INVALSI

PROGETTO FINVALI 2005

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

Report dicembre 2006: Analisi ANOVA dei questionari di valutazione – Modello e interpretazioni generali

Umberto Amato[(1)], **Claudia Angelini**[(1)]

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

Abstract: Il report contiene l'attività svolta nel primo semestre del progetto riguardo l'analisi dei questionari di valutazione mediante modelli ANOVA. Nella prima parte vengono discussi i fondamenti teorici dei modelli ANOVA, con riferimento specifico a questionari di valutazione INVALSI. Nella seconda parte vengono presentate le analisi preliminari di tipo generale effettuate sull'intero database INVALSI 2004-2005 relativo alle scuole elementari e medie inferiori, per le quali cui sussisteva l'obbligo della valutazione.

Indice

1. Introduzione

Nel presente report verranno discussi gli strumenti statistici che sono stati utilizzati nella prima fase del progetto e che ne costituiranno l'asse portante anche nel prosieguo.

Nella prima parte del report, oltre a descrivere i fondamenti della teoria, in particolare verranno ripetutamente descritte le ricadute specifiche sul presente progetto, evidenziando le modalità con cui esse si esplicano e fornendo pertanto piena giustificazione alle scelte metodologiche operate nell'analisi dei dati del Sistema Scolastico.

La seconda parte del report contiene le analisi generali preliminari delle risposte degli studenti effettuate con la metodologia ANOVA. Il database di riferimento è quello dell'anno scolastico 2004-2005 relativamente alle scuole elementari e medie inferiori per cui sussisteva l'obbligo della valutazione (tale database è descritto nel report "Lettura dati INVALSI e sviluppo di software").

L'analisi è stata effettuata a partire da tutti i dai disponibili degli studenti senza analisi di qualità ed è da intendersi come uno studio preliminare tendente a valutare l'efficacia dei modelli ANOVA nel descrivere i dati. Il confronto dettagliato con l'anno scolastico 2005-2006, la rielaborazione dei dati alla luce dell'analisi di qualità dei dati che emergerà dal progetto, l'analisi disaggregata e l'interpretazione dei dati saranno oggetto del prosieguo della ricerca.

2. Modelli ANOVA

I modelli di Analisi della Varianza costituiscono un potente strumento statistico per studiare la relazione esistente tra una variabile (detta *responso*) e una o più variabili (dette *predittori*). Tale metodologia è fortemente connessa a quella della *Regressione* che si pone obiettivi analoghi e che verrà considerata in dettaglio nel prosieguo del progetto. In questa fase verranno comunque discusse le analogie e differenze tra le due metodologie.

Il punto fondamentale è che i modelli ANOVA non richiedono assunzioni circa la natura statistica della relazione che intercorre tra la variabile *responso* e le variabili *predittori*; inoltre esse non richiedono neanche l'ipotesi che le variabili *predittori* siano quantitative. Tale mancanza di assunzioni distingue notevolmente ANOVA dalla Regressione, che, al contrario, nella sua formulazione più comune si prefigge lo scopo di descrivere la natura della relazione statistica tra la variabile *responso* e le variabili *predittori*. Risulta tuttavia possibile dimostrare che riformulando il problema della regressione in modo opportuno anche considerando variabili qualitative, è possibile ricondursi ad un problema equivalente ad ANOVA e che fornirà pertanto esattamente le stesse soluzioni. Ciò nonostante la metodologia statistica ANOVA esiste come metodologia statistica distinta principalmente perché sfruttando la particolare struttura del problema è possibile addivenire a semplificazioni di tipo computazionale che rendono più veloce il calcolo della soluzione e più evidenti le proprietà della soluzione che interessano.

I modelli di Analisi della Varianza al pari di quelli di Regressione tendono a valutare mediante opportuni studi se alcune variabili (*predittori*) hanno effetto su un'altra variabile (*responso*). Tuttavia per ANOVA viene spesso utilizzata una particolare terminologia: le variabili *predittori* assumono il nome di *fattori* oppure *trattamenti*. Una prima categorizzazione dei modelli di Analisi della Varianza relativamente agli studi cui sono destinati comprende gli studi *Sperimentali* e quelli basati su *Osservazioni*. Gli studi sperimentali sono quelli in cui lo sperimentatore controlla completamente l'esperimento, nel senso che decide a priori quali *fattori* prendere in considerazione, i valori, quantitativi o qualitativi, che possono assumere ed anche il numero di dati sperimentali da rilevare

per ogni *fattore* e ogni valore possibile del *fattore*. In questo modo il controllo risulta totale e pertanto è possibile organizzare l'esperimento in modo da sfruttare a pieno le caratteristiche statistiche di ANOVA e rispettare le ipotesi di validità della metodologia. Viceversa negli studi basati sulle *Osservazioni* non è possibile un controllo preventivo di tutti gli aspetti dell'esperimento e, in particolare, non è possibile prevedere a priori il numero di dati rilevati per ogni *fattore* o valore assunto dal singolo *fattore*. Risulta ovvio che gli studi di tipo *Sperimentale* sono da preferire rispetto a quelli basati sulle *Osservazioni*, tuttavia il loro utilizzo è destinato a quelle applicazioni, tipiche nella statistica medica, dove, ad esempio, si vuole valutare l'effetto di una particolare terapia, si dispone di un'ampia popolazione e si può decidere a priori il numero di pazienti da sottoporre ai diversi trattamenti . È evidente che il presente progetto ricade nella categoria degli studi basati sull'osservazione (almeno per quanto concerne la Scuola Elementare e quella Secondaria di primo grado), in quanto non sussiste un controllo sul numero di studenti e/o scuole appartenenti a ciascun potenziale fattore (per esempio sesso, regione, tipologia della suola, ecc.). Nella terminologia ANOVA accanto al fattore si definisce il *livello* associato a quel *fattore* come la particolare forma (o valore) che può assumere il *fattore*. Nella fattispecie del presente progetto esempi di livelli associati ai fattori sono la tipologia `Maschio` oppure `Femmina` associata al fattore `Sesso`, oppure l'elenco delle Regioni associata al fattore `Regione`, oppure ancora la natura `Privata` o `Pubblica` associata alla `Tipologia` dell'Istituto, e così via.

Gli studi ANOVA differiscono anche in base al numero di *fattori* che sono oggetti di studio. Alcuni sono a singolo *fattore*, dove la variabile *responso* si presuppone sia legata ad un unico *predittore* (o *fattore*). Analogamente gli studi sono detti multifattore se tale dipendenza è legata a più *fattori*. Anche se lo studio relativo al Sistema Scolastico è certamente di tipo multifattore, tuttavia diverse considerazioni di tipo sia pratico, sia teorico, sia computazionale suggeriscono di isolare particolari fattori ed eseguire le corrispondenti analisi a *fattore* singolo.

Una successiva categorizzazione prevede la distinzione tra *fattore qualitativo* e *quantitativo*. Nel primo caso i livelli differiscono per qualche attributo di tipo qualitativo; nel secondo il livello è descritto da una quantità numerica definita su una certa scala. Nel caso del presente progetto sono presenti sia fattori di tipo qualitativo, sia quantitativo.

Costituiscono un esempio del primo tipo la regione di appartenenza delle scuole, il sesso, ecc.; nel secondo caso possiamo considerare, ad esempio, le risorse degli istituti scolastici (laboratori, biblioteca, disponibilità finanziaria, ecc.).

Infine negli studi a singolo fattore il trattamento corrisponde ad un livello del fattore. Per esempio un trattamento può essere costituito dal Molise in uno studio comprendente come unico fattore la Regione. Negli studi a fattore multiplo un trattamento si riferisce ad una combinazione di livelli dei fattori; ad esempio gli studenti maschi appartenenti a Istituti privati della Regione Puglia.

Uno studio che richiede gli strumenti di Analisi della Varianza va progettato accuratamente. In particolare vanno definiti i trattamenti da includere, che devono essere in grado di fornire indicazioni sui meccanismi che stanno alla base del fenomeno che si sta studiando. Va osservato che è bene che gli studi iniziali non tentino di esaminare l'intero meccanismo in tutti i dettagli, ma piuttosto devono mirare ad individuare i principali fattori coinvolti e a stimare l'entità degli effetti che producono. È compito di ulteriori studi fornire indicazioni maggiormente dettagliate sul sistema. Dunque la scelta dei trattamenti risulta un aspetto delicato della progettazione di uno studio. Allo stesso modo è possibile pensare di includere tra i trattamenti alcuni di controllo che consistono nell'applicare procedure identiche ad altri trattamenti, eccetto che per un effetto che si vuole analizzare. Un trattamento di controllo è necessario quando non ne è nota a priori l'efficacia. In questo caso il trattamento di controllo fornisce informazioni sull'efficacia del livello di un fattore rispetto alla condizione di controllo. Esso risulta indispensabile per esempio in quegli studi medici in cui si vuole valutare l'efficacia di diverse terapie nel curare una certa malattia: allora la presenza di un trattamento di controllo costituito da "nessuna terapia" rappresenta la base con cui confrontarsi per valutare l'efficacia delle varie terapie. Nel caso del progetto in esame, la natura completamente osservazionale dello studio non lascia spazio a scelte particolari sui trattamenti.

La scelta tra modelli a singolo fattore e a fattori multipli è legata all'analisi che si intende effettuare sui fattori dallo studio della variabile responso. In concreto, sudi a singolo fattore mirano a confrontare gli effetti dei diversi livelli del fattore e spesso ad individuare qual è il fattore considerato migliore. Gli studi a fattori multipli tendono a determinare se i diversi fattori interagiscono tra loro, quali fattori sono quelli chiave,

quali combinazioni di fattori risultano migliori. Nel presente studio del Sistema Scolastico sono presenti tutte queste componenti, per cui verranno condotto sia studi a fattore singolo, sia a fattori multipli.

Un'ulteriore distinzione dei modelli ANOVA (indicati come Modelli I e Modelli II) riguarda il ruolo che i livelli assumono rispetto alla popolazione possibile di livelli. Nei modelli di tipo I i livelli considerati sono quelli relativi all'intera popolazione e di fatto costituiscono una caratteristica intrinseca del fattore. Al contrario nei modelli di tipo II i livelli sono solo rappresentativi dell'intera popolazione e ne costituiscono pertanto solo un campione, le cui indicazioni desunte dall'analisi vengono poi estese agli altri livelli possibili dell'intera popolazione. Il presente studio sul Sistema Scolastico si configura come un modello di tipo I.

2.1 Modello ANOVA a 1 fattore

Gli ingredienti di un modello ANOVA ad 1 fattore sono piuttosto semplici. Associata a ciascun livello del fattore vi è una distribuzione di probabilità del responso. Il modello ANOVA assume che

1. Ciascuna densità di probabilità è normale
2. Le densità di probabilità dei diversi livelli hanno la stessa varianza
3. Il dati rilevati del responso per ciascun livello del fattore costituiscono una campione casuale estratto dalla corrispondente densità di probabilità del livello e sono, inoltre, indipendenti dal responso di ogni altro livello del fattore.

L'analisi di un modello ANOVA generalmente viene eseguita in due passi successivi

a) determinare se i valori medi dei livelli del fattore sono uguali tra loro;

b) nel caso in cui i valori medi differiscano, esaminare in che misura e quali sono le implicazioni delle differenze riscontrate.

Introduciamo ora la notazione rilevante.

Sia r il numero di livelli del fattore oggetto di studio. Indichiamo il numero di dati dell'i-mo livello del fattore con n_i a il numero di dati complessivo con n_T, dove ovviamente

$$n_T = \sum_{i=1}^{r} n_i \, .$$

Indichiamo poi con Y_{ij} il valore della variabile responso per l'*i*-mo livello del fattore e riferita al *j*-mo campione.

Il modello ANOVA di tipo I si scrive come

$$Y_{ij} = \mu_i + \varepsilon_{ij} \, , \quad (1)$$

dove μ_i, $i=1,\ldots,r$ sono i parametri del modello ANOVA e ε_{ij} sono variabili Gaussiane indipendenti a media nulla e varianza σ^2. Un siffatto modello è chiamato a *medie per cella*.

Dal modello si evince immediatamente che il valore osservato del predittore *Y* per l'*i*-mo livello del fattore nel *j*-mo campione è la somma di due componenti: un termine costante, μ_i, ed un termine di errore casuale, ε_{ij}. Poiché $E[\varepsilon_{ij}]=0$, ne segue che $E[Y_{ij}] = \mu_i$.

Pertanto tutte le osservazioni Y_{ij} relative allo stesso *i*-mo livello del fattore hanno lo stesso valore medio μ_i che si configura quindi come il responso medio per l'i-mo livello del fattore. Poiché i μ_i sono costanti, ne consegue che $\sigma^2(Y_{ij}) = \sigma^2(\varepsilon_{ij}) = \sigma^2$, vale a dire si assume che tutte le osservazioni abbiano la stessa varianza, indipendentemente dal livello del fattore. Inoltre, poiché si assume che gli ε_{ij} hanno una distribuzione Gaussiana, lo stesso vale anche per gli Y_{ij}. L'ipotesi di indipendenza dei termini di errore significa che una qualunque campione non influisce sul termine di errore di qualunque altro campione dello stesso livello del fattore o per una altro livello. Ne consegue che anche i responsi Y_{ij} sono indipendenti.

Il modello ANOVA è evidentemente lineare. Esso può essere riscritto in una forma matriciale che si rileverà utile nel caso di analisi a più fattori. Siano $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$ e $\boldsymbol{\varepsilon}$ vettori definiti come

$$\mathbf{Y}=\begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1,n_1} \\ \vdots \\ Y_{r1} \\ Y_{r2} \\ \vdots \\ Y_{r,n_r} \end{bmatrix},\quad \mathbf{X}=\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & 0 & \vdots \\ 1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix},\quad \boldsymbol{\beta}=\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_r \end{bmatrix},\quad \boldsymbol{\varepsilon}=\begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1,n_1} \\ \vdots \\ \varepsilon_{r1} \\ \varepsilon_{r2} \\ \vdots \\ \varepsilon_{r,n_r} \end{bmatrix}. \quad (2)$$

Allora il modello ANOVA potrà essere scritto come

$$\mathbf{Y}=\begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1,n_1} \\ \vdots \\ Y_{r1} \\ Y_{r2} \\ \vdots \\ Y_{r,n_r} \end{bmatrix}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}=\begin{bmatrix} \mu_1 \\ \mu_1 \\ \vdots \\ \mu_1 \\ \vdots \\ \mu_r \\ \mu_r \\ \vdots \\ \mu_r \end{bmatrix}+\begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1,n_1} \\ \vdots \\ \varepsilon_{r1} \\ \varepsilon_{r2} \\ \vdots \\ \varepsilon_{r,n_r} \end{bmatrix}. \quad (3)$$

In tale formulazione appare evidente la forte interrelazione esistente con i modelli di regressione, che sarà discussa in maggiore dettaglio nel seguito. In particolare per completare il modello di regressione osserviamo che i termini di errore del modello ANOVA presentano la stessa struttura dei modelli di regressione lineare generali (indipendenza, varianza costante), vale a dire la matrice di varianza-covarianza dei termini di errore risulta la stessa:

$$\sigma^2(\varepsilon)=\begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix},$$

per cui

$$\sigma^2(\mathbf{Y}) = \sigma^2\mathbf{I}.$$

In uno studio sperimentale le medie dei livelli del fattore μ_i assumono il significato del responso medio che si otterrebbe se l'i-mo trattamento (livello) fosse applicato all'intera popolazione. Allo stesso modo la varianza σ^2 si riferisce alla variabilità del responso se un qualunque trattamento (livello) fosse applicato all'intera popolazione.

