Studio sperimentale sull'assegnazione dei punteggi nelle prove scritte dell'esame di stato

R. Bolletta

Premessa

La legge di riforma 425 del 10 dicembre 1997 e il DPR 323 del 23 luglio 1998 hanno radicalmente modificato gli esami di Stato conclusivi del ciclo secondario eliminando alcuni aspetti del precedente esame di maturità che nel tempo avevano subìto un progressivo degrado soprattutto nella loro capacità di realizzare un'affidabile valutazione dei livelli di apprendimento dei candidati.

L'esame previsto dalla riforma del '97 intendeva realizzare una valutazione finale maggiormente centrata sulla preparazione scolastica dei candidati, sulle conoscenze, le competenze e le capacità, superando quella nozione di 'maturità' troppo spesso ascrivibile prevalentemente a doti e a tratti di personalità dei giovani esaminati. Anche a questo scopo la valutazione finale si esprime mediante la somma di punteggi specificamente legati a singole prestazioni, a partire dalla carriera scolastica degli ultimi tre anni fino alla prova orale e al bonus. Per ottenere una valutazione più attendibile l'introduzione di una terza prova ha aumentato il peso delle prove scritte che venivano proposte sia come saggi complessi (la prima e la seconda), sia come prove strutturate. Con ciò la riforma ha inteso realizzare un esame finale più affidabile, più giusto, più legato alle prestazioni effettivamente rilevate e meno dipendente dalla casualità della circostanze che possono influire sulle prestazioni dei candidati, meno influenzato dalla variabilità dei criteri dei singoli commissari. La stessa presenza paritetica dei docenti interni nella commissione[1], oltre a rassicurare i candidati rispetto a un esame più impegnativo, intendeva migliorare la capacità di valutazione degli elaborati con commissari che conoscevano da lungo tempo i candidati e ne sapevano quindi interpretare meglio le prestazioni in un rapporto dialettico con i commissari esterni.

Va inoltre rilevato che l'introduzione della terza prova scritta lanciava un segnale forte alla scuola italiana poiché rendeva istituzionale l'uso di prove strutturate, anche a risposta chiusa, rispetto alle quali molti docenti avevano espresso in passato una pregiudiziale opposizione. L'assegnazione del punteggio alla terza prova ha posto due problemi fondamentali:

- quali dovevano essere i criteri di sufficienza visto che la prova era locale ed originale e quindi priva di una taratura preventiva? La definizione della soglia di sufficienza difficilmente si poteva basare su una esperienza condivisa dai commissari circa la prestazione attesa e prevedibile dei candidati.
- come si potevano rendere i punteggi delle altre prove scritte e orali, per le quali esisteva una esperienza più consolidata, omogenei al punteggio della terza prova?

Per comprendere meglio la rilevanza di tali problemi occorre anche considerare che la riforma aveva adottato scale numeriche diverse per le varie prove (quindicesimi per le

[1] Ricordiamo che prima della sessione 1999 solo una docente della classe partecipava ai lavori come membro interno mentre a partire dalla sessione 2002 tutta la commissione tranne il presidente è formata dai docenti della stessa classe.

prove scritte e trentacinquesimi per l'orale) e un valore di soglia per la sufficienza che non corrispondeva proporzionalmente al sei della scala in decimi.

Nel dibattito tra i docenti e nella formazione realizzata in occasione della riforma si è largamente diffusa una attenzione nuova per il miglioramento della attendibilità e precisione delle varie fasi della valutazione anche attraverso l'assegnazione più 'oggettiva' dei punteggi nelle prove di italiano e nelle seconde prove. Questa esigenza si è presto tradotta nell'adozione sistematica di griglie di correzione e/o di valutazione e in procedimenti di correzione più analitici.

Lo stesso DPR 323, che ha attuato la riforma, ha istituito un Osservatorio nazionale permanente sugli effetti dell'innovazione con il duplice scopo di facilitare l'attività delle commissioni che dovevano mettere a punto la terza prova e di monitorare gli andamenti dei risultati. L'Osservatorio ha centrato l'attenzione del monitoraggio prevalentemente sulla distribuzione statistica degli esiti attraverso la rilevazione analitica su tutta la popolazione dei punteggi delle singole prove. Si è trattato di uno sforzo notevolissimo che non aveva avuto uguali in passato, quando le rilevazioni sugli esiti degli esami di maturità erano state fatte in forma aggregata per classi di voto a partire dai singoli istituti scolastici. E' stata costruita una serie storica triennale di dati che, oltre ad una lettura di tipo censimentario, resa possibile dai repertori statistici annuali[2], si presta ad analisi più approfondite della qualità delle valutazioni operate dalle commissioni.

A tali archivi di dati sono stati affiancati archivi di elaborati raccolti da campioni rappresentativi di commissioni sui quali sono state condotte anche analisi approfondite di tipo qualitativo.

In tale contesto di lavoro si inquadra lo studio sperimentale i cui risultati sono alla base delle riflessioni condotte in questo volume.

Come nasce il problema

L'uso di scale numeriche per la formalizzazione degli esiti degli esami finali, dopo un lungo periodo in cui nel precedente esame di maturità sono stati usati giudizi articolati di tipo descrittivo, corredati da un voto complessivo espresso in forma sintetica e globale, ha riproposto il problema di una valutazione che si fondasse su operazioni di 'misura' valide ed affidabili. Un punteggio espresso in 15-simi richiede una discriminazione delle prestazioni rilevate molto più precisa e fine di quella necessaria per esprimere un giudizio qualitativo su tre o su cinque livelli. Tale situazione può essere direttamente osservata nell'attività di qualsiasi commissione: vi è una maggiore difficoltà a trovare l'accordo tra commissari che valutano la stessa prova mediante delle scale numeriche rispetto alla più facile convergenza su pochi livelli di tipo qualitativo.

L'attenzione delle commissioni si è quindi spostata dal momento valutativo, cioè dal momento in cui un fatto viene giudicato in base ad un criterio, a quello della '*misura*'

[2] Osservatorio Nazionale sugli Esami di Stato,*Gli esami in numeri. Sessione 1999* . Franco Angeli, 2000

cioè a quello della discriminazione quantitativa mediante delle procedure che assegnano le stesse quantità a parità di prestazione osservate e indipendentemente dal soggetto che misura o rileva la prestazione. Nei documenti diffusi dall'Osservatorio si raccomanda di tener distinti questi due momenti e si suggerisce l'uso di scale numeriche diverse da quelle usate per valutare le prove (v. allegato 1 in cui si insiste sulla distinzione tra punteggio grezzo e punteggio votato). Ma pur distinguendo nettamente la fase della '*misurazione*' da quella della valutazione, l'imprecisione e l'incertezza propri della fase della misura sono ineliminabili. Nell'ambito dell'educazione, l'esistenza di errori di misura, l'imprecisione di scale quantitative tende ad essere rifiutata da chiunque voglia di associare tali valori alle prestazioni di persone che devono essere giudicate.

