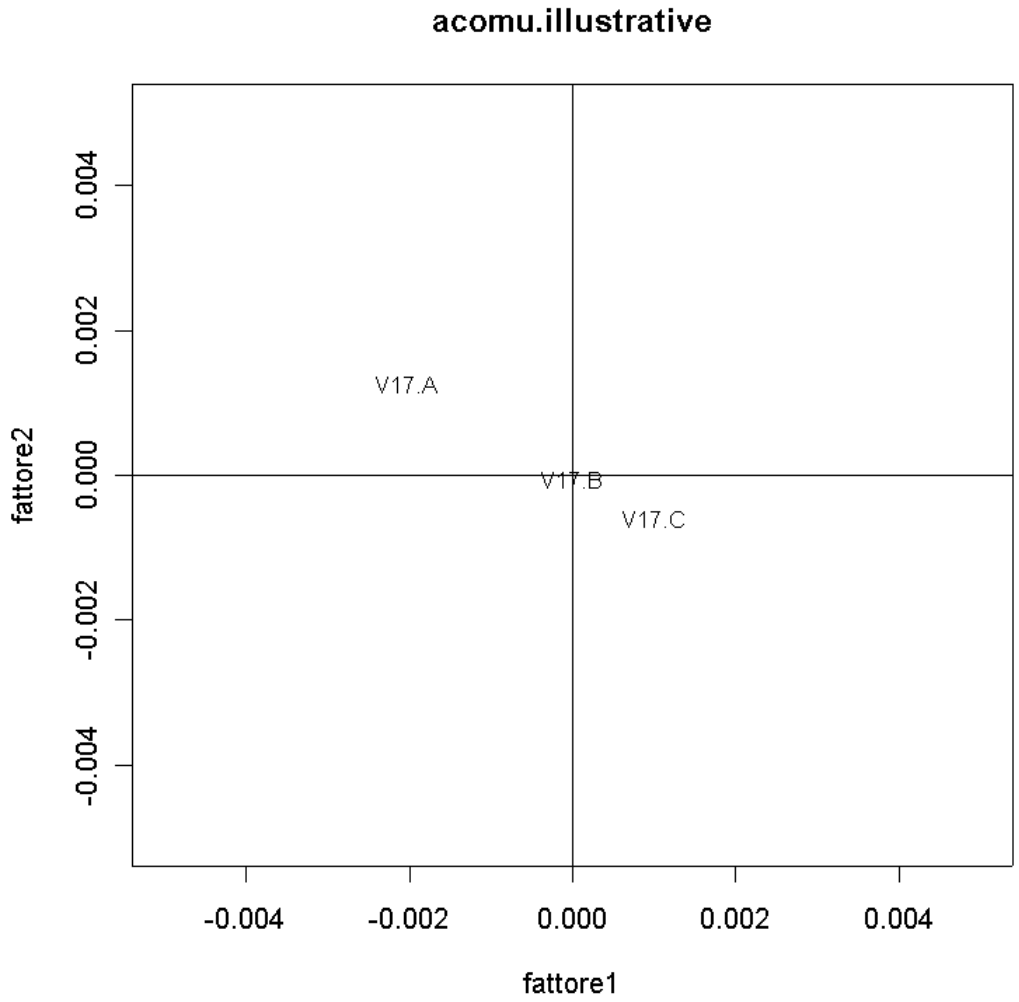
Grafico 13:



Anche in questo grafico le modalità sono concentrate tutte vicino all'origine. Fa eccezione la modalità 4 (corrispondente alla scuola di Camisano) che si allontana leggermente dall'origine verso le modalità di tipo "fisico". Questo fatto è probabilmente dovuto al fatto che i professori di questa scuola sono tutti di sesso maschile e quindi, come detto in precedenza, tendono a privilegiare gli aspetti "fisici" dell'insegnamento.

Infine è stata proiettata la variabile V17 corrispondente alla frequenza pratica dell'attività fisica (Grafico 14).