È molto importante osservare che nei casi pratici è estremamente probabile che lo studio sperimentale condotto non soddisfi a pieno le ipotesi di validità del modello ANOVA. Va tuttavia rilevato che il modello ANOVA è sufficientemente robusto alle ipotesi, per cui l'approssimazione dell'analisi statistica è generalmente adeguata agli scopi degli studi.

2.1.2 ANOVA e regressione

Si è già accennato alla possibilità di considerare il modello ANOVA come un modello di regressione. Faremo ora vedere che in effetti le soluzioni ottenute tramite le metodologie specifiche di ANOVA e quelle generali di regressione coincidono.

Introduciamo preliminarmente la notazione specifica.

Sia come in precedenza Y_{ij} il valore della variabile responso per l'i-mo livello del fattore e riferita al j-mo campione. Indichiamo con $Y_{i\cdot}$ il totale delle osservazioni relative all'i-mo livello del fattore,

$$Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}$$

Dove la notazione del punto indica una somma sull'indice corrispondente (j in questo caso) relativamente agli altri indici (il livello i del fattore nel presente caso). Indichiamo la media campionaria dell'i-mo livello del fattore con

$$\overline{Y}_{i\cdot} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} = \frac{Y_{i\cdot}}{n_i}.$$

Allo stesso modo il totale delle osservazioni dello studio sarà indicato con $Y_{\cdot\cdot}$, dove

$$Y_{\cdot\cdot} = \sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{ij} ,$$

con la convenzione che i doppi punti indicano una somma su ambedue gli indici cui si riferiscono. Infine la media globale dell'intero campione di responsi sarà indicata con $\overline{Y}_{\cdot\cdot}$,

$$\overline{Y}_{\cdot\cdot} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{ij}}{n_T} .$$

Nella tecnica dei minimi quadrati per la regressione si minimizza la somma dei quadrati delle deviazioni delle osservazioni rispetto ai valori medi attesi rispetto ai parametri incogniti. Nei modelli ANOVA sappiamo, come visto in precedenza, che il valore atteso delle osservazioni è proprio il valor medio dei livelli ($E[Y_{ij}]=\mu_i$). Pertanto la quantità da minimizzare risulta essere

$$Q = \sum_i \sum_j \left(Y_{ij} - \mu_i\right)^2$$

Che può essere scritta come

$$Q = \sum_j (Y_{1j} - \mu_1)^2 + \sum_j (Y_{2j} - \mu_2)^2 + \cdots \sum_j (Y_{rj} - \mu_r)^2 .$$

Poiché ciascun parametro incognito appare in un solo termine della sommatoria, allora Q può essere minimizzato minimizzando indipendentemente ciascun termine della sommatoria. Poiché risulta ben noto (e si dimostra facilmente) che la media campionaria minimizza la somma delle deviazioni quadratiche, ne risulta che lo stimatore ai minimi quadrati di μ_i, indicato con $\hat{\mu}_i$, è

$$\hat{\mu}_i = \overline{Y}_i .$$

Pertanto il valore stimato (fit) dell'osservazione Y_{ij}, indicato con $\hat{Y}_{ij}$, è semplicemente la media campionaria del corrispondente livello del fattore:

$$\hat{Y}_{ij} = \overline{Y}_i .$$

Allo stesso modo si può dimostrare che la soluzione ottenuta mediante il metodo dei minimi quadrati coincide con la soluzione che si otterrebbe applicando il metodo della massima verosimiglianza.

Una menzione speciale va riservata ai residui. Essi sono utilissimi nel valutare l'idoneità di un modello ANOVA a descrivere i dati dell'esperimento. I residui e_{ij} sono definiti come

$$e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \overline{Y}_{i\cdot}$$

Una importante proprietà dei residui del modello ANOVA è che la loro somma è nulla per ciascun livello del fattore:

$$\sum_j e_{ij} = 0, \quad i = 1, \ldots, r\,.$$

Inoltre l'analisi dei residui risulta molto importante per verificare l'appropriatezza del modello ANOVA. A tale scopo introduciamo una notazione molto classica nei modelli ANOVA.

Sia $Y_{ij} - \hat{Y}_{\cdot\cdot}$ la variabilità delle osservazioni Y_{ij} indipendentemente dal livello. Allo stesso modo utilizzando le informazioni sui livelli la deviazione dovuta all'incertezza sui dati rispetto al valore medio stimato del singolo livello è $Y_{ij} - \overline{Y}_{i\cdot}$ La differenza tra ambedue le deviazioni rappresenta ovviamente la differenza tra la media del livello del fattore e la media globale:

$$\left(Y_{ij} - \overline{Y}_{\cdot\cdot}\right) - \left(Y_{ij} - \overline{Y}_{i\cdot}\right) = \left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right).$$

In altre parole la deviazione totale si compone di due termini:

$$Y_{ij} - \overline{Y}_{\cdot\cdot} = \overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot} \quad + \quad Y_{ij} - \overline{Y}_{i\cdot}$$

a) la deviazione della media del livello rispetto alla media globale

b) la deviazione di Y_{ij} rispetto alla media del livello, in pratica il residuo e_{ij}.

Elevando al quadrato e sommando sui campioni si ottiene

$$\underbrace{\sum_i \sum_j \left(Y_{ij} - \overline{Y}_{\cdot\cdot}\right)^2}_{\text{SSTO}} = \underbrace{\sum_i n_i \left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2}_{\text{SSTR}} + \underbrace{\sum_i \sum_j \left(Y_{ij} - \overline{Y}_{i\cdot}\right)^2}_{\text{SSE}}\,. \quad (4)$$

Il termine a sinistra è denominato *somma totale dei quadrati* (SSTO), il primo termine a destra *somma dei quadrati dei trattamenti* (SSTR) ed il secondo *somma dei quadrati degli errori* (SSE), con SSTO = SSTR + SSE.

- SSE. SSE è una misura delle variazione casuale delle osservazioni rispetto alla media (stimata) del livello del fattore: minore è la variazione tra le osservazioni

per ciascun livello del fattore, più piccolo sarà SSE. Nel caso limite di SSE = 0, le osservazioni relative ad un determinato livello saranno tutte nulle e questo risulterà vero per tutti i livelli. Al contrario quanto più le osservazioni differiranno rispetto ai valori medi (stimati) di livello, tanto più altro sarà il valore di SSE.

- SSRT. È una misura delle differenze tra le medie (stimate) dei livelli del fattore, basata sulle deviazioni delle medie dei livelli stimate rispetto alla media globale. Se le medie stimate dei livelli fossero tutte uguali, risulterebbe SSTR = 0; al contrario, quanto più le medie dei livelli differiscono, tanto più alto risulterà SSTR.

In base alla decomposizione $\underbrace{\sum_i \sum_j \left(Y_{ij} - \overline{Y}_{\cdot\cdot}\right)^2}_{\text{SSTO}} = \underbrace{\sum_i n_i \left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2}_{\text{SSTR}} + \underbrace{\sum_i \sum_j \left(Y_{ij} - \overline{Y}_{i\cdot}\right)^2}_{\text{SSE}}$.

(4) è possibile stabilire il numero di gradi di libertà posseduti da ciascun termine:

- SSTO possiede n_T-1 gradi di libertà associati. Essi provengono dal fatto che il termine è basato su n_T deviazioni $Y_{ij} - \hat{Y}_{\cdot\cdot}$, ma un grado di libertà viene perso perché la loro somma è nulla: $\sum_i \sum_j Y_{ij} - \hat{Y}_{\cdot\cdot} = 0$.
- SSTR possiede r-1 gradi di libertà associati. Infatti vi sono r deviazioni delle medie dei livelli dalla media globale $Y_{ij} - \overline{Y}_{\cdot\cdot}$, tuttavia un grado di libertà viene perso perché le deviazioni non sono indipendenti, ma la loro somma pesata è nulla: $\sum_i n_i \left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right) = 0$.
- SSE possiede $n_T - r$ gradi di libertà. Infatti per il livello i-mo del fattore la componente di SSE risulta essere

$$\sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\cdot}\right)^2,$$

per cui vi sono n_i-1 gradi di libertà associati con il livello i-mo. Poichè SSE è la somma su tutti i livelli, ne consegue che il numero di gradi di libertà totale è ovviamente $n_T - r$.

Associati ai termini di errore vi sono poi i termini di medie quadratiche, ottenute dividendo i termini di errori per i rispettivi gradi di libertà:

$$\text{MSTR} = \frac{\text{SSTR}}{r-1} \quad \text{MSE} = \frac{\text{SSE}}{n_T - r}.$$

Tutte le informazioni sulle deviazioni e le medie quadratiche vengono tradizionalmente racchiuse in una Tabella ANOVA composta nel seguente modo:

Sorgente di variazione	SS	Gradi di libertà	MS	E[MS]
Tra i trattamenti	$\text{SSTR} = \sum_i n_i \left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2$	r -1	$\text{MSTR} = \frac{\text{SSTR}}{r-1}$	$\sigma^2 + \frac{\sum n_i (\mu_i - \mu_\cdot)^2}{r-1}$
Errore (all'interno dei trattamenti)	$\text{SSE} = \sum_i \sum_j \left(Y_{ij} - \overline{Y}_{i\cdot}\right)^2$	$n_T - r$	$\text{MSE} = \frac{\text{SSE}}{n_T - r}$	σ^2
Totale	$\text{SSTO} = \sum_i \sum_j \left(Y_{ij} - \overline{Y}_{\cdot\cdot}\right)^2$	n_T -1		

Tabella 1: Tavola ANOVA

I termini dell'ultima colonna rappresentano i valori attesi degli errori medi quadratici. In particolare MSE risulta uno stimatore corretto di σ^2, la varianza dei termini di errore ε_{ij}, e questo vale indipendentemente dal fatto che le medie dei livelli μ_i siano o non siano uguali tra loro. Ciò dipende dal fatto che la variabilità delle osservazioni all'interno di ciascun livello del fattore non è influenzata dal valore delle medie dei livello per popolazioni normali. Inoltre quando le medie dei livelli μ_i sono uguali tra loro, allora E[MSTR]= σ^2 in quanto il secondo termine della formula nella tavola ANOVA si annulla. Pertanto sia NSTR, sia MSE stimano la varianza dell'errore σ^2 quando tutte le medie dei livelli del fottore sono uguali. Tuttavia nel caso in cui le medie dei livelli non siano tutte uguali tra loro, allora MSTR tende in media ad essere maggiore di MSE a causa dell'influenza del secondo termine della formula. Tale proprietà è alla base della costruzione del test statistico per determinare se le medie dei livelli del fattore sono uguali tra loro: se MSTR e MSE sono dello stesso ordine di grandezza, si può ragionevolmente desumere che le medie μ_i dei livelli del fattore sono uguali. Se MSTR è

significativamente maggiore di MSE, allora si può pensare che i μ_i non sono tutti uguali tra loro.

2.1.3 F-test per l'eguaglianza delle medie dei livelli

L'obiettivo principale di un'analisi ANOVA a singolo fattore è determinare se in uno studio le medie μ_i dei livelli del fattore sono uguali tra loro. In pratica le ipotesi alternative che noi andiamo a considerare sono

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$$
$$H_a : \text{non tutti gli } \mu_1 \text{ sono uguali}$$

Il test statistico utilizzato per scegliere tra le alternative H_0 *e* H_a è

$$F^* = \frac{\text{MSTR}}{\text{MSE}}. \quad (5)$$

Valori alti di F^* supportano l'ipotesi H_a, poiché MSTR tenderà ad essere maggiore di MSE quando vale l'ipotesi H_a; valori di F^* attorno a 1 supporteranno invece l'ipotesi H_0, dal momento che sia MSTR, sia MSE hanno lo stesso valore medio atteso quando vale l'ipotesi H_0. Ne consegue che il test è a una coda (quella superiore).

Quando tutte le medie dei trattamenti μ_i sono uguali tra loro, ciascun responso Y_{ij} ha lo stesso valore atteso. Per l'additività delle somme dei quadrati dal teorema di Cochran consegue che quando vale H_0, SSE/σ^2 e SSTR/σ^2 sono variabili indipendenti distribuite come una distribuzione χ^2 e pertanto in questo caso F^* è distribuito come una distribuzione F (Fisher) con r-1 e n_T - r gradi di libertà, $F(r-1, n_T - r)$. In caso contrario, quando vale l'ipotesi alternativa H_a, vale a dire le medie μ_i non sono tutte uguali tra loro, allora F^* non segue la distribuzione F (in effetti si può dimostrare che segue la cosiddetta distribuzione F non centrale).

Osserviamo inoltre che SSTR e SSE sono indipendenti tra loro anche nel caso in cui tutte le medie mi non siano uguali tra loro. Infatti SSTR è basato esclusivamente sulla stima delle medie dei livelli $\overline{Y}_{i\cdot}$. Allo stesso modo SSE tiene conto della variabilità del

campione all'interno dei livelli del fattore, che non è influenzata dalla media del livello quando gli errori sono distribuiti secondo una normale.

Per utilizzare operativamente la statistica F^* è necessario costruire una regola di decisione, che difatti controlla il rischio di commettere un errore di Tipo I. Successive analisi permettono poi di controllare anche gli errori di Tipo II. Poiché sappiamo che quando vale l'ipotesi H_0 la statistica F^* è distribuita secondo una distribuzione di Fisher $F(r-1, n_T - r)$ e che valori alti di F^* portano ad accettare l'ipotesi alternativa H_a, allora la regola di decisione appropriata è costituita da

$$\begin{array}{l} Se\ F^* \le F(1-\alpha; r-1, n_T - r),\ l'ipotesi\ H_0\ è\ vera \\ Se\ F^* > F(1-\alpha; r-1, n_T - r),\ l'ipotesi\ H_a\ è\ vera \end{array},$$

dove $F(1-\alpha; r-1, n_T - r)$ è il 100(1-α) percentile della distribuzione F con r - 1 e $n_T - r$ gradi di libertà.