Esorcizzare l'errore di misura negandone l'esistenza non migliora però la situazione anzi la peggiora in quanto non si assume l'atteggiamento di chi cerca di aumentare la precisione delle misure effettuate ma piuttosto di chi difende come indiscutibile il valore puntuale accertato in una singola misura. Nell'ambito delle scienze sperimentali tutti sanno che i dati prodotti da una misurazione sono affetti da errori casuali ineliminabili e la possibilità di apprezzare l'intensità di tali errori e di poterne ridurre gli effetti affinando i metodi e gli strumenti di misura consente di procedere nella conoscenza e di operare sulla realtà con un'efficacia ed una precisione sempre crescenti. Nell'ambito della valutazione scolastica una discriminazione quantitativa delle prestazioni di una persona troppo spesso è rifiutata perché ideologicamente inaccettabile o è, all'opposto, assunta come un giudizio assoluto difficilmente discutibile.

Queste considerazioni valgono per tutta la valutazione scolastica ma assumono un rilievo particolare nell'esame di Stato finale della scuola secondaria che formalizza una valutazione sommativa senza appelli e che lascia una segno forte per tutta la successiva carriera di lavoro o di studio.

Gli effetti di questa situazione sui casi singoli sono ben evidenti e sono alla base sia delle difficoltà di accordo tra i correttori di prove sia di ingiustizie, vere o presunte, denunciate da numerosi candidati e studenti che non ritengono equa la valutazione ricevuta.

Ma l'imprecisione nella assegnazione dei punteggi produce degli effetti riscontrabili anche sul complesso della popolazione? Possiamo trovare una traccia empirica che abbia una significatività statistica nelle distribuzioni degli esiti? Ci sono effetti sistematici legati alla struttura della popolazione degli studenti o alla composizione delle commissioni o alle varie tipologie degli indirizzi di studio che spostano significativamente i valori assegnati? E' possibile saggiare attraverso le distribuzioni dei punteggi assegnati l'intensità degli errori di misura e ricostruire una stima attendibile dei valori veri delle prestazioni rilevate?

L'analisi dei dati delle prime due sessioni 1999 e 2000 ci ha fornito indizi piuttosto chiari della rilevanza del problema, indizi che sono stati tempestivamente resi di pubblico dominio mediante la disponibilità di alcuni grafici sul sito Internet del Cede.

Nel grafico della figura 1 sono rappresentate le distribuzioni dei punteggi delle tre prove scritte assegnati nella sessione 1999.

In ordinata sono rappresentate le frequenze relative percentuali di tutta la popolazione registrata, circa 400.000 casi. Trattandosi di una popolazione estremamente vasta fornita di competenze complesse, possiamo supporre che la effettiva distribuzione della padronanza del possesso delle competenze accertate dalle prove sia distribuita normalmente come accade per tutte quelle variabili statistiche che dipendono da un gran numero di fattori indipendenti, nessuno dei quali è preponderante. Quindi la distribuzione teoricamente attesa, più adatta a rappresentare il voto vero per le tre prove, dovrebbe essere una classica distribuzione gaussiana. Osserviamo invece che le distribuzioni effettivamente osservate presentano delle irregolarità chiaramente spiegabili:

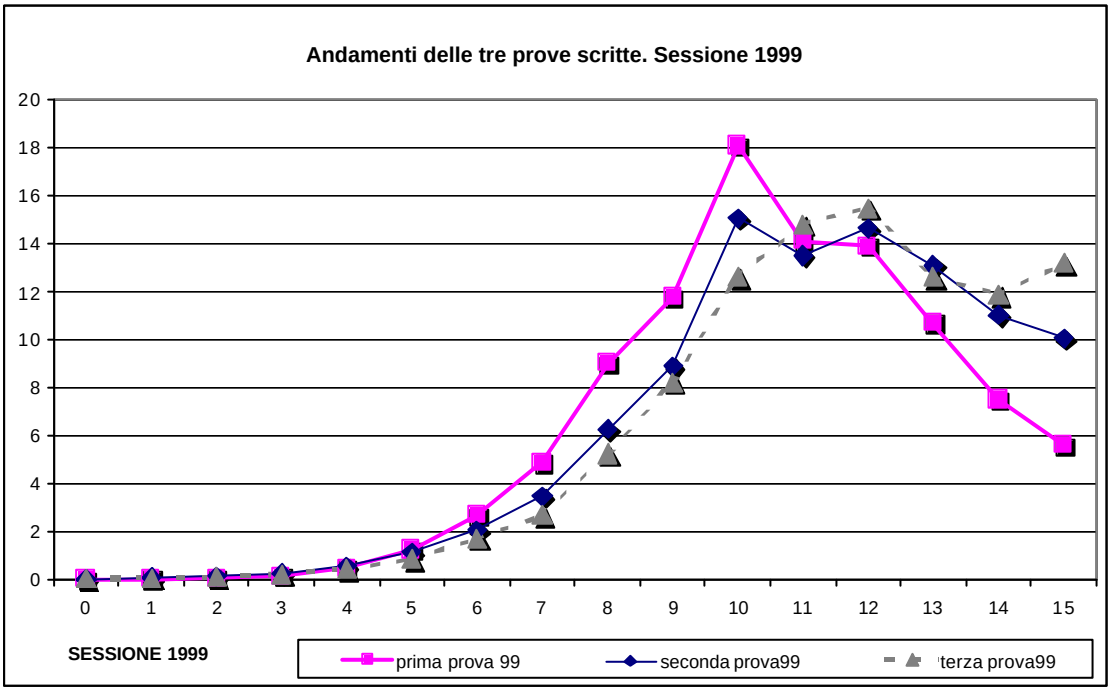


Fig.1 Punteggi nelle prove scritte sessione 1999

- in corrispondenza al valore 10, che è la soglia di sufficienza nella prima e nella seconda prova, compaiono due frequenze modali, due picchi che rompono la regolarità della distribuzione;
- nella terza prova un'analoga situazione si riscontra invece in 15, che è il punteggio massimo.

La spiegazione di questi due fatti è piuttosto semplice:

- sulla soglia della sufficienza avvengono probabilmente arrotondamenti verso l'alto dei punteggi insufficienti,
- la difficoltà media della terza prova è stata più bassa delle altre due determinando sul punteggio massimo un accumulo delle frequenze che dovevano trovarsi in una coda superiore al massimo della scala.

