INVALSI

PROGETTO FINVALI 2005

Progetto 032: Il sistema scolastico come sistema complesso: qualità delle rilevazioni e modelli di interpretazione dei risultati

Istituto per le Applicazioni del Calcolo 'Mauro Picone' del Consiglio Nazionale delle Ricerche, Sede di Napoli

Dipartimento di Statistica e Matematica per la Ricerca Economica, Università degli Studi di Napoli 'Parthenope'

Con la partecipazione di

Dipartimento di Statistica, Università degli Studi di Milano Bicocca

Dipartimento di Scienze della Terra, Università degli Studi di Napoli 'Federico II'

Dipartimento di Matematica e Applicazioni, Università degli Studi di Napoli 'Federico II'

INVALSI

PROGETTO FINVALI 2005

Progetto 032: Il sistema scolastico come sistema complesso: qualità delle rilevazioni e modelli di interpretazione dei risultati

Report maggio 2007: Analisi ANOVA e CLUSTER dei questionari di apprendimento

Umberto Amato[(1)], Claudia Angelini[(1)], Silvia Razzano[(1,2)]

[(1)]Istituto per le Applicazioni del Calcolo 'Mauro Picone' del Consiglio Nazionale delle Ricerche, Sede di Napoli

[(2)] Dipartimento di Statistica e Matematica per la Ricerca Economica, Università degli Studi di Napoli 'Parthenope'

Abstract: Vengono riportati i risultati ottenuti mediante Cluster Analysis ed ANOVA per raggruppare le abilità degli studenti secondo clusters omogenei. L'analisi è riferita al questionario degli apprendimenti per gli anni scolastici 2004-2005 e 2005-2006.

Indice

Introduzione

Come preventivato il secondo semestre della ricerca è stato dedicato, in prima battuta, all'estensione dell'analisi, svolta con i modelli ANOVA, all'ordine scolastico inizialmente non considerato (Scuole Medie Superiori) ed all'anno scolastico 2005-2006.

A partire, infatti, dai dati, precedentemente acquisiti, presenti nel database INVALSI dei questionari di Sistema e Valutazione, sono stati rivisitati ed ampliati sia gli algoritmi messi a punto per la lettura e l'elaborazione dei dati in ambiente Matlab sia i relativi script implementati per l'esecuzione dell'analisi ANOVA.

L'analisi dei risultati ottenuti ha mostrato la significatività di tutti i parametri presi in considerazione (tipo di scuola, sesso, regione, ordine di scuola). Pertanto allo scopo di ridurre drasticamente il numero di variabili in gioco si ricorrerà a metodologie di clustering per l'individuazione di gruppi omogenei. Ultimo passo di questa fase è stato poi eseguire nuovamente l'analisi ANOVA sui gruppi individuati con la procedura di clustering.

Va osservato che l'analisi presentata in questo report è riferita ai dati non corretti secondo la procedura descritta nel Report "Una procedura di qualità basata sulla fuzzy clustering per l'individuazione e la correzione dei dati anomali nell'ambito del Servizo Nazionale di Valutazione scolastica degli apprendimenti (SNV)". Vi sono diverse motivazioni alla base di tale scelta. Innanzitutto la correzione apportata nel report citato (ed in particolare i pesi stimati di qualità del dato) sono riferiti alla singola classe, mentre il presente report prende in considerazione l'abilità media calcolata sul singolo Istituto. È in corso al momento una analisi simile a quella riportata nel report citato basata sull'Istituto anziché sulla classe, a valle della quale si potrebbe ripetere l'analisi descritta nel presente report. Ciò nonostante non sono attese variazioni estremamente significative tra le analisi effettuate con i due tipi diversi di correzione dei dati (per classe e per Istituto), per cui la difficoltà evidenziata nel presente report di difficoltà di raggruppamento in cluster omogenei permane. Si tenga inoltre conto che per le Scuole Medie e Medie Superiori non si prevedono variazioni significative delle abilità dovute ai fattori di ponderazione, per cui l'analisi presentata nel presente report è da considerarsi valida.

1. Analisi ANOVA sui questionari di valutazione INVALSI, a.s. 2004/2005

Di seguito vengono riportate le analisi, eseguite mediante la metodologia ANOVA, sui dati dei questionari di valutazione che, in questa seconda fase del progetto, si riferiscono alle Scuole Elementari, Medie Inferiori e Medie Superiori ed agli anni scolastici 2004~2005 e 2005~2006.

Come per le analisi precedenti sono stati considerati quattro fattori fissi:

- Tipo, presenta due livelli: Scuola Pubblica e Scuola privata;
- Regione, cui sono associati i 20 livelli corrispondenti al numero delle regioni italiane;
- Ordine, con tre livelli: Scuole Elementari, Scuole Medie Inferiori, Scuole Medie Superiori;
- Sesso, con due livelli: Maschi e Femmine.

Ancora una volta le analisi per la valutazione degli Istituti sono state effettuate prendendo in considerazione sia l'insieme delle domande relative al complesso delle tre discipline (Italiano, Matematica e Scienze), sia disciplina per disciplina; ed ancora in qualità di variabile responso è stata scelta sia la percentuale di risposte esatte fornite dagli studenti sia l'abilità stimata con il metodo Item Response Theory (IRT) a tre parametri.

Inoltre, così come fatto per il precedente report, anche in questo caso i risultati saranno forniti in forma sia tabellare che grafica. In particolare nella tabella verranno riportati i seguenti dati:

- Media generale stimata dal modello ANOVA ottenuta mediante risoluzione ai minimi quadrati (Grandmean)

- Valore dello statistic F (F-test) e relaitvo livello di significatività in base ai gradi di libertà del problema (Liv. Signif.): valori alti indicano una significatività del fattore

- Media e deviazione standard stimate per ognuno dei livelli del fattore. In particolare questì'ultima quantità è anche rappresentata graficamente: per ogni livello del fattore il valore medio è riportato con un simbolo circolare e la deviazione standard con una barra: in modo "grezzo" (ma non lontano dall'indagine statistica rigorosa) i livelli sono omogenei tra loro quando le barre (corrispondenti a una deviazione standard) si intersecano tra loro guardandole verticalmente.

Nel caso del fattore Regione, visto il numero (relativamente) alto di livelli, è stato anche prodotto un grafico che rappresenta in maniera visiva le differenze di abilità esistenti tra le diverse regioni: su ascissa ed ordinata sono riportate le Regioni italiane numerate da 1 a 20 secondo il seguente ordine: Valle d'Aosta, Piemonte, Liguria, Lombardia, Trentino, Veneto, Friuli Venezia Giulia, Emilia Romagna, Toscana, Umbria, Marche, Lazio, Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria, Sicilia, Sardegna. Un elemento della matrice rappresenta la differenza di percentuale esistente tra la regione indicata nella riga e quella nella colonna. La colorazione (riportata nella barra a destra del grafico) indica visivamente questa differenza: colori giallo-verde-azzurro chiaro indicano piccole differenze, mentre rosso e blu indicano forti differenze (in positivo e negativo, rispettivamente). Pertanto regioni omogenee sono rappresentate nel grafico con colorazioni giallo-verde-azzurro chiaro.

1.1. Percentuale di risposte esatte: Italiano, Matematica e Scienze

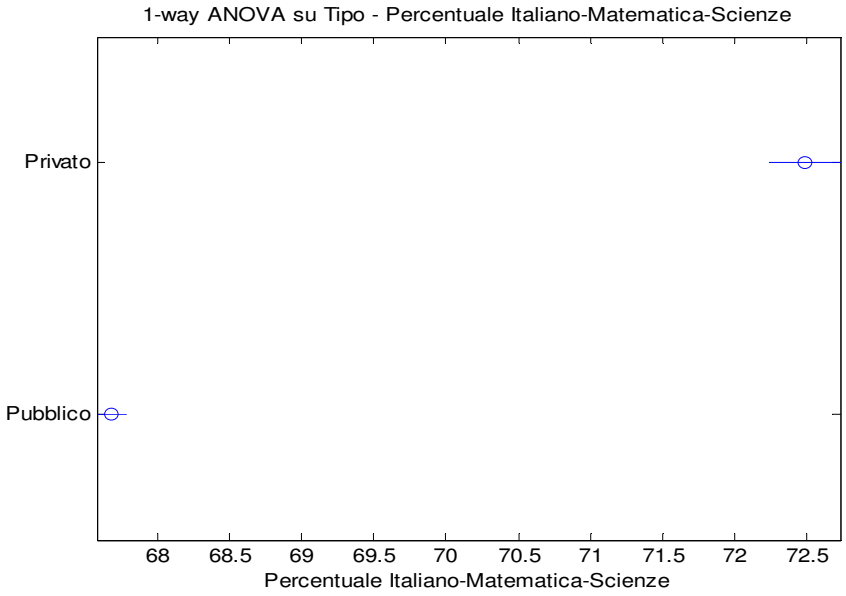
1.1.1 Fattore Tipo

Grandmean: 70.09

F-test: 3.282082e+002 - Liv. signif.: 0

Tipo 1: Pubblico - Estimated mean: 67.68 +/- 0.10

Tipo 2: Privato - Estimated mean: 72.49 +/- 0.24



Anche dopo l'inserimento delle Scuole Medie Superiori si può osservare che le Scuole Private continuano ad avere una valutazione significativamente migliore rispetto alle Scuole Pubbliche.

1.1.2 Fattore Regione

Grandmean: 67.82

F-test: 3.184483e+001 - Liv. signif.: 0

Regione 1: Valle d'Aosta - Estimated mean: 62.09 +/- 2.48

Regione 2: Piemonte - Estimated mean: 67.04 +/- 0.38

Regione 3: Liguria - Estimated mean: 68.49 +/- 0.65

Regione 4: Lombardia - Estimated mean: 65.92 +/- 0.25

Regione 5: Trentino - Estimated mean: 63.66 +/- 0.80

Regione 6: Veneto - Estimated mean: 66.42 +/- 0.35

Regione 7: Friuli - Estimated mean: 68.04 +/- 0.73

Regione 8: Emilia - Estimated mean: 67.11 +/- 0.41

Regione 9: Toscana - Estimated mean: 68.12 +/- 0.43

Regione 10: Umbria - Estimated mean: 67.43 +/- 0.81

Regione 11: Marche - Estimated mean: 68.26 +/- 0.59

Regione 12: Lazio - Estimated mean: 69.31 +/- 0.32

Regione 13: Abruzzo - Estimated mean: 68.53 +/- 0.63

Regione 14: Molise - Estimated mean: 69.15 +/- 1.01

Regione 15: Campania - Estimated mean: 72.19 +/- 0.27

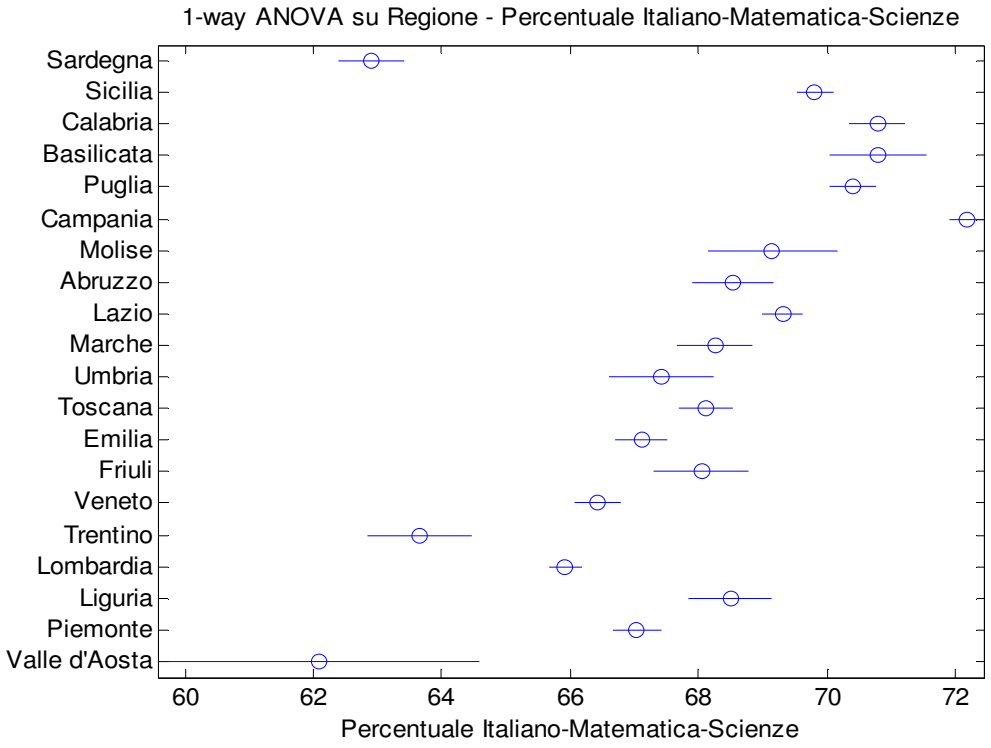
Regione 16: Puglia - Estimated mean: 70.39 +/- 0.36

Regione 17: Basilicata - Estimated mean: 70.79 +/- 0.76

Regione 18: Calabria - Estimated mean: 70.78 +/- 0.43

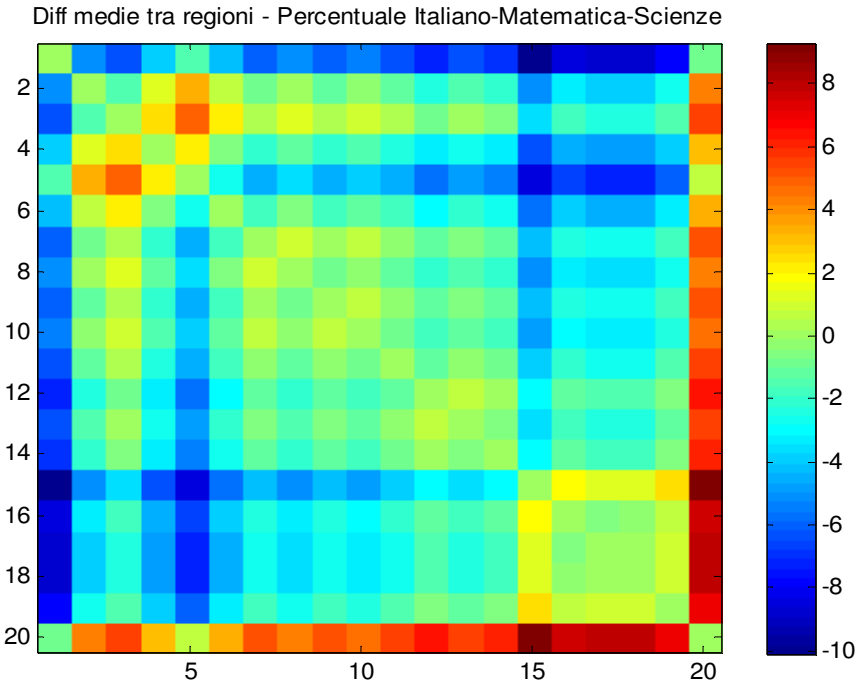
Regione 19: Sicilia - Estimated mean: 69.81 +/- 0.29

Regione 20: Sardegna - Estimated mean: 62.89 +/- 0.51



Il fattore Regione risulta anche questa volta significativo e si conferma il particolare trend geografico in base al quale è possibile osservare un miglioramento della valutazione quando ci si sposta dal Nord al Centro ed ancora al Sud.

Se si osserva anche il grafico successivo, su cui vengono riportate le differenza fra le percentuali di risposta ottenute dalle regioni sulle righe meno quelle ottenute dalle regioni sulle colonne, è possibile individuare ancora due aree ben distinte, corrispondenti in linea di massima alle regioni del Centro e del Sud, che presentano al proprio interno una colorazione simile elemento questo che indica la presenza di un elevato grado di omogeneità del valore medio.



1.1.3 Fattore Ordine

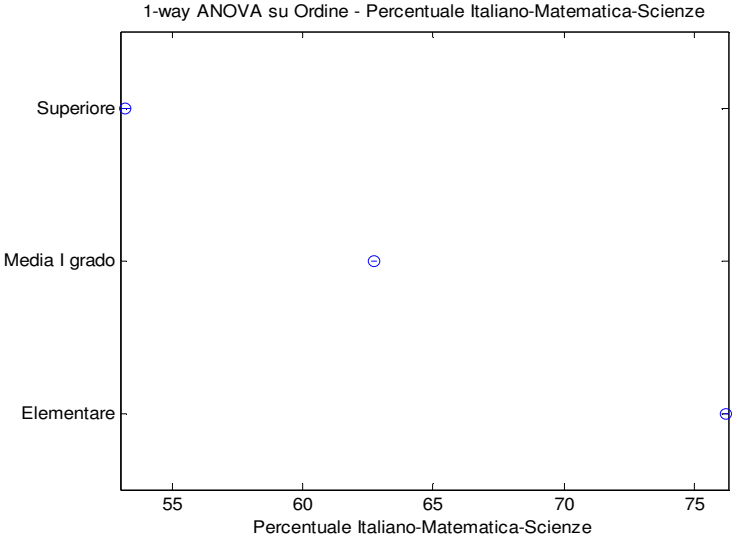
Grandmean: 64.05

F-test: 7.082857e+003 - Liv. signif.: 0

Ordine 1: Elementare - Estimated mean: 76.20 +/- 0.10

Ordine 2: Media I grado - Estimated mean: 62.73 +/- 0.11

Ordine 3: Superiore - Estimated mean: 53.23 +/- 0.21



Come è possibile osservare dalla tabella e dal grafico qui riportati, il fattore Ordine è certamente significativo, ed inoltre fanalino di coda, per quel che riguarda appunto la percentuale di risposte esatte sulle tre materie in funzione dell'Ordine scolastico, spetta sicuramente alle Scuole Medie Superiori, mentre le Scuole Elementari conserva ancora una percentuale di risposte esatte decisamente maggiore rispetto alle altre scuole.

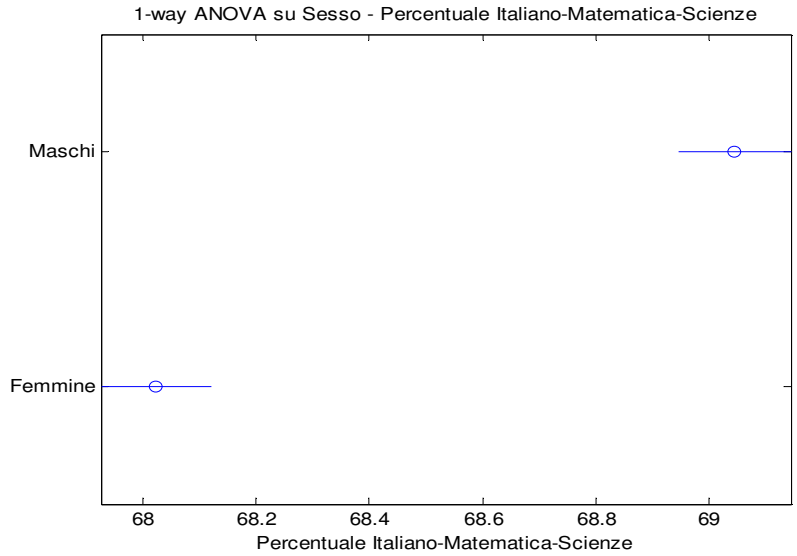
1.1.4 Fattore Sesso

Grandmean: 68.54

F-test: 5.394615e+001 - Liv. signif.: 2.113865e-013

Sesso 1: Femmine - Estimated mean: 68.02 +/- 0.10

Sesso 2: Maschi - Estimated mean: 69.05 +/- 0.10



Altro fattore significativo risulta essere il Sesso per il quale è possibile notare una percentuale di risposte esatte fornite dai maschi leggermente maggiore rispetto a quelle fornite dalle femmine (all'incirca 1 punto percentuale).

Da quanto fin qui visto l'inserimento delle Scuole Medie Superiori non ha messo in discussione i risultati delle analisi preliminari eseguite nella prima fase del progetto anche se ha comportato una notevole diminuzione dei valori medi percentuali per tutti i fattori presi in considerazione.

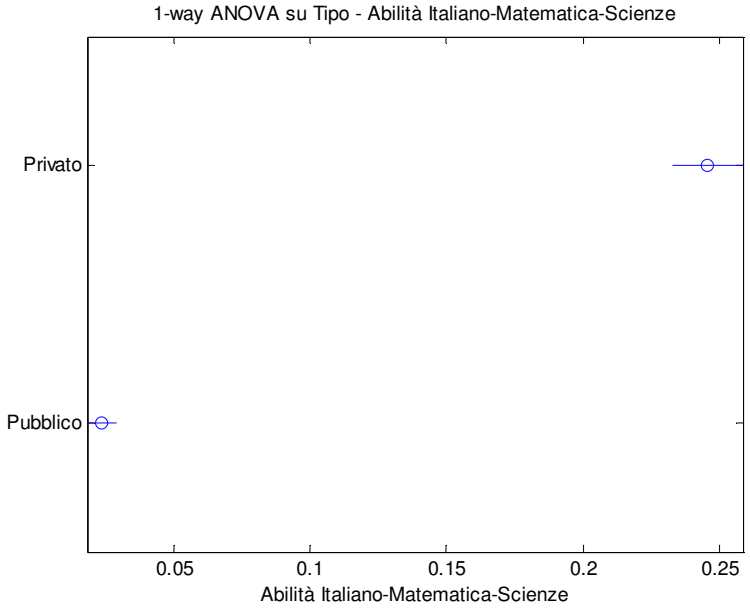
1.2. Abilità: Italiano, Matematica e Scienze

Come già accennato precedente mente, proponiamo, di seguito, la stessa analisi eseguita prendendo in considerazione, però, in qualità di variabile responso le abilità stimate con la IRT. Dal momento che le indicazioni ottenute nei due casi sono molto simili

non saranno analizzate nel dettaglio ed inoltre, nelle analisi successive, specifiche per singola disciplina, non si riporteranno i risultati ottenuti utilizzando, appunto, le abilità calcolate con la IRT.

1.2.1 Fattore Tipo

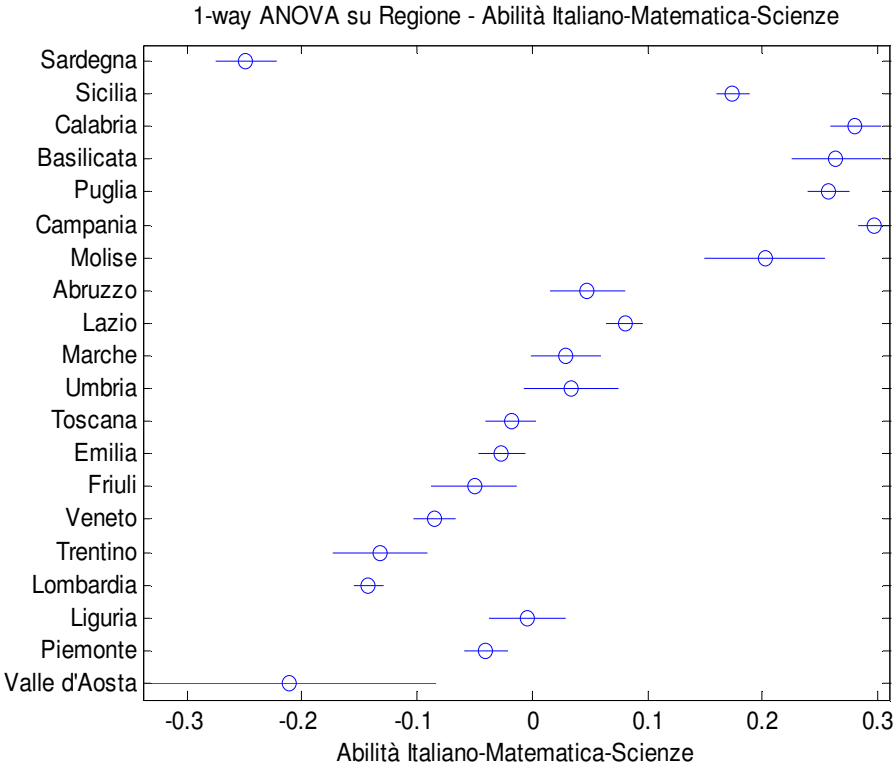
Grandmean: 0.13
F-test: 2.547382e+002 - Liv. signif.: 0
Tipo 1: Pubblico - Estimated mean: 0.02 +/- 0.01
Tipo 2: Privato - Estimated mean: 0.25 +/- 0.01

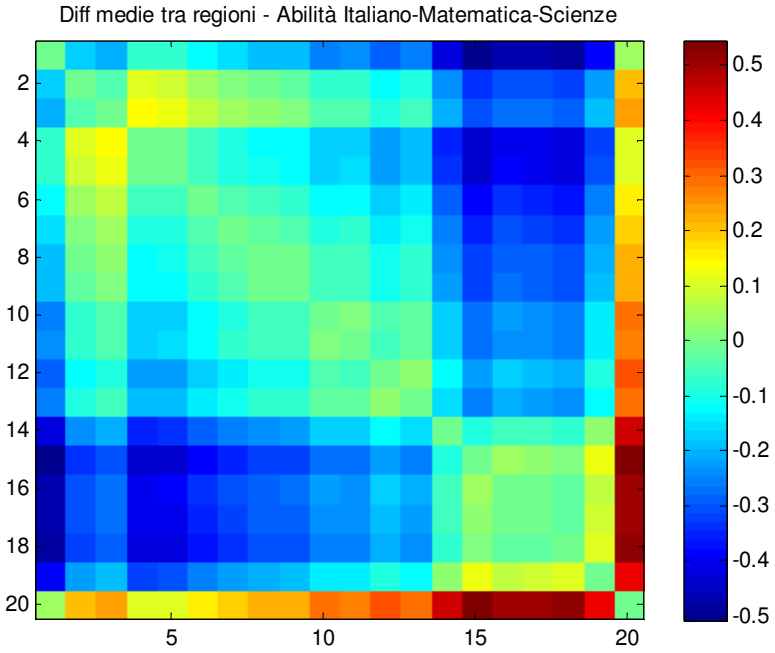


1.2.2 Fattore Regione

Grandmean: 0.04
F-test: 5.965635e+001 - Liv. signif.: 0
Regione 1: Valle d'Aosta - Estimated mean: -0.21 +/- 0.13
Regione 2: Piemonte - Estimated mean: -0.04 +/- 0.02
Regione 3: Liguria - Estimated mean: -0.00 +/- 0.03
Regione 4: Lombardia - Estimated mean: -0.14 +/- 0.01
Regione 5: Trentino - Estimated mean: -0.13 +/- 0.04
Regione 6: Veneto - Estimated mean: -0.08 +/- 0.02
Regione 7: Friuli - Estimated mean: -0.05 +/- 0.04
Regione 8: Emilia - Estimated mean: -0.03 +/- 0.02
Regione 9: Toscana - Estimated mean: -0.02 +/- 0.02
Regione 10: Umbria - Estimated mean: 0.03 +/- 0.04
Regione 11: Marche - Estimated mean: 0.03 +/- 0.03
Regione 12: Lazio - Estimated mean: 0.08 +/- 0.02
Regione 13: Abruzzo - Estimated mean: 0.05 +/- 0.03
Regione 14: Molise - Estimated mean: 0.20 +/- 0.05
Regione 15: Campania - Estimated mean: 0.30 +/- 0.01
Regione 16: Puglia - Estimated mean: 0.26 +/- 0.02