Grafico 14:



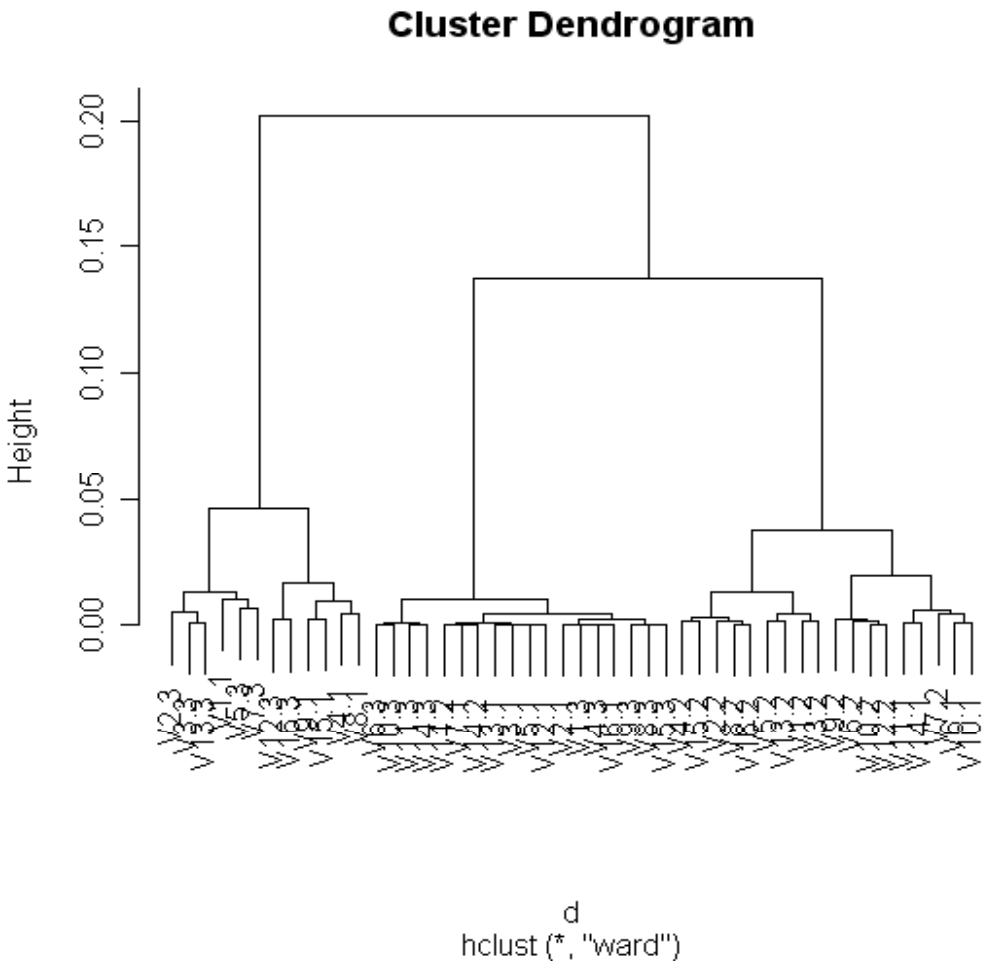
Come nei casi precedenti, anche questa variabile illustrativa non dà molte informazioni per l'interpretazione degli assi. La modalità A (che rappresenta la

bassa frequenza pratica), però, si scosta leggermente dall'origine in direzione delle modalità che rappresentano gli individui che hanno risposto con il valore centrale della scala Likert. Questo risultato è abbastanza ragionevole in quanto i bambini che fanno poca attività fisica hanno più difficoltà a rispondere a delle domande sul gradimento del professore di educazione fisica rispetto a quelli che fanno molta attività fisica.