Ma mentre il secondo fenomeno è solo una spia dell'influenza che ha il livello di difficoltà della prova sulla distribuzione degli esiti, l'arrotondamento sulla soglia di sufficienza è l'indizio evidente dell'imprecisione con cui vengono assegnati i punteggi. Quanto più la stima del punteggio vero è imprecisa tanto più sono vistosi gli effetti sistematici del desiderio dei commissari di non danneggiare nessuno.

Da notare che questa distorsione è più forte nella prima prova, meno accentuata nella seconda prova (dove però, come vedremo, appare un altro evento sistematico), non compare nella terza prova. Possiamo supporre (anche ciò è stato oggetto di verifica nello studio sperimentale) che l'imprecisione della stima

- sia più ampia nella prima prova,
- si riduca nella seconda prova, in cui le varie prestazioni richieste sono meglio identificate e circoscritte entro le competenze specifiche dell'indirizzo di studi,
- non compare nella terza prova in cui l'assegnazione del punteggio viene effettuata mediante il conteggio di elementi riscontrabili più oggettivamente delle prime due prove scritte.

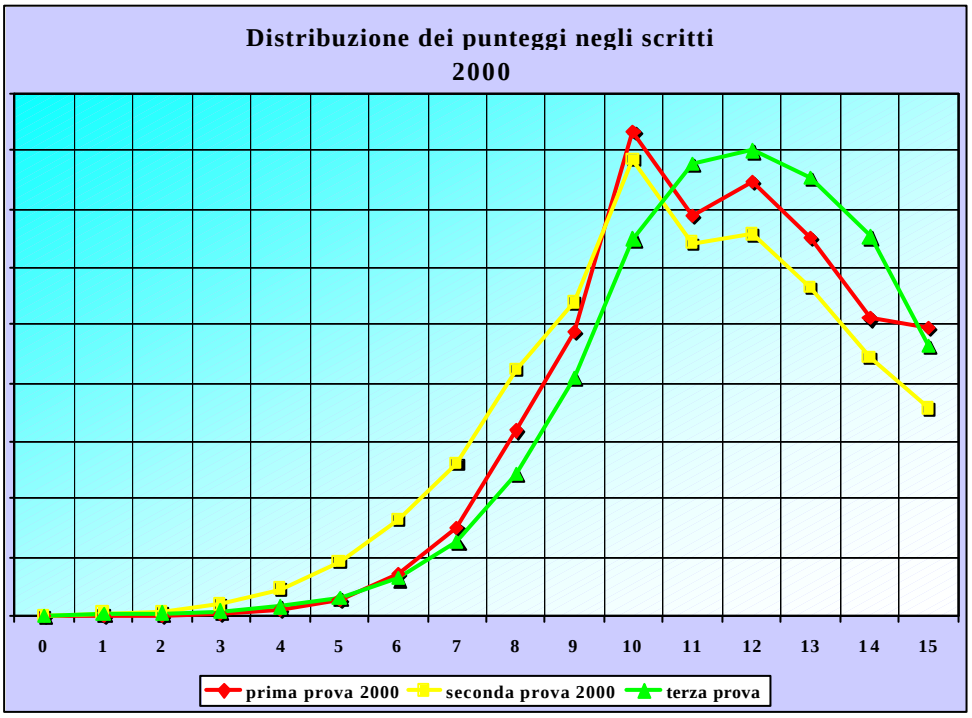


Fig.2 Punteggi delle prove scritte sessione 2000

Nella figura 2 si può osservare la situazione verificatasi nella sessione successiva del 2000 e trarre immediata conferma dei problemi ora segnalati. Qui è ancora più evidente sul valore del livello soglia il diverso comportamento della terza prova rispetto alle prime due. Sulla terza prova si nota inoltre che scompare l'accumulazione della frequenza sul valore 15 per effetto di un migliore adattamento dei livelli di difficoltà alle situazioni effettive e all'uso di un maggior numero di quesiti rispetto al primo anno di attuazione della riforma.

Confrontando i grafici delle due sessioni è possibile osservare un ulteriore effetto dell'incertezza insita nell'assegnazione dei punteggi: l'incidenza delle caratteristiche dei correttori. Ovviamente, possiamo effettuare un'analisi solo rispetto all'unica caratteristica disponibile dei commissari, ovvero l'essere docenti interni o esterni.

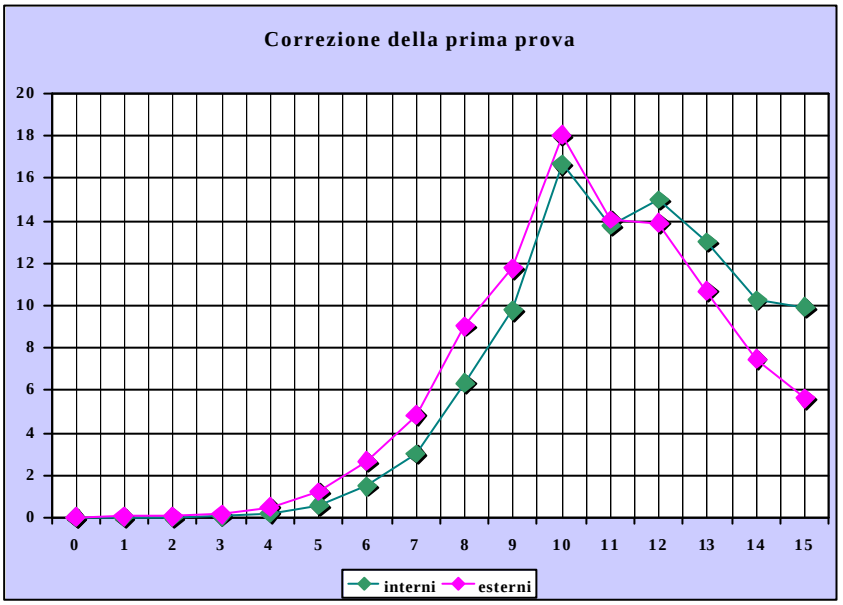


Fig.3 Comparazione dei punteggi rispetto al tipo di correttore (I prova)