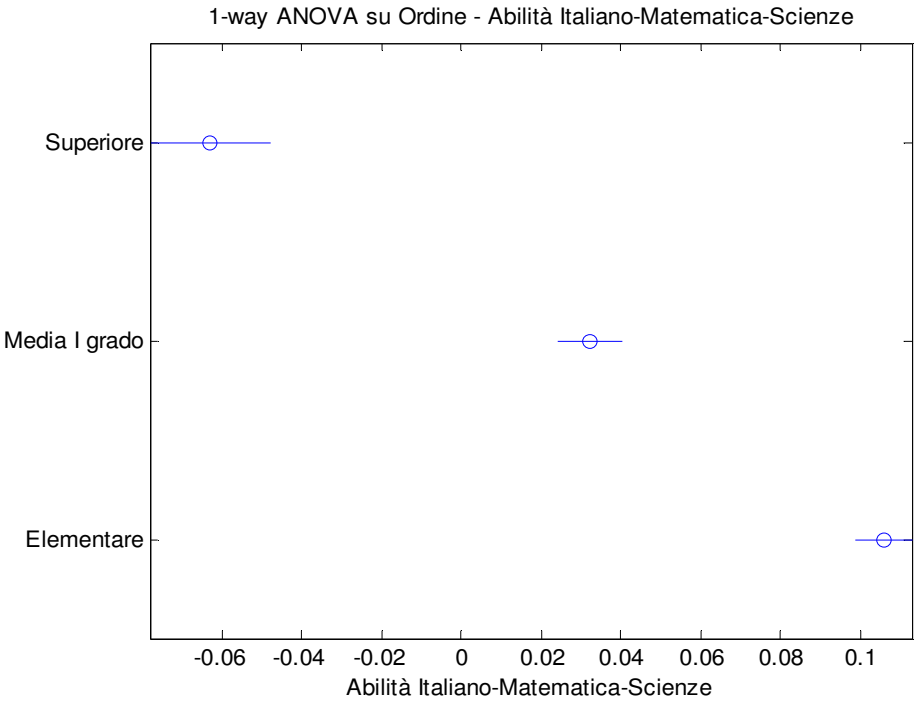
Regione 17: Basilicata - Estimated mean: 0.26 +/- 0.04
Regione 18: Calabria - Estimated mean: 0.28 +/- 0.02
Regione 19: Sicilia - Estimated mean: 0.17 +/- 0.01
Regione 20: Sardegna - Estimated mean: -0.25 +/- 0.03





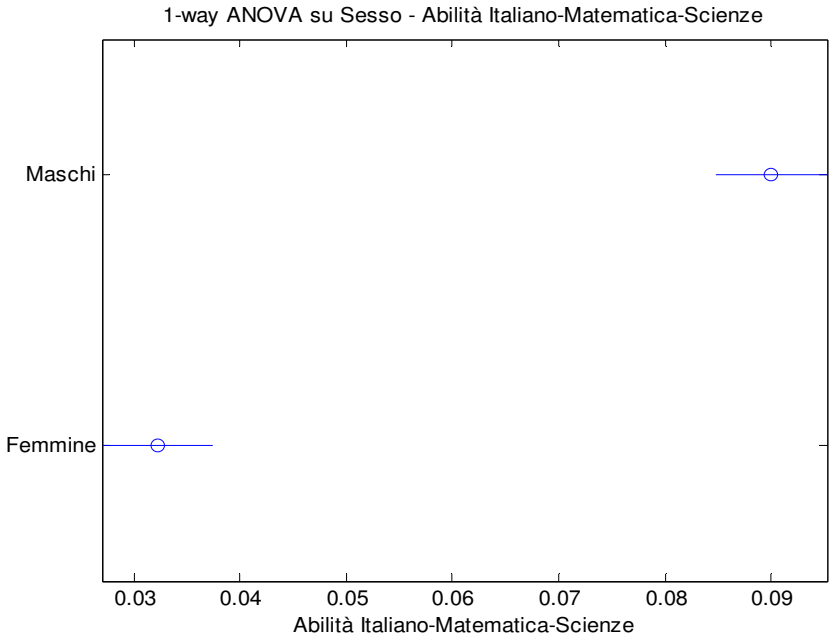
1.2.3 Fattore Ordine

Grandmean: 0.02
F-test: 5.939716e+001 - Liv. signif.: 0
Ordine 1: Elementare - Estimated mean: 0.11 +/- 0.01
Ordine 2: Media I grado - Estimated mean: 0.03 +/- 0.01
Ordine 3: Superiore - Estimated mean: -0.06 +/- 0.02



1.2.4 Fattore Sesso

Grandmean: 0.06
F-test: 6.191252e+001 - Liv. signif.: 3.663736e-015
Sesso 1: Femmine - Estimated mean: 0.03 +/- 0.01
Sesso 2: Maschi - Estimated mean: 0.09 +/- 0.01



1.3. Percentuale di risposte esatte: Italiano

Il linea di massima le percentuali di risposte esatte per la singola disciplina "Italiano" sono leggermente inferiori rispetto al caso generale (all'incirca un punto percentuale).

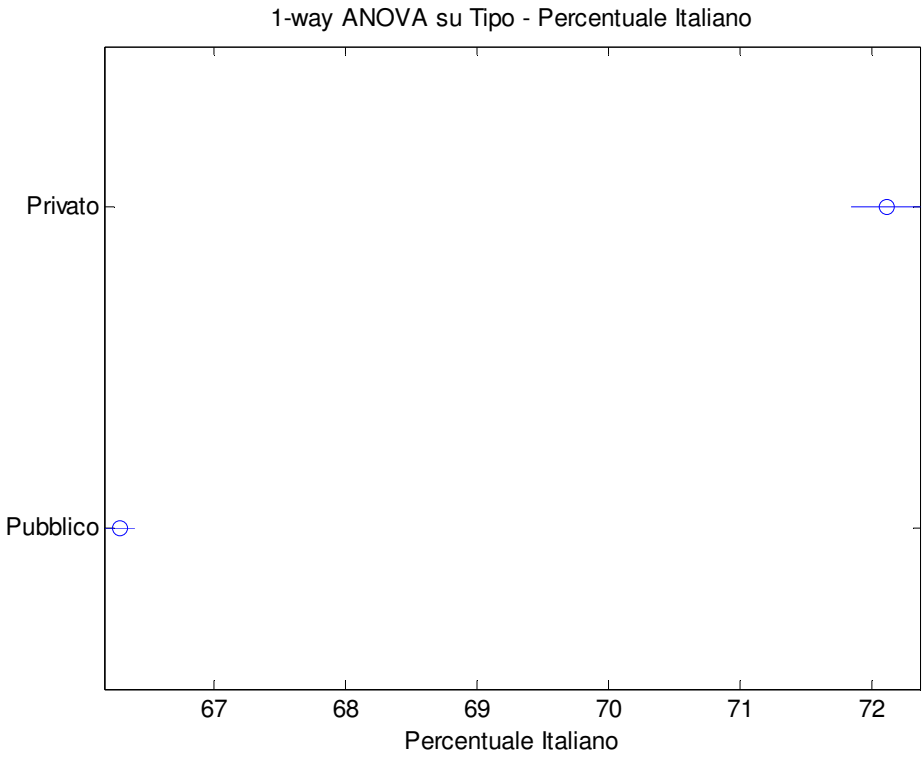
1.3.1 Fattore Tipo

Grandmean: 69.20

F-test: 4.074439e+002 - Liv. signif.: 0

Tipo 1: Pubblico - Estimated mean: 66.29 +/- 0.11

Tipo 2: Privato - Estimated mean: 72.11 +/- 0.26



Con l'inserimento nell'analisi delle Scuole Medie Superiori il divario fra le Scuole Pubbliche e le Scuole Private risulta leggermente superiore rispetto al caso generale, (circa 5,8%).

1.3.2 Fattore Regione

Grandmean: 66.63

F-test: 1.413706e+001 - Liv. signif.: 0

Regione 1: Valle d'Aosta - Estimated mean: 62.58 +/- 2.74

Regione 2: Piemonte - Estimated mean: 66.73 +/- 0.41

Regione 3: Liguria - Estimated mean: 67.50 +/- 0.72

Regione 4: Lombardia - Estimated mean: 66.10 +/- 0.28

Regione 5: Trentino - Estimated mean: 62.40 +/- 0.89

Regione 6: Veneto - Estimated mean: 66.04 +/- 0.39

Regione 7: Friuli - Estimated mean: 67.46 +/- 0.81

Regione 8: Emilia - Estimated mean: 66.37 +/- 0.45

Regione 9: Toscana - Estimated mean: 67.24 +/- 0.47

Regione 10: Umbria - Estimated mean: 66.20 +/- 0.89

Regione 11: Marche - Estimated mean: 67.65 +/- 0.65

Regione 12: Lazio - Estimated mean: 68.38 +/- 0.35

Regione 13: Abruzzo - Estimated mean: 67.35 +/- 0.69

Regione 14: Molise - Estimated mean: 67.08 +/- 1.12

Regione 15: Campania - Estimated mean: 69.97 +/- 0.30

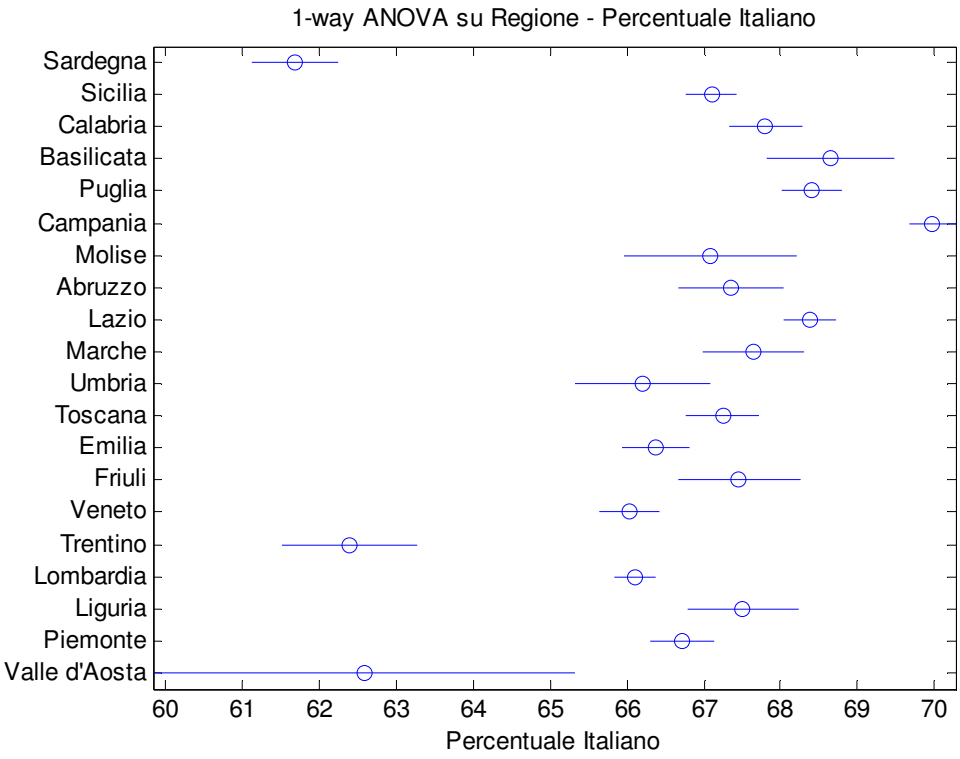
Regione 16: Puglia - Estimated mean: 68.40 +/- 0.40

Regione 17: Basilicata - Estimated mean: 68.64 +/- 0.84

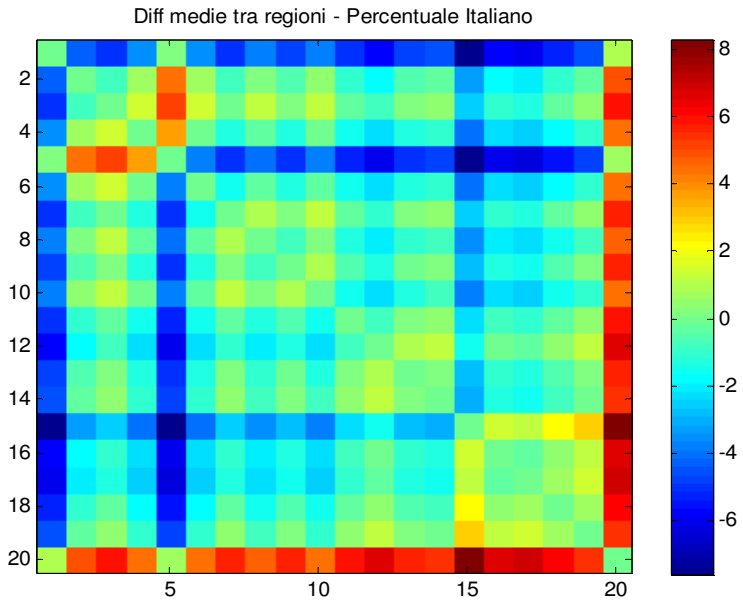
Regione 18: Calabria - Estimated mean: 67.80 +/- 0.47

Regione 19: Sicilia - Estimated mean: 67.10 +/- 0.32

Regione 20: Sardegna - Estimated mean: 61.68 +/- 0.56



L'andamento generale osservato per tutte le materie è qui confermato ed inoltre si registra una diminuzione generalizzata delle percentuali di risposte esatte di circa 1.5 – 2.5 punti percentuali per regione.



1.3.3 Fattore Ordine

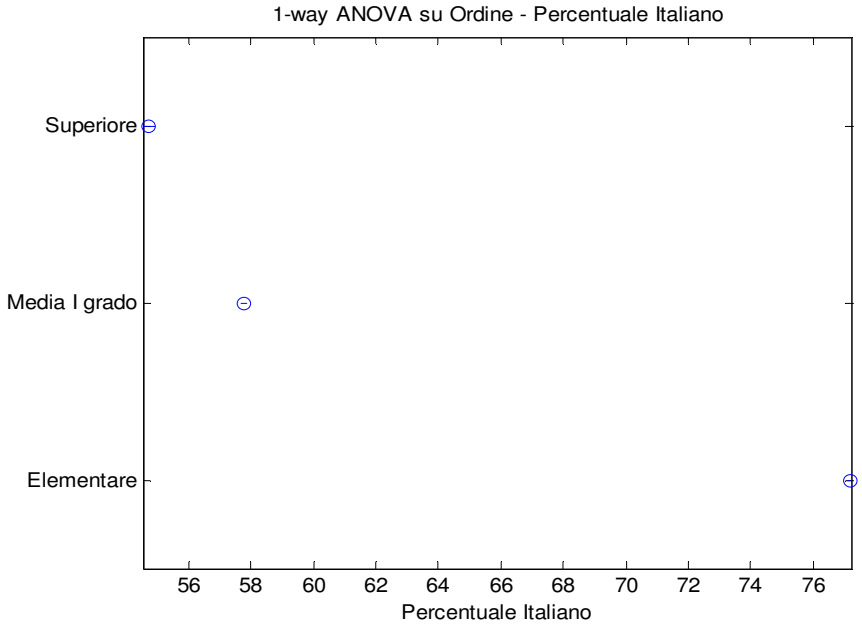
Grandmean: 63.24

F-test: 1.151525e+004 - Liv. signif.: 0

Ordine 1: Elementare - Estimated mean: 77.18 +/- 0.09

Ordine 2: Media I grado - Estimated mean: 57.79 +/- 0.11

Ordine 3: Superiore - Estimated mean: 54.74 +/- 0.20



Si riduce leggermente il divario fra le Scuole Elementari e le Scuole Medie Superiori, assestatosi a circa il 19,5% in favore delle prime, mentre aumenta il divario ancora fra le Scuole Elementari e le Scuole Medie Inferiori, che passa da un circa 13% nel caso generale a quasi il 19% per la singola materia Italiano.

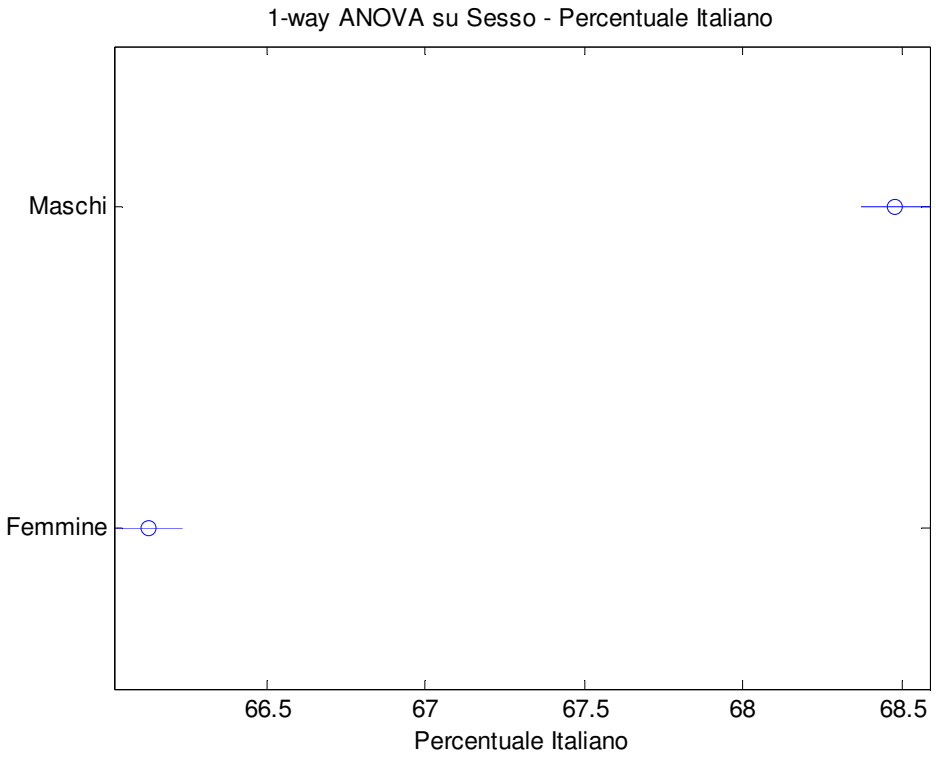
1.3.4 Fattore Sesso

Grandmean: 67.30

F-test: 2.393025e+002 - Liv. signif.: 0

Sesso 1: Femmine - Estimated mean: 66.13 +/- 0.11

Sesso 2: Maschi - Estimated mean: 68.48 +/- 0.11



La significatività del fattore Sesso è confermata anche per la singola materia "Italiano" con un sensibile aumento del margine fra i due livelli (Maschi e Femmine) pari a circa 2,5 punti percentuali ancora a favore dei Maschi.

1.4. Percentuale di risposte esatte: Matematica

Anche dopo l'inserimento dei dati relativi alle Scuole Medie Superiori le percentuali di risposte esatte per le domande di matematica continuano ad essere le più basse rispetto alle altre materie, con una media del 65.2% rispetto, ad esempio, al 67.6% riscontrato per il caso generale su tutte e tre le materie.

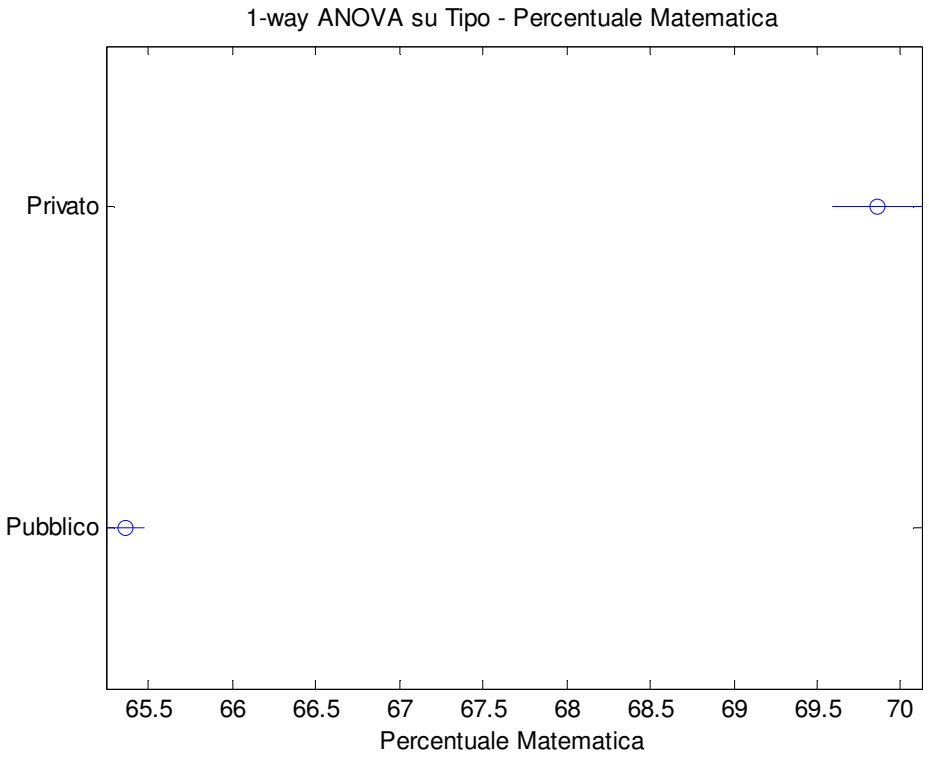
1.4.1 Fattore Tipo

Grandmean: 67.61

F-test: 2.385940e+002 - Liv. signif.: 0

Tipo 1: Pubblico - Estimated mean: 65.36 +/- 0.11

Tipo 2: Privato - Estimated mean: 69.86 +/- 0.27



Pur riscontrando una diminuzione dei valori percentuali di circa 3 punti rispetto al caso generale va segnalato che il divario fra Scuole Pubbliche e Private, sempre in favore delle Scuole Private, resta praticamente costante.

1.4.2 Fattore Regione

Grandmean: 65.43

F-test: 4.647536e+001 - Liv. signif.: 0

Regione 1: Valle d'Aosta - Estimated mean: 58.31 +/- 2.69

Regione 2: Piemonte - Estimated mean: 64.05 +/- 0.41

Regione 3: Liguria - Estimated mean: 65.69 +/- 0.71

Regione 4: Lombardia - Estimated mean: 62.46 +/- 0.27

Regione 5: Trentino - Estimated mean: 61.98 +/- 0.87

Regione 6: Veneto - Estimated mean: 63.27 +/- 0.38

Regione 7: Friuli - Estimated mean: 64.93 +/- 0.79

Regione 8: Emilia - Estimated mean: 64.52 +/- 0.44

Regione 9: Toscana - Estimated mean: 65.68 +/- 0.46

Regione 10: Umbria - Estimated mean: 65.10 +/- 0.87

Regione 11: Marche - Estimated mean: 66.01 +/- 0.64

Regione 12: Lazio - Estimated mean: 66.79 +/- 0.35

Regione 13: Abruzzo - Estimated mean: 66.50 +/- 0.68

Regione 14: Molise - Estimated mean: 67.24 +/- 1.10

Regione 15: Campania - Estimated mean: 70.84 +/- 0.30

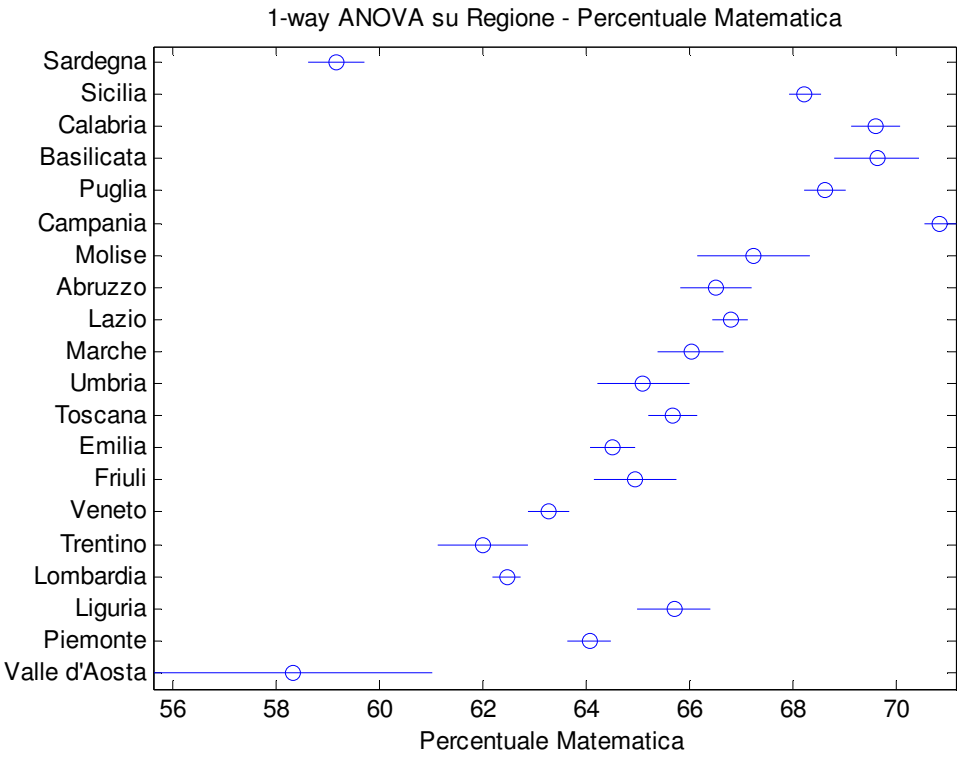
Regione 16: Puglia - Estimated mean: 68.61 +/- 0.39

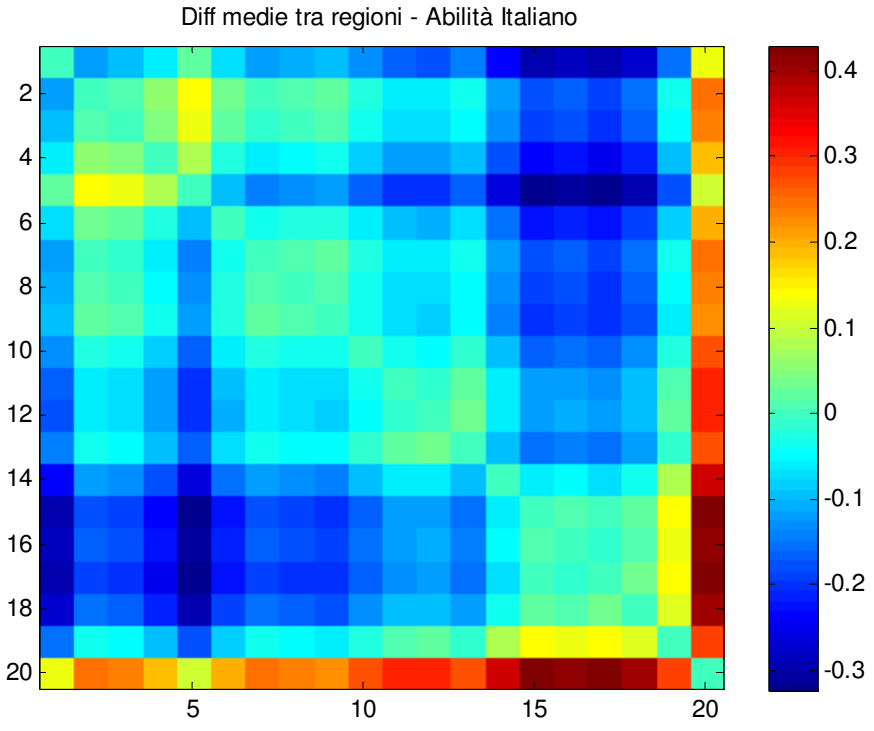
Regione 17: Basilicata - Estimated mean: 69.61 +/- 0.82

Regione 18: Calabria - Estimated mean: 69.59 +/- 0.47

Regione 19: Sicilia - Estimated mean: 68.23 +/- 0.31

Regione 20: Sardegna - Estimated mean: 59.16 +/- 0.56





In sostanziale accordo con il caso generale possiamo confermare anche per la singola disciplina "Matematica" gli andamenti visti in precedenza.