Si dimostra che nel caso il fattore sia composto di due soli livelli, allora il test F^* (che diventa con 1 e n_T - 2 gradi di libertà) è equivalente al t-test a due code con n_T - 2 gradi di libertà.

2.1.4 Formulazione alternativa del Modello ANOVA – Modello a effetti

Spesso viene utilizzata per gli studi una formulazione diversa ma completamente equivalente per i modelli ANOVA a singolo fattore. Tale formulazione prende il nome di modello a effetti di fattore. In essa le medie dei trattamenti mi sono espresse in una formulazione equivalente mediante l'identità

$$\mu_i = \mu_\cdot + (\mu_i - \mu_\cdot),$$

dove la costante $\mu_\cdot$ sarà definita in seguito. Indichiamo con τ_i la differenza $\tau_i = \mu_i - \mu_\cdot$, in modo che il valore medio μ_i può essere espresso come $\mu_i = \mu_\cdot + \tau_i$; la differenza τ_i viene chiamata l'effetto dell'i-mo livello del fattore oppure direttamente effetto dell'i-mo trattamento. Pertanto in definitiva il modello ANOVA può essere riscritto in maniera equivalente come

$$Y_{ij} = \mu_\cdot + \tau_i + \varepsilon_{ij} \quad (6)$$

dove

- $\mu_{\cdot}$ è una componente costante comune a tutte le osservazioni
- τ_i è l'effetto dell'i-mo livello del fattore (costante per ogni livello del fattore)
- ε_{ij} sono variabili indipendenti distribuite come Gaussiane $N(0,\sigma^2)$
- $i = 1,\ldots,r\,;\ j = 1,\ldots,n_i\,.$

Il modello $Y_{ij} = \mu_{\cdot} + \tau_i + \varepsilon_{ij}$ **(6)** è chiamato ad effetti di fattore proprio perché è espresso direttamente in termini di effetti del fattore τ_i, in modo da distinguerlo dal modello originario presentato in precedenza, denominato a cella in quanto espresso in termini del valor medio del trattamento (cella) μ_i. Anche il modello a effetti di fattore è di tipo lineare, al pari di quello a cella.

Rimane da determinare la definizione di $\mu_{\cdot}$, che può essere fatta in diversi modi.

Media non pesata. In questo caso risulta

$$\mu_{\cdot} = \frac{\sum_{i=1}^{r} \mu_i}{r}, \quad (7)$$

da cui si ottiene

$$\sum_{i=1}^{r} \tau_i = 0 \quad (8)$$

Media pesata. In questo caso

$$\mu_{\cdot} = \sum_{i=1}^{r} w_i \mu_i \ , \ \text{con} \ \sum_{i=1}^{r} w_i = 1$$

dove i w_i sono pesi opportuni. Ne consegue che

$$\sum_{i=1}^{r} w_i \tau_i = 0$$

La scelta dei pesi w_i può essere dettata da diversi criteri. Un primo criterio è costituito dall'importanza relativa che possono assumere i diversi livelli del fattore, se nota. Un criterio più diffuso è legato alla taglia del campione sperimentale di ogni livello del campione:

$$w_i = \frac{n_i}{n_T}$$

per cui

$$\mu_{\cdot} = \sum_{i=1}^{r} \frac{n_i}{n_T} \mu_i$$

la cui stima si dimostra facilmente essere data da $\overline{Y}_{\cdot\cdot}$:

$$\hat{\mu}_{\cdot} = \sum_{i=1}^{r} \frac{n_i}{n_T} \overline{Y}_{i\cdot} = \overline{Y}_{\cdot\cdot}.$$

Data l'equivalenza dei modelli a cella e a fattori, il test di eguaglianza delle medie è basato sulla stessa statistica F^*. L'unica differenza è che la formulazione del test diventa

$$H_0 : \tau_1 = \tau_2 = \cdots \tau_r = 0$$
$$H_a : \text{non tutti i } \tau_i \text{ sono uguali a } 0$$

2.1.5 Analisi della varianza a fattore singolo mediante regressione

Torniamo a considerare il modello di analisi della varianza visto come modello di regressione. Nonostante le formulazioni siano equivalenti, tuttavia vedremo che nel caso di più fattori il modello basato sulla regressione assumerà caratteristiche specifiche.

Consideriamo nuovamente la formulazione a effetti di fattore $Y_{ij} = \mu_{\cdot} + \tau_i + \varepsilon_{ij}$ **(6)** e assumiamo che $\mu_{\cdot}$ sia definito con pesi uguali (Equazione $\mu_{\cdot} = \frac{\sum_{i=1}^{r} \mu_i}{r}$, **(7)**. Poiché la somma dei fattori sui livelli è nulla (Equazione $\sum_{i=1}^{r} \tau_i = 0$ **(8)**, ne consegue che uno degli *r* parametri t_i (per esempio l'*r*-mo) non è necessario nel modello ANOVA, in quanto può essere espresso in termini dei restanti *r* -1 parametri mediante il vincolo ($\sum_{i=1}^{r} \tau_i = 0$ **(8)**:

$$\tau_r = -\tau_1 - \tau_2 - \cdots - \tau_{r-1}.$$

Pertanto i parametri incogniti del modello ANOVA risultano essere $\mu_\cdot$, $\tau_1,\ldots,\tau_{r-1}$, cosicché nella formulazione matriciale del modello ANOVA

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1,n_1} \\ \vdots \\ Y_{r1} \\ Y_{r2} \\ \vdots \\ Y_{r,n_r} \end{bmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \begin{bmatrix} \mu_1 \\ \mu_1 \\ \vdots \\ \mu_1 \\ \vdots \\ \mu_r \\ \mu_r \\ \vdots \\ \mu_r \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1,n_1} \\ \vdots \\ \varepsilon_{r1} \\ \varepsilon_{r2} \\ \vdots \\ \varepsilon_{r,n_r} \end{bmatrix}.$$

(3) risulta

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1,n_1} \\ \vdots \\ Y_{r1} \\ Y_{r2} \\ \vdots \\ Y_{r,n_r} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & \vdots & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & \ddots & \vdots \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu_\cdot \\ \tau_2 \\ \vdots \\ \tau_{r-1} \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1,n_1} \\ \vdots \\ \varepsilon_{r1} \\ \varepsilon_{r2} \\ \vdots \\ \varepsilon_{r,n_r} \end{bmatrix}.$$

Per i valori attesi risulta

$$\mathrm{E}[\mathbf{Y}] = \begin{bmatrix} \mathrm{E}[Y_{11}] \\ \mathrm{E}[Y_{12}] \\ \vdots \\ \mathrm{E}[Y_{1,n_1}] \\ \vdots \\ \mathrm{E}[Y_{r1}] \\ \mathrm{E}[Y_{r2}] \\ \vdots \\ \mathrm{E}[Y_{r,n_r}] \end{bmatrix} = \mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & \vdots & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & \ddots & \vdots \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \mu_\cdot \\ \tau_2 \\ \vdots \\ \tau_{r-1} \end{bmatrix} = \begin{bmatrix} \mu_1 + \tau_1 \\ \mu_1 + \tau_1 \\ \vdots \\ \mu_1 + \tau_1 \\ \vdots \\ \mu_\cdot - \tau_1 - \cdots - \tau_{r-1} \\ \mu_\cdot - \tau_1 - \cdots - \tau_{r-1} \\ \vdots \\ \mu_\cdot - \tau_1 - \cdots - \tau_{r-1} \end{bmatrix};$$

poiché $\tau_r = -\tau_1 - \tau_2 - \cdots - \tau_{r-1}$, allora $\mathrm{E}[Y_{rj}] = \mu_. + \tau_r$, per cui la rappresentazione matriciale in termini di modello a fattori fornisce per tutti i livelli *i*

$$\mathrm{E}[Y_{ij}] = \mu_. + \tau_i$$

come per il modello a cella.

Osserviamo inoltre che nel caso del modello a fattori la matrice **X**, pur semplice, ha una struttura diversa dal modello a cella, in quanto sono presenti valori pari a -1, 0 e 1. Formalmente pertanto il modello a regressione multipla si scrive come

$$Y_{ij} = \mu_. + \tau_1 X_{ij1} + \tau_2 X_{ij2} + \cdots + \tau_{r-1} X_{ij,r-1} + \varepsilon_{ij} = \mu_. + \sum_{k=1}^{r-1} \tau_k X_{ijk} + \varepsilon_{ij},$$

dove

$$X_{ijk} = \begin{cases} 1 \text{ se } k = i \\ -1 \text{ se } i = r \\ 0 \text{ altrimenti} \end{cases}.$$

Si osservi infine come nella presente formulazione mediante regressione multipla sia presente il termine noto da stimare $\mu_.$, contrariamente al modello di regressione a cella

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1,n_1} \\ \vdots \\ Y_{r1} \\ Y_{r2} \\ \vdots \\ Y_{r,n_r} \end{bmatrix} = \mathbf{X\beta} + \boldsymbol{\varepsilon} = \begin{bmatrix} \mu_1 \\ \mu_1 \\ \vdots \\ \mu_1 \\ \vdots \\ \mu_r \\ \mu_r \\ \vdots \\ \mu_r \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1,n_1} \\ \vdots \\ \varepsilon_{r1} \\ \varepsilon_{r2} \\ \vdots \\ \varepsilon_{r,n_r} \end{bmatrix}.$$

(si confronti con l'Equazione **(3)**).

L'approccio della regressione multipla generalmente non viene utilizzato per l'analisi della varianza a 1 fattore. Il motivo sta nel fatto che la matrice **X** ha una struttura molto semplice che consente semplificazioni computazionali in grado di ottenere la soluzione in modo estremamente semplice e con un numero di operazioni molto minore. Tuttavia, aldilà della generalizzazione dei modelli ANOVA offerta dai modelli lineari, il motivo principale per considerare modelli di regressione multipla lineari sta nel fatto che nel caso il numero di fattori sia maggiore di 1, la struttura della matrice si complica, anche perché

il tipo di analisi che si intende compiere dipende molto dalle scelte dell'investigatore (per esempio quali fattori considerare e quali interazioni tra essi); ne consegue che non risulta sempre possibile ottenere le soluzioni in modo semplificato, mentre il modello di regressione multipla rappresenta una struttura generale che si adatta alle varie configurazioni possibili del modello ANOVA.

2.2 Modello ANOVA a due fattori

Analizziamo ora il caso di studi a più fattori, prendendo come riferimento il caso di due fattori. Tali situazioni sono pertinenti al problema del Sistema Scolastico. L'approccio che seguiremo considererà dapprima il caso in cui il numero di osservazioni coincide per ciascuna coppia di livelli (uno per ogni fattore). In seguito considereremo il caso in cui i livelli dei fattori hanno un numero di campioni che differisce in generale tra i livelli stessi. Tale situazione è tipica dei dati rilevati dai questionari. Essa distrugge l'ortogonalità della decomposizione ANOVA, per cui le formule che abbiamo visto sulla decomposizione delle somme dei quadrati non sono più valide.

Siano A e B i due fattori; i e j gli indici dei livelli corrispondenti; a il numero di livelli del fattore A e b il numero di livelli del fattore B; n_{ij} la taglia del campione relativo al trattamento corrispondente al livello i del fattore A e al livello j del fattore B. Il numero totale di casi per il livello i del fattore A verrà indicato come

$$n_{i\cdot} = \sum_j n_{ij} ;$$

analogamente il numero totale di casi per il j-mo livello del fattore B sarà denotato come

$$n_{\cdot j} = \sum_i n_{ij} .$$

Ovviamente inoltre il numero totale di casi sarà indicato con

$$n_T = \sum_i \sum_j n_{ij} .$$

La stima della media del trattamento relativo al livello i del fattore A e al livello j del fattore B si ottiene come

$$\overline{Y}_{ij\cdot} = \frac{Y_{ij\cdot}}{n_{ij}}$$

dove con notazione consistente con il caso monofattoriale

$$Y_{ij\cdot} = \sum_{k=1}^{n_{ij}} Y_{ijk} \ .$$

In generale uno studio ANOVA a due fattori può essere concettualmente rappresentato mediante una matrice in cui il fattore *A* è rappresentato per righe ed il fattore *B* per colonne:

Fattori-livelli	*B* 1	*B* 2	…	*B* b	Medie per il fattore *A*
A 1	Y_{111} Y_{112} … $Y_{11,n_{11}}$	Y_{121} Y_{122} … $Y_{12,n_{12}}$	… … … …	Y_{1b1} Y_{1b2} … $Y_{1b,n_{1b}}$	$\overline{Y}_{1\cdot\cdot}$
	$\overline{Y}_{11\cdot}$	$\overline{Y}_{12\cdot}$	…	$\overline{Y}_{1b\cdot}$	
A 2	Y_{211} Y_{212} … $Y_{21,n_{21}}$	Y_{221} Y_{222} … $Y_{22,n_{22}}$	… … … …	Y_{2b1} Y_{2b2} … $Y_{2b,n_{2b}}$	$\overline{Y}_{2\cdot\cdot}$
	$\overline{Y}_{21\cdot}$	$\overline{Y}_{22\cdot}$	…	$\overline{Y}_{2b\cdot}$	
…	… … …	… … …	… … …	… … …	…
A a	Y_{a11} Y_{a12} … $Y_{a1,n_{a1}}$	Y_{a21} Y_{a22} … $Y_{a2,n_{a1}}$	… … … …	Y_{ab1} Y_{ab2} … $Y_{ab,n_{a1}}$	$\overline{Y}_{a\cdot\cdot}$
	$\overline{Y}_{a1\cdot}$	$\overline{Y}_{a2\cdot}$	…	$\overline{Y}_{ab\cdot}$	
Medie per il fattore *B*	$\overline{Y}_{\cdot 1\cdot}$	$\overline{Y}_{\cdot 2\cdot}$	…	$\overline{Y}_{\cdot b\cdot}$	$\overline{Y}_{\cdots}$

dove con notazione ovvia ed analoga al caso di 1 fattore i termini *Y* con una barra e un punto · rappresentano le medie sull'indice contraddistinto dal punto · (in questo caso le osservazioni della cella), vale a dire le medie su ciascun trattamento di livello *i* del primo fattore e *j* del secondo fattore; i termini *Y* con una barra e due punti · rappresentano le medie sui due indici contraddistinti dagli stessi simboli ·, vale a dire la media riferita a ciascun indice non rappresentato dal punto calcolata rispetto a tutti i campioni di tutti gli altri livelli dell'altro fattore; infine i termini *Y* con tre punti · rappresentano la media calcolata su tutti i campioni di tutti i livelli dei due fattori.