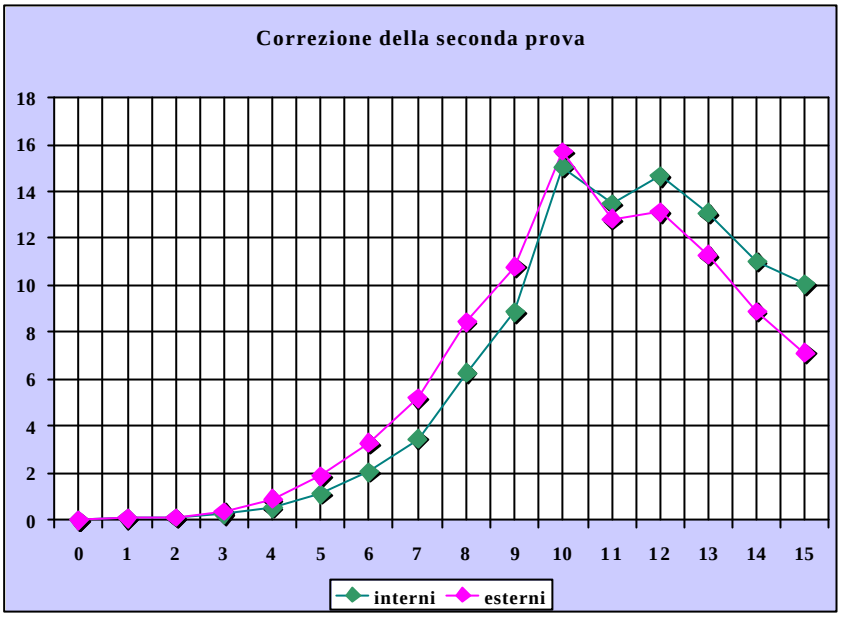


Fig.4 Comparazione dei punteggi rispetto al tipo di correttore (II prova)

Ebbene, poiché possiamo supporre che la distribuzione delle competenze rilevate dalle prove sia sostanzialmente stabile in due anni successivi, e che quindi le distribuzioni complessive delle due sessioni siano sostanzialmente identiche, possiamo dedurre che la differenza di andamento osservabile nelle figure 3 e 4[3] dipenda da arrotondamenti sistematici operati in modo più favorevole ai candidati da parte dei commissari interni e forse da un meccanismo contrario da parte dei commissari esterni. Si potrebbe supporre anche l'esistenza di un diverso criterio di valutazione, l'uso di una *'unità di misura'* diversa, ma, ancora una volta, il fatto che le differenze appaiano più ampie nella prima prova rispetto alla seconda ci induce a ipotizzare una maggiore incertezza nella stima dei punteggi nella prima prova e che le differenze nelle distribuzioni siano effetto degli arrotondamenti delle stime più che della presenza di criteri di valutazione sistematicamente diversi.

[3] Nelle prime due sessioni vi è stato uno scambio di ruoli: nel 1999 gli esterni hanno corretto la prima prova e gli interni la seconda, mentre nel 2000 gli interni hanno corretto la prima lasciando agli esterni la seconda. Le fig.3 e 4 consentono di mettere a confronto gli esiti delle due sessioni 1999 e 2000 per ciascuna prova

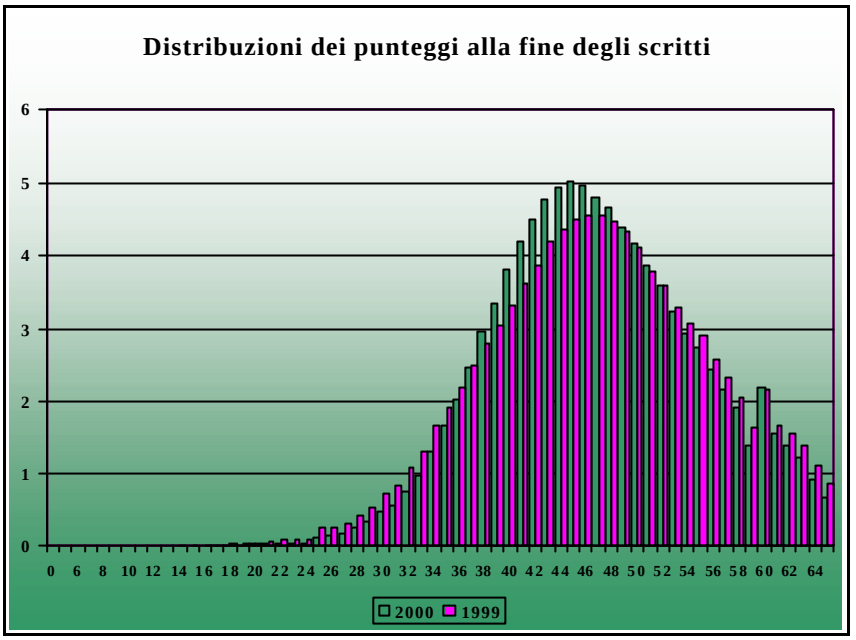


Fig.5 Punteggio totale delle prove scritte

I grafici delle figure 5 e 6 completano il quadro problematico da cui parte lo studio sperimentale. Osserviamo che la distribuzione del punteggio totalizzato alla fine delle prove scritte (credito + prove scritte) ha un andamento del tutto regolare e l'effetto soglia, evidente nelle singole prove, sparisce nella somma poiché agisce in modo indipendente tra le due prime prove (pochi sono i candidati che trovandosi leggermente al di sotto della soglia di sufficienza beneficiano dell'arrotondamento correttivo verso l'alto su entrambe le prove) e l'errore sistematico in una singola misura ha un effetto relativo più ridotto se la misura è sommata ad altre tre misure. Rimane una leggera irregolarità intorno al 60 che dipende, anche in questo caso, da un arrotondamento positivo e intenzionale verso tale soglia per consentire eventualmente di assegnare il bonus nei casi di eccellenza.

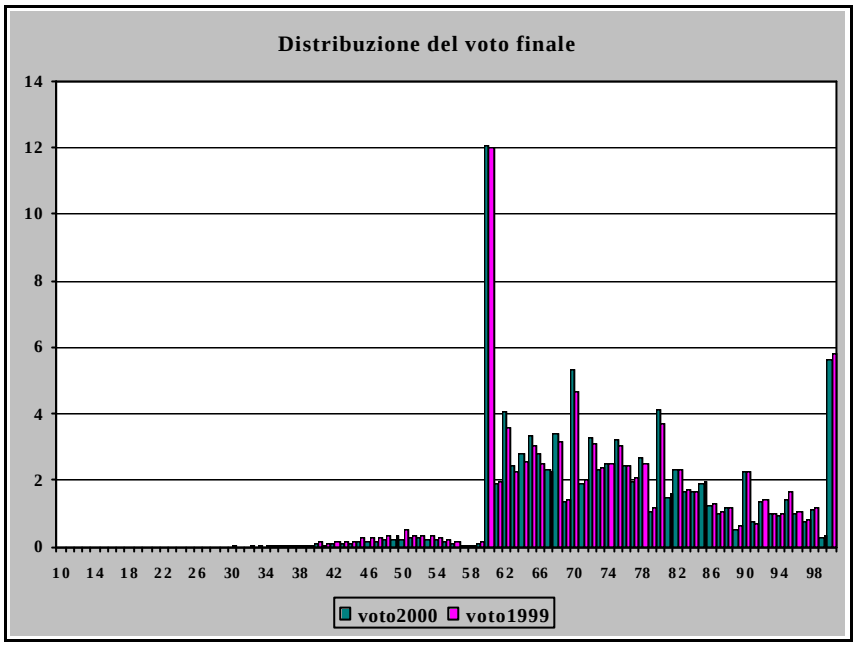


Fig.6 Distribuzione del punteggio finale (1999 e 2000)