1.4.3 Fattore Ordine

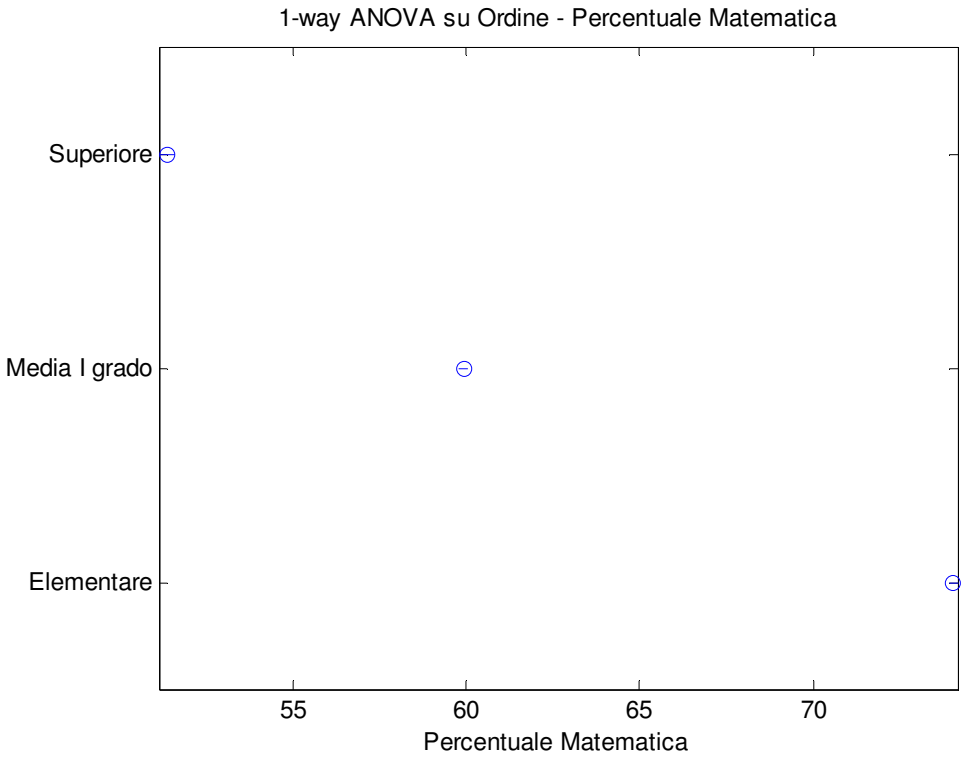
Grandmean: 61.78

F-test: 5.383263e+003 - Liv. signif.: 0

Ordine 1: Elementare - Estimated mean: 74.05 +/- 0.11

Ordine 2: Media I grado - Estimated mean: 59.92 +/- 0.13

Ordine 3: Superiore - Estimated mean: 51.37 +/- 0.24



Il divario fra Scuole Elementari e Scuole Medie Superiori rispecchia i valori ottenuti nel caso generale mentre cresce leggermente la distanza fra Scuole Elementari e Scuole Medie Inferiori (circa un punto percentuale).

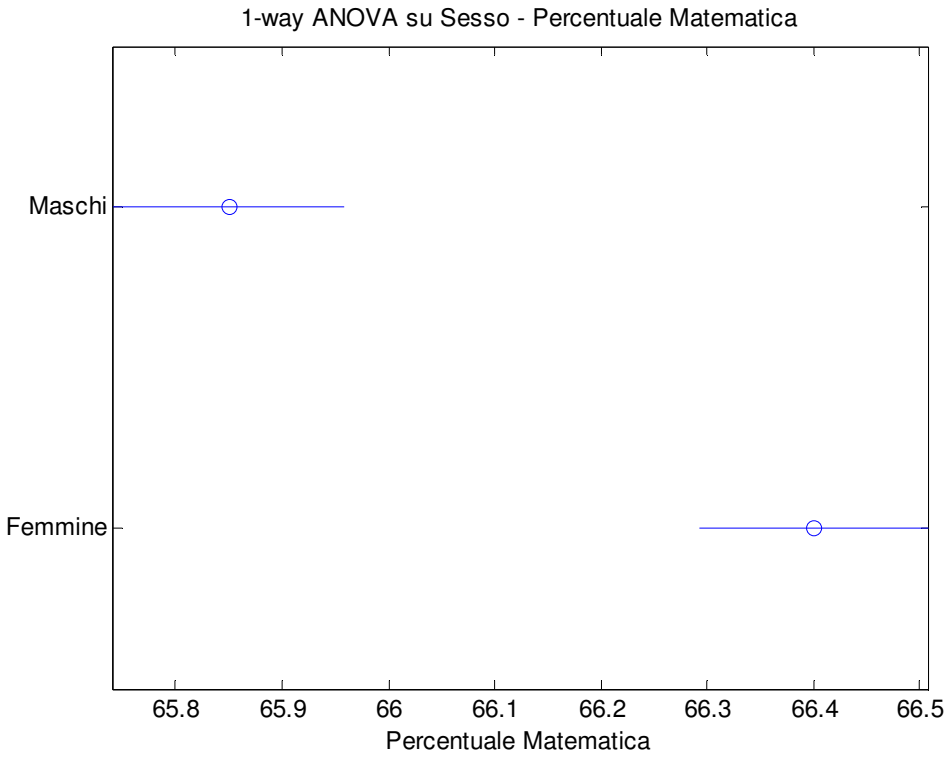
1.4.4 Fattore Sesso

Grandmean: 66.13

F-test: 1.290585e+001 - Liv. signif.: 3.280864e-004

Sesso 1: Femmine - Estimated mean: 66.40 +/- 0.11

Sesso 2: Maschi - Estimated mean: 65.85 +/- 0.11



Per le prove di Matematica si conferma, anche a seguito dell'inserimento nell'analisi dei dati relativi alle Scuole Medie Superiori,una inversione di tendenza, nel senso che le percentuali di risposte esatte delle Femmine risultano essere lievemente (circa 0.6%), ma significativamente maggiori rispetto a quelle fornite dai Maschi.

1.5. Percentuale di risposte esatte: Scienze

Le percentuali di risposte esatte fornite alle domande di Scienze si confermano essere costantemente più elevate rispetto alle altre materie, valutate singolarmente, ed al caso generale.

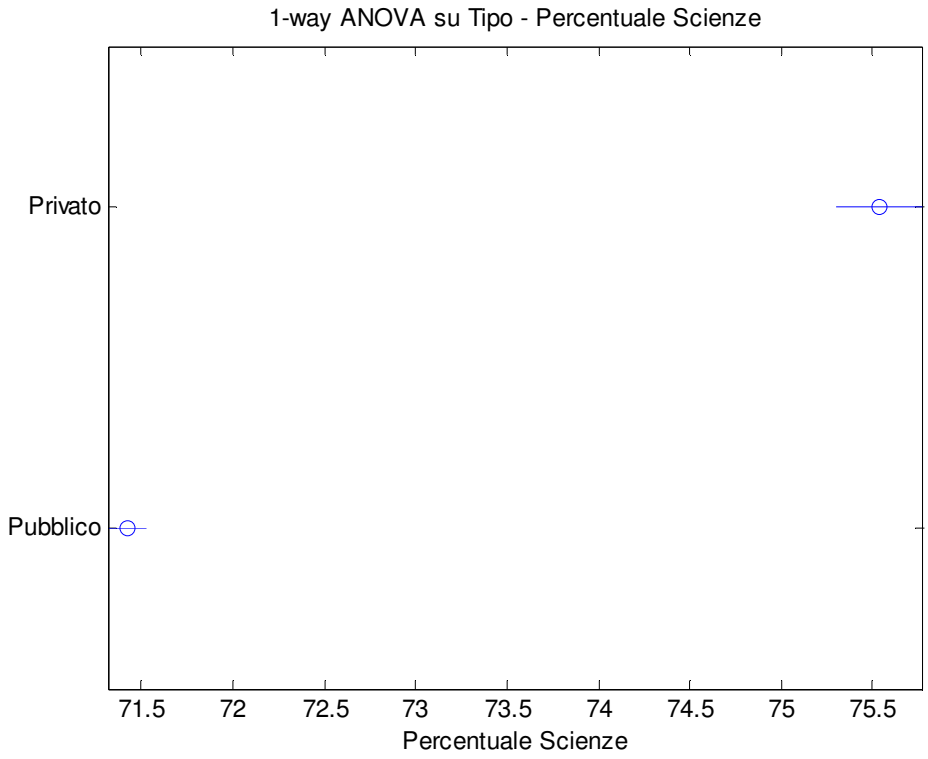
1.5.1. Fattore Tipo

Grandmean: 73.48

F-test: 2.473094e+002 - Liv. signif.: 0

Tipo 1: Pubblico - Estimated mean: 71.42 +/- 0.10

Tipo 2: Privato - Estimated mean: 75.54 +/- 0.24



L'andamento generale è ancora una volta confermato con un distacco fra Scuole Pubbliche e Scuole Private, valutato intorno al 5% a favore di queste ultime, in linea sia con il valore generale che con quello riscontrabile per le singole materie.

1.5.2. Fattore Regione

Grandmean: 71.42

F-test: 3.601464e+001 - Liv. signif.: 0

Regione 1: Valle d'Aosta - Estimated mean: 65.52 +/- 2.44

Regione 2: Piemonte - Estimated mean: 70.39 +/- 0.37

Regione 3: Liguria - Estimated mean: 72.28 +/- 0.64

Regione 4: Lombardia - Estimated mean: 69.20 +/- 0.25

Regione 5: Trentino - Estimated mean: 66.60 +/- 0.79

Regione 6: Veneto - Estimated mean: 69.96 +/- 0.35

Regione 7: Friuli - Estimated mean: 71.71 +/- 0.72

Regione 8: Emilia - Estimated mean: 70.47 +/- 0.40

Regione 9: Toscana - Estimated mean: 71.42 +/- 0.42

Regione 10: Umbria - Estimated mean: 70.96 +/- 0.79

Regione 11: Marche - Estimated mean: 71.13 +/- 0.58

Regione 12: Lazio - Estimated mean: 72.77 +/- 0.31

Regione 13: Abruzzo - Estimated mean: 71.76 +/- 0.62

Regione 14: Molise - Estimated mean: 73.11 +/- 1.00

Regione 15: Campania - Estimated mean: 75.80 +/- 0.27

Regione 16: Puglia - Estimated mean: 74.16 +/- 0.36

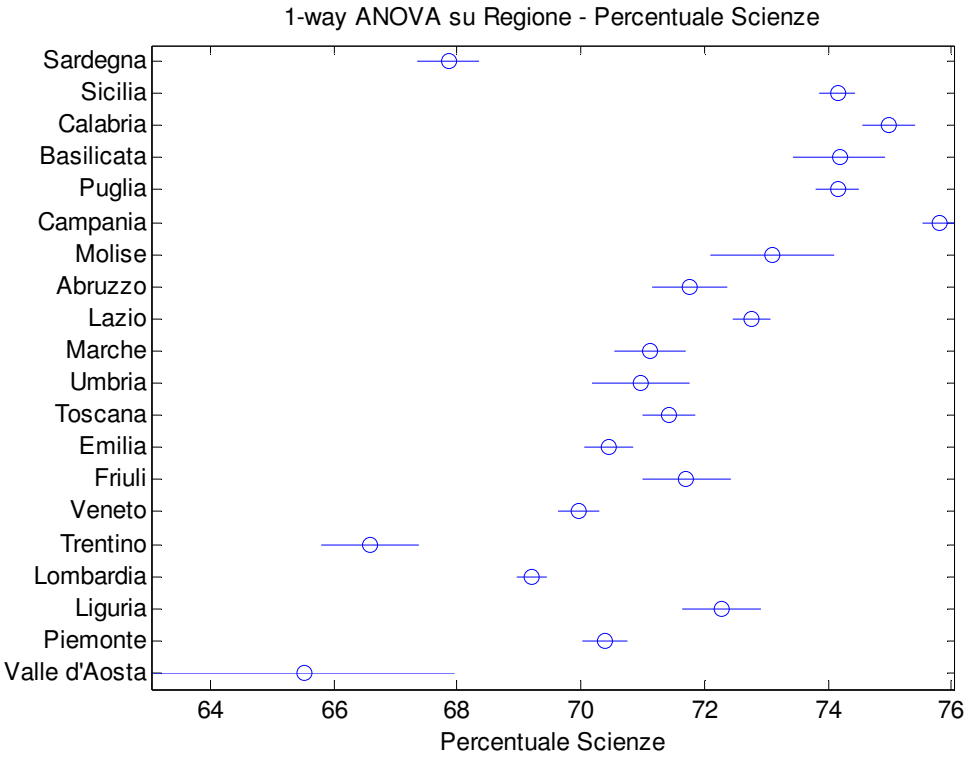
Regione 17: Basilicata - Estimated mean: 74.18 +/- 0.74

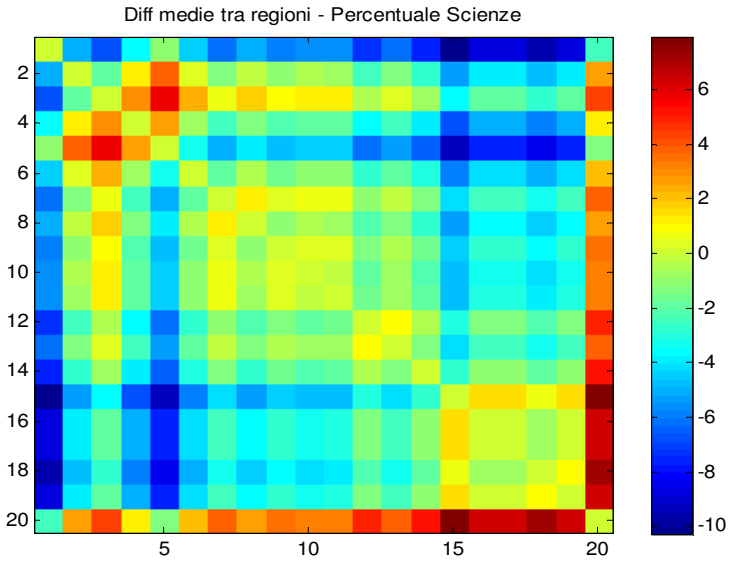
Regione 18: Calabria - Estimated mean: 74.99 +/- 0.42

Regione 19: Sicilia - Estimated mean: 74.15 +/- 0.29

Regione 20: Sardegna - Estimated mean: 67.85 +/- 0.50

Si conferma l'andamento omogeneo per materia





1.5.3. Fattore Ordine

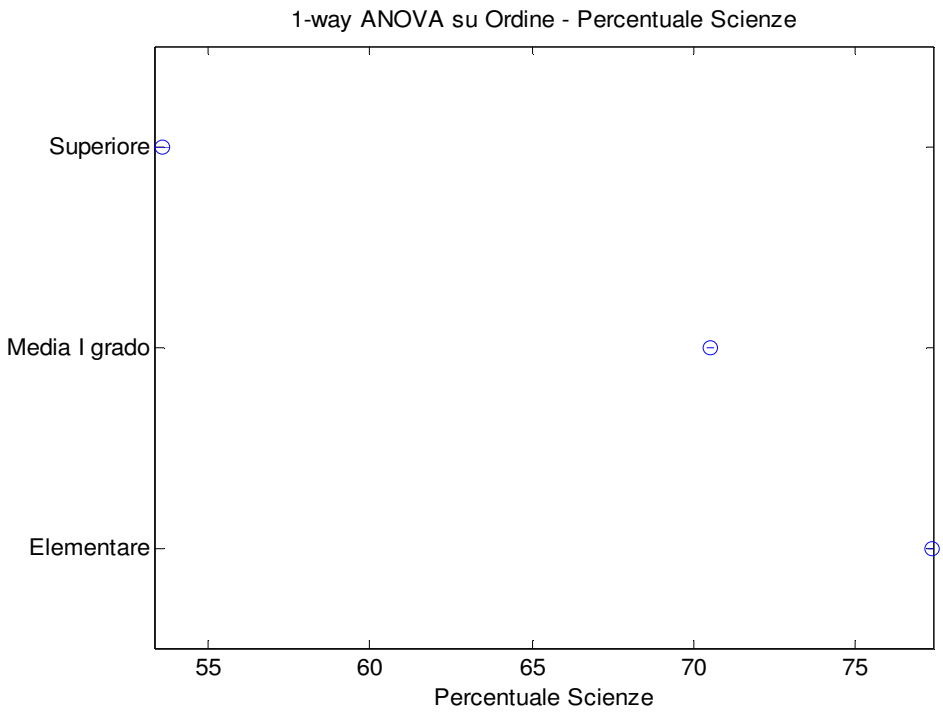
Grandmean: 67.16

F-test: 4.771918e+003 - Liv. signif.: 0

Ordine 1: Elementare - Estimated mean: 77.37 +/- 0.11

Ordine 2: Media I grado - Estimated mean: 70.51 +/- 0.12

Ordine 3: Superiore - Estimated mean: 53.59 +/- 0.22



Per la singola disciplina "Scienze" è possibile osservare una riduzione del divario fra Scuole Elementari e Scuole Medie Inferiori, attestatosi intorno all'8%a favore sempre delle prime, mentre aumenta la distanza fra Scuole Elementari e Scuole Medie Superiori che raggiunge un valore pari al 23.8%.

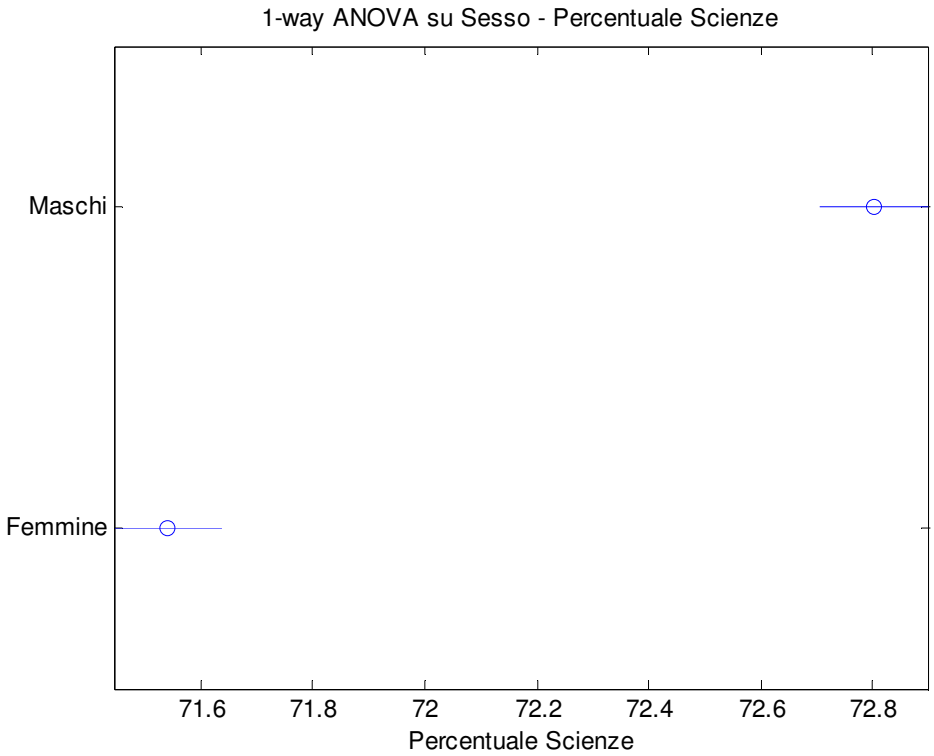
1.5.4. Fattore Sesso

Grandmean: 72.17

F-test: 8.419185e+001 - Liv. signif.: 0

Sesso 1: Femmine - Estimated mean: 71.54 +/- 0.10

Sesso 2: Maschi - Estimated mean: 72.80 +/- 0.10



Come per le valutazioni fatte per il caso generale e per la singola materia "Italiano" anche nelle Scienze i Maschi presentano una percentuale di risposte esatte maggiore rispetto alle Femmine, con una differenza di circa un punto percentuale.

2. Analisi ANOVA sui questionari di valutazione INVALSI, a.s. 2005/2006

Lo stesso tipo di analisi fin qui presentate è stato eseguito sui dati relativi all'anno scolastico 2005/2006 ma purtroppo in questo caso non è possibile presentare la medesima

valutazione grafico – analitica. In particolare non è stato possibile eseguire una corretta valutazione delle scuole in base all'indice di abilità stimato con il modello Item Response Theory (IRT). Nel corso di un'accurata analisi dei dati presenti nel Database INVALSI, sono stati, infatti, individuati, nel file SPSS originale relativo alle risposte fornite dagli studenti delle Scuole Medie Superiori ai quesiti di Scienze, un notevole numero di valori anomali per la variabile "mle", ossia l'indice di abilità, molto probabilmente causati da errori di digitazione avvenuti a monte, e non essendo stato possibile recuperare i dati di partenza si è dovuto procedere all'esclusione dall'analisi di tali unità.

Inoltre, non tutte le scuole cui è stato somministrato il questionario nell'a.s. 2005/2006 sono entrate a far parte delle analisi eseguite in quanto non è stato possibile recuperare alla fonte il database contenente tutte le informazioni geografico – anagrafiche di tali scuole, cosa che ha comportato ovviamente l'inserimento nell'analisi solo degli istituti che avevano partecipato all'indagine anche l'anno precedente e di cui quindi già si possedevano tali dati.

Pertanto ci limiteremo a presentare, di seguito, solo le valutazioni degli Istituti eseguite considerando come variabile responso la percentuale di risposte esatte fornite dagli studenti.

2.1. Percentuale di risposte esatte: Italiano, Matematica e Scienze

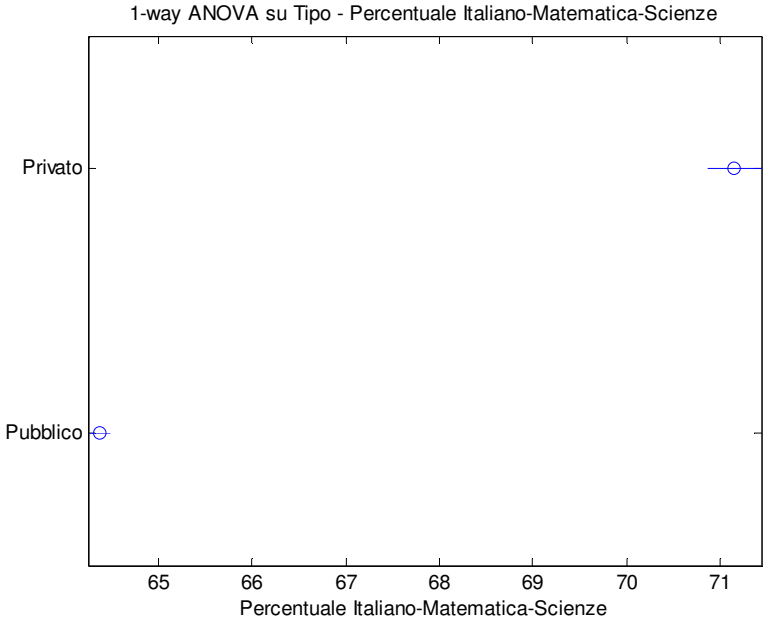
2.1.1. Fattore Tipo

Grandmean: 67.76

F-test: 4.738263e+002 - Liv. signif.: 0

Tipo 1: Pubblico - Estimated mean: 64.37 +/- 0.12

Tipo 2: Privato - Estimated mean: 71.15 +/- 0.29



Com'è possibile osservare dai valori presentati in tabella e dal grafico il fattore Tipo risulta sicuramente significativo indicando quindi l'esistenza di una differenza rilevante fra Scuole Pubbliche e Private, per le quali le percentuali di risposte esatte fornite si discostano mediamente di circa 7 punti percentuali in favore delle Scuole Private.