3.3 Analisi cluster

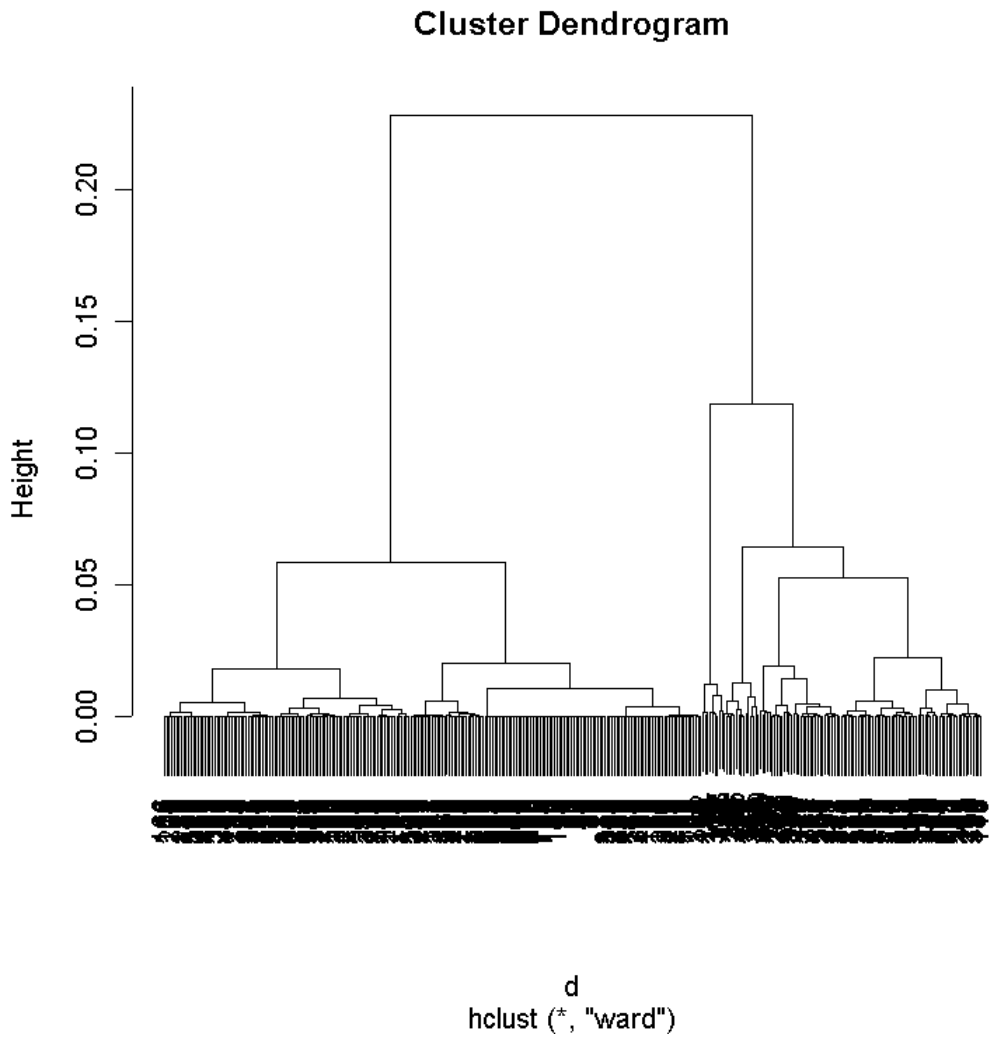
Come già anticipato nel secondo capitolo, spesso è utile accompagnare l'analisi delle corrispondenze con un'analisi dei gruppi in modo tale da compattare l'insieme di unità in pochi gruppi ed ottenere una rappresentazione più essenziale.
I risultati ottenuti applicando l'analisi cluster ai punteggi fattoriali ottenuti con l'analisi delle corrispondenze sono rappresentati nel Grafico 15.

Grafico 15:



Guardando il dendrogramma si nota che è possibile ridurre le modalità in soli tre gruppi. In questo modo si ottengono 3 cluster omogenei al loro interno rispetto alla soddisfazione del professore di educazione fisica, che corrispondono ai tre gruppi emersi dal grafico dell'analisi delle corrispondenze.
È possibile inoltre applicare l'analisi cluster alle modalità riga, invece che alle modalità colonna, cioè ai singoli individui. Il dendrogramma che risulta in questo caso è quello rappresentato nel Grafico 16.