Ma la distribuzione della figura 5 viene completamente modificata: sommando il punteggio della prova orale si ottiene una distribuzione del voto finale del tutto *'intenzionale'*. Anche per il voto finale valgono le stesse considerazioni applicate al punteggio complessivo ottenuto alla fine delle prove scritte: la preparazione complessiva effettivamente accertata dall'esame ha un andamento regolare di tipo gaussiano, e non a dente di sega. Le irregolarità che osserviamo nella distribuzione effettiva dipendono da un'assegnazione del punteggio dell'orale che, data l'imprecisione della stima, ne ha arrotondato in modo intenzionale il valore, tenendo conto del punteggio accumulato alla fine degli scritti e a volte anche per compensarne alcuni livelli troppo scadenti. L'effetto di tali aggiustamenti è evidente nell'alta frequenza del sessanta, soglia minima per ottenere la promozione, che assorbe probabilmente casi che dovevano trovarsi al di sotto, se la valutazione dell'orale non fosse stata aggiustata tenendo conto dell'esito degli scritti e di altre informazioni globali disponibili. L'arrotondamento è visibile anche negli effetti soglia presenti in tutte le decine successive che determinano una distribuzione a dente di sega.

La frequenza del 100 dipende sia dall'effetto del *bonus* che è assegnabile solo ai casi di eccellenza sia da un generale effetto di trascinamento legato al valore simbolico di tale voto rispetto alla qualità complessiva della classe esaminata o della scuola.

Riassumendo il quadro problematico da cui parte lo studio, possiamo dire che

- sono emersi effetti statisticamente significativi di distorsioni sistematiche dei punteggi legate alla variabilità propria di misure affette da errori casuali,

- la qualità complessiva della valutazione operata dal nuovo esame dipende dalla precisione delle operazioni di assegnazione del punteggio numerico alle singole prove di esame.

Obiettivi dello studio sperimentale

Data la rilevanza del problema dell'affidabilità nell'assegnazione dei punteggi sia per gli effetti diretti che ha sui singoli candidati, sia per il successo stesso dell'innovazione indotta dalla riforma degli esami sia infine per il miglioramento della qualità della valutazione scolastica corrente, è stato realizzato uno studio sperimentale volto a:

- quantificare l'errore di misura delle operazioni di assegnazione dei punteggi nelle prove scritte dell'esame di stato,
- determinare i fattori che influenzano l'ampiezza di tale errore,
- individuare strategie di miglioramento della precisione delle valutazioni compatibili con le modalità di esecuzione degli esami.

La problematica generale che abbiamo descritto è stata, per comprensibili ragioni di fattibilità, limitata alle sole prove scritte. Infatti solo per queste si disponeva di elaborati autentici da valutare mentre per i colloqui è praticamente impossibile acquisire una documentazione autentica del loro svolgimento su un vasto numero di casi senza turbarne il normale svolgimento. D'altra parte gli stessi grafici di figura 6 come pure la diretta esperienza di tutti coloro che hanno condotto colloqui mostrano che l'errore di misura dei punteggi assegnabili nell'orale è certamente maggiore di quelli assegnabili nelle prove scritte.

La ricerca ha altresì individuato i seguenti obiettivi specifici che ne hanno ispirato e guidato lo svolgimento:

- documentare empiricamente l'esistenza degli errori di misura casuali non eliminabili;
- diffonderne la consapevolezza tra coloro che valutano, per migliorare l'accuratezza della fase di *'misura'* delle prestazioni legate al profitto scolastico;
- analizzare in che modo le varie prove scritte (saggi, problemi, progetti, prove strutturate) contribuiscono alla formulazione di un punteggio finale attendibile;
- ricostruire con apposite simulazioni possibili distribuzioni 'vere' depurate dagli effetti di errori sistematici.

Disegno sperimentale

Come abbiamo detto, è facile riscontrare una certa discordanza tra correttori dello stesso elaborato scritto, soprattutto se si usa una scala numerica con una gamma piuttosto ampia, come accade negli attuali esami di Stato. Per riuscire a valutare il grado di

accuratezza dei punteggi, occorrerebbe ripetere la correzione della stesso elaborato per un numero di volte praticamente infinito, ripetere questa stessa procedura per molti altri elaborati della stessa prova e, infine, vedere se le cose cambiano variando il tipo di prova. Se tutti i punteggi assegnati allo stesso elaborato fossero uguali, e ciò fosse vero per ogni elaborato, potremmo dire che il nostro procedimento non sia affetto da errori; se invece i punteggi assegnati sono diversi, l'errore per ogni misura è la differenza tra il punteggio assegnato dal singolo correttore e il punteggio *'vero'*.

Ma quale tra i tanti assegnati è il punteggio *vero*? Dopo aver variato opportunamente tutti i fattori che potrebbero provocare degli errori sistematici (correttori più o meno severi, particolari tecniche di correzione più o meno condivise ecc.), potremo assumere come **stima** puntuale del punteggio *vero* la media aritmetica di tutti i punteggi assegnati. Allora l'errore di misura sarà la differenza tra ciascun punteggio assegnato e la media aritmetica di tutti i punteggi assegnati.

E' evidente che tale procedura è realisticamente attuabile se le correzioni sono ripetute in un numero economicamente sostenibile. Per studiare i fattori che influiscono sull'intensità dell'errore occorre inoltre correggere ripetutamente lo stesso elaborato variandone opportunamente le condizioni, ovvero il tipo di correttore.

I fattori che abbiamo tenuto presente in questo esperimento e che vanno opportunamente incrociati sono stati:

il tipo di prova scritta: prima, seconda e terza

Per la prima occorre distinguere la traccia

Per ogni traccia occorre distinguere l'ordine scolastico (licei tecnici e professionali)

Per la seconda distinguere la materia

Per la terza distinguere l'ordine scolastico

Il tipo di correttore: da solo o in commissione

Se da solo distinguere il tipo di istituto di provenienza

Per lo stesso istituto distinguere per genere, età e territorio

Se in commissione distinguere per istituto.