2.1.2. Fattore Regione

Grandmean: 64.74

F-test: 1.198785e+001 - Liv. signif.: 0

Regione 1: Valle d'Aosta - Estimated mean: 58.48 +/- 2.95

Regione 2: Piemonte - Estimated mean: 65.03 +/- 0.45

Regione 3: Liguria - Estimated mean: 65.98 +/- 0.78

Regione 4: Lombardia - Estimated mean: 63.68 +/- 0.30

Regione 5: Trentino - Estimated mean: 62.77 +/- 1.05

Regione 6: Veneto - Estimated mean: 63.45 +/- 0.41

Regione 7: Friuli - Estimated mean: 65.72 +/- 0.86

Regione 8: Emilia - Estimated mean: 64.54 +/- 0.48

Regione 9: Toscana - Estimated mean: 65.18 +/- 0.50

Regione 10: Umbria - Estimated mean: 64.01 +/- 0.94

Regione 11: Marche - Estimated mean: 65.05 +/- 0.69

Regione 12: Lazio - Estimated mean: 66.73 +/- 0.37

Regione 13: Abruzzo - Estimated mean: 65.94 +/- 0.74

Regione 14: Molise - Estimated mean: 63.94 +/- 1.19

Regione 15: Campania - Estimated mean: 68.04 +/- 0.32

Regione 16: Puglia - Estimated mean: 65.94 +/- 0.42

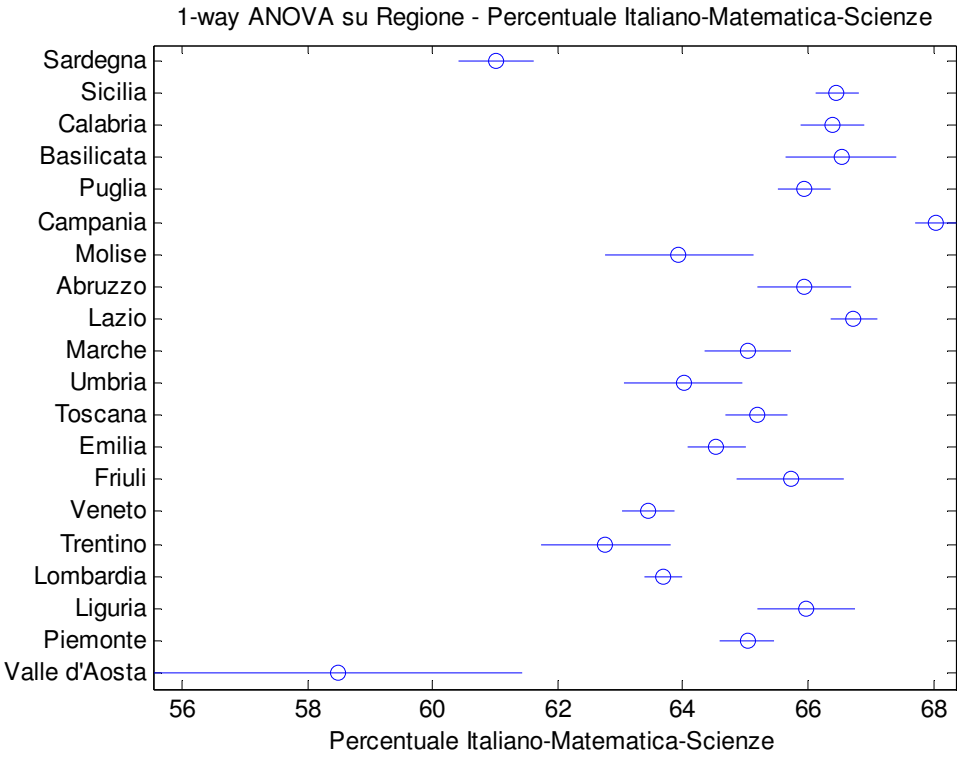
Regione 17: Basilicata - Estimated mean: 66.54 +/- 0.89

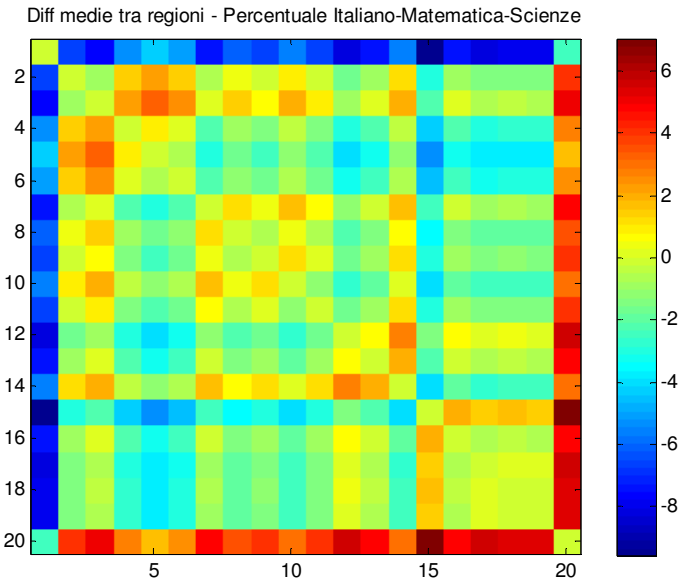
Regione 18: Calabria - Estimated mean: 66.39 +/- 0.51

Regione 19: Sicilia - Estimated mean: 66.47 +/- 0.34

Regione 20: Sardegna - Estimated mean: 61.01 +/- 0.60

Anche il fattore Regione è significativo ed inoltre si conferma, per l'anno scolastico 2005/2006, un andamento crescente, anche se meno netto rispetto all'anno scolastico 2004/2005, del valore delle percentuali di risposte esatte fornite man mano che ci si sposta dalle regioni del Nord a quelle del Centro e poi del Sud, fermo restando la dovuta cautela indispensabile, come accennato precedentemente, nell'attuare questo tipo di confronti.





2.1.3. Fattore Ordine

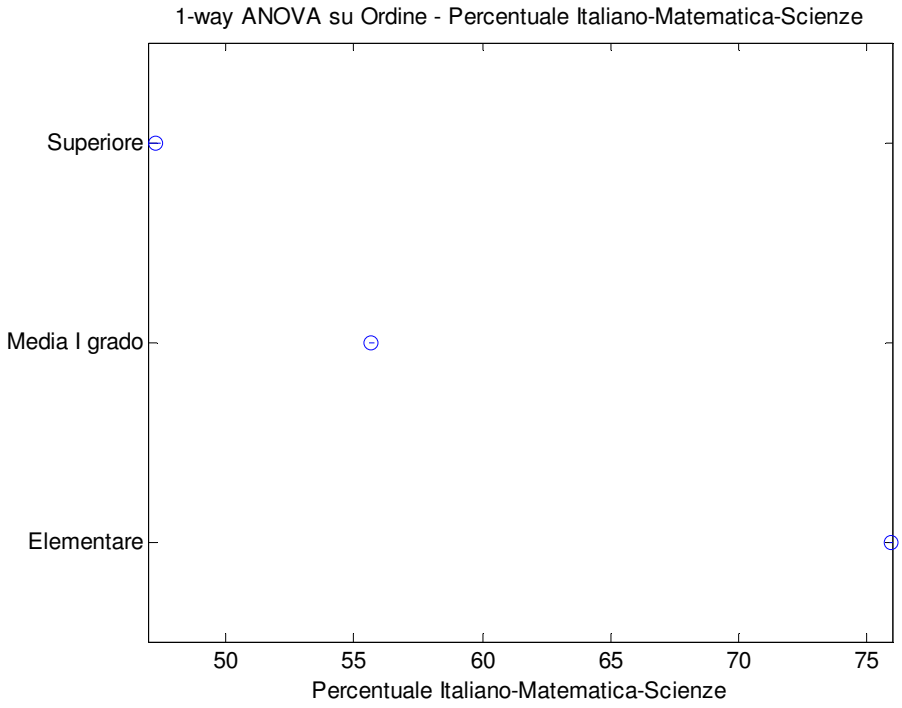
Grandmean: 59.61

F-test: 1.529123e+004 - Liv. signif.: 0

Ordine 1: Elementare - Estimated mean: 75.96 +/- 0.09

Ordine 2: Media I grado - Estimated mean: 55.67 +/- 0.10

Ordine 3: Superiore - Estimated mean: 47.22 +/- 0.21



La valutazione in funzione dell' Ordine scolastico mostra la presenza di differenze significative fra le Scuole Elementari, Medie Inferiori e Medie Superiori, con un distacco fra le prime e le ultime che sfiora il 30%.

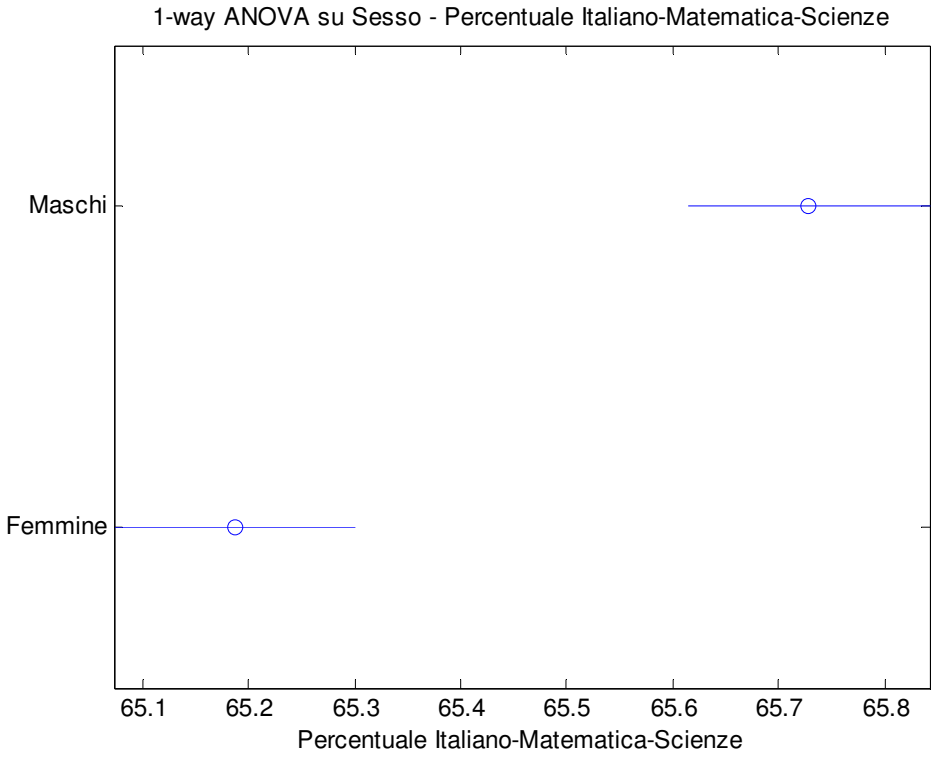
2.1.4. Fattore Sesso

Grandmean: 65.46

F-test: 1.133322e+001 - Liv. signif.: 7.623208e-004

Sesso 1: Femmine - Estimated mean: 65.19 +/- 0.11

Sesso 2: Maschi - Estimated mean: 65.73 +/- 0.11



La significatività del fattore sesso è confermata anche se il divario tra i due livelli non supera, mediamente, l'1%.

2.2. Percentuale di risposte esatte: Italiano

Spostandosi dal caso generale all'analisi per la singola disciplina "Italiano" è possibile osservare una generale, lieve, diminuzione delle percentuali di successo pari a circa l'1%.

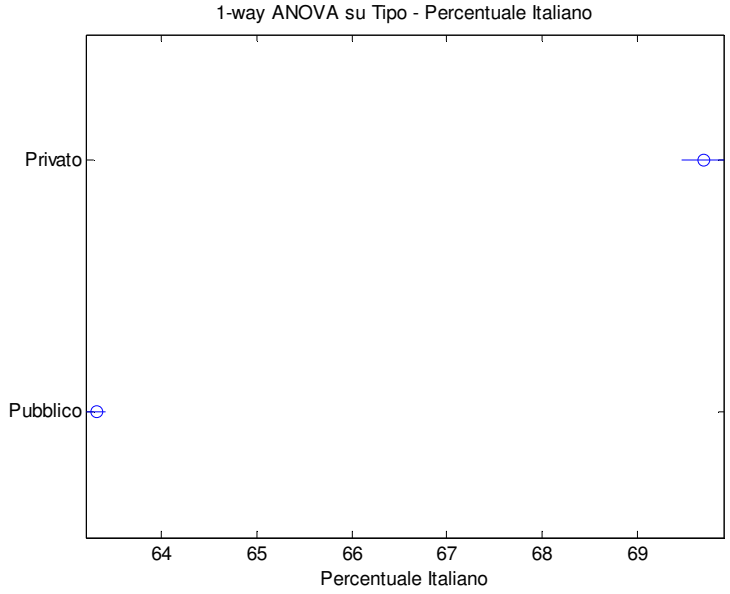
2.2.1. Fattore Tipo

Grandmean: 66.50

F-test: 6.827928e+002 - Liv. signif.: 0

Tipo 1: Pubblico - Estimated mean: 63.31 +/- 0.09

Tipo 2: Privato - Estimated mean: 69.69 +/- 0.23



Per la singola disciplina Italiano si osserva una diminuzione generalizzata dei valori medi di circa un punto percentuale ma è confermato l'andamento osservato per il caso generale in cui la valutazione è stata eseguita considerando le tre materie nel loro complesso.

2.2.2. Fattore Regione

Grandmean: 63.73

F-test: 1.091743e+001 - Liv. signif.: 0

Regione 1: Valle d'Aosta - Estimated mean: 59.18 +/- 2.33

Regione 2: Piemonte - Estimated mean: 64.15 +/- 0.35

Regione 3: Liguria - Estimated mean: 64.42 +/- 0.61

Regione 4: Lombardia - Estimated mean: 63.51 +/- 0.23

Regione 5: Trentino - Estimated mean: 61.85 +/- 0.83

Regione 6: Veneto - Estimated mean: 63.06 +/- 0.33

Regione 7: Friuli - Estimated mean: 64.66 +/- 0.68

Regione 8: Emilia - Estimated mean: 63.78 +/- 0.38

Regione 9: Toscana - Estimated mean: 63.57 +/- 0.40

Regione 10: Umbria - Estimated mean: 63.44 +/- 0.74

Regione 11: Marche - Estimated mean: 64.46 +/- 0.54

Regione 12: Lazio - Estimated mean: 65.63 +/- 0.30

Regione 13: Abruzzo - Estimated mean: 64.62 +/- 0.59

Regione 14: Molise - Estimated mean: 63.10 +/- 0.94

Regione 15: Campania - Estimated mean: 66.31 +/- 0.25

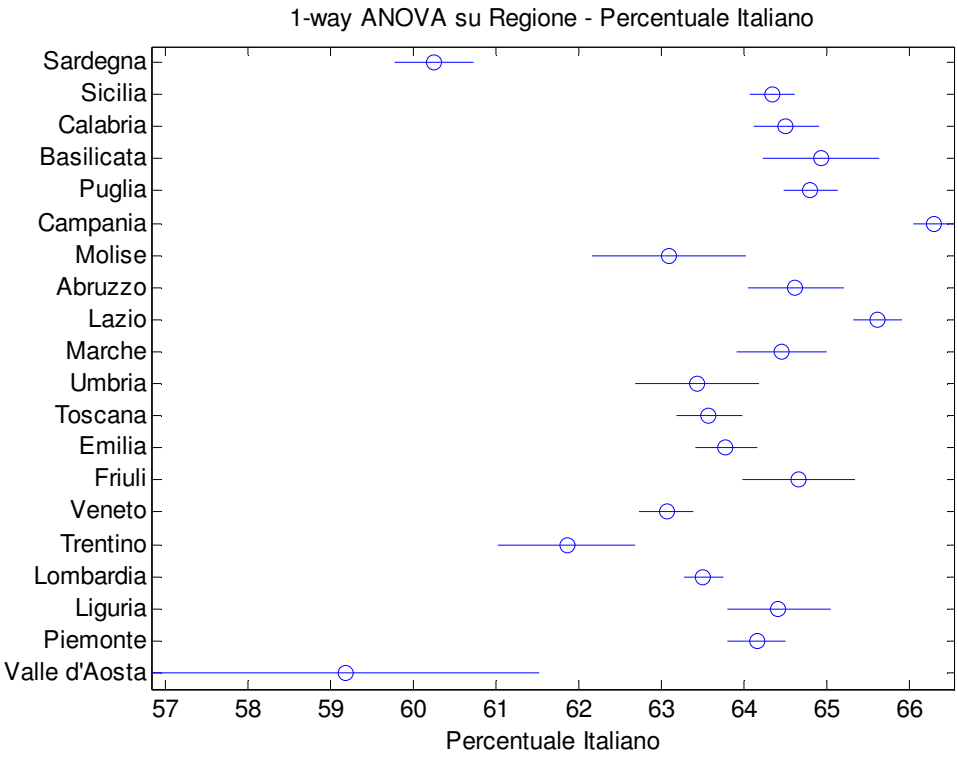
Regione 16: Puglia - Estimated mean: 64.81 +/- 0.34

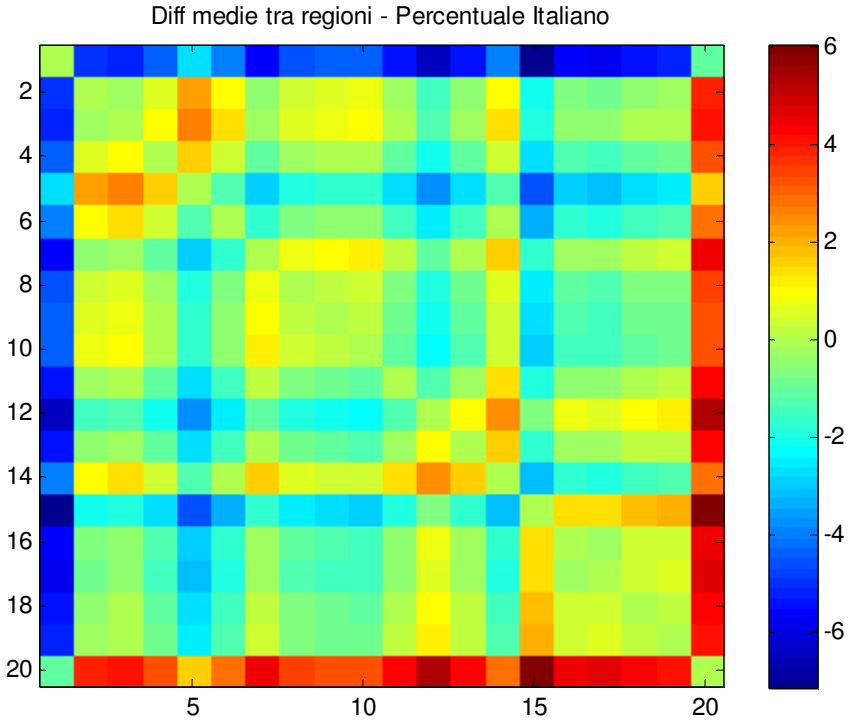
Regione 17: Basilicata - Estimated mean: 64.94 +/- 0.70

Regione 18: Calabria - Estimated mean: 64.51 +/- 0.40

Regione 19: Sicilia - Estimated mean: 64.35 +/- 0.27

Regione 20: Sardegna - Estimated mean: 60.25 +/- 0.48





I divari fra le varie regioni risultano significativi e mediamente confermano, ancora una volta, la presenza di percentuali di risposte esatte maggiori, anche se non eccessivamente, per gli Istituti del Cento e Sud Italia.

2.2.3. Fattore Ordine

Grandmean: 60.67

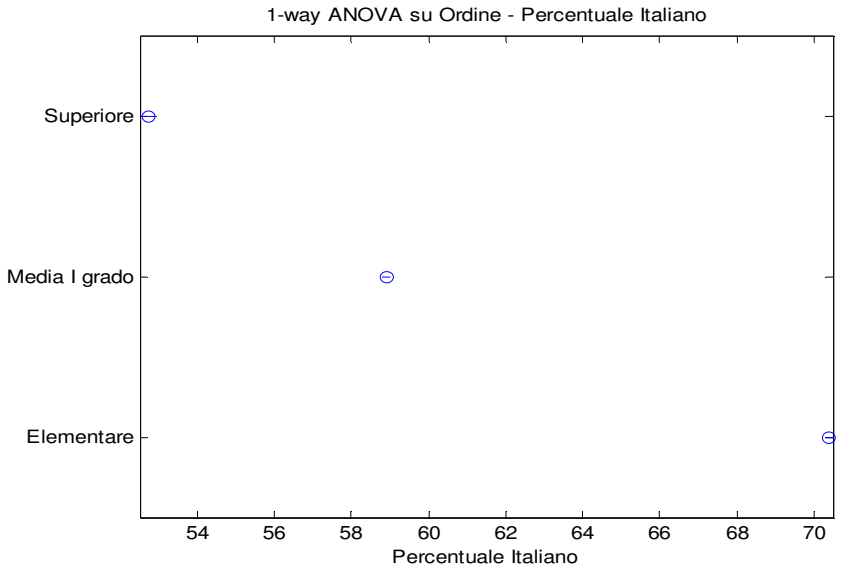
F-test: 4.331304e+003 - Liv. signif.: 0

Ordine 1: Elementare - Estimated mean: 70.39 +/- 0.10

Ordine 2: Media I grado - Estimated mean: 58.90 +/- 0.11

Ordine 3: Superiore - Estimated mean: 52.74 +/- 0.23

Rispetto al caso generale è possibile osservare, per la singola disciplina "Italiano" una riduzione dei margini esistenti fra i tre livelli del fattore Ordine (Elementari, Medie Inferiori, Medie Superiori) che tuttavia resta pienamente significativo.



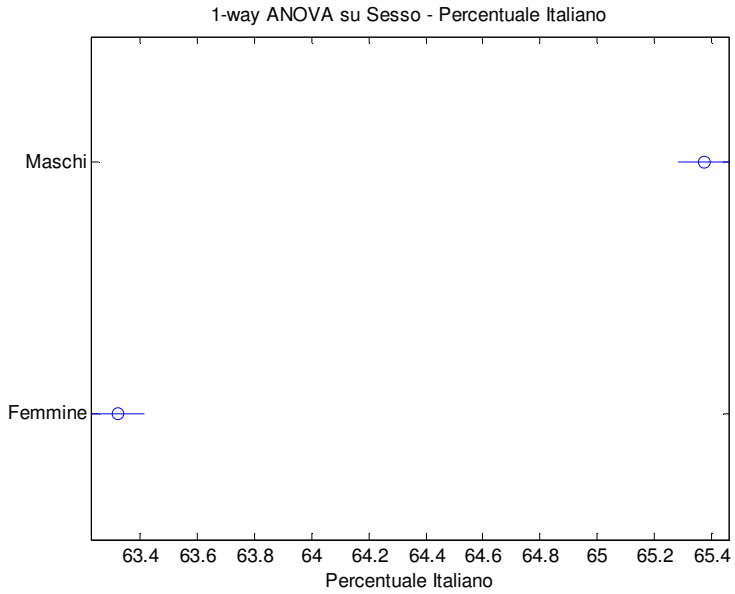
2.2.4. Fattore Sesso

Grandmean: 64.35

F-test: 2.545826e+002 - Liv. signif.: 0

Sesso 1: Femmine - Estimated mean: 63.32 +/- 0.09

Sesso 2: Maschi - Estimated mean: 65.37 +/- 0.09



Anche per l'Italiano è confermata la significatività del fattore Sesso e si osserva che pur essendo diminuita la media generale di circa l'1% è al tempo stesso aumentato il divario fra i due livelli, Maschi e Femmine, che si assesta a circa 2 punti percentuali in favore dei primi.

2.3. Percentuale di risposte esatte: Matematica

Le percentuali di risposte esatte per i quesiti di Matematica risultano ancora le più basse tra le tre materie considerate (circa il 59.8% in media contro il 64.3% generalmente su tutte le materie).

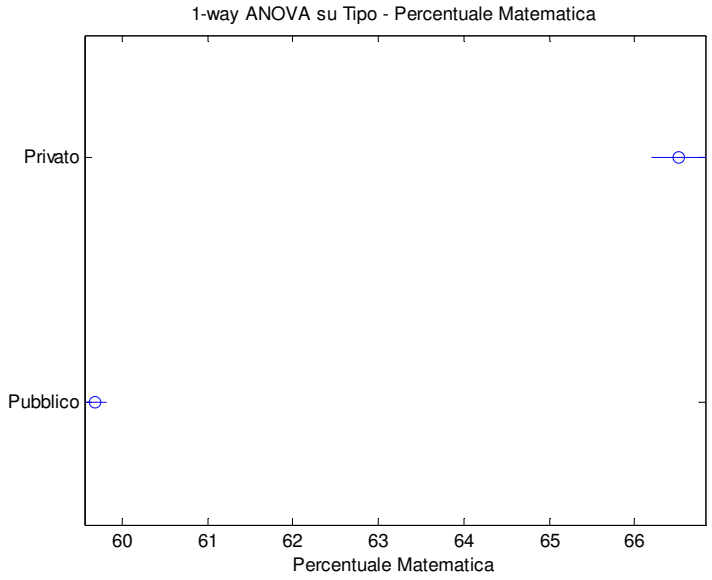
2.3.1. Fattore Tipo

Grandmean: 63.10

F-test: 3.992760e+002 - Liv. signif.: 0

Tipo 1: Pubblico - Estimated mean: 59.69 +/- 0.13

Tipo 2: Privato - Estimated mean: 66.51 +/- 0.31



Gli Istituti Privati conservano una percentuale di risposte esatte significativamente maggiore rispetto alle Scuole Pubbliche, con un divario presso che costante rispetto al caso generale.