Osserviamo subito le principali caratteristiche che differenziano il modello a due fattori da quello a 1 fattore. Innanzitutto vi sono due tipologie di effetti principali anziché una, legati ai livelli del fattore *A* e a quelli del fattore *B*. Inoltre esiste un effetto di interazione tra i fattori che esprime l'eventualità che il responso *Y* dipenda oltre oppure invece che dai singoli livelli dei fattori in maniera indipendente, anche congiuntamente dalla coppie di livelli formate da un livello del primo fattore e da un livello del secondo; ovviamente tali termini di interazione corrispondono al prodotto incrociato tra gli indici dei livelli dei due fattori. In pratica questo nuovo tipo di variabile (ovviamente non presente nell'analisi a 1 fattore) tiene conto di tutte quelle situazioni in cui l'effetto combinato di una coppia di livelli corrispondenti ai due fattori ha un ruolo significativo nella spiegazione della variabile responso, come ad esempio il caso di due fattori che potrebbero non avere influenza sulla variabile responso per nessun livello preso singolarmente, ma potrebbero invece spiegare la varianza del responso in maniera significativa mediante una particolare combinazione di livelli dei due fattori. Un esempio tipico viene dalla farmacologia, dove due medicine prese singolarmente potrebbero non produrre effetti terapeutici, mentre la loro combinazione risulterebbe invece efficace.

L'analisi della varianza a due fattore consente di verificare tre tipi di ipotesi nulle:

1) Le medie dei livelli del fattore *A* sono uguali:
$$\mu_{1\cdot\cdot} = \mu_{2\cdot\cdot} = \cdots \mu_{a\cdot\cdot}$$
2) Le medie dei livelli del fattore *B* sono uguali:
$$\mu_{\cdot 1\cdot} = \mu_{\cdot 2\cdot} = \cdots \mu_{\cdot b\cdot}$$
3) Gli effetti dei livelli di un fattore sono consistenti con quelli dei livelli dell'altro fattore; in pratica i fattori *A* e *B* sono indipendenti tra loro.

Consideriamo ora la somma totale dei quadrati (SSTO) già analizzata nel caso di modello a 1 fattore. Supponiamo in questa fase per semplicità che il numero di osservazioni per ogni cella sia costante e indichiamo tale costante con *n*. Risulta

$$\text{SSTO} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk} - \overline{Y}_{\cdots}\right)^2 .$$

Possiamo considerare la deviazione del singolo dato Y_{ijk} dalla media generale $\overline{\overline{Y}}_{\cdots}$ come somma di quattro contributi indipendenti:

a) la deviazione della media di una riga (livello *i* del fattore *A*) dalla media generale che è significativa quando la prima ipotesi nulla è falsa:

$$\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}$$

b) la deviazione della media di una colonna (livello *j* del fattore *B*) dalla media generale che è significativa quando la seconda ipotesi nulla è falsa:

$$Y_{\cdot j\cdot} - \overline{Y}_{\cdots}$$

c) la deviazione della singola osservazione dalla media della cella cui appartiene; essa non dipende dalla riga o dalla colonna (cioè né dal livello, né dal fattore):

$$Y_{ijk} - \overline{Y}_{ij\cdot}$$

Sottraendo le tre componenti dalla deviazione della singola osservazione dalla media generale $Y_{ijk} - \overline{\overline{\overline{Y}}}_{\cdots}$ otteniamo

$$Y_{ijk} - \overline{Y}_{\cdots} - \left[\left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}\right) + \left(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}\right) + \left(Y_{ijk} - \overline{Y}_{ij\cdot}\right)\right] = \left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}\right).$$

Il termine al lato destro identifica l'effetto dell'interazione per la cella; esso à basato sulla deviazione della media della cella dalle corrispondenti medie di riga e colonna (rispettivamente fattori A e B) con inclusa una correzione basata sulla media generale. Il termine riflette quegli effetti di cella che si verificano quando i corrispondenti fattori vengono combinati nei livelli riferiti alla cella. Il termine risulta significativo solo quando la terza ipotesi nulla è falsa.

In base all'equazione appena ottenuta possiamo scrivere la somma SSTO come

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk} - \overline{Y}_{\cdots}\right)^2}_{\text{SSTO}} = \underbrace{bn\sum_{i=1}^{a}\left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}\right)^2}_{\text{SSA}} + \underbrace{an\sum_{i=1}^{a}\left(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}\right)^2}_{\text{SSB}} + \underbrace{n\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}\right)^2}_{\text{SSAB}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk} - \overline{Y}_{ij\cdot}\right)^2}_{\text{SSE}}$$

Ciascun termine della somma ha associati dei gradi di libertà che consentono di stimare le corrispondenti medie quadratiche per effetti di riga (fattore A), colonna (fattore B), interazione secondo la seguente tabella:

Termine	SS	Gradi di libertà	Medie quadratiche

SSA	$bn\sum_{i=1}^{a}\left(\overline{Y}_{i\cdot\cdot}-\overline{Y}_{\cdots}\right)^2$	a-1	SSA/(a-1)
SSB	$an\sum_{i=1}^{a}\left(\overline{Y}_{\cdot j\cdot}-\overline{Y}_{\cdots}\right)^2$	b-1	SSB/(b-1)
SSAB	$n\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{Y}_{ij\cdot}-\overline{Y}_{i\cdot\cdot}-\overline{Y}_{\cdot j\cdot}+\overline{Y}_{\cdots}\right)^2$	(a-1)(b-1)	SSAB/((a-1)(b-1))
SSE	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk}-\overline{Y}_{ij\cdot}\right)^2$	ab(n-1)	SSE/(ab(n-1))
SSTO	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk}-\overline{Y}_{\cdots}\right)^2$	abn-1	SSTO/(abn-1)

La verifica pratica delle tre ipotesi nulle viene effettuata in maniera analoga a quanto visto per il caso di 1 fattore in base al test F e alla tabella sopra:

Termine dell'ipotesi nulla	SS	Gradi di libertà	Media quadratica	Statistic F-test	Distribuzione	Media quadratica attesa
Fattore A	SSA	a-1	$\frac{\text{SSA}}{a\text{-}1}$	$\frac{\text{MSA}}{\text{MSAB}}$	F(a-1; ab(n-1))	$\sigma^2+bn\sigma^2_A$
Fattore B	SSB	b-1	$\frac{\text{SSB}}{b\text{-}1}$	$\frac{\text{MSB}}{\text{MSAB}}$	F(b-1; ab(n-1))	$\sigma^2+bn\sigma^2_B$
Interazione AB	SSAB	(a-1)(b-1)	$\frac{\text{SSAB}}{(a\text{-}1)(b\text{-}1)}$	$\frac{\text{MSAB}}{\text{MSE}}$	F((a-1)(b-1); ab(n-1))	$\sigma^2+n\sigma^2_{AB}$
	SSE	ab(n-1)	$\frac{\text{SSE}}{ab(n\text{-}1)}$			σ^2
Totale	SSTO	abn-1	$\frac{\text{SSTO}}{abn\text{-}1}$			

Consideriamo ora nuovamente il modello generale con numero di osservazioni per ogni cella non necessariamente uguale. In particolare considereremo il modello di regressione del tipo a effetti perché nella valutazione degli studenti del Sistema Scolastico le osservazioni di ciascuna cella sono generalmente non coincidenti tra le celle; in questo caso il modello di regressione unito al metodo dei minimi quadrati consente di trovare la soluzione del problema ANOVA.

Nella formulazione a effetti il modello ANOVA per 2 fattori si scrive come

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},\ i = 1,\ldots,a;\ j = 1,\ldots,b$$

dove rispetto al modello a 1 fattore troviamo due tipologie di effetti principali anziché una, legati ai livelli del fattore *A* (α_i) e a quelli del fattore *B* (β_i). Inoltre i termini di interazione tra i fattori sono indicati con $(\alpha\beta)_{ij}$, che esprimono l'eventualità che il responso *Y* dipenda oltre oppure invece che dai singoli livelli dei fattori in maniera indipendente, anche congiuntamente dalla coppie di livelli formate da un livello del primo fattore e da un livello del secondo; ovviamente tali termini di interazione corrispondono al prodotto incrociato tra gli indici dei livelli dei due fattori.

Nel modello a effetti vanno rispettati i seguenti vincoli in analogia con il modello a 1 fattore:

$$\begin{cases} \sum_{i=1}^{a} \alpha_i = 0 \\ \sum_{j=1}^{b} \beta_j = 0 \\ \sum_{i=1}^{a} (\alpha\beta)_{ij} = 0, \quad j = 1,\ldots,b \\ \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0, \quad i = 1,\ldots,a \end{cases}$$

Le proprietà delle osservazioni Y_{ijk} sono identiche a quelle del modello a 1 fattore:

$$\begin{cases} E[Y_{ijk}] = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} = \mu_{ij} \\ \sigma^2(Y_{ijk}) = \sigma^2 \end{cases}$$

Nella formulazione a effetti del modello ANOVA a due fattori le ipotesi da verificare si scrivono in maniera equivalente:

per il fattore A

$$\begin{cases} H_0 : \alpha_1 = \cdots = \alpha_a = 0 \\ H_a : \text{non tutti gli } \alpha_i \text{ sono uguali a } 0 \end{cases};$$

per il fattore B

$$\begin{cases} H_0 : \beta_1 = \cdots = \beta_b = 0 \\ H_b : \text{non tutti i } \beta_j \text{ sono uguali a } 0 \end{cases};$$

per le interazioni

$$\begin{cases} H_0 : \text{tutti gli } (\alpha\beta)_{ij} \text{ sono uguali a } 0 \\ H_a : \text{non tutti gli } (\alpha\beta)_{ij} \text{ sono uguali a } 0 \end{cases};$$

In analogia con il caso a 1 fattore, il modello di regressione a effetti per 2 fattori si scrive come

$$Y_{ijk} = \mu_{..} + \underbrace{\sum_{\ell=1}^{a-1} \alpha_\ell X^A_{ijk\ell}}_{\text{Effetto A}} + \underbrace{\sum_{m=1}^{b-1} \beta_m X^B_{ijkm}}_{\text{Effetto B}} + \underbrace{\sum_{\ell=1}^{a-1}\sum_{m=1}^{b-1} (\alpha\beta)_{\ell m} X^A_{ijk\ell} X^B_{ijkm}}_{\text{Effetto delle interazioni AB}} + \varepsilon_{ijk},\ i = 1,\ldots,a;\ j = 1,\ldots,b,$$

dove in analogia con il caso ad 1 fattore

$$X^A_{ijk\ell} = \begin{cases} 1 \text{ se } i = \ell \\ -1 \text{ se } i = a \\ 0 \text{ altrimenti} \end{cases} \text{ e } X^B_{ijkm} = \begin{cases} 1 \text{ se } i = m \\ -1 \text{ se } j = b \\ 0 \text{ altrimenti} \end{cases}.$$

Si osservi che per ciascun fattore la somma degli effetti continua ad esser nulla:

$$\sum_{\ell=1}^{a} \alpha_\ell = 0, \quad \sum_{m=1}^{b} \beta_m = 0;$$

pertanto assumiamo per convenzione che α_a e β_b siano determinati dai restanti α_ℓ e β_m come

$$\alpha_a = \sum_{\ell=1}^{a-1} \alpha_\ell \quad \text{e} \quad \beta_b = \sum_{m=1}^{b-1} \beta_m$$

e quindi non rientrano nel modello di regressione come variabili incognite.

Inoltre simili relazioni valgono anche per i termini degli effetti di interazione di ogni livello di un fattore con i livelli degli altri fattori:

$$\sum_{\ell=1}^{a} (\alpha\beta)_{k\ell} = 0, \quad k = 1,\ldots,a; \quad \sum_{k=1}^{a} (\alpha\beta)_{k\ell} = 0, \quad \ell = 1,\ldots,b$$

La corrispondenza tra il modello a effetti e quello a celle è data dalle seguenti relazioni:

$$\begin{cases} \mu_{\cdot\cdot} \\ \alpha_\ell = \mu_{\ell\cdot} - \mu_{\cdot\cdot} \\ \beta_m = \mu_{\cdot m} - \mu_{\cdot\cdot} \\ (\alpha\beta)_{\ell m} = \mu_{\ell m} - \mu_{\ell\cdot} - \mu_{\cdot m} + \mu_{\cdot\cdot} \end{cases}$$

In definitiva la soluzione del modello ANOVA a effetti viene ottenuta risolvendo il seguente problema ai minimi quadrati

$$\min \quad (Y_{ijk} - \mu_{\cdot\cdot} - \sum_{\ell=1}^{a-1} \alpha_\ell X^A_{ijk\ell} - \sum_{m=1}^{b-1} \beta_m X^B_{ijkm} - \sum_{\ell=1}^{a-1}\sum_{m=1}^{b-1} (\alpha\beta)_{\ell m} X^A_{ijk\ell} X^B_{ijkm})^2$$

rispetto alle variabili $\mu_{\cdot\cdot}; \alpha_\ell; \beta_m; (\alpha\beta)_{\ell m}$, $\ell = 1,\ldots,a-1; m = 1,\ldots,b-1$.

2.3 Modello ANOVA a più fattori

Il modello ANOVA a due fattori si presta bene ad essere generalizzato a più fattori. Innanzitutto la nomenclatura delle variabili si estende in questo modo: i fattori vengono indicati con le lettere maiuscole *A*, *B*, *C*, *D*, …; il numero di livelli per ogni fattore è contraddistinto dalle rispettive lettere minuscole *a*, *b*, *c*, *d*, …; le osservazioni avranno ora un numero di indici pari al numero di fattori +1: i primi indici saranno contraddistinti dalle lettere *i*, *j*, *k*, *l*, … e si riferiscono nell'ordine ai fattori corrispondenti (*i*=1,…,*a* per il fattore *A*; *j*=1,…,*b* per il fattore *B*; *k*=1,…,*c* per il fattore *C*; ecc.); l'ultimo indice (per il quale utilizzeremo la lettera *z*) indicherà le osservazioni del campione presente nella cella *ijkl*…, il cui numero sarà denotato con $n_{ijkl\ldots}$.