La figura 7 illustra in modo procedurale tali criteri di scelta che caratterizzano i singoli fascicoli di prove che sono stati sottoposti a correzione ripetuta.

La figura 8 rappresenta complessivamente in che modo il tipo di prova , il tipo di correttore e il tipo di istituto concorrono alla classificazione del corpo degli elaborati.

Inoltre occorre studiare la stabilità della correzione prevedendo che lo stesso elaborato sia corretto due volte dallo stesso correttore in tempi diversi.

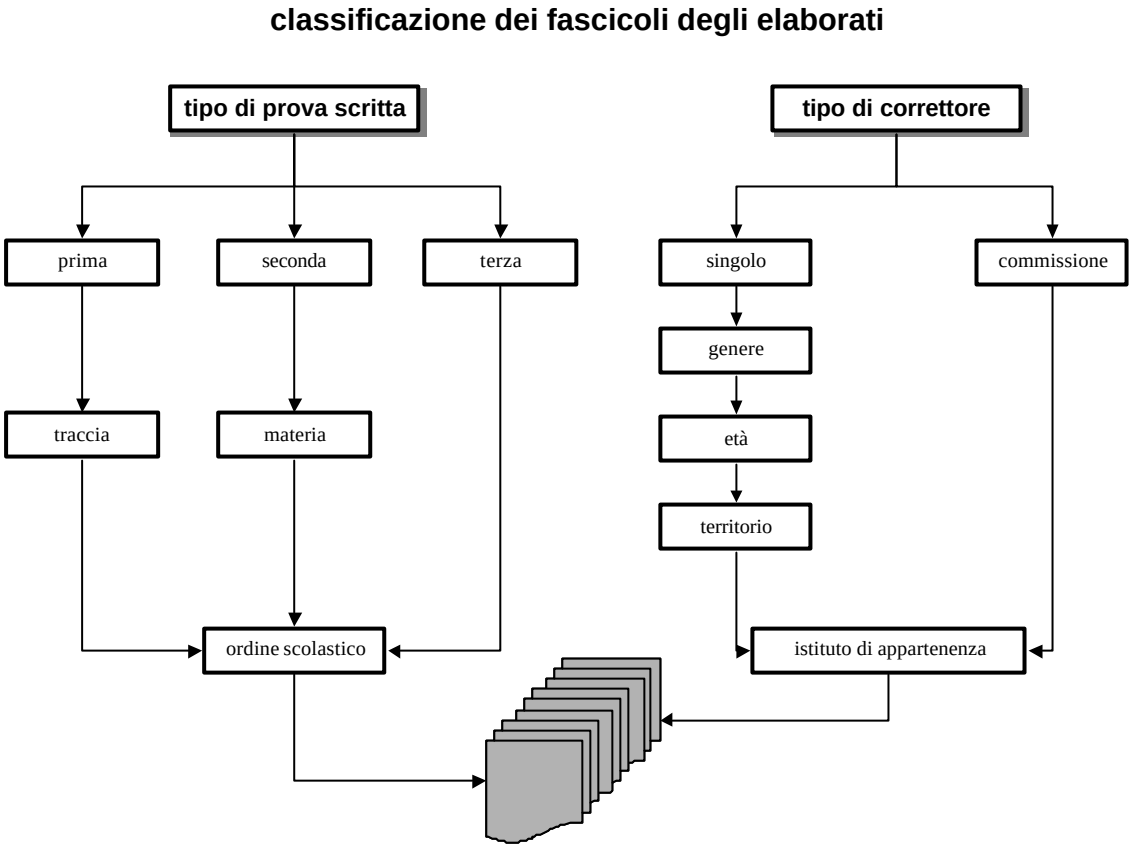


Fig.7 Classificazione dei fascicoli degli elaborati per la correzione ripetuta

Per poter infine analizzare l'effetto degli errori di misura sulle tre prove dello stesso candidato, ferma restando la casualizzazione rispetto ai precedenti fattori, si è cercato di massimizzare il numero dei candidati di cui si correggevano tutte e tre le prove. Per questo nel campionamento casuale degli elaborati si è partiti dalla estrazione delle seconde prove cui venivano associate le terze prove dello stesso candidato e la prima prova, se questa rientrava nei limiti numerici previsti dalle rotazioni dei fattori della prima prova.

In pratica, ogni elaborato della seconda prova scritta è stato corretto 11 volte: 4 volte da altrettanti docenti della disciplina in 15-simi, 1 volta da uno dei cinque correttori dopo 15 giorni, 2 volte da altrettanti commissioni in modo collegiale, 2 volte da una commissione di due docenti della stessa materia, 2 volte da altrettanti docenti usando i voti in decimi.

La tabella 1 illustra la struttura del piano di correzione nel caso della prova di matematica del liceo scientifico. Ogni correttore è stato identificato da un codice univoco che riporta la tipologia: MT* per i correttori singoli, MTCP* per i correttori in coppia e COMLS* per le due commissioni del liceo scientifico. La tabella 1 ha consentito di pianificare l'uniforme distribuzione dei correttori sul territorio e l'identificazione del genere.