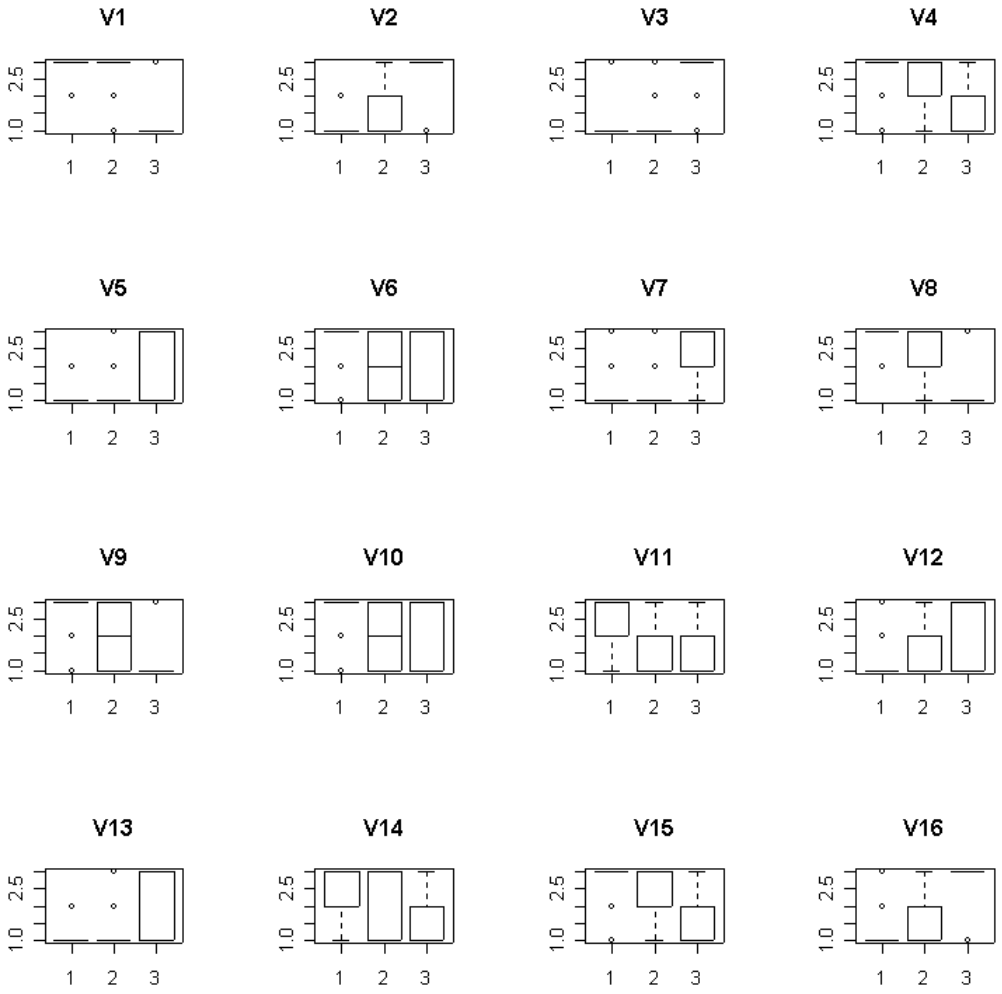
Grafico 16:



Tagliando il dendrogramma all'altezza 0.07, cioè all'altezza del massimo salto tra livelli di somiglianza, si ottengono tre gruppi, il primo di numerosità 225, il secondo di numerosità 108 ed il terzo di numerosità 9.

Per caratterizzare meglio questi gruppi sono stati fatti i box-plot (Grafico 17) di ciascuna variabile rispetto ai tre gruppi trovati con l'analisi cluster.

Grafico 17:



Come si può notare, il primo gruppo è quello dei soddisfatti dell'attività fisica svolta, in quanto la hanno risposto con valori alti alle domande "positive" (per esempio la V1: mi diverto), mentre hanno risposto con valori bassi alle risposte "negative" (per esempio la V2: mi annoio). Un'altra cosa che si vede da questi box-plot è che la variabilità di questo gruppo è molto bassa (i tre quartili coincidono), come era gia stato visto nel grafico delle corrispondenze.

Il secondo gruppo ha una variabilità maggiore rispetto al primo, ma le risposte sono abbastanza concordanti con quelle del primo gruppo. Questi sono i cosiddetti "neutri", cioè quelle persone che sostanzialmente sono soddisfatte dell'attività svolta, ma che rispondono con maggiore variabilità alle domande del questionario.

Il terzo gruppo, infine, ha una variabilità ancora maggiore ma le risposte sono l'opposto di quelle date dai componenti del primo gruppo: questi sono gli "insoddisfatti".

Confrontando i box-plot con la numerosità dei gruppi, si nota che gli scontenti dell'attività svolta sono molto pochi (3%) rispetto a quelli pienamente contenti (65%) e quelli essenzialmente contenti (32%).