Il modello di regressione a effetti per più fattori si scrive come

$$
\begin{aligned}
Y_{ijk\ldots,z} &= \mu_{..} \\
&+ \underbrace{\sum_{\ell_A=1}^{a-1} \alpha_{\ell_A} X^{A}_{ijk\ldots z,\ell_A}}_{\text{Effetto A}} + \underbrace{\sum_{\ell_B=1}^{b-1} \beta_{\ell_B} X^{B}_{ijk\ldots z,\ell_B}}_{\text{Effetto B}} + \underbrace{\cdots\cdots}_{\substack{\text{Effetto altre}\\ \text{interazioni}\\ \text{a 1 fattore}}} \\
&+ \underbrace{\sum_{\ell_A=1}^{a-1}\sum_{\ell_B m=1}^{b-1} (\alpha\beta)_{\ell_A\ell_B} X^{A}_{ijk\ldots z,\ell_A} X^{B}_{ijk\ldots z,\ell_B}}_{\text{Effetto delle interazioni AB}} + \underbrace{\sum_{\ell_A=1}^{a-1}\sum_{\ell_C m=1}^{c-1} (\alpha\gamma)_{\ell_A\ell_C} X^{A}_{ijk\ldots z,\ell_A} X^{C}_{ijk\ldots z,\ell_C}}_{\text{Effetto delle interazioni AC}} + \underbrace{\cdots\cdots}_{\substack{\text{Effetto altre}\\ \text{interazioni}\\ \text{a 2 fattori}}} \\
&+ \underbrace{\sum_{\ell_A=1}^{a-1}\sum_{\ell_B m=1}^{b-1}\sum_{\ell_C=1}^{c-1} (\alpha\beta\gamma)_{\ell_A\ell_B\ell_C} X^{A}_{ijk\ldots z,\ell_A} X^{B}_{ijk\ldots z,\ell_B} X^{C}_{ijk\ldots z,\ell_C}}_{\text{Effetto delle interazioni ABC}} + \underbrace{\sum_{\ell_A=1}^{a-1}\sum_{\ell_B m=1}^{b-1}\sum_{\ell_D=1}^{d-1} (\alpha\beta\delta)_{\ell_A\ell_B\ell_D} X^{A}_{ijk\ldots z,\ell_A} X^{B}_{ijk\ldots z,\ell_B} X^{D}_{ijk\ldots z,\ell_D}}_{\text{Effetto delle interazioni ABD}} \\
&+ \underbrace{\cdots\cdots\cdots}_{\substack{\text{Effetto altre}\\ \text{interazioni}\\ \text{a 3 fattori}}} + \underbrace{\cdots\cdots\cdots\cdots}_{\substack{\text{Effetto tutte}\\ \text{interazioni}\\ \text{a 4 fattori}}} \\
&+ \varepsilon_{ijk}, \qquad i = 1,\ldots,a;\ j = 1,\ldots,b;\ k = 1,\ldots,c;\ l = 1,\ldots,d;\ \ldots
\end{aligned}
$$

dove la prima riga al lato destro dell'equazione contiene la media generale dell'intera popolazione; la seconda riga contiene i termini degli effetti di ciascun fattore sulla variabile responso; la terza riga contiene gli effetti dell'interazione tra livelli appartenenti a coppie di fattori; la riga successiva contiene gli effetti dell'interazione tra livelli di triplette di fattori, e così via. In pratica il modello a più fattori è un'estensione immediata del modello a due fattori, purché si tenga conto delle interazioni aggiuntive tra fattori di ordine superiore (a triplette di fattori, quadruplette, e così via).

Ovviamente il modello ANOVA a più fattori si complica notevolmente all'aumentare del numero di fattori e non risulta agevole ottenere formule analitiche compatte per la risoluzione del modello (come quelle descritte nel presente report nel caso di 1 fattore) già da 4 fattori in poi, anche nel caso di numero di osservazioni costante per ogni cella.

Va comunque osservato che nel caso dei questionari INVALSI due motivazioni rendono la formulazione del modello ANOVA mediante un modello di regressione risolvibile con il metodo dei minimi quadrati l'unica strada percorribile:

- la presenza di un numero di osservazioni fortemente disomogeneo sulle celle;
- l'interesse a valutare l'importanza di numerosi fattori; questo oltre a rendere impossibile il calcolo analitico della soluzione del modello ANOVA, richiede anche capacità di flessibilità (per esempio nello scegliere in maniera selettiva solo

particolari interazioni tra fattori) che sono possibili solo con il modello di regressione.

3. Analisi ANOVA sui questionari di valutazione INVALSI

Vengono riportate di seguito le analisi effettuate mediante il modello ANOVA dei dati relativi ai questionari di valutazione per le scuole elementari e medie inferiori. Sono stati considerati quattro fattori fissi: Tipo (con due livelli, Scuola Pubblica e Privata); Regione (con 20 livelli corrispondenti alle regioni); Ordine (con due livelli: Scuole elementari e medie inferiori); Sesso (due livelli: Maschi e Femmine). Le analisi preliminari hanno mostrato che tutti i fattori considerati sono significativi; pertanto non vengono mostrati i risultati dell'analisi di varianza a più fattori che risulterebbero poco significativi.

Le analisi sono state effettuate considerando la valutazione degli Istituti sia nel loro complesso (mettendo insieme le domande di tutte le tre discipline, Italiano, Matematica e Scienze), sia disciplina per disciplina. Inoltre è stata considerata come variabile responso sia la percentuale di risposte esatte fornite dagli studenti, sia l'abilità stimata con il modello Item Response Theory (IRT) a 3 parametri.

I risultati sono forniti mediante una tabella che riassume gli elementi presenti nella tabella completa discussa nella Sezione 2.1.2, in particolare la Grandmean (stima della media generale della popolazione $\mu_{..}$ mediante il metodo dei minimi quadrati a partire dal modello di regressione); il valore dello statistic F ed il corrispondente livello di significatività calcolato in base ai gradi di libertà del modello ANOVA. Inoltre vengono riportati i valori delle medie stimate per ciascun livello del fattore, insieme con la stima della deviazione standard. Queste ultime informazioni sono importanti per stabilire in primo luogo le motivazioni della significatività del fattore (vale a dire quali livelli presentano una media differente dagli altri) e in secondo luogo per raggruppare i livelli in gruppi omogenei (cioè aventi lo stesso valore medio). A tale proposito vengono forniti anche due tipi di rappresentazioni grafiche che perseguono lo scopo in maniera più evidente. Nel primo le medie di ciascun livello vengono riportate in un grafico a barre insieme con le corrispondenti deviazioni standard: approssimativamente i gruppi di livelli omogenei sono quelli per cui tutte le barre si intersecano tra loro. Per alcuni fattori viene fornita anche una visualizzazione matriciale in cui le differenze incrociate tra tutti i livelli

di un fattore sono mostrate codificate con un'opportuna mappa di colori: i gruppi di livelli omogenei possono essere individuati da zone della matrice aventi colorazione più o meno uniforme.

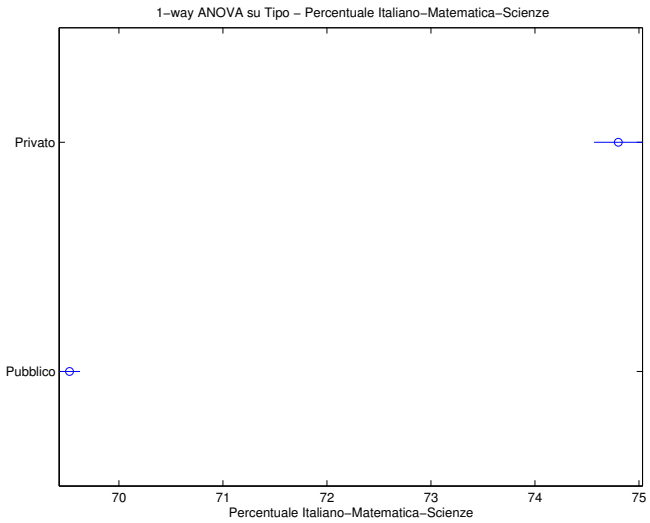
3.1 Percentuale di risposte esatte: Italiano, Matematica e Scienze

3.1.1 Fattore tipo

Grandmean: 72.16

F-test: 4.330542e+002 - Liv. signif.: 0

Tipo 1: Pubblico - Estimated mean: 69.53 +/- 0.10
Tipo 2: Privato - Estimated mean: 74.80 +/- 0.23



Il dato indica che le Scuole Private hanno una valutazione significativamente migliore di quella delle Scuole pubbliche.

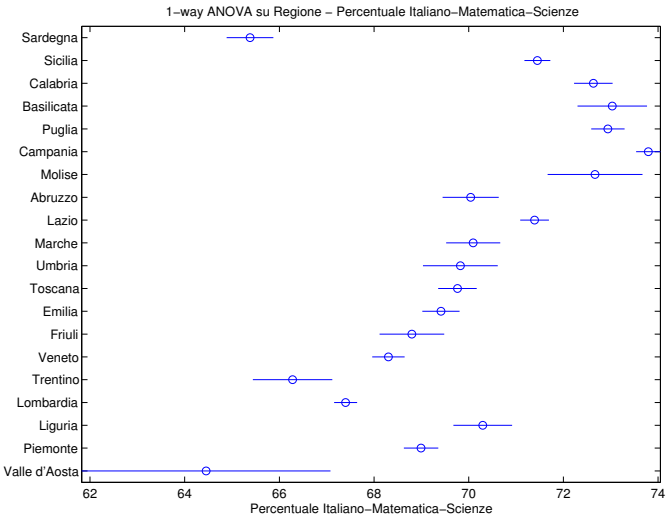
3.1.2 Fattore regione

Grandmean: 69.85

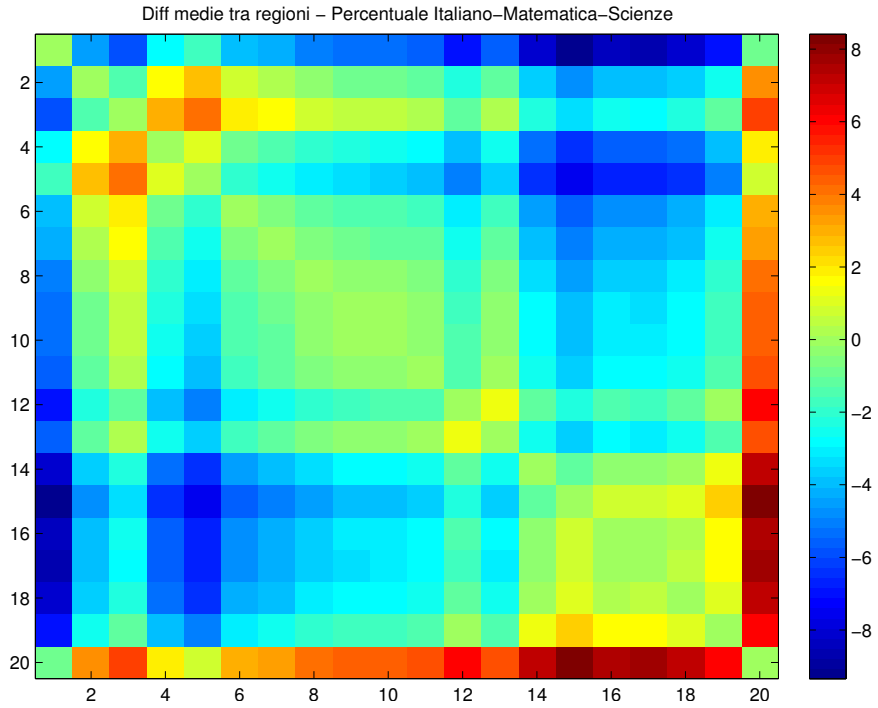
F-test: 3.443813e+001 - Liv. signif.: 0

Regione 1: Valle d'Aosta - Estimated mean: 64.45 +/- 2.63
Regione 2: Piemonte - Estimated mean: 68.99 +/- 0.36
Regione 3: Liguria - Estimated mean: 70.30 +/- 0.62
Regione 4: Lombardia - Estimated mean: 67.40 +/- 0.24

Regione 5: Trentino - Estimated mean: 66.28 +/- 0.84
Regione 6: Veneto - Estimated mean: 68.30 +/- 0.34
Regione 7: Friuli - Estimated mean: 68.80 +/- 0.68
Regione 8: Emilia - Estimated mean: 69.41 +/- 0.39
Regione 9: Toscana - Estimated mean: 69.76 +/- 0.41
Regione 10: Umbria - Estimated mean: 69.82 +/- 0.79
Regione 11: Marche - Estimated mean: 70.09 +/- 0.57
Regione 12: Lazio - Estimated mean: 71.39 +/- 0.30
Regione 13: Abruzzo - Estimated mean: 70.04 +/- 0.59
Regione 14: Molise - Estimated mean: 72.67 +/- 1.00
Regione 15: Campania - Estimated mean: 73.79 +/- 0.26
Regione 16: Puglia - Estimated mean: 72.94 +/- 0.35
Regione 17: Basilicata - Estimated mean: 73.03 +/- 0.73
Regione 18: Calabria - Estimated mean: 72.63 +/- 0.41
Regione 19: Sicilia - Estimated mean: 71.45 +/- 0.27
Regione 20: Sardegna - Estimated mean: 65.38 +/- 0.49



Anche in questo caso il fattore regione risulta significativo. Tuttavia il trend geografico risulta estremamente interessante, in quanto è evidente un miglioramento della valutazione quando ci si sposta dal Nord al Centro al Sud. Per valutare meglio tale fenomeno, consideriamo anche il grafico seguente, dove è riportata la differenza tra le percentuali di risposte esatte ottenute per le regioni sulla righe meno quelle ottenute per le regioni sulle colonne. È evidente la presenza di due aree, corrispondenti grosso modo alle regioni del centro e a quelle del Sud (incluso la Sicilia) a colorazione simile all'interno e pertanto omogenee per quanto riguarda il valore medio.



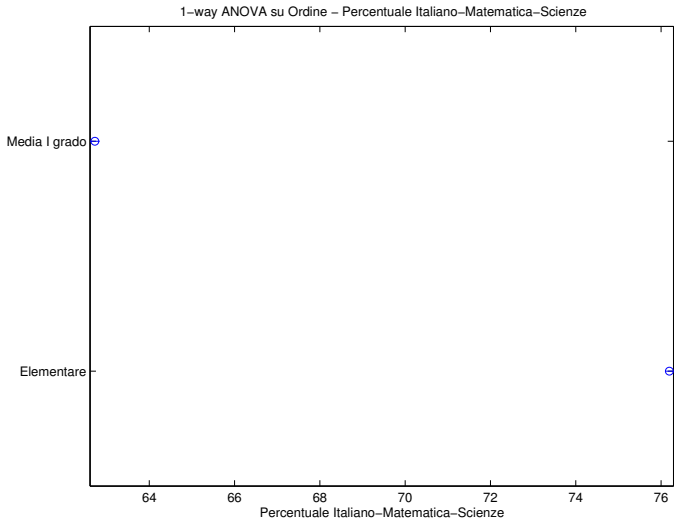
3.1.3 Fattore ordine

Grandmean: 69.46

F-test: 8.299524e+003 - Liv. signif.: 0

Ordine 1: Elementare - Estimated mean: 76.20 +/- 0.10
Ordine 2: Media I grado - Estimated mean: 62.73 +/- 0.11

La tabella ed il grafico seguente indicano una forte significatività del fattore ordine di scuola, con le scuole elementari aventi una percentuale di risposte esatte nettamente maggiore rispetto alle scuole medie.