2.3.2. Fattore Regione

Grandmean: 59.93

F-test: 2.420139e+001 - Liv. signif.: 0

Regione 1: Valle d'Aosta - Estimated mean: 50.83 +/- 3.20

Regione 2: Piemonte - Estimated mean: 59.79 +/- 0.48

Regione 3: Liguria - Estimated mean: 60.65 +/- 0.84

Regione 4: Lombardia - Estimated mean: 57.64 +/- 0.32

Regione 5: Trentino - Estimated mean: 57.80 +/- 1.14

Regione 6: Veneto - Estimated mean: 57.70 +/- 0.45

Regione 7: Friuli - Estimated mean: 60.16 +/- 0.93

Regione 8: Emilia - Estimated mean: 59.76 +/- 0.51

Regione 9: Toscana - Estimated mean: 60.20 +/- 0.55

Regione 10: Umbria - Estimated mean: 59.37 +/- 1.02

Regione 11: Marche - Estimated mean: 60.54 +/- 0.74

Regione 12: Lazio - Estimated mean: 61.92 +/- 0.41

Regione 13: Abruzzo - Estimated mean: 61.39 +/- 0.80

Regione 14: Molise - Estimated mean: 60.11 +/- 1.29

Regione 15: Campania - Estimated mean: 64.59 +/- 0.35

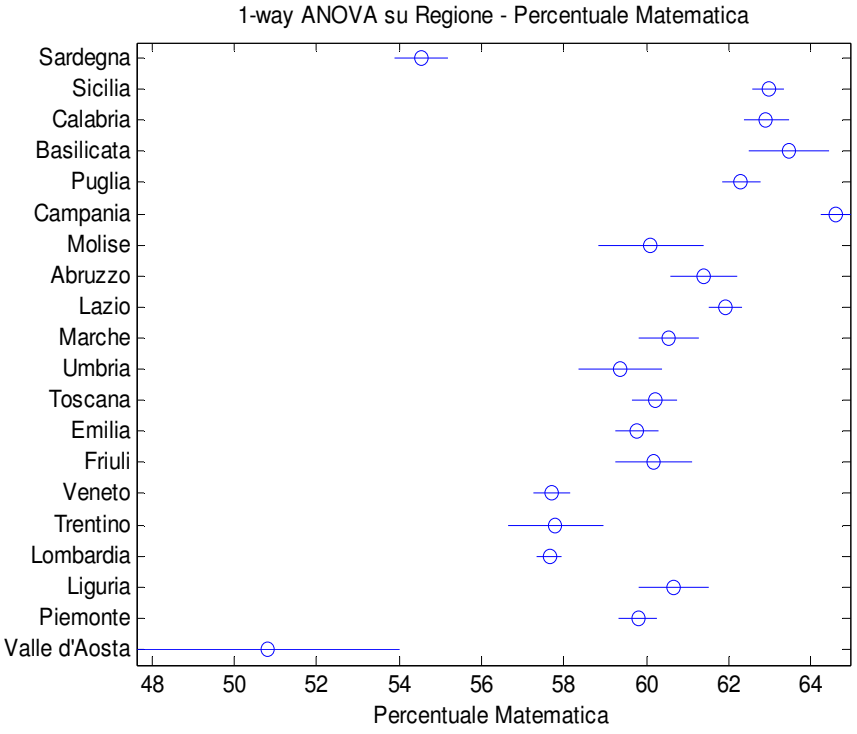
Regione 16: Puglia - Estimated mean: 62.29 +/- 0.46

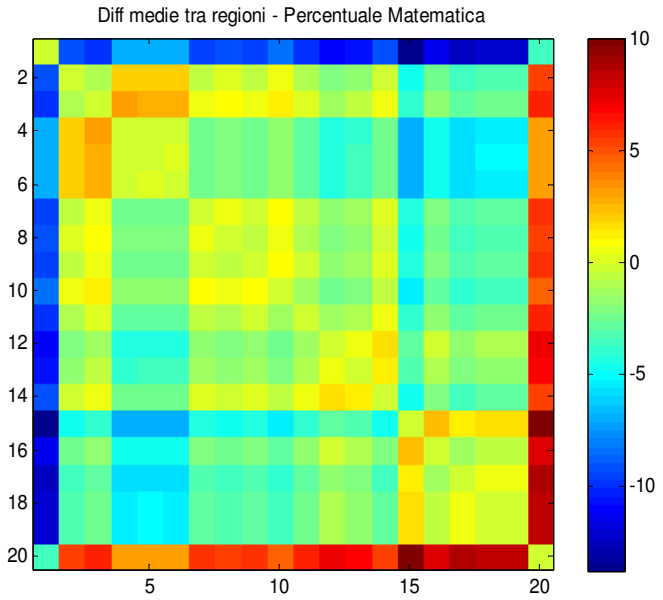
Regione 17: Basilicata - Estimated mean: 63.45 +/- 0.96

Regione 18: Calabria - Estimated mean: 62.91 +/- 0.55

Regione 19: Sicilia - Estimated mean: 62.95 +/- 0.37

Regione 20: Sardegna - Estimated mean: 54.55 +/- 0.65





Per quanto attiene il fattore Regione sono confermati gli andamenti osservati per il caso generale.

2.3.3. Fattore Ordine

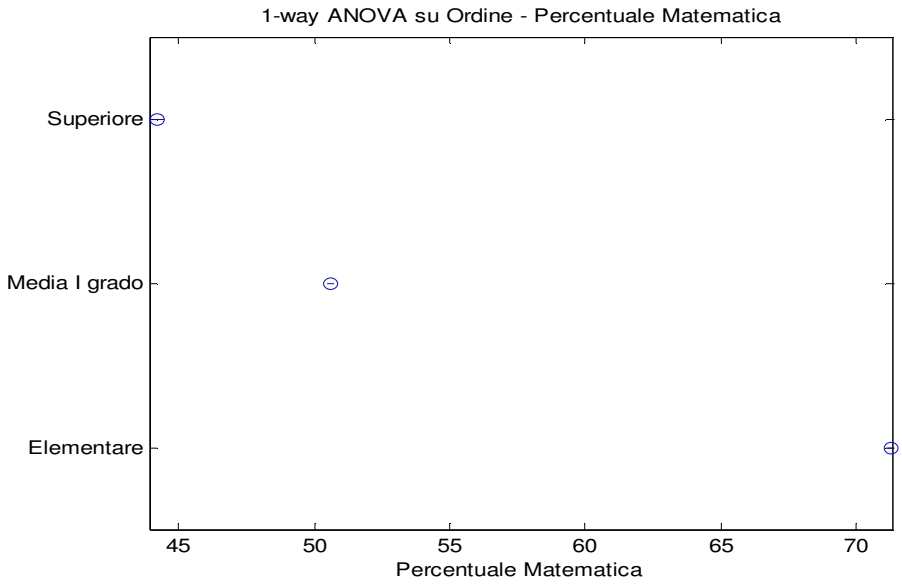
Grandmean: 55.36

F-test: 9.088545e+003 - Liv. signif.: 0

Ordine 1: Elementare - Estimated mean: 71.26 +/- 0.11

Ordine 2: Media I grado - Estimated mean: 50.63 +/- 0.13

Ordine 3: Superiore - Estimated mean: 44.19 +/- 0.27



I divari fra le Scuole Elementari e Medie, Inferiori e Superiori, assumono valori prossimi a quelli riscontrati nel caso generale (superiori al 20%) sempre a favore delle Scuole Elementari.

2.3.4. Fattore Sesso

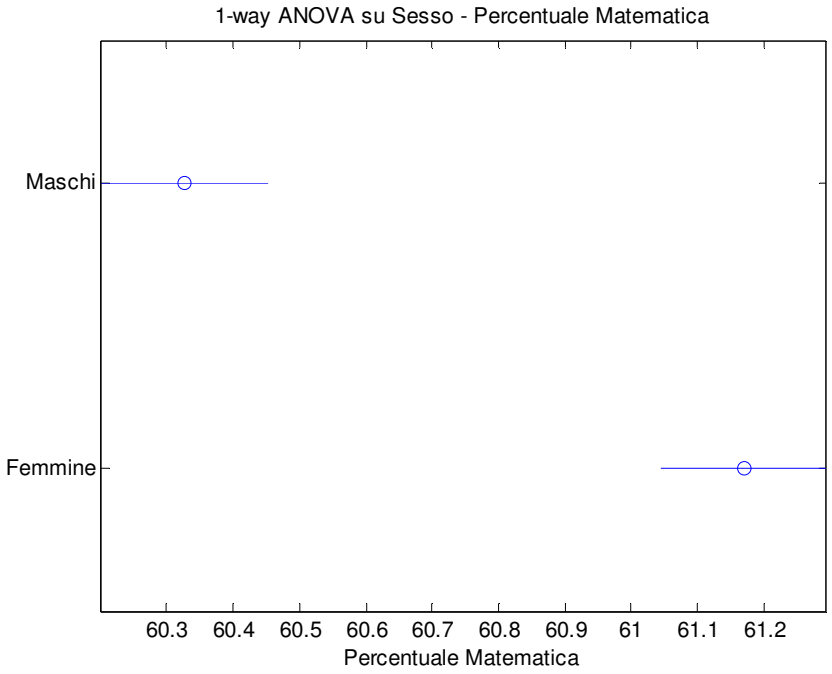
Grandmean: 60.75

F-test: 2.275813e+001 - Liv. signif.: 1.846206e-006

Sesso 1: Femmine - Estimated mean: 61.17 +/- 0.12

Sesso 2: Maschi - Estimated mean: 60.33 +/- 0.13

Per le prove di Matematica riscontriamo nuovamente un'inversione di tendenza, registrando, per le Femmine, percentuali di risposte esatte leggermente (0.8%) ma significativamente maggiori rispetto ai Maschi.



2.4. Percentuale di risposte esatte: Scienze

Le percentuali di risposte esatte fornite dagli studenti ai quesiti di Scienze risultano costantemente più alte rispetto alle altre materie.

2.4.1. Fattore Tipo

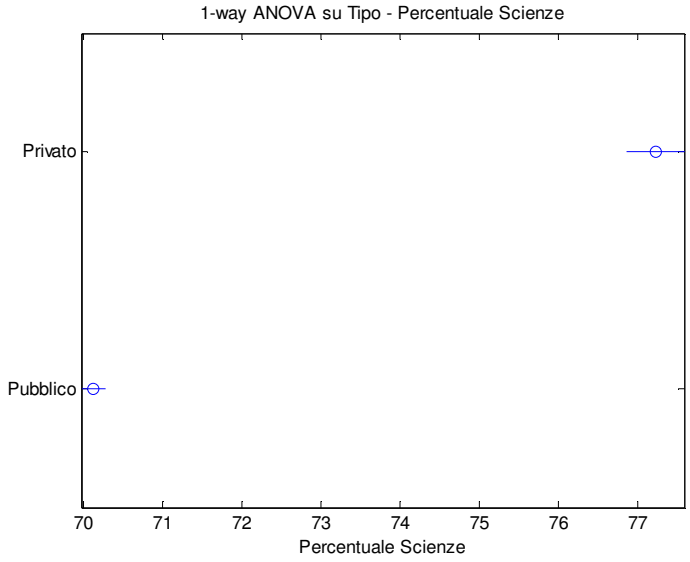
Grandmean: 73.67

F-test: 3.165299e+002 - Liv. signif.: 0

Tipo 1: Pubblico - Estimated mean: 70.12 +/- 0.15

Tipo 2: Privato - Estimated mean: 77.22 +/- 0.37

Per la singola disciplina "Scienze" è ancora una volta confermato l'andamento generale con un divario fra Scuole Pubbliche e Scuole Private, valutato intorno al 7% a favore di queste ultime, leggermente superiore sia al valore ottenuto per il caso generale che a quello riscontrabile per le singole materie.



2.4.2. Fattore Regione

Grandmean: 70.59

F-test: 5.065718e+000 - Liv. signif.: 2.835066e-012

Regione 1: Valle d'Aosta - Estimated mean: 65.41 +/- 3.78

Regione 2: Piemonte - Estimated mean: 71.14 +/- 0.57

Regione 3: Liguria - Estimated mean: 72.75 +/- 1.00

Regione 4: Lombardia - Estimated mean: 69.90 +/- 0.38

Regione 5: Trentino - Estimated mean: 68.81 +/- 1.34

Regione 6: Veneto - Estimated mean: 69.54 +/- 0.53

Regione 7: Friuli - Estimated mean: 72.28 +/- 1.10

Regione 8: Emilia - Estimated mean: 70.15 +/- 0.61

Regione 9: Toscana - Estimated mean: 71.72 +/- 0.65

Regione 10: Umbria - Estimated mean: 69.25 +/- 1.20

Regione 11: Marche - Estimated mean: 70.09 +/- 0.88

Regione 12: Lazio - Estimated mean: 72.67 +/- 0.48

Regione 13: Abruzzo - Estimated mean: 71.89 +/- 0.95

Regione 14: Molise - Estimated mean: 68.68 +/- 1.52

Regione 15: Campania - Estimated mean: 73.21 +/- 0.41

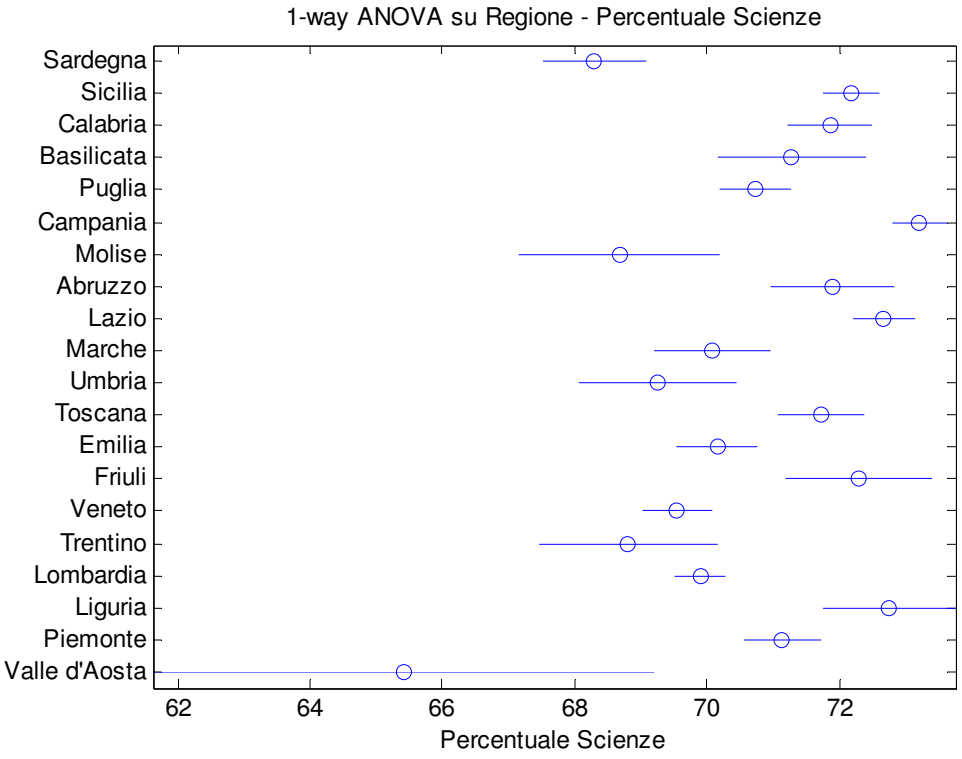
Regione 16: Puglia - Estimated mean: 70.73 +/- 0.54

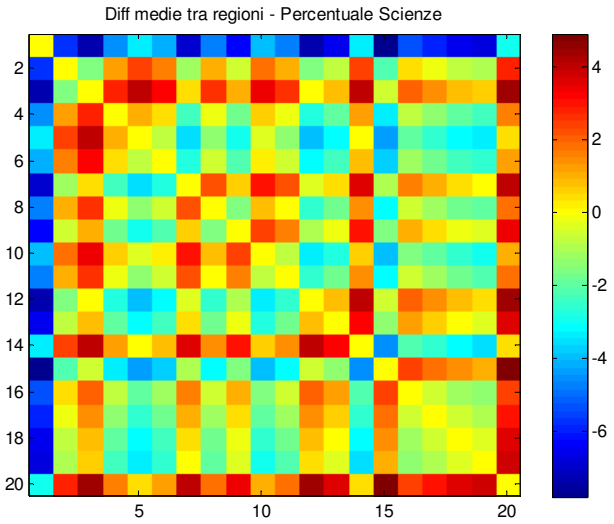
Regione 17: Basilicata - Estimated mean: 71.28 +/- 1.13

Regione 18: Calabria - Estimated mean: 71.85 +/- 0.65

Regione 19: Sicilia - Estimated mean: 72.17 +/- 0.44

Regione 20: Sardegna - Estimated mean: 68.30 +/- 0.77





Il fattore Regione risulta, anche in questo caso, significativo ed inoltre, come è possibile osservare più facilmente dai grafici, evidenzia la possibilità di suddividere, idealmente, le regioni in due gruppi, ossia uno formato da quelle per cui le percentuali di risposte esatte sono inferiori al valore della media generale (70.59%) e l'altro in cui confluiscono le regioni che invece tale valore lo superano.

2.4.3. Fattore Ordine

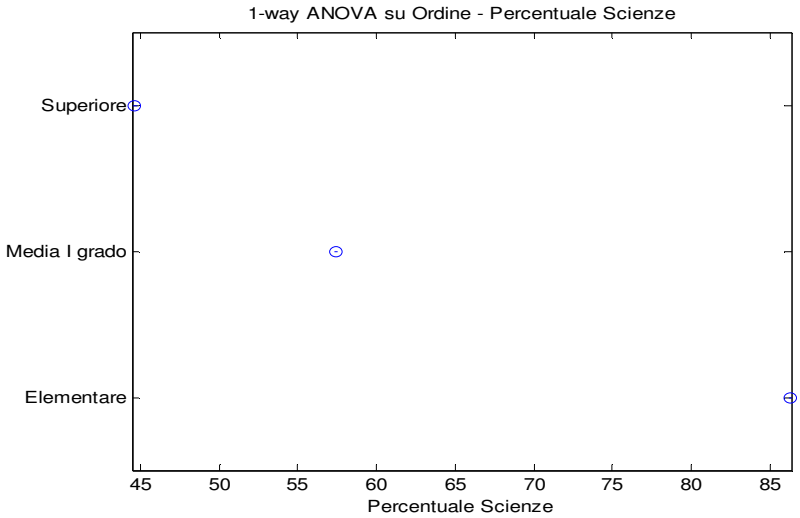
Grandmean: 62.80

F-test: 4.337348e+004 - Liv. signif.: 0

Ordine 1: Elementare - Estimated mean: 86.28 +/- 0.08

Ordine 2: Media I grado - Estimated mean: 57.47 +/- 0.09

Ordine 3: Superiore - Estimated mean: 44.65 +/- 0.18



Si conferma la significatività del fattore Ordine e l'andamento omogeneo per materia, osservando che nel caso della singola disciplina "Scienze" registriamo il massimo divario fra Scuole Elementari e Scuole Medie Superiori, superiore al 40%.

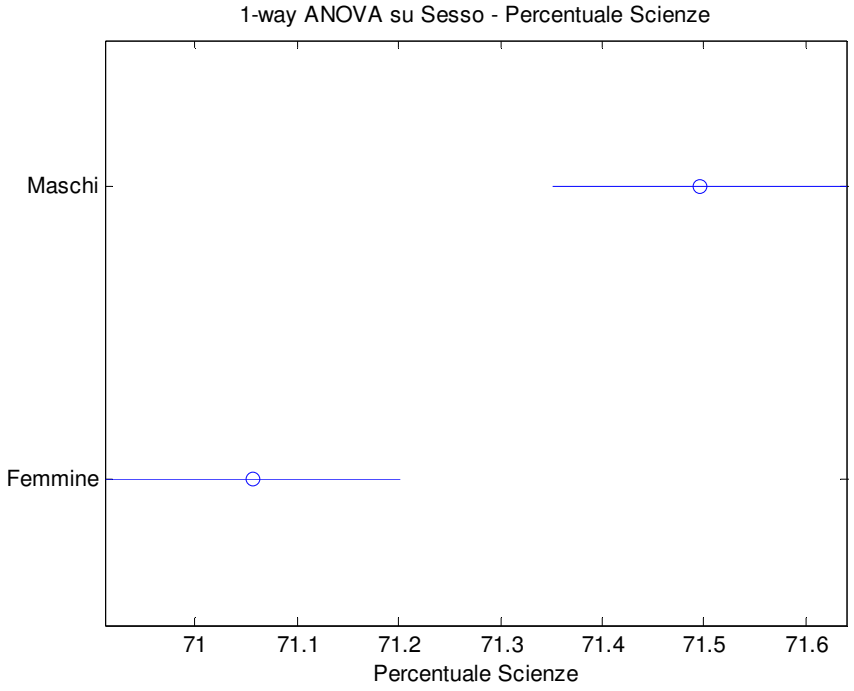
2.4.4 Fattore Sesso

Grandmean: 71.28

F-test: 4.621212e+000 - Liv. signif.: 3.158719e-002

Sesso 1: Femmine - Estimated mean: 71.06 +/- 0.14

Sesso 2: Maschi - Estimated mean: 71.50 +/- 0.14



Nelle Scienze il rendimento dei Maschi ritorna ad essere significativamente migliore rispetto a quello delle Femmine, con una differenza che si attesta attorno allo 0.5%, ossia la minore fin qui registrata.

3. Cluster Analysis

I risultati ottenuti dalle prime analisi generali effettuate con metodologia ANOVA sulle risposte fornite dagli studenti ai questionari di Valutazione INVALSI, hanno evidenziato la significatività di tutti i fattori considerati rispetto alle abilità dello studente aprendo così la strada ad ulteriori analisi volte ad indagare la possibilità di sfruttare queste significatività per cercare di ridurre il numero di variabili in campo.

Esistono diversi metodi di analisi atti a perseguire tale scopo, primo fra tutti probabilmente l'Analisi Fattoriale, progenitore dei metodi di analisi multivariata, o ancora l'Analisi Discriminante, ma in entrambi i casi esistono assunzioni di base spesso troppo specifiche o restrittive che non ne consentono o, per meglio dire, non ne suggeriscono un'applicazione immediata.

È sembrato quindi naturale, in questa fase del progetto, puntare l'attenzione sull'individuazione di eventuali omogeneità riscontrabili nei livelli dei fattori considerati affidandosi ad opportune tecniche di clustering per individuare gruppi omogenei effettuando così una riduzione del numero di unità di analisi.

In letteratura con il termine Cluster Analysis, o Analisi dei raggruppamenti si è soliti indicare varie tecniche di analisi multivariata dei dati volte ad assegnare unità di analisi a gruppi non definiti in partenza.

L'Analisi dei raggruppamenti è, quindi, presentata, solitamente, come un metodo essenzialmente esplorativo proprio perché non si assume alcuna classificazione a priori ma ci si attende che le relazioni fra le unità siano evidenziate dall'analisi stessa.

L'idea principale alla base dell'analisi dei gruppi è molto semplice: individuare k gruppi tali che gli oggetti appartenenti allo stesso gruppo siano tra loro simili (gruppi omogenei al loro interno) mentre gli oggetti appartenenti a gruppi differenti siano dissimili (gruppi tra loro eterogenei).

Per avere la possibilità di individuare un gruppo, o più nello specifico, per poter effettuare una distinzione tra vari tipi di gruppi, occorre ricordarne alcune caratteristiche, come, ad esempio, la dimensione, la densità, la forma e la separazione.

Per quel che attiene alla dimensione di un gruppo essa può essere definita, in relazione del suo raggio, solo nei casi in cui il gruppo ha forma regolare, cosa che accade ad esempio per una ipersfera. Un gruppo, invece, si dice denso quando in esso è possibile trovare un agglomerato di punti più fitto rispetto a quello presente in altri gruppi; a questo

proposito, pur non essendoci una misura assoluta della densità, si è soliti identificarla con la varianza.

La forma di un gruppo è data, poi, dalla disposizione dei punti-unità nello spazio, disposizione che consente di identificare eventuali regolarità nella struttura dei dati, mentre la separazione è il modo in cui i gruppi, disposti nello spazio, risultano distinti o, al contrario, si sovrappongono.

Il primo ad affrontare lo studio della classificazione da un punto di vista statistico fu, verso la fine XIX secolo, il matematico e statistico Karl Pearson ma è solo a partire dalla seconda metà del secolo scorso che, grazie agli sviluppi delle tecnologie di calcolo, si è iniziata a porre una maggiore attenzione agli aspetti algoritmici delle tecniche di raggruppamento.

Svariati sono i campi di applicazione dell'Analisi dei gruppi e fra questi è vasto l'apporto fornito a studi di sociologia, economia, medicina, antropologia e di varie altre discipline; così come diversi sono i motivi che possono giustificarne l' utilizzo, in quanto, la cluster analysis ha rilevanza, ad esempio, per ridurre i dati in una forma grafica che sia semplice e parsimoniosa, per generare ipotesi di ricerca, identificare i tipi, stratificare popolazioni da sottoporre a campionamento, trovare dati validi per sostituire valori mancanti o una modalità con cui confrontare una risposta elusiva ed ancora stimare la probabilità che si verifichi un certo evento in campioni di numerosità esigua.

3.1. Fasi operative della Cluster Analysis

Un'analisi dei gruppi si basa essenzialmente su di una serie di decisioni preliminari, quali:

1. identificazione delle variabili di classificazione;
2. selezione di una misura di prossimità tra le unità;
3. scelta della tecnica di raggruppamento più adatta alla struttura ed all'obiettivo dell'analisi;
4. identificazione del numero di gruppi entro i quali classificare le unità;
5. scelta, facoltativa, di utilizzare in modo integrato altre tecniche di analisi multivariata per la lettura dei risultati dell'analisi.

La scelta delle variabili dipende, in modo quasi esclusivo, dalle conoscenze del ricercatore rispetto al fenomeno oggetto di studio, da cui derivano le finalità assegnate all'analisi dei raggruppamenti.

Per quanto concerne il progetto FINVALI, le analisi preliminari hanno individuato, come accennato precedentemente, la presenza di differenze significative rispetto a diversi fattori considerati, ossia Tipo di scuola (Pubblica o Privata), Regione, Ordine scolastico (Elementari, Medie, Superiori), Sesso, ma in questa prima fase dell'analisi dei raggruppamenti si è deciso di puntare l'attenzione soprattutto su due di tali fattori: Regione ed Ordine scolastico.

Il concetto di cluster è, a ben vedere, inscindibile da quello di prossimità (esprimibile in termini di similarità o distanza) tra una coppia di unità statistiche e la scelta di una metrica che esprima, quindi, tale misura è uno degli elementi necessari per poter effettuare l'Analisi dei gruppi.