Per vedere la composizione di ciascuno dei tre gruppi, sono state fatte delle tabelle di frequenza delle distribuzioni tra i gruppi delle variabili esplicative riportate di seguito.

Tabella 5: Variabile V17 (frequenza pratica)

	GRUPPI		
	1	2	3
A	31 (48%)	30 (46%)	4 (6%)
B	96 (66%)	45 (31%)	4 (3%)
C	98 (74%)	33 (25%)	1 (1%)

Tabella 6: Variabile V18 (sesso del bambino)

	GRUPPI		
	1	2	3
F	110 (65%)	54 (32%)	6 (3%)
M	115 (67%)	54 (31%)	3 (2%)

Tabella 7: Variabile V19 (scuola)

	GRUPPI		
	1	2	3
ARCUGNANO	50 (80%)	11 (18%)	1 (2%)
TORRI	47 (67%)	22 (31%)	1 (2%)
MONTICELLO	26 (63%)	14 (34%)	1 (3%)
CAMISANO	61 (57%)	43 (40%)	3 (3%)
SCAMOZZI	41 (66%)	18 (29%)	3 (5%)

Tabella 8: Variabile V20 (sesso professore)

	GRUPPI		
	1	2	3
F	130 (66%)	60 (31%)	6 (4%)
M	95 (65%)	48 (33%)	3 (2%)

Dalla Tabella 5 si nota che chi pratica più frequentemente attività fisica (modalità C) è più contento delle attività svolte, mentre chi fa poca educazione fisica (modalità A) è meno contento.

Dalla Tabella 6, invece, si vede che i maschi sono leggermente più soddisfatti dell'attività fisica rispetto alle femmine.

La Tabella 7 mostra che i bambini della scuola di Arcugnano sono i più contenti dell'educazione fisica svolta, mentre quelli della scuola di Scamozzi sono i meno contenti.

La Tabella 8, infine, fa vedere che il sesso del professore non influenza la soddisfazione, lo stesso risultato era emerso proiettando questa variabile nel grafico delle corrispondenze.

Conclusioni

Lo scopo dello studio è valutare il gradimento del professore di educazione fisica da parte degli allievi di alcune scuole elementari del Veneto.
I dati analizzati si riferiscono a due questionari con scala di Likert a 5 gradi cui sono stati sottoposti i bambini.
Questo studio mette in evidenza che il gradimento dei professori di educazione fisica è in generale molto buono, in quanto nell'analisi delle corrispondenze i punti modalità con le risposte positive sul gradimento dei professori sono concentrati nell'origine degli assi ed in questo punto è concentrata anche la maggioranza delle unità statistiche. L'analisi delle corrispondenze mostra inoltre che ci sono tre gruppi di individui: uno molto numeroso che gradisce notevolmente l'attività fisica svolta, un secondo gruppo leggermente meno numeroso che è fondamentalmente contento anche se in maniera inferiore rispetto al primo e un terzo molto esiguo che non è soddisfatto dell'educazione fisica praticata.
I risultati ottenuti fanno vedere, ancora, che il sesso del bambino, quello del professore, la frequenza dell'attività pratica e la scuola di appartenenza non influenzano in modo sostanziale il gradimento del professore.
L'analisi cluster effettuata, infine, porta agli stessi risultati ottenuti con l'analisi delle corrispondenze, ma specifica meglio la numerosità dei tre gruppi trovati e mostra le caratteristiche degli individui in ciascun gruppo; in particolare questa analisi dice che chi pratica più attività fisica è anche più contento dell'insegnante.

Riferimenti bibliografici

Fabbris L. (1983), *Analisi esplorativa di dati multidimensionali*, Cleup, Padova.

Bolasco S. (1999), *Analisi multidimensionale dei dati*, Carocci, Roma.

Ercolani A., Areni A., Leone L. (2001), *Statistica per la psicologia: i fondamenti di psicometria e statistica descrittiva*, Il Mulino, Bologna.

Ercolani A., Areni A., Leone L. (2002), *Statistica per la psicologia: statistica inferenziale e analisi dei dati*, Il Mulino, Bologna.

Brunoro G. (1994), *Analisi delle corrispondenze*, Cedam, Padova.