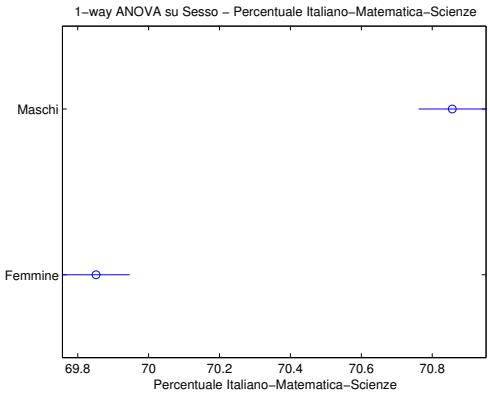


3.1.4 Fattore sesso

Grandmean: 70.35

F-test: 5.588713e+001 - Liv. signif.: 7.915890e-014

Sesso 1: Femmine - Estimated mean: 69.85 +/- 0.09
Sesso 2: Maschi - Estimated mean: 70.86 +/- 0.10



Anche il fattore sesso è significativo, con una percentuale di rispote esatte fornite dai maschi leggermente maggiore rispetto ai maschi (circa 1 punto)

3.2 Abilità: Italiano, Matematica e Scienze

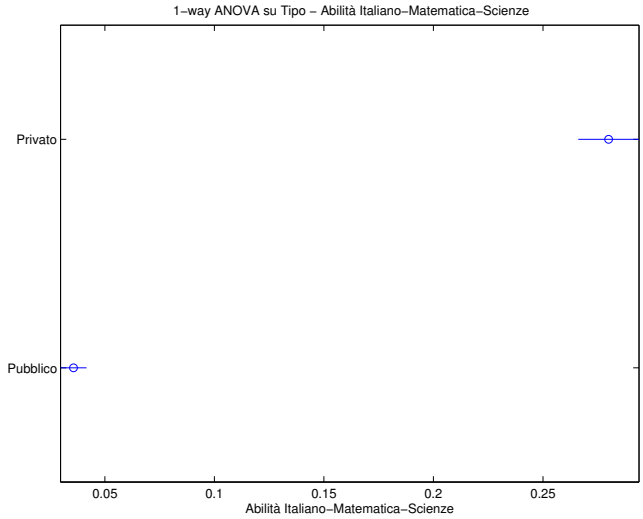
Viene ora proposta la stessa analisi effettuata analizzando le abilità stimate con la IRT come variabile responso anziché la percentuale di risposte esatte. Le indicazioni che si ottengono sono molto simili tra i due casi e non vengono discusse in dettaglio. Pertanto nelle successive analisi dettagliate per singola materia i risultati ottenuti utilizzando le abilità calcolate con IRT non verranno mostrati.

3.2.1 Fattore tipo

Grandmean: 0.16

F-test: 2.627200e+002 - Liv. signif.: 0

Tipo 1: Pubblico - Estimated mean: 0.04 +/- 0.01
Tipo 2: Privato - Estimated mean: 0.28 +/- 0.01



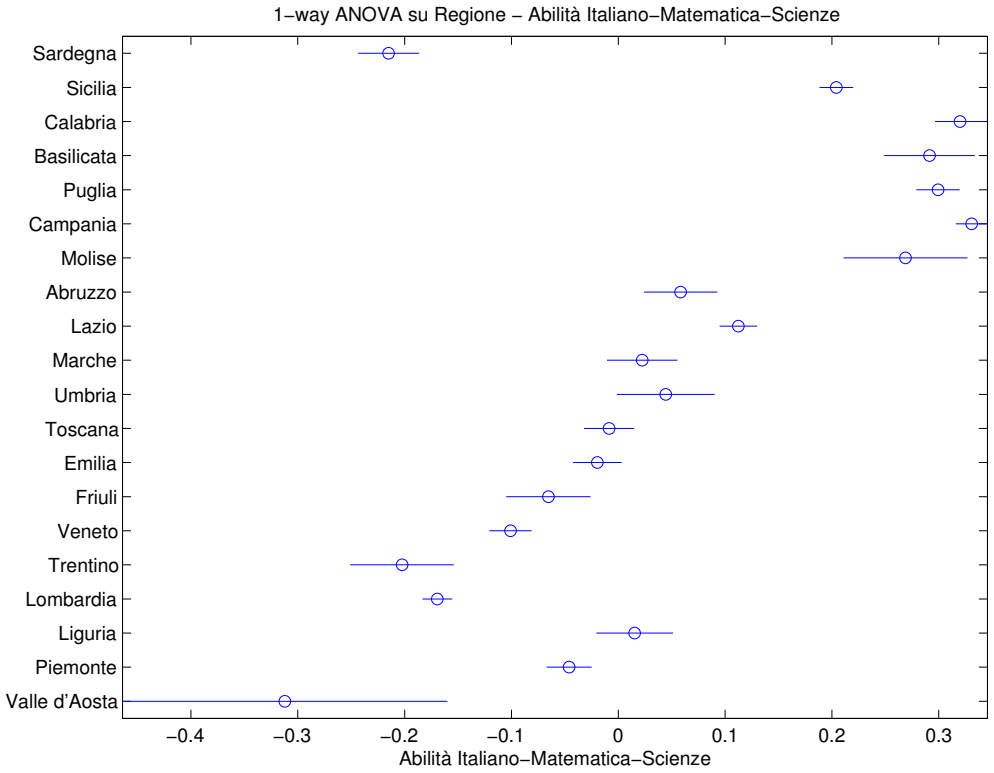
3.2.2 Fattore regione

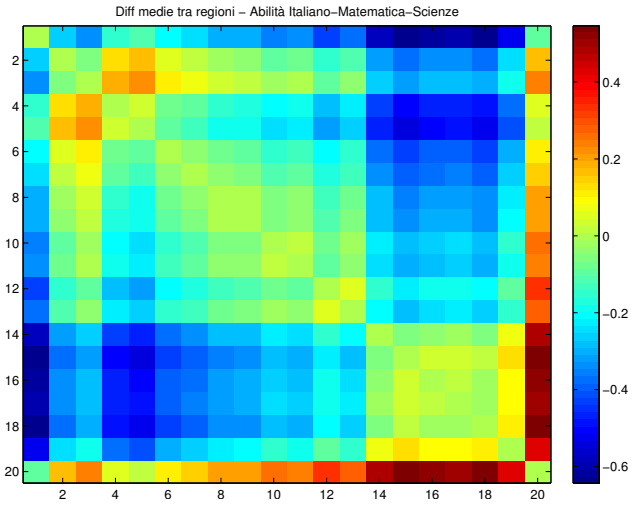
Grandmean: 0.04

F-test: 6.533032e+001 - Liv. signif.: 0

Regione 1: Valle d'Aosta - Estimated mean: -0.31 +/- 0.15
Regione 2: Piemonte - Estimated mean: -0.05 +/- 0.02
Regione 3: Liguria - Estimated mean: 0.02 +/- 0.04
Regione 4: Lombardia - Estimated mean: -0.17 +/- 0.01

Regione 5: Trentino - Estimated mean: -0.20 +/- 0.05
Regione 6: Veneto - Estimated mean: -0.10 +/- 0.02
Regione 7: Friuli - Estimated mean: -0.07 +/- 0.04
Regione 8: Emilia - Estimated mean: -0.02 +/- 0.02
Regione 9: Toscana - Estimated mean: -0.01 +/- 0.02
Regione 10: Umbria - Estimated mean: 0.04 +/- 0.05
Regione 11: Marche - Estimated mean: 0.02 +/- 0.03
Regione 12: Lazio - Estimated mean: 0.11 +/- 0.02
Regione 13: Abruzzo - Estimated mean: 0.06 +/- 0.03
Regione 14: Molise - Estimated mean: 0.27 +/- 0.06
Regione 15: Campania - Estimated mean: 0.33 +/- 0.01
Regione 16: Puglia - Estimated mean: 0.30 +/- 0.02
Regione 17: Basilicata - Estimated mean: 0.29 +/- 0.04
Regione 18: Calabria - Estimated mean: 0.32 +/- 0.02
Regione 19: Sicilia - Estimated mean: 0.20 +/- 0.02
Regione 20: Sardegna - Estimated mean: -0.22 +/- 0.03



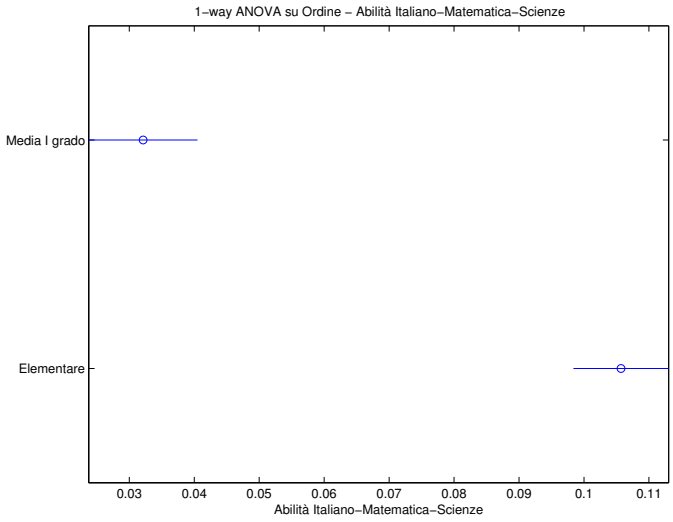


3.2.3 Fattore ordine

Grandmean: 0.07

F-test: 4.376156e+001 - Liv. signif.: 3.851630e-011

Ordine 1: Elementare - Estimated mean: 0.11 +/- 0.01
Ordine 2: Media I grado - Estimated mean: 0.03 +/- 0.01

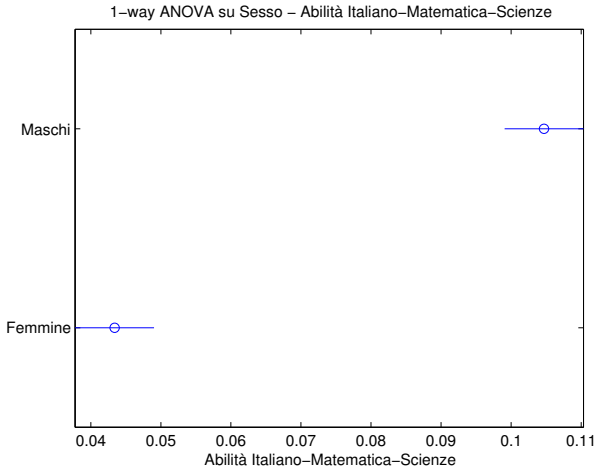


3.2.4 Fattore sesso

Grandmean: 0.07

F-test: 5.936094e+001 - Liv. signif.: 1.354472e-014

Sesso 1: Femmine - Estimated mean: 0.04 +/- 0.01
Sesso 2: Maschi - Estimated mean: 0.10 +/- 0.01



3.3 Percentuale di risposte esatte: Italiano

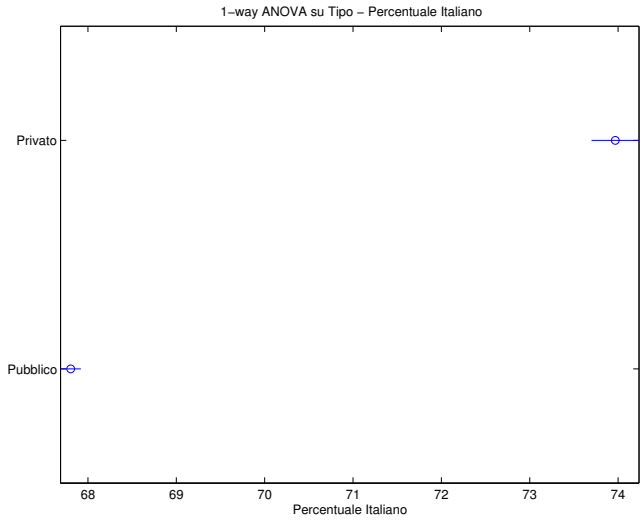
Le percentuali di successo per l'Italiano sono leggermente inferiori (circa 1 punto e mezzo percentuale) rispetto al caso generale.

3.3.1 Fattore tipo

Grandmean: 70.89

F-test: 4.457924e+002 - Liv. signif.: 0

Tipo 1: Pubblico - Estimated mean: 67.80 +/- 0.12
Tipo 2: Privato - Estimated mean: 73.97 +/- 0.27



Il divario tra scuole pubbliche e private risulta leggermente superiore al caso generale (circa 6%).