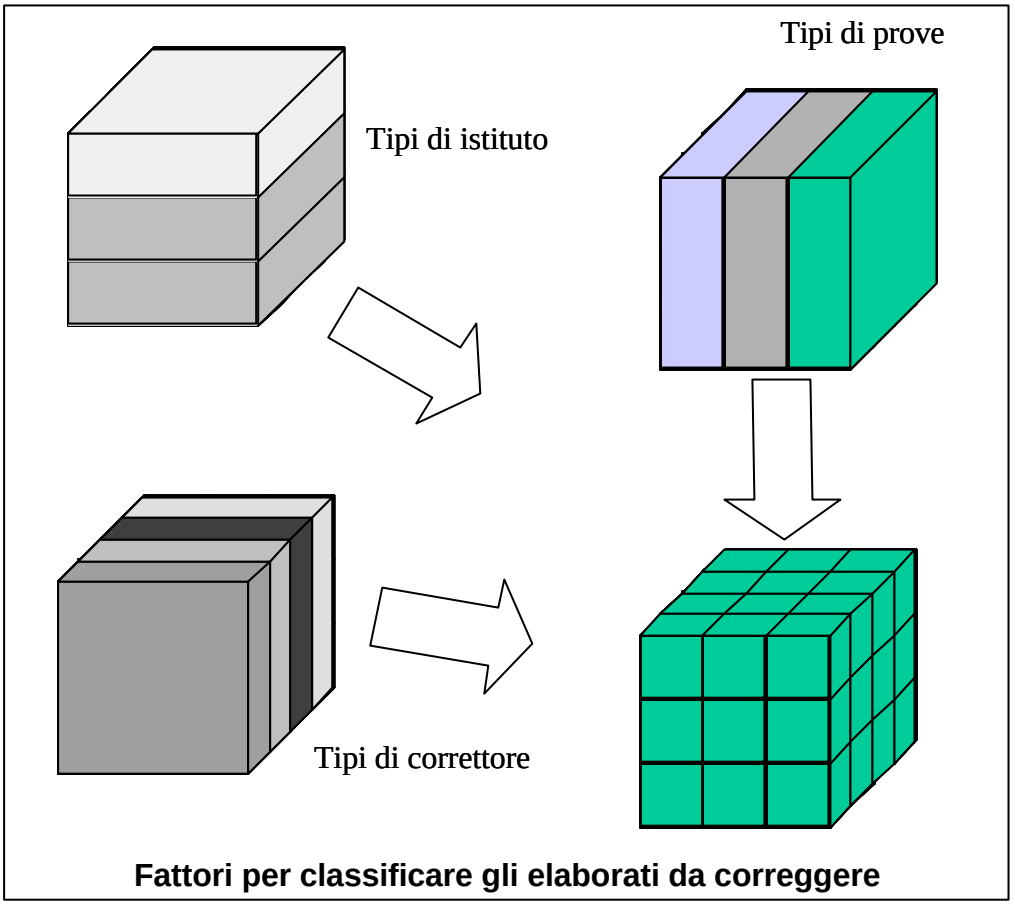


Fig.8 Dimensioni rispetto alle quali sono classificati gli elaborati

Nella testata della tabella 1 compare il nome dei fascicoli che raccolgono gli elaborati e il numero degli elaborati da correggere per ciascun correttore. La tabella mostra inoltre come ai correttori MT1 MT2, MT3 e MT4 sono stati assegnati i fascicoli per la seconda correzione dopo 15 giorni. Si può vedere infine che il totale degli elaborati utilizzati nell'esperimento sono 20 per un totale di 220 correzioni. Il piano della tabella 1 è simile per le quattro seconde prove utilizzate nell'esperimento.

Correttori				Disciplina				correzioni
				N elaborati per fascicolo				
Codice	N	Genere	STRATO	MAT1	MAT2	MAT3	MAT4	
MT1	1	M	nord	5	5	5	5	20
MT2	1	F	centro	5	5	5	5	20
MT3	1	M	sud	5	5	5	5	20
MT4	1	F	nord	5	5	5	5	20

COMLS	6		sud	5	5	5	5	20
COMLS1	6		nord	5	5	5	5	20
MTCP1	1		centro	5	5	5	5	20
MTCP2	1		centro	5	5	5	5	20
MT1		M	nord	*5*				5
MT2		F	centro		*5*			5
MT3		M	sud			*5*		5
MT4		F	nord				*5*	5
MT5	1	M	centro	5	5	5	5	20
MT6	1	F	sud	5	5	5	5	20
Totale dati								220

Tab.1 Esempio di pianificazione del campionamento per gli elaborati di matematica

Anche per le prime prove è stato utilizzato un piano di correzione analogo ma con qualche complicazione in più: 3 correzioni da parte di singoli docenti di italiano in 15-simi (docenti provenienti ciascuno da un diverso ordine scolastico, licei, tecnici e professionali), 1 volta da uno dei tre dopo 15 giorni, 2 volte da altrettante commissioni specifiche dell'indirizzo di studio in modo collegiale, 2 volte da una commissione di due docenti della stessa materia presi nell'ordine scolastico da cui proviene l'elaborato, 1 volta da altrettanti docenti che usano i voti in decimi, infine 1 correzione utilizzando una griglia.

traccia	Licei	Tecnici	Profess.	N.elaborati
A	LIC1	TEC1	PRF1	15
B1	LIC2	TEC2	PRF2	15
B2	LIC3	TEC3	PRF3	15
B3	LIC4	TEC4	PRF4	15
B4	LIC5	TEC5	PRF5	15
C	LIC6	TEC6	PRF6	15
D	LIC7	TEC7	PRF7	15
N. elaborati	35	35	35	105

Tab.2 Struttura dei fascicoli degli elaborati della prima prova.

Correttori				A	B1	B2	B3	B4	C	D	A	B1	B2	B3	B4	C	D	A	B1	B2	B3	B4	C	D	correzioni
Codice	N	Genere	Strato	LIC1	LIC2	LIC3	LIC4	LIC5	LIC6	LIC7	TEC1	TEC2	TEC3	TEC4	TEC5	TEC6	TEC7	PRF1	PRF2	PRF3	PRF4	PRF5	PRF6	PRF7	
ITALC1	1	M	nord	5				5				5				5				5				5	30
ITALC2	1	F	centro		5				5				5				5				5				25
ITALC3	1	M	sud			5				5	5								5			5			25
ITALC4	1	F	nord				5							5	5			5					5		25
ITATC1	1	M	centro	5				5				5				5				5				5	30
ITATC2	1	F	sud		5				5				5				5				5				25
ITATC3	1	M	nord			5				5	5								5			5			25
ITATC4	1	F	centro				5							5	5			5					5		25
ITAPR1	1	M	sud	5				5				5				5				5				5	30
ITAPR2	1	F	nord		5				5				5				5				5				25
ITAPR3	1	M	centro			5				5	5								5			5			25
ITAPR4	1	F	sud				5							5	5			5					5		25
COMLC	6		nord	5	5	5	5	5	5	5															35
COMLS	6		sud	5	5	5	5	5	5	5															35
COMTEC	6		nord								5	5	5	5	5	5	5								35
COMRAG	6		sud								5	5	5	5	5	5	5								35
COMPROF	6		nord															5	5	5	5	5	5	5	35
COMPROF1	6		sud															5	5	5	5	5	5	5	35
ITALCCP5	1		nord	5	5	5	5	5	5	5															35
ITALCCP6	1		nord	5	5	5	5	5	5	5															35
ITATCCP5	1		sud								5	5	5	5	5	5	5								35
ITATCCP6	1		sud								5	5	5	5	5	5	5								35
ITAPRCP5	1		centro															5	5	5	5	5	5	5	35
ITAPRCP6	1		centro															5	5	5	5	5	5	5	35
ITALC1		M	nord	5				5																	10
ITALC2		F	centro		5				5																10
ITALC3		M	sud			5				5															10
ITALC4		F	nord				5																		5
ITATC1		M	centro									5				5									10
ITATC2		F	sud										5				5								10
ITATC3		M	nord								5														5
ITATC4		F	centro											5	5										10
ITAPR1		M	sud																	5				5	10