Bortot P., Ventura L., Salvan A, (2000), *Inferenza statistica: applicazioni con S-Plus e R*, Cedam, Padova.

Appendice 1

i / j	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16
V1	-	0	0	4.22^{-11}	5.09^{-11}	2.29^{-13}	2.21^{-6}	0	2.80^{-12}	2.54^{-11}	5.35^{-15}	4.76^{-6}	1.99^{-14}	0	0	1.55^{-7}
V2		-	0	5.08^{-15}	0	9.62^{-16}	6.42^{-11}	0	8.04^{-12}	7.60^{-9}	6.31^{-10}	2.55^{-10}	0	4.26^{-14}	3.15^{-12}	7.65^{-9}
V3			-	4.03^{-11}	0	1.45^{-14}	9.68^{-8}	0	1.22^{-10}	5.35^{-7}	1.18^{-7}	2.17^{-8}	0	4.41^{-10}	5.93^{-13}	3.63^{-8}
V4				-	1.89^{-12}	3.75^{-13}	6.47^{-5}	0	1.15^{-14}	4.79^{-9}	2.45^{-8}	0.005	8.88^{-8}	1.26^{-9}	1.54^{-14}	1.94^{-6}
V5					-	5.66^{-6}	1.42^{-13}	2.47^{-12}	7.40^{-8}	4.09^{-6}	0.0007	5.69^{-15}	0	3.67^{-6}	1.29^{-9}	6.18^{-6}
V6						-	7.11^{-7}	0	0	3.12^{-12}	2.46^{-10}	0.003	3.38^{-8}	4.25^{-14}	0	8.27^{-6}
V7							-	6.02^{-7}	2.01^{-6}	0.006	0.03	6.55^{-11}	1.39^{-10}	0.005	4.85^{-6}	2.10^{-7}
V8								-	0	3.99^{-11}	0	8.42^{-6}	1.85^{-11}	0	0	1.01^{-11}
V9									-	0	8.83^{-13}	0.001	8.38^{-10}	1.01^{-13}	0	0.005
V10										-	3.99^{-10}	0.09	2.11^{-6}	0	0	0.01
V11											-	0.15	5.99^{-6}	0	1.19^{-10}	0.001
V12												-	1.25^{-12}	0.002	0.001	2.34^{-8}
V13													-	2.65^{-6}	4.82^{-12}	8.59^{-13}
V14														-	0	0.004
V15															-	6.03^{-9}
V16																-

Appendice 2

In questa sezione vengono riportati i comandi del software R per l'analisi delle corrispondenze e per l'analisi cluster.

La funzione per ottenere l'analisi delle corrispondenze si trova all'interno della libreria *MASS*, quindi è necessario caricarla con il comando:
>library(MASS) .

Le variabili che utilizzo per fare l'analisi sono qualitative quindi è necessario fattorizzarle tramite l'istruzione:
>dataset$nomevariabile<-factor(dataset$nomevariabile) .

Fatto ciò si può cominciare a fare l'analisi delle corrispondenze vera e propria, scegliendo le variabili attive e formando un dataframe (qui chiamato dati) con tali variabili:
>dati<-data.frame(variabili) ,
dove tra parentesi vanno inseriti i nomi delle variabili attive separate da una virgola (in questo caso le variabili da V1 a V16).

In seguito, per calcolare l'analisi delle corrispondenze, utilizzo il comando:
> acomu<-mca(dati,nf=x,abbrev=FALSE) ,
dove dati è il dataframe appena creato, x è il numero di fattori scelto per l'analisi.

Per rappresentare graficamente i risultati ottenuti, sono necessarie due istruzioni:
> plot(acomu$cs,type="n")
> text(acomu$cs,rownames(acomu$cs),cex=0.7)
dove cs è la matrice delle coordinate dei punti modalità sui fattori.

Poiché in questo caso le variabili sono ordinabili è possibile unire i punti modalità delle variabili con una spezzata per interpretare meglio gli assi fattoriali. Ciò e possibile con il comando:
> lines(acomu$cs[i:j,z1],acomu$cs[i:j,z2]) ,

in cui i e j sono le righe della matrice cs relative alla variabile che prendo in esame, mentre z1 e z2 sono le colonne della matrice cs.