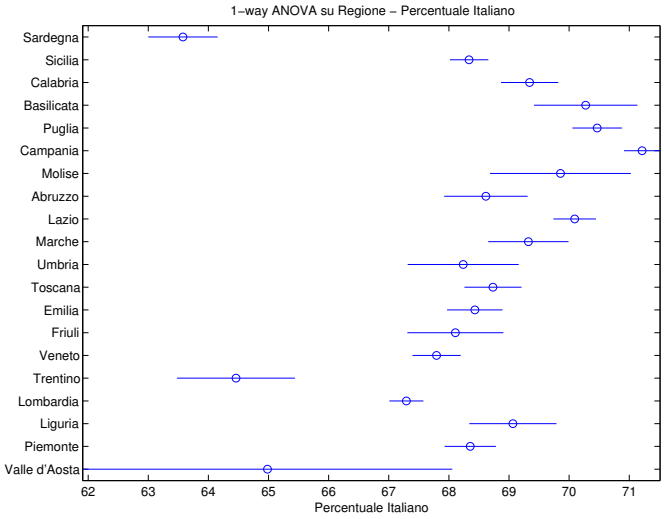
3.3.2 Fattore regione

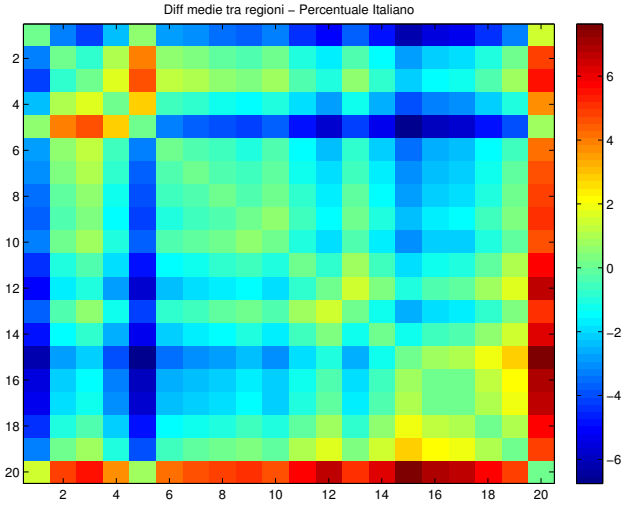
Grandmean: 68.33

F-test: 1.277675e+001 - Liv. signif.: 0

Regione 1: Valle d'Aosta - Estimated mean: 64.98 +/- 3.07
Regione 2: Piemonte - Estimated mean: 68.36 +/- 0.43
Regione 3: Liguria - Estimated mean: 69.06 +/- 0.73
Regione 4: Lombardia - Estimated mean: 67.29 +/- 0.28
Regione 5: Trentino - Estimated mean: 64.46 +/- 0.98
Regione 6: Veneto - Estimated mean: 67.80 +/- 0.40
Regione 7: Friuli - Estimated mean: 68.11 +/- 0.80
Regione 8: Emilia - Estimated mean: 68.43 +/- 0.46
Regione 9: Toscana - Estimated mean: 68.73 +/- 0.48
Regione 10: Umbria - Estimated mean: 68.24 +/- 0.92
Regione 11: Marche - Estimated mean: 69.32 +/- 0.67
Regione 12: Lazio - Estimated mean: 70.09 +/- 0.35
Regione 13: Abruzzo - Estimated mean: 68.62 +/- 0.69
Regione 14: Molise - Estimated mean: 69.86 +/- 1.17
Regione 15: Campania - Estimated mean: 71.21 +/- 0.30
Regione 16: Puglia - Estimated mean: 70.47 +/- 0.41
Regione 17: Basilicata - Estimated mean: 70.28 +/- 0.86
Regione 18: Calabria - Estimated mean: 69.34 +/- 0.48
Regione 19: Sicilia - Estimated mean: 68.34 +/- 0.32
Regione 20: Sardegna - Estimated mean: 63.58 +/- 0.58

Si conferma l'andamento generale già visto per tutte le materie, con diminuzione generalizzata delle percentuali di successo di circa 1-2 punti per le regioni.



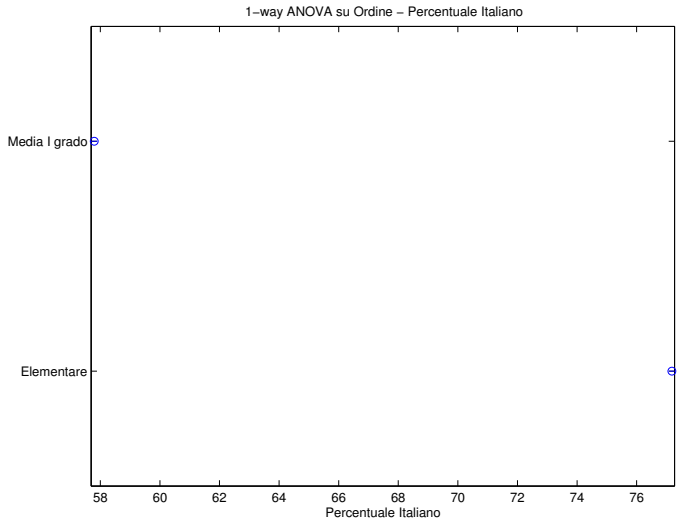


3.3.3 Fattore ordine

Grandmean: 67.49

F-test: 1.997236e+004 - Liv. signif.: 0

Ordine 1: Elementare - Estimated mean: 77.18 +/- 0.09
Ordine 2: Media I grado - Estimated mean: 57.79 +/- 0.10



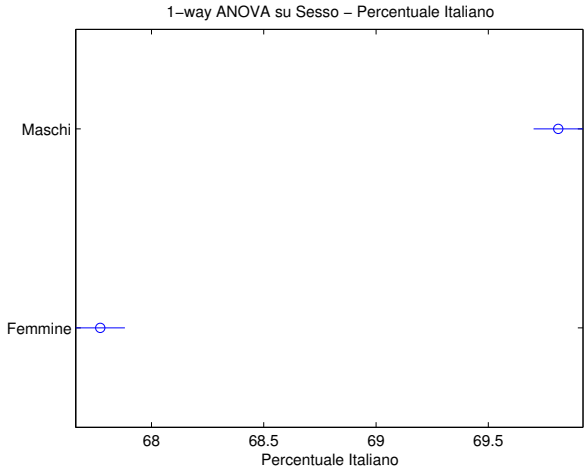
Cresce sensibilmente il divario tra Scuole elementari e medie, che si attesta a quasi il 20% a favore delle prime.

3.3.4 Fattore sesso

Grandmean: 68.79

F-test: 1.726323e+002 - Liv. signif.: 0

Sesso 1: Femmine - Estimated mean: 67.77 +/- 0.11
Sesso 2: Maschi - Estimated mean: 69.81 +/- 0.11



Si conferma anche per l'Italiano la significatività del fattore Sesso, con una margine tra i due livelli (Maschio e Femmina) che aumenta a circa due punti percentuali sempre a favore dei Maschi.

3.4 Percentuale di risposte esatte: Matematica

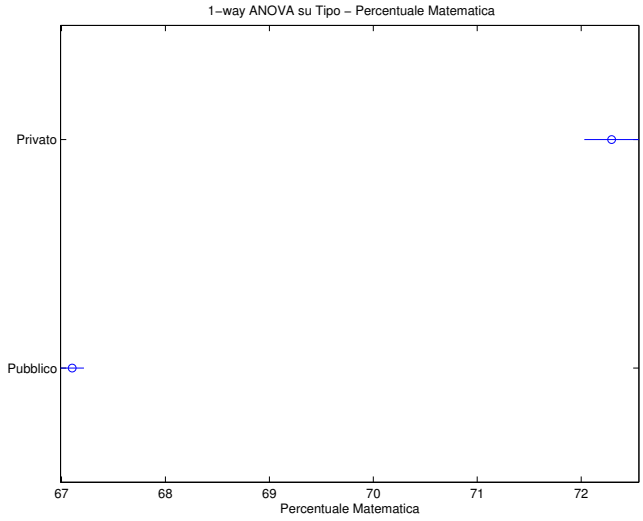
Le percentuali di risposte esatte per le domande di matematica sono le più basse tra tutte le materie (67.9% in media contro il 70.3% generalmente su tutte le materie).

3.4.1 Fattore tipo

Grandmean: 69.70

F-test: 3.291022e+002 - Liv. signif.: 0

Tipo 1: Pubblico - Estimated mean: 67.10 +/- 0.11
Tipo 2: Privato - Estimated mean: 72.29 +/- 0.26



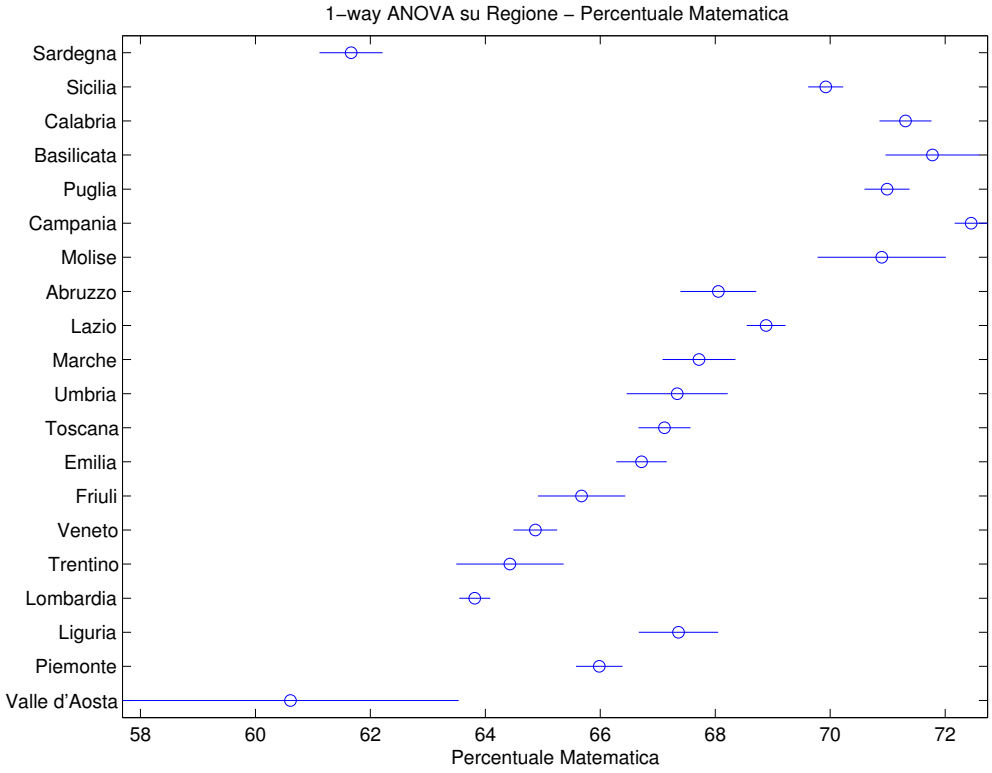
Il fattore tipo continua ad essere significativo, con una margine che si mantiene attorno al 5% sempre a favore delle Scuole private

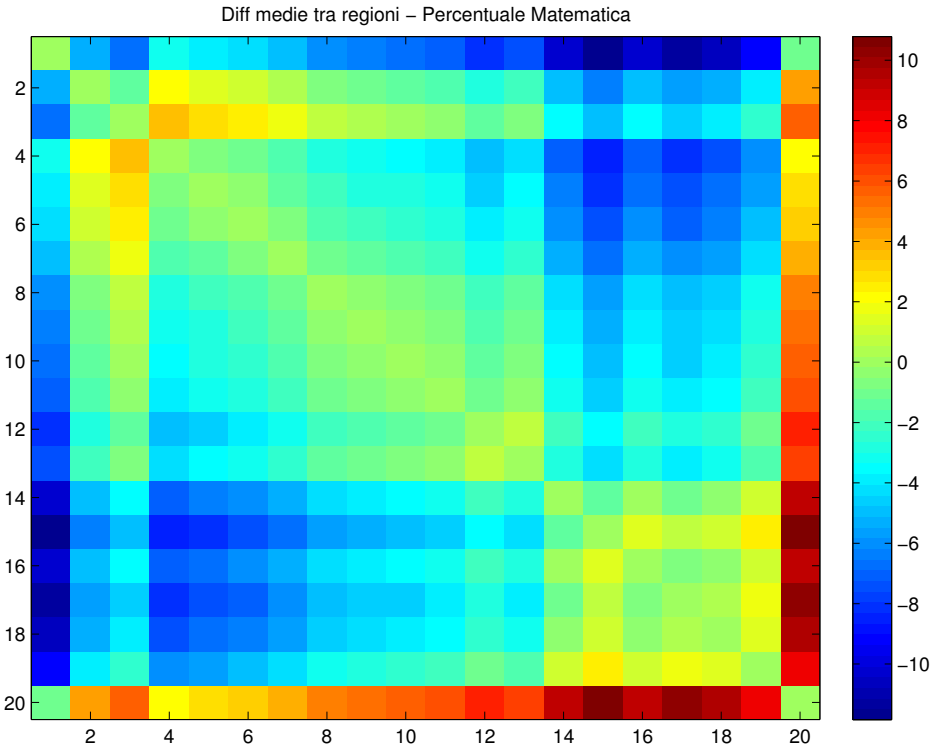
3.4.2 Fattore regione

Grandmean: 67.38

F-test: 4.946572e+001 - Liv. signif.: 0

Regione 1: Valle d'Aosta - Estimated mean: 60.61 +/- 2.92
Regione 2: Piemonte - Estimated mean: 65.98 +/- 0.41
Regione 3: Liguria - Estimated mean: 67.36 +/- 0.69
Regione 4: Lombardia - Estimated mean: 63.82 +/- 0.27
Regione 5: Trentino - Estimated mean: 64.43 +/- 0.93
Regione 6: Veneto - Estimated mean: 64.87 +/- 0.38
Regione 7: Friuli - Estimated mean: 65.68 +/- 0.76
Regione 8: Emilia - Estimated mean: 66.72 +/- 0.44
Regione 9: Toscana - Estimated mean: 67.12 +/- 0.45
Regione 10: Umbria - Estimated mean: 67.34 +/- 0.88
Regione 11: Marche - Estimated mean: 67.72 +/- 0.64
Regione 12: Lazio - Estimated mean: 68.88 +/- 0.34
Regione 13: Abruzzo - Estimated mean: 68.05 +/- 0.66
Regione 14: Molise - Estimated mean: 70.90 +/- 1.12
Regione 15: Campania - Estimated mean: 72.45 +/- 0.29
Regione 16: Puglia - Estimated mean: 70.99 +/- 0.39
Regione 17: Basilicata - Estimated mean: 71.78 +/- 0.82
Regione 18: Calabria - Estimated mean: 71.31 +/- 0.45
Regione 19: Sicilia - Estimated mean: 69.92 +/- 0.30
Regione 20: Sardegna - Estimated mean: 61.67 +/- 0.55





Sono confermati gli andamenti visti per il caso generale.

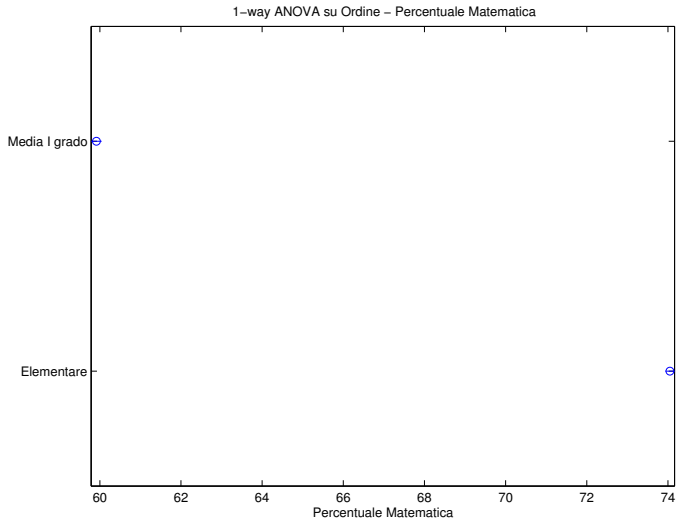
3.4.3 Fattore ordine

Grandmean: 66.98

F-test: 6.688000e+003 - Liv. signif.: 0

Ordine 1: Elementare - Estimated mean: 74.05 +/- 0.11
Ordine 2: Media I grado - Estimated mean: 59.92 +/- 0.13

Il divario tra le scuole elementari e medie torna a livelli medi (quasi il 15%) a favore delle scuole elementari.