Attualmente si dispone di molteplici soluzioni alternative per effettuare una Analisi dei gruppi e quasi tutte le tecniche hanno come punto di partenza la creazione di una matrice di prossimità tra le diverse unità statistiche, matrice, questa, che può essere ottenuta da una valutazione soggettiva delle prossimità ottenuta attraverso il giudizio diretto dei soggetti, oppure tramite opportune trasformazioni della matrice dei **X**, (*n x p*).

In linea generale si può dire che una matrice di prossimità **P** è una matrice quadrata (*n x n*) i cui elementi sono misure di similarità o dissimilarità tra i membri dell'insieme delle unità statistiche. Nel momento in cui ci si sposta dalle relazioni tra variabili alle misure di prossimità gli oggetti principali di riferimento diventano i *profili* della matrice dei dati.

Il generico profilo *i* di **X** altro non è che un vettore $\mathbf{x}^i = [x_{i1} x_{i2} x_{i3} .... x_{ip}]$ il che significa che possiamo scrivere:

$$\mathrm{X} = \begin{bmatrix} x^1 \\ x^2 \\ \dots \\ x^n \end{bmatrix}$$

Il questo caso, quindi, avremo che una matrice di prossimità **P** è una matrice quadrata (*n x n*) che sintetizza il grado di similarità, o dissimilarità, tra le possibili coppie di profili ($\mathbf{x}^r$, $\mathbf{x}^s$) di **X**, (con *r, s = 1,2,3,…,n*), e che il generico elemento p^{rs} denota la misura di prossimità tra l'oggetto *r* e l'oggetto *s* (o equivalentemente tra il profilo $\mathbf{x}^r$ e il profilo $\mathbf{x}^s$).

Fin qui si è parlato, quasi indifferentemente, di somiglianza o di distanza ma in realtà va chiarito che questi, pur essendo concetti analoghi, sono tuttavia opposti, ed infatti quanto più è minore la distanza fra due unità tanto più risulterà maggiore la loro similarità.

Due esempi classici di misure di prossimità sono:

- Il coefficiente di correlazione tra due profili $\mathbf{x}^r$ e $\mathbf{x}^s$ come misura di similarità tra l'oggetto *r* e l'oggetto *s*;
- La distanza euclidea calcolata su due profili $\mathbf{x}^r$ e $\mathbf{x}^s$ come misura di dissimilarità tra l'oggetto *r* e l'oggetto *s*.

In generale, comunque, le misure di similarità e dissimilarità, per essere tali, devono soddisfare alcune proprietà formali.

- Proprietà delle misure di similarità

  Dati due oggetti r ed s la misura di prossimità p_{rs} è una misura di similarità se soddisfa le seguenti condizioni:

  1. $0 \leq p_{rs} \leq 1$ per tutti gli oggetti *r* ed *s*;
  2. $p_{rs} = 1$ se e solo se *r* ed *s* sono identici;
  3. $p_{rs} = p_{sr}$.

La misura di similarità più comune è il coefficiente di correlazione di Pearson $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$. In realtà però questo coefficiente varia tra *-1* ed *1* per cui non dovrebbe essere preso in considerazione in quanto non soddisfa la prima condizione. A questo tuttavia si rimedia utilizzando, in alternativa, il suo valore assoluto o aggiungendo *1,0* al valore del coefficiente e dividendo poi il risultato per *2*. Particolari tipi di misure di associazione che, come il coefficiente di correlazione di Pearson, soddisfano la seconda e la terza condizione ma non la prima, sono chiamate similarità di tipo Q (*Q-type*).

- Proprietà delle misure di dissimilarità

  Dati due oggetti r ed s la misura di prossimità p_{rs} è una misura di dissimilarità se soddisfa le seguenti condizioni:

  1. $p_{rs} \leq 0$ per tutti gli oggetti *r* ed *s*;
  2. $p_{rs} = 0$ se e solo se *r* ed *s* sono identici;;
  3. $p_{rs} = p_{sr}$.

La misura di dissimilarità più comunemente utilizzata è la distanza Euclidea.

Se si indicano, rispettivamente, con $(x_{r1}, x_{r2}, ..., x_{rp})$ e $(x_{s1}, x_{s2}, ..., x_{sp})$ i due profili di **X**, $\mathbf{x}^r$ ed $\mathbf{x}^s$, la distanza Euclidea al quadrato, d_{rs}^2, tra i due profili è data dalla somma dei quadrati delle differenze fra i singoli elementi dei profili, ossia: $d_{rs}^2 = \sum_{j=1}^{p} (x_{rj} - x_{sj})^2$, che espressa in forma vettoriale diventa:

$d_{rs}^2 = (\mathbf{x}^r - \mathbf{x}^s)' (\mathbf{x}^r - \mathbf{x}^s)$, con r,s=1, 2, ….., n.

Se al posto di **X** si utilizza la matrice dei dati corretti dalla media $\mathbf{X}^C$ il valore di d_{rs}^2 resta ovviamente immutato, mentre se si utilizza la matrice dei dati standardizzati $\mathbf{X}^S$ (cosa che può rendersi necessaria vista l'elevata sensibilità della distanza Euclidea nei confronti delle scale di misurazione) allora $d_{r_S s_S}^2 = \sum \left(\frac{1}{s_j^2} \right) (x_{rj} - x_{sj})^2$.

Un'altra misura di dissimilarità spesso utilizzata è la distanza Manhattan o City Block. In questo caso, dati due profili $\mathbf{x}^r$ ed $\mathbf{x}^s$ di X la distanza b_{rs} tra $\mathbf{x}^r$ ed $\mathbf{x}^s$ è definita dalla formula: $b_{rs} = \sum_{j=1}^{p} | x_{rj} - x_{sj} |$. Questa distanza si basa sui valori assoluti delle differenze fra le coordinate e fu chiamata Manhattan perché indica l'ipotetica distanza calcolabile se ci si spostasse da un punto all'altro di questa città avente una tipica conformazione a griglia.

E' sempre possibile, comunque, costruire misure di similarità a partire da misure di dissimilarità in quanto esiste una funzione che lega tali misure: $p_{rs}^s = \frac{1}{1 + p_{rs}^d}$ dove p_{rs}^s e p_{rs}^d indicano rispettivamente una generica misura di similarità e di dissimilarità. Data questa funzione è possibile calcolare la funzione inversa di conversione dissimilarità – similarità:

$$p_{rs}^s = \frac{1}{1 + p_{rs}^d} \Leftrightarrow (1 + p_{rs}^d) p_{rs}^s = 1$$

$$\Leftrightarrow p_{rs}^s + p_{rs}^s p_{rs}^d = 1$$

$$\Leftrightarrow p_{rs}^d p_{rs}^s = 1 - p_{rs}^s$$

$$\Leftrightarrow p_{rs}^d = \frac{1 - p_{rs}^s}{p_{rs}^s}.$$

La scelta della tecnica di raggruppamento parte, come accennato precedentemente, dall'obiettivo basilare di individuare, ove possibile, gruppi "naturali" di unità, ossia gruppi che soddisfino le proprietà di coesione interna, cioè le unità appartenenti al medesimo gruppo devono essere simili tra loro, e di separazione esterna, che equivale a richiedere che i gruppi siano il più possibile distinti l'uno dall'altro.

Nello scegliere, poi, il numero di gruppi occorrerebbe porsi come obiettivo principale la creazione di una suddivisione dei dati nello spazio che sia quanto più possibile significativa e semplice da analizzare per far si che i miglioramenti eventualmente ottenuti passando dalla struttura dei dati originali (non modificati) a quella ottenuta dalla clusterizzazione sia tangibile.

Per il progetto FINVALI, dopo numerose prove sperimentali, si è deciso di richiedere la creazione di sette differenti gruppi in quanto, avendo scelto come variabile il fattore Regione, essi consentivano di ottenere l'omogeneità all'interno secondo il modello ANOVA.

In generale, quindi,, si può dire che un metodo di classificazione è caratterizzato da due fattori:

1. una misura del grado di prossimità tra coppie di unità;
2. un algoritmo con cui procedere alla ricerca dei cluster.

Modificando l'uno o l'altro di questi fattori si può ottenere un gran numero di metodi diversi dei quali in letteratura sono state proposte diverse classificazioni alcune basate sul tipo di algoritmo utilizzato dal metodo altre basate sul tipo di risultato da esso fornito. La classificazione più diffusa è però quella che si basa sul tipo di algoritmo e che distingue, essenzialmente, i metodi in:

a) Metodi gerarchici;
b) Metodi non gerarchici.

Come risulterà chiaro nel prosieguo ai fini della ricerca si è utilizzato un metodo non gerarchico, nello specifico il ***metodo delle k medie***, grazie al quale ottenere gruppi omogenei all'interno e ben distinti l'uno dall'altro, in quanto, avendo scelto come variabili i fattori "Regione" ed "Ordine" non era ipotizzabile un'influenza di tipo gerarchico fra i gruppi.

3.2. Metodi gerarchici

In un'analisi gerarchica dei gruppi ogni classe fa parte di una classe più ampia la quale è contenuta, a sua volta, in una classe di ampiezza superiore, e così in progressione fino a giungere alla classe che contiene l'intero insieme di entità analizzate.

I metodi gerarchici sono utilizzati soprattutto quando occorre investigare la struttura dei dati a differenti livelli, ed infatti un metodo di formazione dei gruppi è detto gerarchico se considera tutti i livelli di distanza e se i gruppi che si ottengono ad un determinato livello di distanza comprendono i gruppi ottenuti ad un livello di distanza inferiore.

Una delle caratteristiche principali delle analisi dei gruppi di tipo gerarchico è l'irrevocabilità dell'assegnazione di un oggetto ad un cluster, ovvero una volta che una unità è entrata a far parte di un gruppo non ne viene più rimossa.

Esistono varie tecniche di analisi gerarchica che possono essere distinte in:

- Tecniche Agglomerative, le quali, partendo da *n* elementi distinti; producono di volta in volta un numero decrescente di clusters di ampiezza crescente, fino ad associare in un unico gruppo tutte le *n* unità di partenza.
- Tecniche divisive o scissorie, che ripartiscono gli stessi *n* elementi, inizialmente compresi in un unico insieme, in gruppi sempre più piccoli e numerosi, fino a quando il numero di clusters viene a coincidere con il numero delle unità.

3.2.1. Tecniche gerarchiche agglomerative

A partire da un collettivo non suddiviso in gruppi si procede per aggregazioni successive generando gruppi sempre più numerosi. Il procedimento di raggruppamento parte dalla matrice simmetrica di prossimità tra elementi e procede iterativamente in due passi:

i. Raggruppando gli elementi più somiglianti;
ii. Calcolando la matrice di prossimità fra gruppi e/o elementi, avendo fissato un criterio per stabilire la distanza dei gruppi dai singoli elementi e/o dagli altri gruppi.

Il procedimento si arresta quando tutti gli elementi sono aggregati in un unico cluster.

Quindi, una volta stabilito l'indice di prossimità o la distanza tra le osservazioni quello che differenzia le varie tecniche gerarchiche di raggruppamento è essenzialmente il criterio utilizzato per stabilire la distanza tra i diversi clusters; la distanza tra l'entità *k* ed il

gruppo (*i*, *j*) si calcola, infatti, combinando le distanze d_{ij}, d_{ik}, d_{jk} con pesi che differiscono in base al criterio di aggregazione scelto.

Metodo del legame singolo o del vicino più prossimo.

La distanza tra l'entità *k* e la nuova fusione (*i, j*) è la distanza minore tra *k* e le due entità aggregate, ossia: $d_{k(i,j)} = \min\{d_{ik}, d_{jk}\}$ con $i \neq j \neq k = 1,.....,n$. Ad ogni successiva aggregazione l'entità che entra nel gruppo si collocherà su di una delle estremità dello spazio occupato dal gruppo generando così un concatenamento tra entità. I gruppi che si ottengono applicando il metodo del legame singolo hanno una forma allungata detta anche "a losanga". L'adozione di questo metodo per la composizione dei gruppi evidenzia in modo netto le similitudini tra gli elementi privilegiando però la differenza tra i gruppi piuttosto che l'omogeneità degli elementi appartenenti ad ogni gruppo.

Metodo del legame completo o del vicino più lontano

Il criterio del legame completo si contrappone, come logica e come risultati, a quello del legame singolo, assumendo che tra l'entità esterna *k* ed il gruppo di nuova formazione (*i, j*) la distanza sia data dal valore più elevato tra d_{ik} e d_{jk}, ossia:

$d_{k(i,j)} = \max\{d_{ik}, d_{jk}\}$, con $i \neq j \neq k = 1,.....,n$.

Applicando questo criterio si ottengono gruppi di forma circolare caratterizzati da notevole omogeneità interna. Geometricamente si ha, quindi, che l'accettazione di una nuova entità in un gruppo porta all'allargamento della sfera multidimensionale che contiene il gruppo.

Metodo del legame medio o della media di gruppo

La distanza fra l'elemento *k* ed il gruppo formatosi dalla fusione di *i* e di *j* è data dalla media aritmetica delle distanze d_{ik} e d_{jk} ponderate con le numerosità degli elementi appartenenti ai gruppi *i* e *j*:

$d_{k(ij)} = \alpha_i d_{ik} + \alpha_j d_{jk}$, $(i \neq j \neq k = 1,.....,n)$, dove:

$\alpha_i = \frac{n_i}{(n_i + n_j)}$ e $\alpha_j = \frac{n_j}{(n_i + n_j)}$.

Essendo un metodo basato sulla media delle distanze i risultati forniti risultano essere più attendibili ed i gruppi più omogenei e ben differenziati tra loro.

Metodo del centroide o della distanza tra centroidi

Il centroide è il punto di incontro delle medie di una distribuzione multivariata, o, come definito da qualche autore, il vettore di medie che rappresenta questo punto.

Operando con il metodo del centroide, quindi, la distanza tra due gruppi è calcolata come la distanza euclidea tra i centroidi dei gruppi, ed il quadrato di tale distanza può essere scritto come:

$$_2 d^2_{k(i,j)} = \sum_v^p (\bar{x}_{hv} - \bar{x}_{kv})^2 = \left\{\alpha_i d^2_{ik} + \alpha_j d^2_{jk} - \alpha_i \alpha_j d^2_{ij}\right\}$$

dove $_2 d_{hk}$ indica la distanza euclidea tra due punti *h* e *k* qualsiasi (*h, k=1, …, n*) e α_h è il peso relativo del gruppo *h* valutato come nella formula del legame medio.

Metodo di Ward

Quando si utilizza il metodo di Ward la coppia di entità da aggregare, in una generica fase dell'analisi, è quella che minimizza la devianza tra i centroidi dei possibili gruppi. La devianza ha un minimo pari a zero quando tutte le unità sono isolate ed un massimo, pari alla somma delle devianze delle variabili di classificazione, quando tutte le unità fanno parte di un unico gruppo.

La distanza euclidea tra un entità *k* ed il gruppo di nuova formazione (*i*, *j*) è la radica quadrata di:

$$\frac{n_k n_{(ij)}}{n_k + n_{(ij)}} d^2_{k(ij)} =$$

$$= \frac{1}{n_i + n_j + n_k}\left[(n_i + n_k) d^2_{ik} + (n_j + n_k) d^2_{jk} - n_k d^2_{ij}\right]$$

dove n_k è il numero di unità che compongono il gruppo k ed $n_{(ij)} = n_i + n_j$.

Il metodo di Ward è stato pensato per applicazioni su distanze euclidea ma può essere utilizzato per ogni tipo di distanza ed è applicabile anche se le entità hanno pesi variabili.

Metodo flessibile di Lance e Williams

Tutti i criteri fin qui esposti possono essere ricondotti ad un'unica formula parametrica di distanza tra *k* e (*i*, *j*), proposta appunto, nel 1967 da Lance e Williams:

$$d_{k(ij)} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \gamma \left| d_{ik} - d_{jk} \right| \qquad \text{con } i \neq j \neq k = 1,.....,n$$

Legame singolo $\Rightarrow \ \alpha_i = \alpha_j = \frac{1}{2}; \qquad \beta = 0; \qquad \gamma = -\frac{1}{2}$

Legame completo $\Rightarrow \ \alpha_i = \alpha_j = \gamma = \frac{1}{2}; \qquad \beta = 0$

Media di gruppo $\Rightarrow \ \alpha_i = \frac{n_i}{(n_i + n_j)}; \qquad \alpha_j = \frac{n_j}{(n_i + n_j)}; \qquad \beta = \gamma = 0$

Centroide $\Rightarrow \ \alpha_i = \frac{n_i}{(n_i + n_j)}; \quad \alpha_j = \frac{n_j}{(n_i + n_j)}; \quad \beta = -\frac{n_i n_j}{(n_i + n_j)^2}; \quad \gamma = 0$

Ward $\Rightarrow \ \alpha_i = \frac{(n_i + n_k)}{(n_i + n_j + n_k)}; \ \alpha_j = \frac{(n_j + n_k)}{(n_i + n_j + n_k)}; \ \beta = -\frac{n_k}{(n_i + n_j + n_k)}; \ \gamma = 0$

Lo schema di aggregazione gerarchica proposto da Lance e Williams è invece basato sui seguenti vincoli tra parametri della distanza:

$$\alpha_i + \alpha_j + \beta = 1; \quad \alpha_i = \alpha_j; \qquad \beta < 1; \quad \gamma = 0.$$

Questo metodo è detto flessibile perché, al variare di β, permette di ottenere schemi di raggruppamento aventi caratteristiche differenti.

3.2.2. Metodi gerarchici divisivi

Le procedure divisive o scissorie si basano sulla suddivisione dell'insieme di entità iniziali. Il processo di suddivisione è concettualmente opposto a quello dell'agglomerazione progressiva delle unità ed infatti si parte dalla situazione in cui le *n* unità fanno parte di un unico gruppo e si perviene, in *n-1* passi, alla formazione di *n* gruppi composti, ognuno, da una sola unità; questo rende i criteri divisivi molto più generali di quelli aggregativi proprio perchè permettono la formazione di un numero qualsiasi di sottogruppi.

Tali metodi non sono usati comunemente per cui in questa sede ci limitiamo a fornirne un metodo generale.

Metodi divisivi basati sulla distanza fra centroidi.

I metodi di analisi di raggruppamento gerarchici che si basano sulla distanza fra i centroidi di due sottogruppi si articolano, genericamente, sulla seguente procedura:

1. il primo passo prevede una suddivisione degli elementi in due gruppi in base alla combinazione di unità che minimizza la devianza interna ai gruppi;

2. ad ogni passo successivo occorre individuare il gruppo che presenta la massima devianza interna (intesa come devianza del singolo elemento dal centroide) e va effettuata un'ulteriore suddivisione dicotomica delle *n* unità del gruppo provando tutte le possibili combinazioni con *1* e (*n - 1*) unità, *2* e (*n - 2*) unità, e così via, fino ad individuare quella che minimizza la funzione:

$$D' = \sum_{g}^{G}\sum_{h}^{n_g}\sum_{i}^{p}\left({}_g x_{hi} - {}_g \bar{x}_i\right)^2 \qquad \text{con} \Rightarrow \begin{matrix} g = 1,2 \\ h = 1,...,n_g \\ i = 1,....., p \end{matrix}$$

dove ${}_g x_{hi}$ è il valore della variabile x_i osservato presso l'unità statistica *h* appartenente al sottogruppo *g*, mentre ${}_g \bar{x}_i$ è il valore medio della variabile *i* nel sottogruppo *g*.

Minimizzare tale funzione equivale a massimizzare la distanza fra i centroidi dei due gruppi:

$$D'' = n\alpha_i\alpha_j\sum_{i}^{p}\left({}_1\bar{x}_i - {}_2\bar{x}_i\right)^2 \qquad \text{dove:}$$

$$\alpha_g = \frac{n_g}{n} \qquad \text{con} \quad g = 1, 2;$$

$D = \sum_{i}^{p}\left(x_{hi} - \bar{x}_i\right)^2 = D' + D'' \Rightarrow$ devianza globale interna al gruppo dicotomizzato;

$\bar{x}_i = {}_1\bar{x}_i\alpha_1 + {}_2\bar{x}_i\alpha_2$ (i = 1, …, p) $\Rightarrow$ media della variabile x_i nello stesso gruppo.

3.3. Metodi non gerarchici.

Si parla di tecniche di raggruppamento non gerarchiche quando l'algoritmo utilizzato produce un'unica suddivisione dell'insieme di partenza considerata ottima rispetto al criterio adottato, il che equivale a dire che genera gruppi non gerarchizzabili, confrontabili soltanto mediante l'utilizzo di indici sintetici della classificazione complessiva e non gruppo per gruppo.

Tra i metodi non gerarchici i più utilizzati sono:

- metodi di programmazione matematica;
- metodi di suddivisione iterativa.

In generale questi metodi partono da una iniziale suddivisione delle unità e procedono spostandole da un gruppo all'altro fino al raggiungimento di una situazione ottimale che non consenta altri spostamenti.

I metodi di programmazione matematica si basano su spostamenti virtuali delle unità che vengono effettuati in relazione alla soluzione di un problema di minimo o di massimo vincolato e non contemplano il calcolo dei centroidi dei gruppi.

I metodi di suddivisione iterativa, invece, eseguono una suddivisione effettiva delle unità; questi metodi hanno come punto di partenza il calcolo dei centroidi dei vari gruppi (oppure la scelta dei nuclei attorno ai quali occorre raggruppare le unità) e l'assegnazione di ogni unità al gruppo più vicino. Ad un secondo passo si calcolano nuovamente i centroidi e si ripete il procedimento fino a quando non è possibile spostare ulteriormente le unità. In alcuni casi i metodi sono dotati anche di una funzione obiettivo che valuta la bontà di una determinata partizione per poter scegliere lo spostamento più conveniente fra quelli possibili.

Anche all'interno dell'ampia gamma di procedure di analisi non gerarchiche è possibile distinguere due specifiche categorie, ossia, i metodi che generano *partizioni*, cioè classi mutuamente esclusive in cui una unità può appartenere ad un unico gruppo, e metodi che generano *classi sovrapposte*, dove è contemplata la possibilità di inserire una unità in più di una classe (metodi "alternativi" di classificazione).

Metodi che generano partizioni

In generale i metodi non gerarchici di questo tipo fondano sulla necessità di collocare le unità all'interno dei gruppi attraverso la specificazione di una funzione obiettivo, che, a seconda dei criteri scelti, potrebbe essere sia quella di minimizzare la devianza interna ai gruppi, sia quella di massimizzare la devianza tra i gruppi; fissato, infatti, il numero g dei gruppi che devono essere costituiti, l'algoritmo classificherà le unità in base al criterio prescelto.

I criteri più noti si rifanno, ovviamente, alla scomposizione della devianza totale, $Dev(T) = \sum_{i=1}^{n}\sum_{k=1}^{p}\left(x_{ik} - \bar{x}_k\right)^2$, scomponibile nella somma della devianza interna ai gruppi e della devianza tra i gruppi:

$$Dev(T) = Dev(W) + Dev(B) = \sum_{g}^{G}\sum_{i}^{n_g}\sum_{k}^{p}\left(x_{ikg} - \bar{x}_{kg}\right)^2 + \sum_{g}^{G}\sum_{k}^{p} n_g\left(\bar{x}_{kg} - \bar{x}_k\right)^2$$

e le tecniche di analisi si reggono, quindi, come scopo o la minimizzazione di W o la massimizzazione di B.

Il criterio più comunemente utilizzato fa riferimento alla minimizzazione della traccia (somma degli elementi diagonale) di W: *min tr*(W), ossia opera minimizzando la somma dei quadrati delle distanze euclidea all'interno dei gruppi.

Adotta tale criterio anche il metodo di classificazione non gerarchica, maggiormente utilizzato nelle analisi reali, proposto da MacQueen e comunemente chiamato ***Metodo delle k-medie***, dove *k* indica il numero dei gruppi che si vuole costruire.

Questa tecnica prevede, dopo aver fissato *k*, la scelta di *k* "poli" iniziali (denominati spesso anche "semi" e generalmente rappresentati da centroidi), che sono punti dello spazio *p-dimensionale* grazie ai quali è possibile costruire una partizione iniziale allocando le unità al cluster avente il polo più vicino. Dopo aver scelto i poli, si calcola, per ogni unità, la distanza dai *k* centroidi e, se l'unità dovesse risultare più vicina al centroide di un altro gruppo, si provvederà a riallocarla in questo ultimo gruppo. Ciò implicherà, ovviamente, il ricalcolo del centroide sia del vecchio che del nuovo gruppo di appartenenza dell'unità spostata. Questo processo viene reiterato fino a quando non è più necessario spostare alcuna unità. L'algoritmo delle *k*-medie, di solito, utilizza una metrica di tipo euclideo che garantisce la convergenza del processo iterativo.

Se si considera, a titolo di esempio, la *t-esima* iterazione, la distanza tra l'unità *i-esima* (*i=1, …, n*) ed il centroide del *g-esimo* gruppo (*g=1, …, k*), è data:

$$d(\mathbf{x}_i, \mathbf{x}_g^{(t)}) = \sum [\sum_k^p (\mathbf{x}_{i,k} - \mathbf{x}_{k,g})^2].$$

Le analisi svolte per il progetto FINVALI si basano, come già accennato precedentemente, appunto sul ***metodo delle k-medie*** per l'attuazione del quale si è deciso di individuare la partizione iniziale tramite scelta casuale, e di utilizzare, come misura di prossimità, il quadrato della distanza Euclidea.

Poiché, inoltre, i maggiori inconvenienti per le tecniche di clustering non gerarchiche sono legati proprio alla necessità di dover scegliere una partizione iniziale per poter dare il via al processo iterativo, ha particolare rilevanza la scelta del numero *g* di gruppi che devono costituire la partizione iniziale, scelta, questa, che può essere effettuata in diversi modi.

Se, come ad esempio accade per la nostra ricerca, le *n* unità di partenza non sono numerose, si può eseguire l'analisi utilizzando diversi valori di *g*, valutando poi le diverse partizioni magari in base ad un indice sintetico quale potrebbe essere R^2 (anche se questo criterio non offre, in realtà, alcuna garanzia di individuare un valore ottimo per *g*), oppure

basandosi sulla conoscenza del fenomeno e cercando di individuare, un po' più empiricamente, il raggruppamento più idoneo agli scopi dell'analisi.

Nel caso in cui, invece, le *n* unità siano numerose, il numero *g* di gruppi può essere individuato a partire dai risultati di una precedente analisi di tipo gerarchico.