Per proiettare sugli assi le variabili non utilizzate nell'analisi vera e propria, cioè le cosidette variabili illustrative o supplementari, si utilizzano le seguenti istruzioni:

```
> plot(acomu$cs, type="n")
> illus<-predict(acomu,newdata=as.data.frame(var illustrativa),type="factor")
> text(illus,rownames(illus),cex=0.8)
```

dove nel secondo comando bisogna mettere tra parentesi il nome della variabile illustrativa che devo proiettare.

Per l'analisi cluster, invece, è stato scelto di svolgere un'analisi gerarchica e quindi è stata utilizzata la funzione **hclust**. Questa funzione però richiede in input una matrice di distanze che deve essere calcolata, a partire dal dataframe, con il comando **dist**.

```
>d<-dist(dati, method="euclidean")
>clu<-hclust(d, method="ward")  .
```

In questo caso sono stati scelti la distanza euclidea, definita da

$$d(x_{i.}, x_{j.}) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \ ,$$

anche se le variabili sono categoriali, perchè le modalità sono state trasformate in numeriche (1-5).

Per applicare l'algoritmo per raggruppare i dati è stato scelto il metodo di *Ward* (metodo che prevede che i gruppi vengano aggregati in modo che l'incremento di varianza nei nuovi gruppi sia più piccolo possibile.

Per diminuire il numero di gruppi si è utilizzato il comando:

```
>cutree(clu,k=n)
```

dove n è il numero di gruppi che si vogliono ottenere.

Appendice 3

MODALITA’	COORDINATA PRIMO FATTORE	COORDINATA SECONDO FATTORE
V1.1	-2.428693e-02	-1.485175e-02
V1.2	-6.689317e-03	1.164909e-02
V1.3	1.182001e-03	-2.326364e-04
V2.1	1.214997e-03	-5.503112e-04
V2.2	-3.385112e-03	1.174969e-02
V2.3	-1.959401e-02	-1.082842e-02
V3.1	9.874904e-04	1.639008e-04
V3.2	-8.714559e-03	1.198457e-02
V3.3	-1.620545e-02	-1.228329e-02
V4.1	-1.498664e-02	-6.774443e-03
V4.2	-4.944821e-03	9.399361e-03
V4.3	1.133582e-03	-4.091354e-04
V5.1	6.012778e-04	1.682623e-04
V5.2	-5.865722e-03	1.506803e-02
V5.3	-1.944472e-02	-2.006225e-02
V6.1	-7.891078e-03	4.015731e-03
V6.2	-3.395386e-04	4.677225e-03
V6.3	1.593177e-03	-2.092194e-03
V7.1	7.506624e-04	3.351553e-04
V7.2	-7.638997e-03	7.200948e-03
V7.3	-1.267306e-02 -	1.819925e-02
V8.1	-1.721027e-02	-2.494581e-03
V8.2	-3.140295e-03	1.057837e-02
V8.3	1.500917e-03	-1.080511e-03
V9.1	-9.829215e-03	-7.528258e-04
V9.2	-1.143008e-03	6.979247e-03
V9.3	1.834238e-03	-1.398251e-03
V10.1	-7.264486e-03	2.908749e-03
V10.2	1.784058e-05	5.546510e-03

V10.3	1.492261e-03	-2.229506e-03
V11.1	-4.678765e-03	3.970741e-03
V11.2	6.650893e-04	6.626454e-04
V11.3	2.015231e-03	-2.413603e-03
V12.1	6.795859e-04	-6.477462e-05
V12.2	1.541529e-04	6.146630e-03
V12.3	-5.853885e-03	-5.416655e-03
V13.1	8.746958e-04	-2.746593e-04
V13.2	-7.260921e-03	1.431636e-02
V13.3	-1.522637e-02	-1.163209e-02
V14.1	-3.904515e-03	4.304849e-03
V14.2	7.445668e-04	4.624475e-04
V14.3	1.862320e-03	-2.779694e-03
V15.1	-1.223304e-02	-1.643604e-03
V15.2	-4.524372e-03	1.078003e-02
V15.3	1.499930e-03	-1.196989e-03
V16.1	1.291322e-03	-3.700729e-04
V16.2	-3.140896e-03	1.013925e-02
V16.3	-6.212462e-03	-7.458726e-03