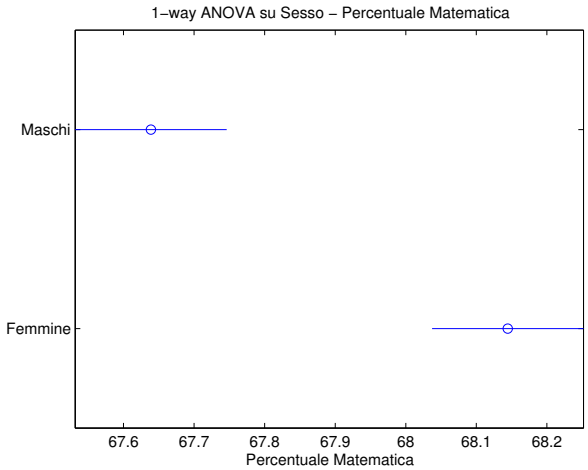


3.4.4 Fattore sesso

Grandmean: 67.89

F-test: 1.109970e+001 - Liv. signif.: 8.646009e-004

Sesso 1: Femmine - Estimated mean: 68.14 +/- 0.11
Sesso 2: Maschi - Estimated mean: 67.64 +/- 0.11



Per le prove di matematica si verifica un'inversione di tendenza, in quanto le percentuali di risposte esatte delle femmine sono leggermente (circa lo 0.5%) ma significativamente maggiori di quelle dei maschi.

3.5 Percentuale di risposte esatte: Scienze

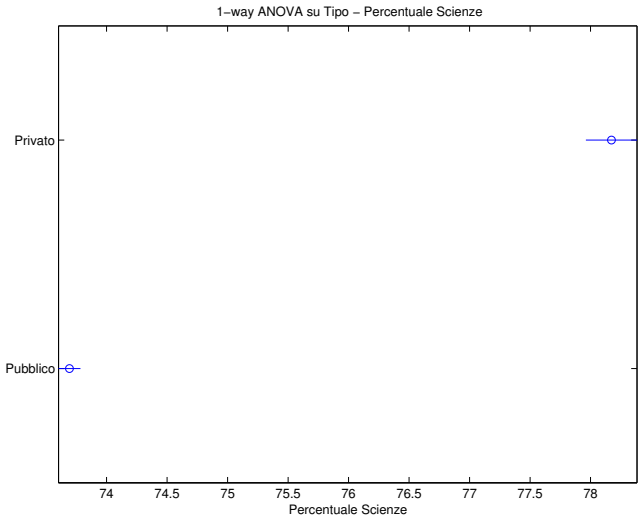
Le percentuali di risposte esatte alle domande di Scienze risultano essere costantemente le più alte tra le tre materie.

3.5.1 Fattore tipo

Grandmean: 75.93

F-test: 3.801739e+002 - Liv. signif.: 0

Tipo 1: Pubblico - Estimated mean: 73.69 +/- 0.09
Tipo 2: Privato - Estimated mean: 78.17 +/- 0.21



Si conferma l'andamento omogeneo per materia.

3.5.2 Fattore regione

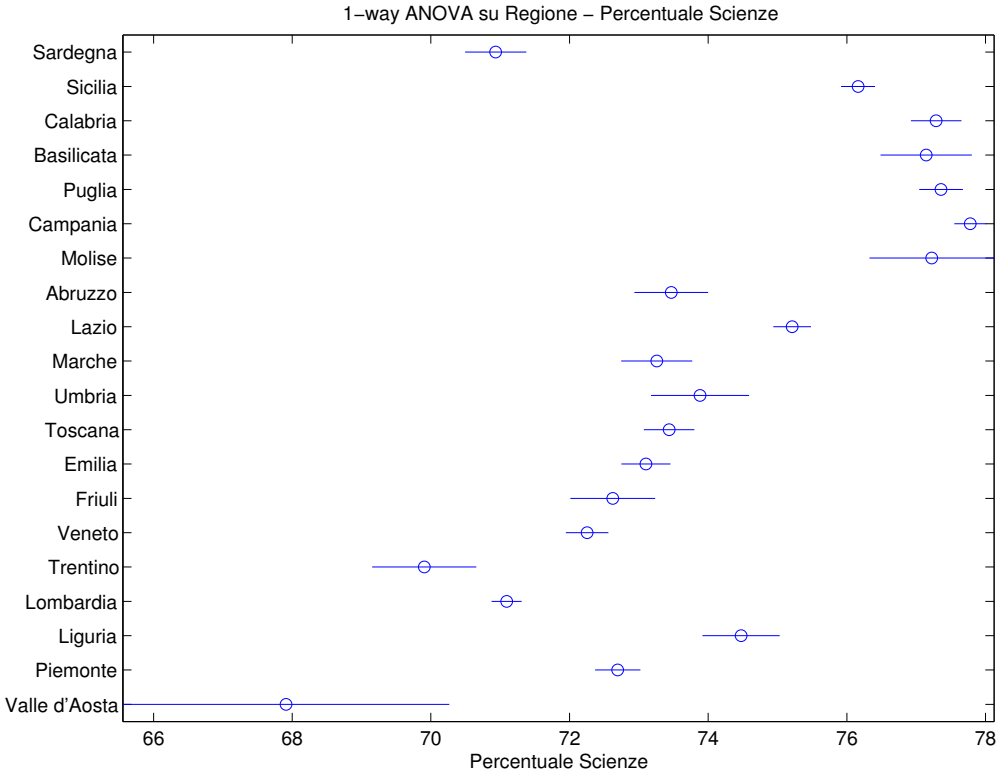
Grandmean: 73.86

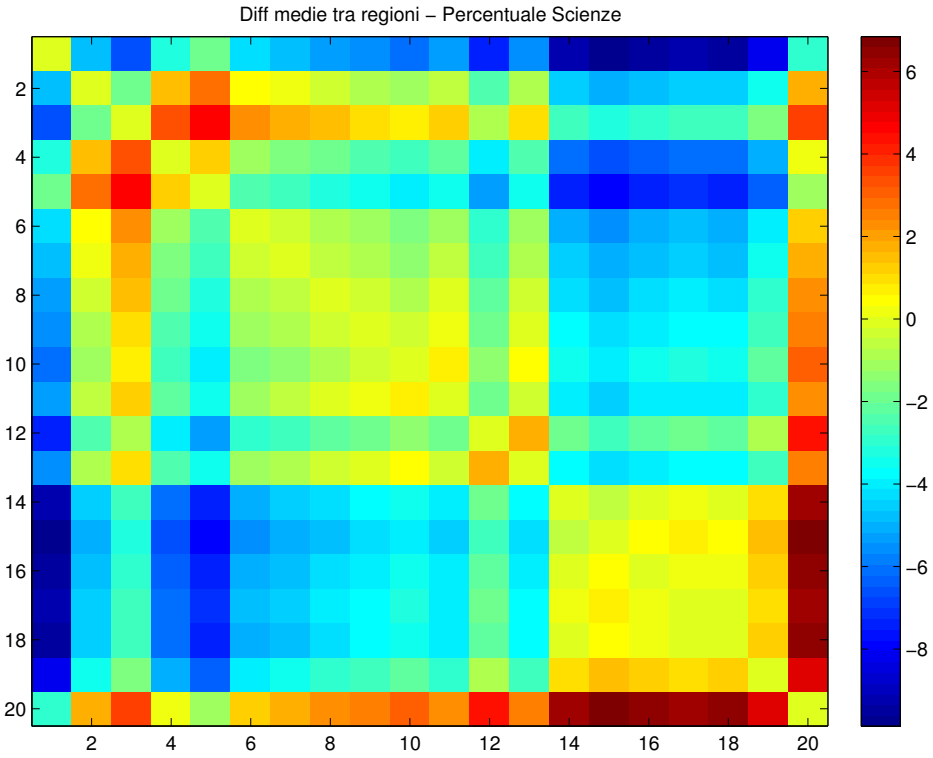
F-test: 4.763529e+001 - Liv. signif.: 0

Regione 1: Valle d'Aosta - Estimated mean: 67.91 +/- 2.36
Regione 2: Piemonte - Estimated mean: 72.69 +/- 0.33
Regione 3: Liguria - Estimated mean: 74.47 +/- 0.56
Regione 4: Lombardia - Estimated mean: 71.09 +/- 0.22
Regione 5: Trentino - Estimated mean: 69.90 +/- 0.75
Regione 6: Veneto - Estimated mean: 72.25 +/- 0.31
Regione 7: Friuli - Estimated mean: 72.62 +/- 0.61

Regione 8: Emilia - Estimated mean: 73.10 +/- 0.35
Regione 9: Toscana - Estimated mean: 73.44 +/- 0.36
Regione 10: Umbria - Estimated mean: 73.88 +/- 0.71
Regione 11: Marche - Estimated mean: 73.26 +/- 0.51
Regione 12: Lazio - Estimated mean: 75.21 +/- 0.27
Regione 13: Abruzzo - Estimated mean: 73.47 +/- 0.53
Regione 14: Molise - Estimated mean: 77.22 +/- 0.90
Regione 15: Campania - Estimated mean: 77.78 +/- 0.23
Regione 16: Puglia - Estimated mean: 77.36 +/- 0.32
Regione 17: Basilicata - Estimated mean: 77.14 +/- 0.66
Regione 18: Calabria - Estimated mean: 77.29 +/- 0.36
Regione 19: Sicilia - Estimated mean: 76.16 +/- 0.24
Regione 20: Sardegna - Estimated mean: 70.93 +/- 0.44

Si conferma l'andamento omogeneo rispetto alla materia.



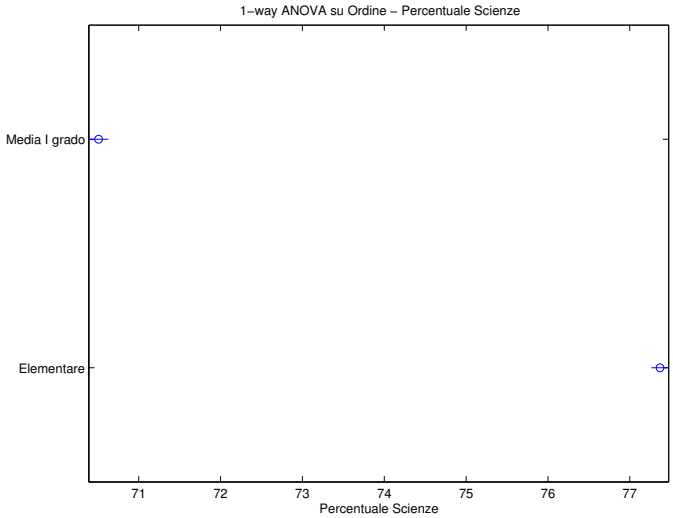


3.5.3 Fattore ordine

Grandmean: 73.94

F-test: 1.842854e+003 - Liv. signif.: 0

Ordine 1: Elementare - Estimated mean: 77.37 +/- 0.11
Ordine 2: Media I grado - Estimated mean: 70.51 +/- 0.12



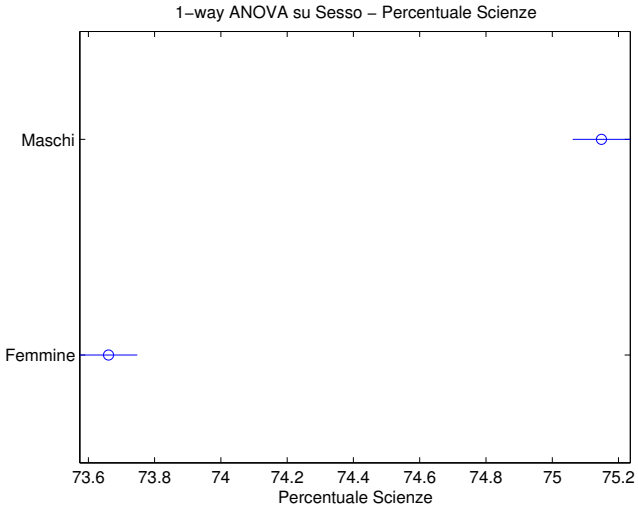
Nel caso delle Scienze si osserva il divario più basso tra le scuole elementari e medie: circa il 7%, sempre a favore delle scuole elementari.

3.5.4 Fattore sesso

Grandmean: 74.40

F-test: 1.469968e+002 - Liv. signif.: 0

Sesso 1: Femmine - Estimated mean: 73.66 +/- 0.09
Sesso 2: Maschi - Estimated mean: 75.15 +/- 0.09



Nelle Scienze i maschi ritornano ad avere un rendimento significativamente migliore delle femmine, con una differenza che si attesta attorno ad un punto e mezzo percentuale.

4. Prospettive

L'analisi svolta con i modelli ANOVA si è dimostrata utile nell'individuare la significatività di alcuni fattori considerati (tipo di scuola, se privata o pubblica; regione; ordine della scuola, se elementare o media; sesso) rispetto all'abilità dello studente, stimata sia mediante la percentuale di risposte esatte al questionario di valutazione, sia mediante un modello di Item Response Theory a tre parametri.

Il lavoro svolto nel corso del primo semestre è da intendersi prevalentemente preparatorio per le attività future, nel senso che sono state individuate le metodologie statistiche per le analisi, è stato acquisito l'intero database INVALSI dei questionari di Sistema e di Valutazione, sono stati messi a punto algoritmi per consentire la lettura ed elaborazione dei dati in ambiente Matlab, sono state implementate script Matlab per le analisi ANOVA. Le elaborazioni finali verranno effettuate alla luce dell'Analisi di Qualità dei dati da svolgere all'interno del progetto. Un obiettivo importante per il secondo semestre sarà costituito dal raggruppamento dei livelli in gruppi omogenei, in modo da consentire una prima significativa riduzione del numero di variabili in gioco. A tale scopo si prevede di ricorrere alla teoria dei test multipli per l'omogeneità di diversi campioni, passo che in maniera naturale segue le analisi ANOVA. Si valuterà anche se ricorrere a metodologie di clustering per l'individuazione dei gruppi omogenei. Inoltre sarà parallelamente incrementato il numero di fattori per valutare l'incidenza di altri parametri strutturali degli Istituti scolastici sulle abilità degli studenti. L'analisi verrà estesa anche all'anno scolastico 2005-2006 e verranno analizzate le differenze riscontrate.

Bibliografia

R.A. Fisher: The correlation between relatives on the supposition of Meldelian law of inheritance. *Trans. R. Soc. Edin.* **54**, 399-433 (1918).

H. Sahai, M.I. Ageel: The analysis of variance. Birkhauser, Berlino (2000)