Metodi "Alternativi"

In alcuni casi ci si potrebbe imbattere in un certo numero di unità che non presentano caratteristiche "assolute" che permettano di ricondurle ad un unico gruppo; in situazioni del genere si può ricorrere a tecniche che giustifichino la presenza di una stessa unità in gruppi differenti, ossia quelle tecniche di analisi non gerarchica che generano classi sovrapposte.

I metodi "alternativi" di classificazione più noti sono:

- metodi di clumping, generano classificazioni non disgiunte in cui è quindi possibile che una unità sia collocata in gruppi differenti creando dei clusters sovrapposti. Tale tecnica è valida soltanto quando non sussiste un numero eccessivo di sovrapposizioni; i vincoli solitamente utilizzati per controllarne il numero si basano sul metodo B_K, (k 0 1, 2, 3, …), per il quale il numero massimo di unità che si possono sovrapporre in ogni coppia di gruppi è pari a (k – 1). Se il numero delle unità che si sovrappongono nei due gruppi dovesse superare tale soglia i due gruppi verranno fusi ed andranno a formare un unico cluster.
- Fuzzy clustering, fa riferimento al concetto di "fuzzy set", o insieme sfocato, in cui l'accostamento di una unità ad un gruppo è legata ad una funzione che indica il grado di appartenenza dell'unità al gruppo considerato. Questa funzione assume valori compresi nell'intervallo [0, 1] e, nello specifico, un valore pari a zero indica l'estraneità dell'unità al gruppo, un valore pari ad uno ne indica l'assoluta appartenenza, mentre i valori intermedi indicano in che misura l'unità è legata al gruppo. Per tale metodo, quindi, due unità saranno tanto più simili tra loro quanto più prossimo ad uno è il valore della loro funzione di appartenenza al medesimo gruppo.

3.4. Linee guida per la scelta del metodo di classificazione

Partendo dal presupposto che strategie di raggruppamento differenti spesso conducono a risultati non dissimili e che i risultati non dipendono solo dalla strategia di analisi ma anche dalle opzioni scelte, fra cui, ad esempio, il tipo di distanza utilizzato o la standardizzazione o meno dei dati di partenza, la qualità di una tecnica per l'analisi dei gruppi può essere valutata in base ai seguenti criteri:

- oggettività, o ripetibilità, della soluzione: è molto importante che una tecnica di raggruppamento sia in grado di generare una soluzione che sia riproducibile da chiunque ripeta l'analisi anche in tempi successivi;
- stabilità della soluzione, nelle analisi reali è facile imbattersi in dati che presentino errori, quindi, la tecnica ottimale deve essere il meno possibile sensibile a piccole variazioni nei dati, così che l'eventuale eliminazione di una unità non modifichi la struttura dei gruppi;
- informatività del risultato, un metodo di classificazione deve essere in grado di produrre risultati che incorporino il maggior numero di informazioni specifiche per l'analisi condotto;
- semplicità è rapidità di esecuzione dell'algoritmo di calcolo, ossia la possibilità di condurre l'analisi agevolmente anche su matrici di dati di elevata dimensione.

Occorre, poi, ricordare che, in generale, le tecniche non gerarchiche sono più informative di quelle gerarchiche perché forniscono anche risultati intermedi e vari indici per la misura della qualità del risultato; inoltre i metodi gerarchici risentono maggiormente della presenza di dati anomali e sono particolarmente disturbati da eventuali errori di misura. Un importante pregio delle tecniche gerarchiche è quello di non richiedere un tempo eccessivo per il calcolo dei risultati, però un'aggregazione impropria effettata nei primi stadi dell'analisi si trascina fino alla fine e rischia di rendere i risultati artificiosi.

Per poter rendere l'analisi il più possibile significativa, spesso, nella pratica si preferisce condurre in successione prima un'analisi di tipo gerarchico e poi un'analisi non gerarchica.

Nel caso del progetto FINVALI questo non è stato necessario, per cui, dopo aver eseguito la cluster analysis, si è deciso di applicare ai gruppi così ottenuti la metodologia ANOVA, utilizzata inizialmente, per poter avere un'ulteriore conferma dell'omogeneità dei gruppi individuati.

4. Cluster Analysis e Analisi ANOVA sui questionari di valutazione INVALSI, per gli anni scolastici 2004/2005 e 2005/2006

L'ultima parte di questo secondo semestre di ricerca è stata dedicata, come già accennato, all'implementazione di script Matlab per l'esecuzione di Cluster Analysis allo scopo di individuare un raggruppamento dei livelli in gruppi omogenei che permetta di raggiungere una significativa riduzione delle variabili in gioco.

Dopo aver effettuato le prime scelte preliminari necessarie per tale tipo di metodologia (nello specifico: individuazione delle variabili di classificazione, di una misura di prossimità e della tecnica di analisi) l'analisi è stata eseguita, separatamente, sui dati relativi ai questionari di valutazione per l'a.s. 2004/2005 e per l'a.s. 2005/2006 utilizzando diversi valori di *g* (numero di gruppi richiesti per l'analisi) ed in base ai risultati preliminari così ottenuti si è cercato di individuare la miglior partizione possibile in funzione delle variabili di classificazione selezionate.

In questa fase la scelta, a nostro giudizio, più coerente e significativa è risultata essere quella che prevede, in funzione di ogni livello del fattore Ordine (ossia Scuole Elementari, Medie Inferiori e Medie Superiori), la partizione dell'insieme delle regioni in sette gruppi.

Consapevoli della necessità di effettuare confronti fra i risultati delle analisi eseguite per i due anni consecutivi di indagine, si è passati a verificare se il valore *g* prescelto risultasse coerente anche quando l'analisi di raggruppamento veniva eseguita contemporaneamente sui dati relativi ai questionari di valutazione di entrambi gli anni scolastici.

Avendo ottenuto, a tale proposito, un riscontro favorevole si è deciso di procedere individuando i gruppi in base, appunto, ai dati relativi ad entrambi gli anni, e di sottoporre, poi, per singolo anno scolastico, i dati partizionati nei gruppi così individuati ad analisi con metodologia ANOVA.

I risultati ottenuti vengono forniti, come in precedenza, sia in forma grafica che tabellare. In particolare per la Cluster Analysis è stato prodotto un grafico ad hoc che riassume i gruppi individuati relativamente alla coppia di anni scolastici presi in esame: in ascissa è riportata la percentuale media nell'a.s. 2004-2005 e in ordinata quella nell'a.s. 2005-2006. Una linea diagonale indica la regione dove le percentuali coincidono per i due anni scolastici. Pertanto le regioni che si porranno al di sopra della linea diagonale avranno

registrato un miglioramento dell'abilità nel 2005-2006 rispetto al 2004-2005; viceversa le regioni che si collocano al di sotto della linea diagonale avranno registrato un peggioramento. La posizione delle regioni nel piano è individuata con un cerchio colorato a seconda del raggruppamento (cluster) individuato mediante la Cluster Analysis. I grafici e le Tabelle per l'analisi ANOVA sono stati già descritti nel presente report.

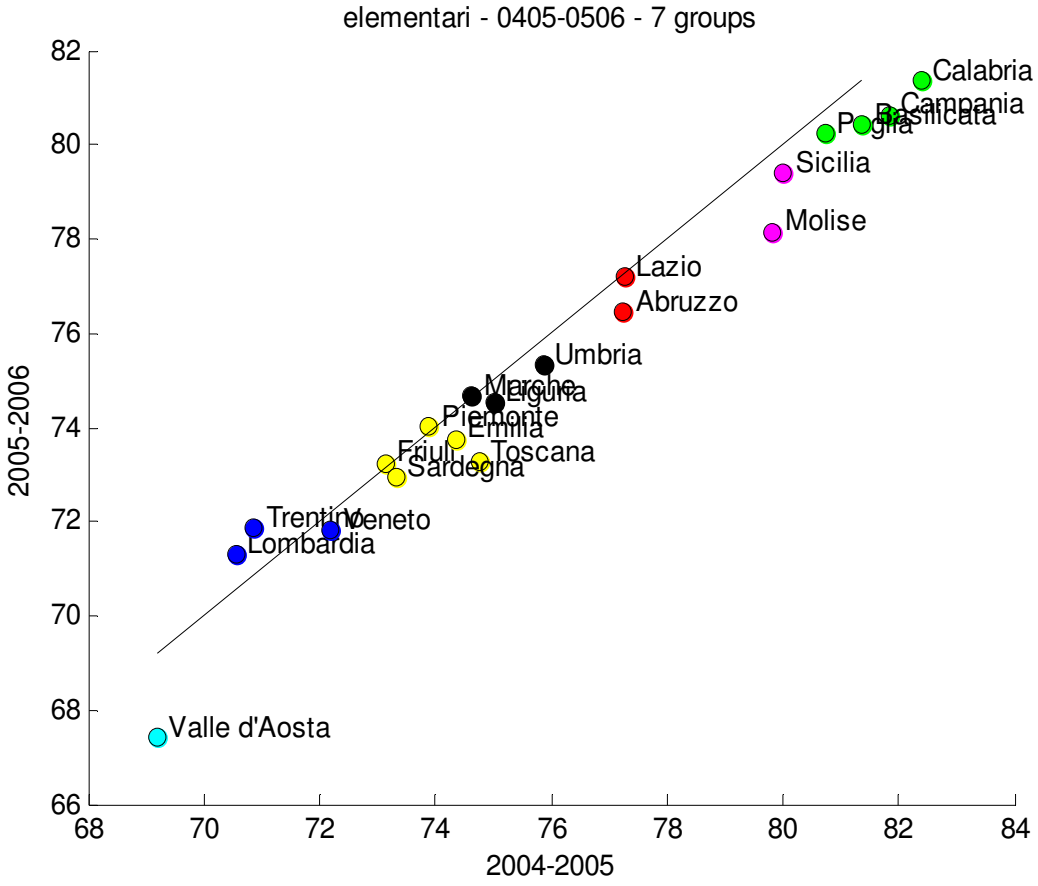
4.1. Scuole Elementari

4.1.1. Cluster Analysis

Il primo grafico che presentiamo mostra i risultati della ottenuti con l'analisi dei cluster applicata ai dati relativi ai questionari di valutazione somministrati alle Scuole Elementari negli anni scolastici 2004/2005 e 2005/2006.

Il grafico presenta in ascissa ed in ordinata le percentuali medie di risposte esatte fornite dagli studenti delle Scuole Elementari, per singola regione, relative, rispettivamente, agli anni scolastici 2004/2005 e 2005/2006 cui è aggiunta una costante fissa delta.

Per poter ottenere, poi, un primo generale confronto fra le valutazioni per i due anni si è deciso di inserire nel grafico una retta, costruita in funzione dei valori minimi e massimi relativi ad entrambi gli anni, che rappresenta una sorta di frontiera della valutazione, nel senso che se le regioni si trovano al si sopra di tale retta significa che la valutazione per lo specifico Ordine scolastico considerato, nel passaggio da un anno all'altro, è migliorata, e, viceversa, per quelle regioni la cui posizione è al di sotto della retta.



Per quel che riguarda le Scuole Elementari , quindi, possiamo affermare che, in media, gli Istituti presenti in Lombardia, Trentino Alto Adige, Friuli, Piemonte e Marche hanno ottenuto una migliore valutazione per l'anno scolastico 2005/2006, mentre tutte le altre regioni va registrato, per quello stesso anno, un rendimento peggiore rispetto all'anno precedente.

Dal grafico, inoltre, si nota immediatamente la grande distanza che intercorre, nelle valutazioni di entrambi gli anni, fra le scuole della Valle d'Aosta ed il resto del territorio italiano cosa che necessariamente porta alla collocazione di tale regione in un gruppo a sé stante.

In base ai gruppi così formatisi è stata poi eseguita l'analisi ANOVA per i singoli anni assumendo come fattore fisso la Regione, ossia l'altra variabile individuata per la Cluster Analysis.

4.1.2. Analisi ANOVA a.s. 2004/2005: Fattore Regione

1° GRUPPO

Grandmean: 76.74

F-test: 2.989545e-001 - Liv. signif.: 5.846755e-001

Regione 1: Lazio - Estimated mean: 76.95 +/- 0.33

Regione 2: Abruzzo - Estimated mean: 76.54 +/- 0.67

2° GRUPPO:

Grandmean: 70.86

F-test: 8.916565e+000 - Liv. signif.: 1.408712e-004

Regione 1: Lombardia - Estimated mean: 70.46 +/- 0.19

Regione 2: Trentino - Estimated mean: 70.30 +/- 0.67

Regione 3: Veneto - Estimated mean: 71.80 +/- 0.27

3° GRUPPO:

Grandmean: 81.04

F-test: 1.816741e+000 - Liv. signif.: 1.420196e-001

Regione 1: Campania - Estimated mean: 81.37 +/- 0.30

Regione 2: Puglia - Estimated mean: 80.34 +/- 0.43

Regione 3: Basilicata - Estimated mean: 80.74 +/- 0.91

Regione 4: Calabria - Estimated mean: 81.71 +/- 0.51

4° GRUPPO:

Grandmean: 74.97

F-test: 1.529213e+000 - Liv. signif.: 2.178326e-001

Regione 1: Liguria - Estimated mean: 74.71 +/- 0.58

Regione 2: Umbria - Estimated mean: 75.92 +/- 0.76

Regione 3: Marche - Estimated mean: 74.29 +/- 0.54

5° GRUPPO:

Grandmean: 79.50

F-test: 2.971435e-002 - Liv. signif.: 8.631776e-001

Regione 1: Molise - Estimated mean: 79.62 +/- 1.42

Regione 2: Sicilia - Estimated mean: 79.37 +/- 0.37

6° GRUPPO:

Grandmean: 73.50

F-test: 3.119050e+000 - Liv. signif.: 1.438824e-002

Regione 1: Piemonte - Estimated mean: 73.51 +/- 0.34

Regione 2: Friuli - Estimated mean: 73.00 +/- 0.65

Regione 3: Emilia - Estimated mean: 74.08 +/- 0.37

Regione 4: Toscana - Estimated mean: 74.40 +/- 0.37

Regione 5: Sardegna - Estimated mean: 72.50 +/- 0.47

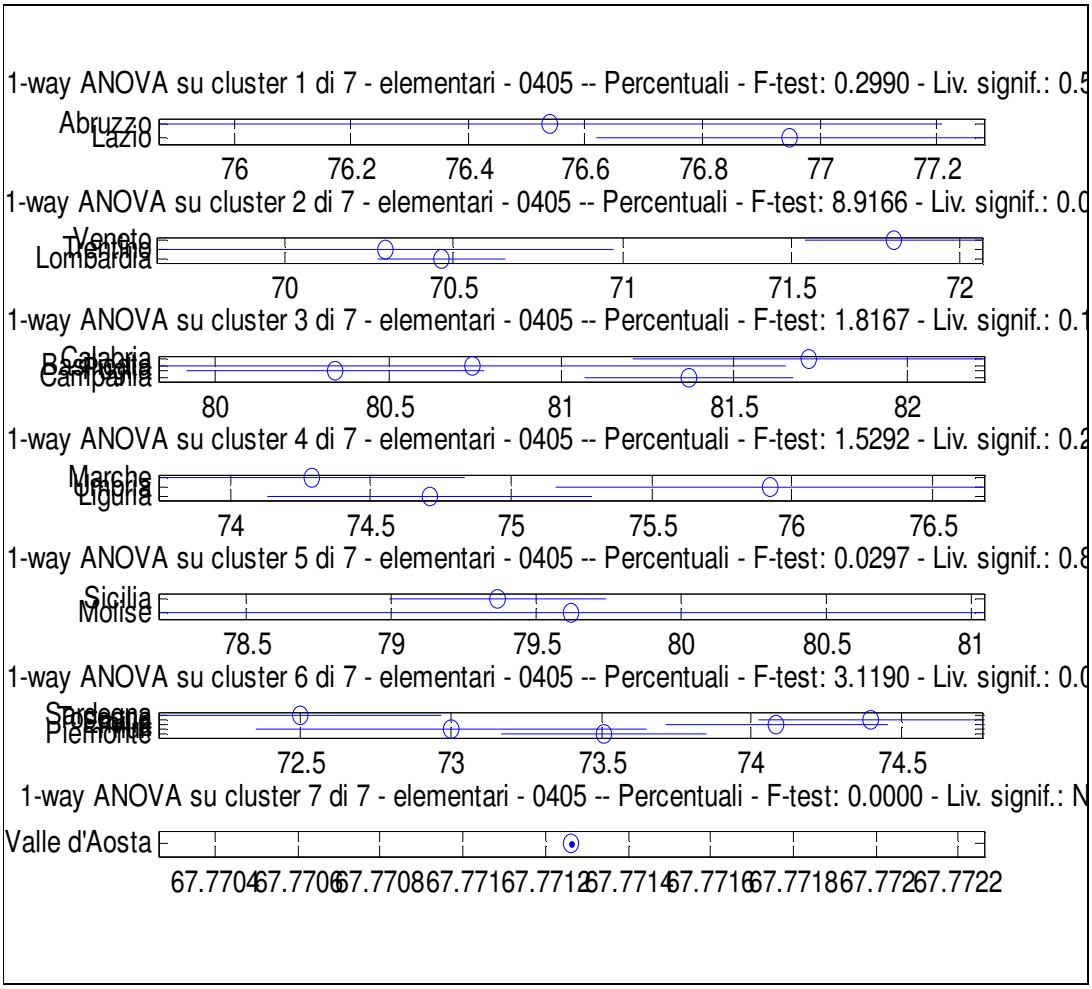
7° GRUPPO:

Grandmean: 67.77

F-test: 0 - Liv. signif.: NaN

Regione 1: Valle d'Aosta - Estimated mean: 67.77 +/- 0.00

I risultati presentati nella precedente tabella e nel grafico che segue mostrano che effettivamente, per quanto riguarda il fattore Regione, da noi preso in considerazione, le differenze all'interno dei singoli gruppi non risultano, nella maggioranza dei casi, significative, il che equivale a dire che i gruppi formatisi grazie alla metodologia di clustering possono ritenersi effettivamente omogenei.



4.1.3. Analisi ANOVA a.s. 2005/2006: Fattore Regione

1° GRUPPO

Grandmean: 76.67

F-test: 1.270269e+000 - Liv. signif.: 2.600180e-001

Regione 1: Lazio - Estimated mean: 77.06 +/- 0.31

Regione 2: Abruzzo - Estimated mean: 76.28 +/- 0.62

2° GRUPPO

Grandmean: 71.33

F-test: 7.350247e-001 - Liv. signif.: 4.796537e-001

Regione 1: Lombardia - Estimated mean: 71.24 +/- 0.17

Regione 2: Trentino - Estimated mean: 71.16 +/- 0.62

Regione 3: Veneto - Estimated mean: 71.59 +/- 0.24

3° GRUPPO

Grandmean: 80.30

F-test: 6.603383e-001 - Liv. signif.: 5.764583e-001

Regione 1: Campania - Estimated mean: 80.43 +/- 0.29

Regione 2: Puglia - Estimated mean: 79.94 +/- 0.41

Regione 3: Basilicata - Estimated mean: 80.06 +/- 0.87

Regione 4: Calabria - Estimated mean: 80.78 +/- 0.49

4° GRUPPO

Grandmean: 74.78

F-test: 7.283502e-001 - Liv. signif.: 4.832728e-001

Regione 1: Liguria - Estimated mean: 74.32 +/- 0.52

Regione 2: Umbria - Estimated mean: 75.36 +/- 0.69

Regione 3: Marche - Estimated mean: 74.65 +/- 0.49

5° GRUPPO

Grandmean: 78.62

F-test: 3.133676e-001 - Liv. signif.: 5.757555e-001

Regione 1: Molise - Estimated mean: 78.24 +/- 1.33

Regione 2: Sicilia - Estimated mean: 79.01 +/- 0.35

6° GRUPPO

Grandmean: 73.20

F-test: 2.786804e+000 - Liv. signif.: 2.529159e-002

Regione 1: Piemonte - Estimated mean: 73.84 +/- 0.30

Regione 2: Friuli - Estimated mean: 73.05 +/- 0.58

Regione 3: Emilia - Estimated mean: 73.72 +/- 0.33

Regione 4: Toscana - Estimated mean: 73.09 +/- 0.33

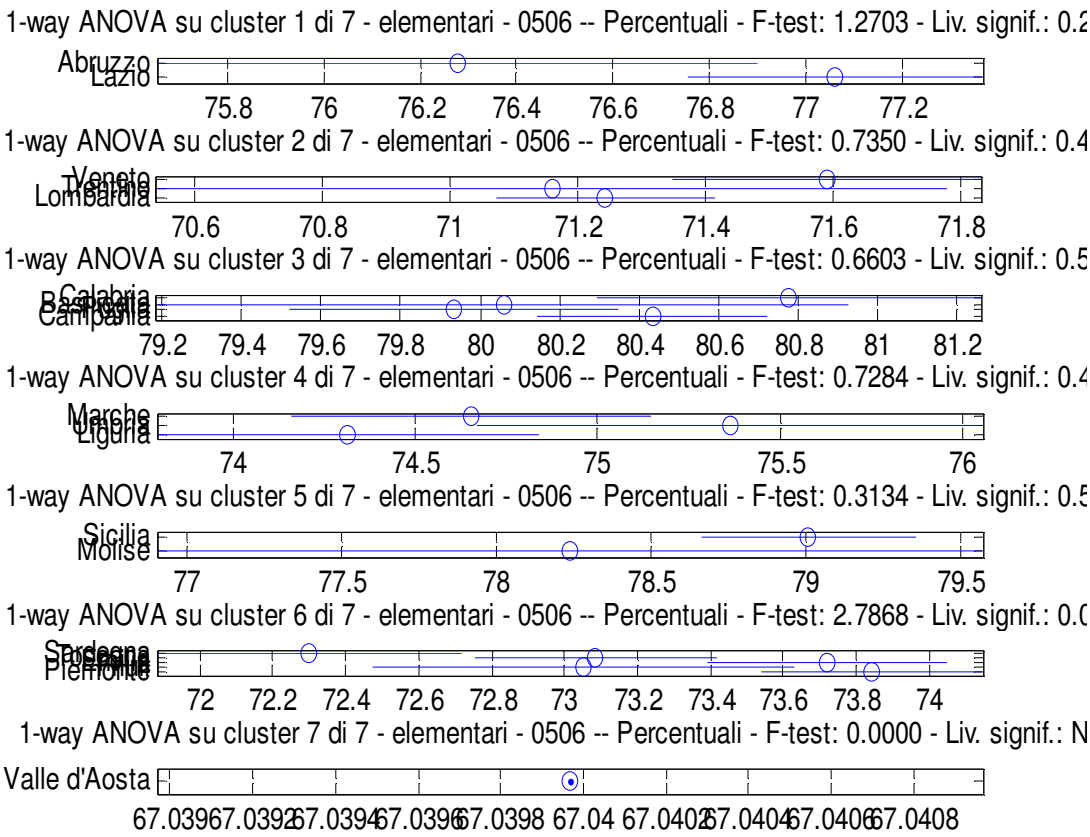
Regione 5: Sardegna - Estimated mean: 72.30 +/- 0.42

7° GRUPPO

Grandmean: 67.04

F-test: 0 - Liv. signif.: NaN

Regione 1: Valle d'Aosta - Estimated mean: 67.04 +/- 0.00

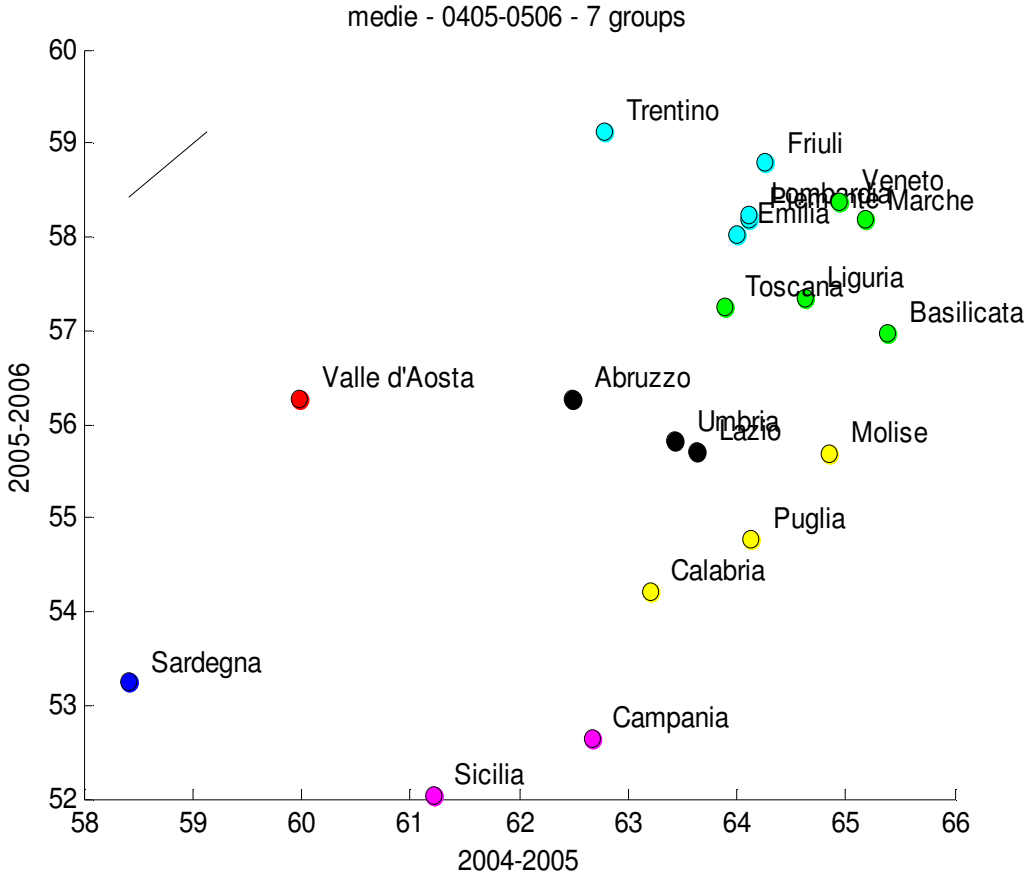


I risultati ottenuti in relazione ai dati del 2005/2006 confermano e probabilmente rafforzano quanto detto in precedenza mostrando gruppi ancora più omogenei al loro interno.

4.2. Scuole Medie Inferiori

Le analisi effettuate per l'Ordine scolastico "Medie Inferiori" hanno evidenziato un netto peggioramento nelle valutazioni degli Istituti in tutte le regioni nel passaggio da un anno all'altro, restituendo quindi, otticamente e non solo, una situazione grafica decisamente più caotica rispetto a quanto osservato per le Scuole Elementari.

4.2.1 Cluster Analysis



Le basse percentuali di risposte esatte registratasi negli Istituti della Sardegna ha portato alla necessità di creare, così come è già avvenuto precedentemente per la Valle d'Aosta e confermato da questa analisi, un gruppo per così dire "dedicato", ossia contenente un'unica unità.

4.2.2 Analisi ANOVA a.s. 2004/2005: Fattore Regione

1° GRUPPO

Grandmean: 61.13

F-test: 0 – Liv. Signif.: NaN

Regione 1: Valle d'Aosta – Estimated mean: 61.13 +/- 0.00

2° GRUPPO

Grandmean: 57.18

F-test: 0 – Liv. Signif.: NaN

Regione 1: Sardegna – Estimated mean: 57.18 +/- 0.00

3° GRUPPO

Grandmean: 64.06

F-test: 2.059647e+000 – Liv. Signif.: 8.402125e-002

Regione 1: Liguria – Estimated mean: 64.06 +/- 0.46

Regione 2: Veneto – Estimated mean: 64.15 +/- 0.24

Regione 3: Toscana – Estimated mean: 63.48 +/- 0.30

Regione 4: Marche – Estimated mean: 64.87 +/- 0.41

Regione 5: Basilicata – Estimated mean: 63.75 +/- 0.52

4° GRUPPO

Grandmean: 62.53

F-test: 3.234152e+000 – Liv. Signif.: 3.997882e-002

Regione 1: Umbria – Estimated mean: 62.60 +/- 0.69

Regione 2: Lazio – Estimated mean: 63.24 +/- 0.28

Regione 3: Abruzzo – Estimated mean: 61.75 +/- 0.53

5° GRUPPO

Grandmean: 61.11

F-test: 1.237702e+001 – Liv. Signif.: 4.500689e-004

Regione 1: Campania – Estimated mean: 62.04 +/- 0.37

Regione 2: Sicilia – Estimated mean: 60.17 +/- 0.38

6° GRUPPO

Grandmean: 63.34

F-test: 3.147063e+000 – Liv. Signif.: 4.354550e-002

Regione 1: Molise – Estimated mean: 64.32 +/- 1.26

Regione 2: Puglia – Estimated mean: 63.62 +/- 0.45

Regione 3: Calabria – Estimated mean: 62.08 +/- 0.51

7° GRUPPO

Grandmean: 63.34

F-test: 2.074110e+000 – Liv. Signif.: 8.179904e-002

Regione 1: Piemonte – Estimated mean: 63.33 +/- 0.25

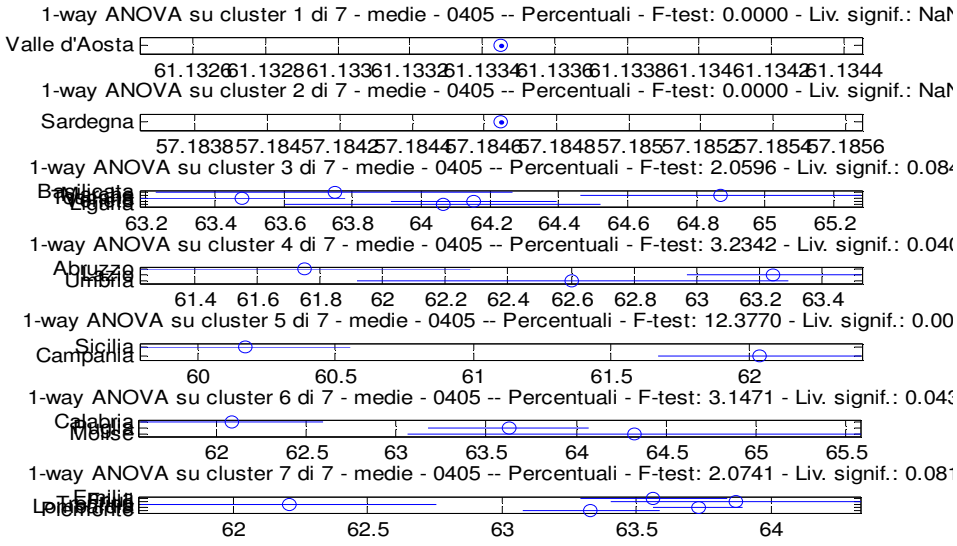
Regione 2: Lombardia – Estimated mean: 63.73 +/- 0.17

Regione 3: Trentino – Estimated mean: 62.21 +/- 0.55

Regione 4: Friuli – Estimated mean: 63.87 +/- 0.47

Regione 5: Emilia – Estimated mean: 63.56 +/- 0.27

Dai dati in tabella si può osservare che, fatta eccezione per il gruppo composto dalla Campania e dalla Sicilia, il livello di significatività dell'analisi non scende mai al di sotto del 3%, cosa che conferma le ipotesi di partenza sulla presenza di omogeneità per il fattore Regione.



4.2.3 Analisi ANOVA a.s. 2005/2006: Fattore Regione

1° GRUPPO

Grandmean: 56.46

F-test: 0 - Liv. signif.: NaN

Regione 1: Valle d'Aosta - Estimated mean: 56.46 +/- 0.00

2° GRUPPO

Grandmean: 52.50

F-test: 0 - Liv. signif.: NaN

Regione 1: Sardegna - Estimated mean: 52.50 +/- 0.00

3° GRUPPO

Grandmean: 57.45

F-test: 6.296885e+000 - Liv. signif.: 5.235091e-005

Regione 1: Liguria - Estimated mean: 57.51 +/- 0.39

Regione 2: Veneto - Estimated mean: 58.33 +/- 0.20

Regione 3: Toscana - Estimated mean: 57.36 +/- 0.25

Regione 4: Marche - Estimated mean: 57.88 +/- 0.34

Regione 5: Basilicata - Estimated mean: 56.15 +/- 0.43

4° GRUPPO

Grandmean: 55.88

F-test: 7.048574e-002 - Liv. signif.: 9.319477e-001

Regione 1: Umbria - Estimated mean: 56.01 +/- 0.53

Regione 2: Lazio - Estimated mean: 55.79 +/- 0.22

Regione 3: Abruzzo - Estimated mean: 55.85 +/- 0.41

5° GRUPPO

Grandmean: 52.04

F-test: 5.057570e+000 - Liv. signif.: 2.469298e-002

Regione 1: Campania - Estimated mean: 52.54 +/- 0.31

Regione 2: Sicilia - Estimated mean: 51.54 +/- 0.32

6° GRUPPO

Grandmean: 54.38

F-test: 3.369913e+000 - Liv. signif.: 3.492272e-002

Regione 1: Molise - Estimated mean: 55.20 +/- 1.02

Regione 2: Puglia - Estimated mean: 54.62 +/- 0.36

Regione 3: Calabria - Estimated mean: 53.33 +/- 0.41

7° GRUPPO

Grandmean: 58.23

F-test: 1.136553e+000 - Liv. signif.: 3.375128e-001

Regione 1: Piemonte - Estimated mean: 57.80 +/- 0.23

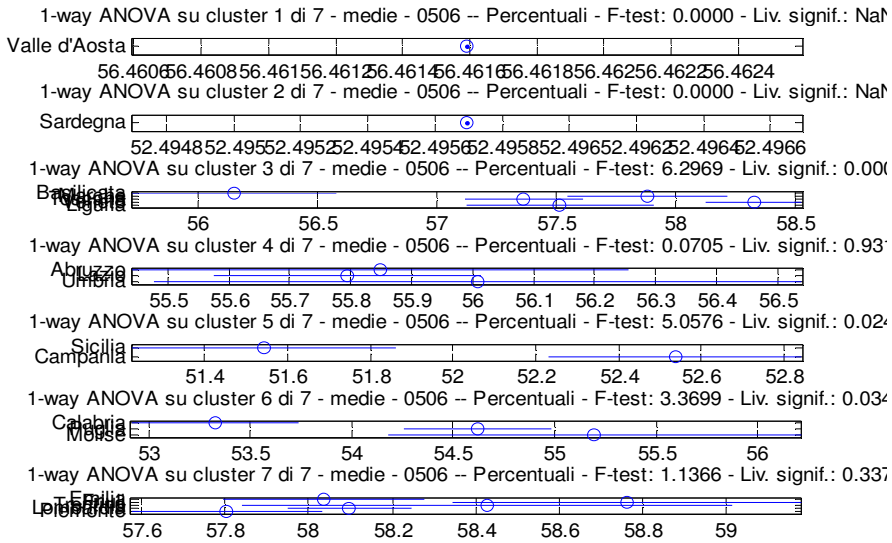
Regione 2: Lombardia - Estimated mean: 58.10 +/- 0.15

Regione 3: Trentino - Estimated mean: 58.43 +/- 0.59

Regione 4: Friuli - Estimated mean: 58.76 +/- 0.42

Regione 5: Emilia - Estimated mean: 58.04 +/- 0.24

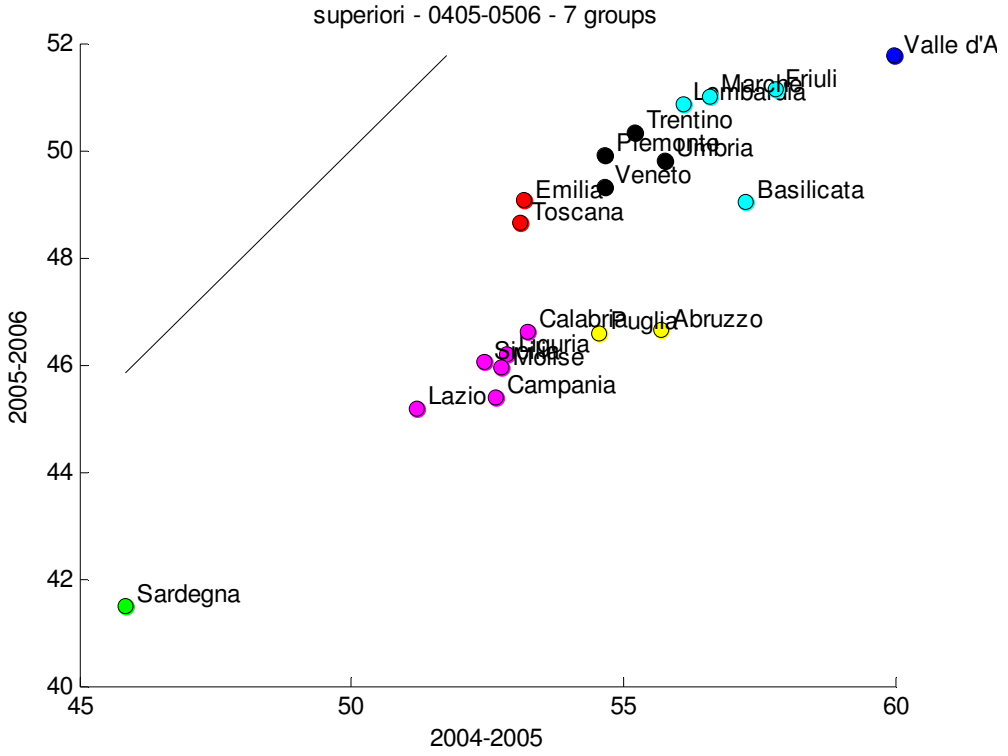
I forti peggioramenti registratisi per tutte le regioni nelle valutazioni dei questionari del 2005/2006 hanno pesato notevolmente, come gia accennato precedentemente, sull'individuazione dei gruppi che spesso, come confermato anche delle successive analisi ANOVA che restituiscono livelli di significatività decisamente bassi, sono risultati essere i migliori possibili in tale situazione e non i migliori in assoluto.



4.3. Scuole Medie Superiori

Quando si passa all'analisi delle valutazioni ottenute dall'Ordine scolastico "Medie Superiori" ritroviamo un quadro più lineare con una suddivisione delle regioni in gruppi di più facile comprensione ed immediato riscontro visivo.

4.3.1. Cluster Analysis



Ancora una volta occorre registrare una netta differenza fra le valutazioni ottenute dagli Istituti nei due anni consecutivi d'indagine, e la conferma di un forte peggioramento anche per le Scuole Medie Superiori nel passaggio dall'anno scolastico 2004/2005 all'a.s. 2005/2006.

Inoltre, i particolari valori delle percentuali medie di risposte esatte registrati negli Istituti della Sardegna e della Valle d'Aosta hanno comportato, anche in questo caso, la creazione di due gruppi formati da un'unica regione

4.3.2. Analisi ANOVA a.s. 2004/2005: Fattore Regione

1° GRUPPO

Grandmean: 52.70

F-test: 2.042960e-001 - Liv. signif.: 6.518077e-001

Regione 1: Emilia - Estimated mean: 53.00 +/- 0.81

Regione 2: Toscana - Estimated mean: 52.40 +/- 1.05

2° GRUPPO

Grandmean: 55.79

F-test: 0 - Liv. signif.: NaN

Regione 1: Valle d'Aosta - Estimated mean: 55.79 +/- 0.00

3° GRUPPO

Grandmean: 44.93

F-test: 0 - Liv. signif.: NaN

Regione 1: Sardegna - Estimated mean: 44.93 +/- 0.00

4° GRUPPO

Grandmean: 54.79

F-test: 5.003561e-001 - Liv. signif.: 6.822639e-001

Regione 1: Piemonte - Estimated mean: 54.35 +/- 0.73

Regione 2: Trentino - Estimated mean: 55.89 +/- 1.13

Regione 3: Veneto - Estimated mean: 54.76 +/- 0.67

Regione 4: Umbria - Estimated mean: 54.15 +/- 1.45

5° GRUPPO

Grandmean: 51.40

F-test: 5.334396e-001 - Liv. signif.: 7.509988e-001

Regione 1: Liguria - Estimated mean: 51.20 +/- 1.72

Regione 2: Lazio - Estimated mean: 50.27 +/- 0.82

Regione 3: Molise - Estimated mean: 51.55 +/- 2.01

Regione 4: Campania - Estimated mean: 51.96 +/- 0.82

Regione 5: Calabria - Estimated mean: 52.04 +/- 1.16

Regione 6: Sicilia - Estimated mean: 51.39 +/- 0.82

6° GRUPPO

Grandmean: 53.52

F-test: 8.710962e-002 - Liv. signif.: 7.682532e-001

Regione 1: Abruzzo - Estimated mean: 53.23 +/- 1.74

Regione 2: Puglia - Estimated mean: 53.80 +/- 0.83

7° GRUPPO

Grandmean: 56.26

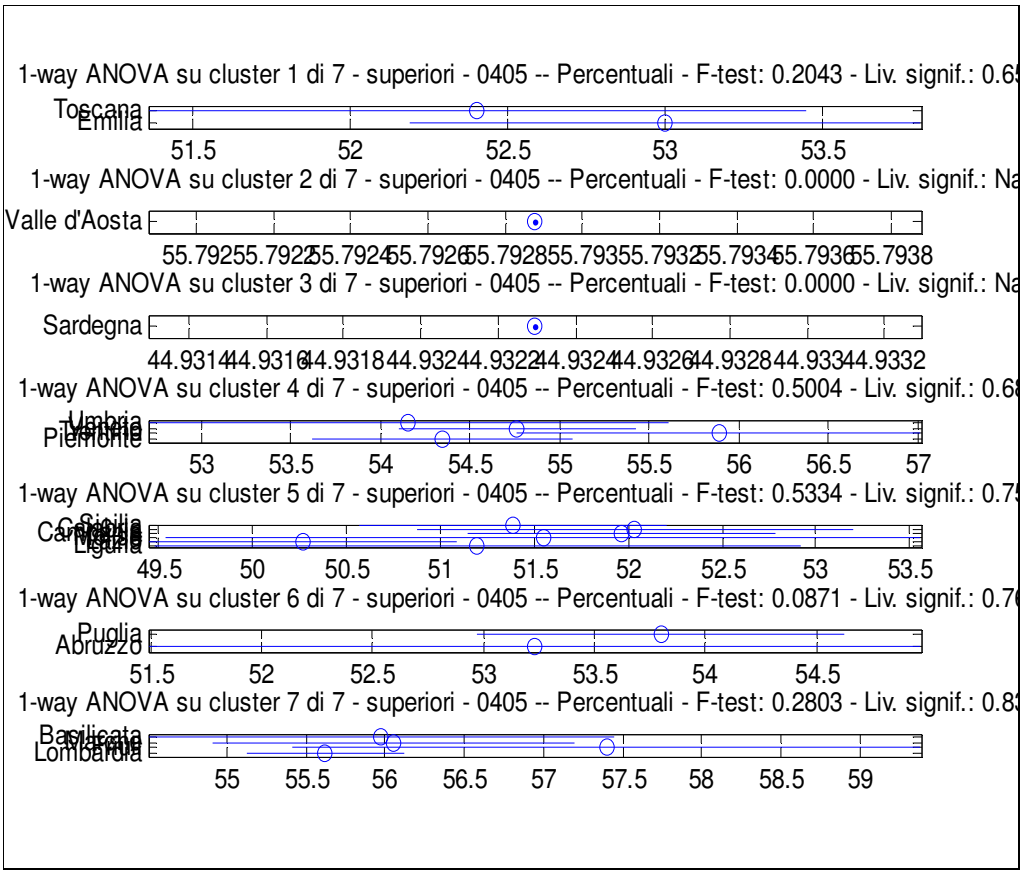
F-test: 2.803225e-001 - Liv. signif.: 8.396016e-001

Regione 1: Lombardia - Estimated mean: 55.62 +/- 0.50

Regione 2: Friuli - Estimated mean: 57.40 +/- 1.99

Regione 3: Marche - Estimated mean: 56.05 +/- 1.15

Regione 4: Basilicata - Estimated mean: 55.97 +/- 1.47



I dati in tabella confermano pienamente la partizione individuata dall'analisi dei gruppi restituendo livelli di significatività oggettivamente elevati.

4.3.3 Analisi ANOVA a.s. 2005/2006: Fattore Regione

1° GRUPPO

Grandmean: 47.87

F-test: 3.139280e-001 - Liv. signif.: 5.761269e-001

Regione 1: Emilia - Estimated mean: 48.23 +/- 0.76

Regione 2: Toscana - Estimated mean: 47.50 +/- 1.06

2° GRUPPO

Grandmean: 49.44

F-test: 0 - Liv. signif.: NaN

Regione 1: Valle d'Aosta - Estimated mean: 49.44 +/- 0.00

3° GRUPPO

Grandmean: 40.21

F-test: 0 - Liv. signif.: NaN

Regione 1: Sardegna - Estimated mean: 40.21 +/- 0.00

4° GRUPPO

Grandmean: 49.42

F-test: 1.732915e-001 - Liv. signif.: 9.144110e-001

Regione 1: Piemonte - Estimated mean: 49.29 +/- 0.75

Regione 2: Trentino - Estimated mean: 50.05 +/- 1.27

Regione 3: Veneto - Estimated mean: 49.03 +/- 0.63

Regione 4: Umbria - Estimated mean: 49.29 +/- 1.36

5° GRUPPO

Grandmean: 44.74

F-test: 3.583211e-001 - Liv. signif.: 8.768167e-001

Regione 1: Liguria - Estimated mean: 45.89 +/- 1.69

Regione 2: Lazio - Estimated mean: 44.82 +/- 0.80

Regione 3: Molise - Estimated mean: 43.48 +/- 1.77

Regione 4: Campania - Estimated mean: 44.95 +/- 0.76

Regione 5: Calabria - Estimated mean: 44.03 +/- 1.10

Regione 6: Sicilia - Estimated mean: 45.26 +/- 0.83

6° GRUPPO

Grandmean: 45.33

F-test: 7.796642e-002 - Liv. signif.: 7.804758e-001

Regione 1: Abruzzo - Estimated mean: 45.07 +/- 1.74

Regione 2: Puglia - Estimated mean: 45.59 +/- 0.69

7° GRUPPO

Grandmean: 49.44

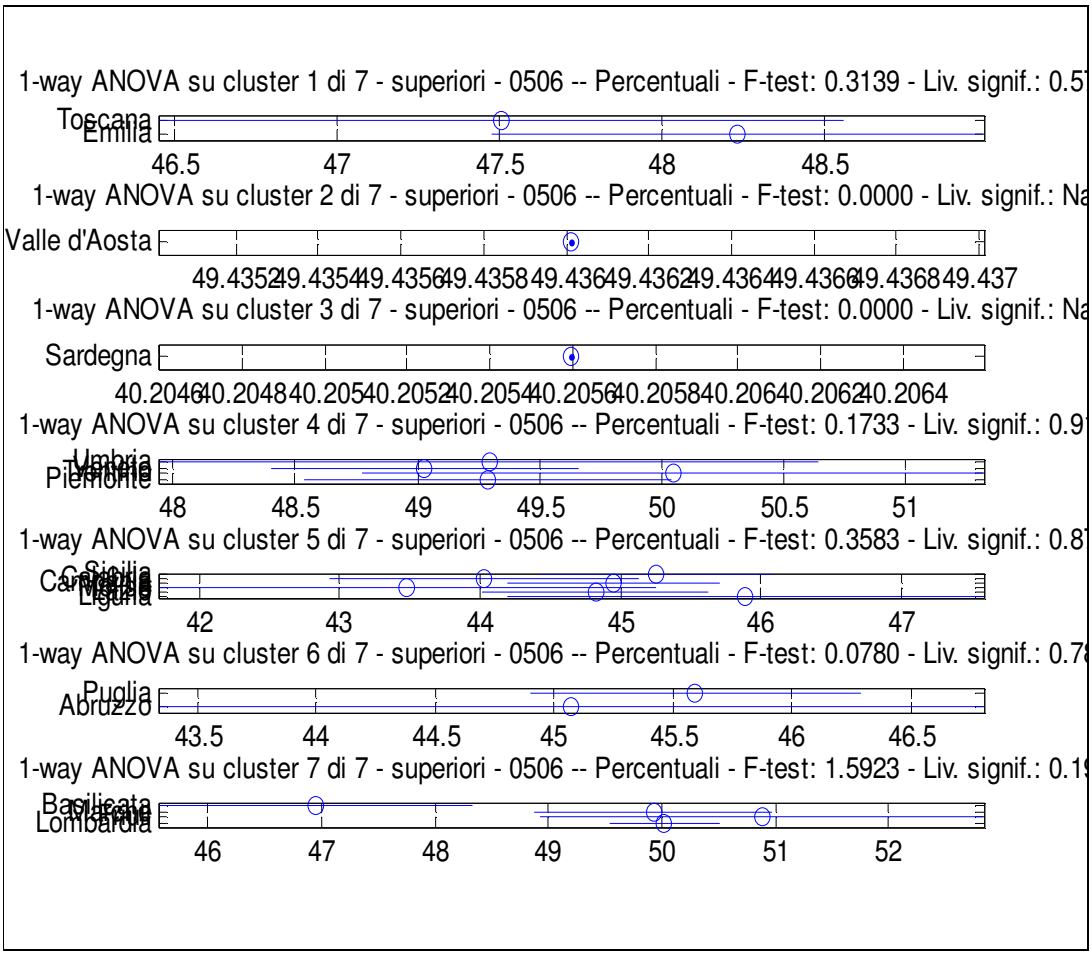
F-test: 1.592280e+000 - Liv. signif.: 1.911536e-001

Regione 1: Lombardia - Estimated mean: 50.02 +/- 0.48

Regione 2: Friuli - Estimated mean: 50.88 +/- 1.96

Regione 3: Marche - Estimated mean: 49.92 +/- 1.05

Regione 4: Basilicata - Estimated mean: 46.95 +/- 1.38



Anche per le analisi eseguite sui dati dell'a.s. 2005/2006 si riscontrano ottimi livelli di significatività.

Conclusioni

Scopo del lavoro alla base del presente report era quello di ridurre in maniera drastica il numero di variabili del sistema (relativamente all'apprendimento), raggruppandole in maniera omogenea. Tale operazione è necessaria in vista del modello di regressione tra la variabile di abilità e quelle di sistema. L'analisi effettuata ha chiaramente mostrato l'estrema difficoltà di una siffatta operazione di raggruppamento. In particolare alcune variabilità non risultano influenti sull'analisi successiva di regressione: in primo luogo il sesso, in quanto gli studenti sono ripartiti in maniera sufficientemente uguale tra i due sessi e in maniera sufficientemente omogenea tra le diverse regioni; inoltre il tipo di scuola (pubblica o privata) verrà trattato analizzando solo le scuole pubbliche che comunque rappresentano la grande maggioranza degli Istituti. Viceversa la variabilità per regione risulta problematica da analizzare per diversi motivi:

- l'incertezza sul dato delle elementari riscontrata nel report "Una procedura di qualità basata sulla fuzzy clustering per l'individuazione e la correzione dei dati anomali nell'ambito del Servizo Nazionale di Valutazione scolastica degli apprendimenti (SNV)", che grava in modo particolare sulle regioni del Mezzogiorno. Un'analisi preliminare effettuata utilizzando i fattori ponderali sviluppati nel Report citato (calcolati per classe), non riportata nel presente report, ha evidentemente modificato la struttura dei clusters individuati, ma non la sostanza della difficoltà di raggruppamento. Si tenga altresì conto che l'analisi di qualità dei dati di apprendimento non ha rivelato particolari problematiche per le Scuole Media inferiori, per cui l'analisi presentata nel presente report è da ritenersi corretta per quest'ordine di scuola.
- Si è osservata una netta variazione tra il 2004-2005 e il 2005-2006 dell'abilità degli studenti espressa in termini di percentuali di successo nelle risposte. Ciò può essere dovuto a fattori non legati alla reale abilità degli studenti (la maggiore difficoltà dei quesiti). L'analisi delle abilità valutate con l'Item Response Theory potrebbe in teoria correggere tale possibile sorgente di bias, tuttavia, come riportato nel Report, sono state evidenziate problematiche nei dati di abilità mediante IRT per le Scuole Medie Superiori che ne inficiano l'utilizzo. Inoltre andrebbe verificato qual è l'insieme di dati su cui sono stati stimati i parametri della IRT, per i quali globalmente l'abilità risulta nulla.

Pertanto nel prosieguo della ricerca si analizzeranno i dati per tutte le regioni globalmente, cercando di spiegarne la variabilità in base ai parametri di Sistema.