Correttori				A	B1	B2	B3	B4	C	D	A	B1	B2	B3	B4	C	D	A	B1	B2	B3	B4	C	D	correzioni
Codice	N	Genere	Strato	LIC1	LIC2	LIC3	LIC4	LIC5	LIC6	LIC7	TEC1	TEC2	TEC3	TEC4	TEC5	TEC6	TEC7	PRF1	PRF2	PRF3	PRF4	PRF5	PRF6	PRF7	
ITAPR2		F	nord																		5				5
ITAPR3		M	centro																5			5			10
ITAPR4		F	sud															5					5		10
GRGLC1	1	M	nord	5		5					5		5					5		5					30
GRGLC2	1	F	centro		5		5					5		5					5		5				30
GRGTC3	1	M	sud			5		5					5		5					5		5			30
GRGTC4	1	F	nord				5		5					5		5					5		5		30
GRGTC5	1	M	centro					5		5					5		5					5		5	30
GRGPR6	1	F	sud	5					5		5					5		5					5		30
GRGPR7	1	M	nord		5					5		5					5		5					5	30
VOTO1	1	M	nord			5				5	5								5			5			25
VOTO2	1	F	sud				5							5	5			5					5		25
	63			10	10	11	11	10	10	11	11	10	10	11	11	10	10	11	11	10	10	11	11	10	1.100

Tab. 3 Pianificazione del campionamento per la prima prova.

Per le prime prove occorreva tener conto delle differenze determinate dalla traccia e dall'ordine scolastico. La tabella 2 mostra i nomi dei fascicoli contenenti le prove da correggere, legati al tipo di traccia e al tipo di scuola. In tal modo è stato possibile ripartire uniformemente i 105 elaborati corretti rispetto alle due caratteristiche considerate.

Ben più complicato è il piano di assegnazione delle correzioni illustrato dalla tabella 3. Per quanto riguarda i correttori occorre infatti tener conto della diversa appartenenza ai vari ordini scolastici e far in modo che la distribuzione degli elaborati sia equiripartita anche rispetto a tale caratteristica. Il totale delle correzioni previste ammonta così a 1.100 valori con un carico di lavoro per correttore da circa 25 a 30 elaborati a testa.

Le correzioni ripetute dello stesso elaborato di terza prova sono state solo tre per ovvi problemi di costi. Infatti è necessario che la terza prova sia comunque corretta collegialmente e ciò implica un alto numero di correttori impegnati per poche correzioni collegiali. Quattro sono stati i tipi di commissioni coinvolte: 4 commissioni che correggono la prima, la seconda e la terza prova, quattro commissioni che correggono solo la seconda e la terza prova e 5 commissioni che correggono solo la terza prova ed infine 2 commissioni che correggono la prima e la terza prova, come emerge dalla tabella 4..

	Prova scritta		
CODICE	prima	seconda	terza
COMLC	**x**	**x**	**x**
COMLS	**x**	**x**	**x**
COMTEC	**x**	**x**	**x**
COMRAG	**x**	**x**	**x**
COMPROF	**x**		**x**
COMPROF1	**x**		**x**
COMLC1		**x**	**x**
COMRAG1		**x**	**x**
COMLS1		**x**	**x**
COMTEC1		**x**	**x**
COMLC2			**x**
COMLS2			**x**
COMTEC2			**x**
COMRAG2			**x**
COMPROF2			**x**

Tab.4 Piano di attribuzione delle prove alle commissioni.

Va notato che, sempre per un criterio di minimizzazione dei costi, vi sono state due sole commissioni che correggono ripetutamente la prima e la seconda prova poiché quelle prove sono corrette anche in altri modi mentre per la terza vi è un solo modo, quello collegiale.

Estrazione degli elaborati da correggere

Gli elaborati usati nell'esperimento sono stati tratti da un campione casuale di 500 commissioni, rappresentativo dell'intera popolazione dei candidati, esaminati nella sessione 2000.

Da un campione casuale di 500 commissioni, durante lo svolgimento degli esami, sono state raccolte le tre prove scritte di tre studenti individudati attraverso l'estrazione casuale delle loro posizioni nella lista ufficiale.

Come abbiamo visto, la correzione ripetuta delle prove scritte ha riguardato tutti i tipi di prime prove, alcuni tipi di seconde prove, e alcune terze prove. Per ogni tipologia di prova sono state individuate dai 15 ai 20 elaborati, ciascuno dei quali è stato corretto ripetutamente e indipendentemente da correttori diversi.

Data la varietà degli indirizzi di studio e quindi delle seconde prove, si è partiti dall'esame del materiale raccolto nel campione e si è verificato che non vi era una quantità sufficiente di elaborati per ogni materia. Ciò ha condotto alla scelta delle discipline riportate in tabella per le quali era disponibile un numero sufficiente di elaborati.

Latino
Matematica
Ragioneria
Elettronica (Tecnici Industriali)

La tabella 5 riporta solo la numerosità delle discipline più rappresentate nel campione. Gli elaborati raccolti nel campione 2000 delle 500 commissioni sono stati a loro volta sorteggiati casualmente.

L'aver assunto nello studio circa 20 elaborati di seconde prove per ciascuna materia consente di poter costituire una ideale 'classe tipo' e un complesso di circa 200 dati da elaborare per ogni traccia.

Data la numerosità degli indirizzi dell'ordine professionale, non è stato possibile individuare una disciplina per la quale fosse disponibile un consistente numero di elaborati della seconda prova.

Indirizzo	commissioni	prove
Scientifico	89	178
Amministrativo	47	94
Classico	35	70
Magistrale	26	52
Elettr e Telecom.	19	38
Socio Psicopedag. (Pr. Brocca)	18	36
Amministrativo (Progetto Igea)	17	34
Geometri	14	28
Tecnico Serv. Turis. (Nuovo Ord.)	12	24
Tecnico Ser. Ristor. (Nuovo Ord.)	11	22
Linguistico Progetto Brocca	10	20
Tecn. Gest. Az. Info. (Nuovo Ord.)	10	20
Programmatori	9	18
P.N.I. Amministrativo	9	18
Tecnico Ind. El. (Nuovo Ord.)	7	14
Tecnico Ind. Meccan. (Nuovo Ord.)	7	14
Elettrot. Autom.	7	14

Tabella 5 Numero delle seconde prove maggiormente rappresentate nel campione 2000

Individuate le seconde prove, sono state prese per l'esperimento le prime e le terze prove degli studenti estratti con la seconda prova. Ciò ha consentito di avere come tipologia di correzione anche quella di commissioni che correggono le tre prove scritte dello stesso studente (situazione realistica) e di analizzare anche le relazioni esistenti tra gli esiti delle tre prove corrette indipendentemente da correttori isolati. Su questi casi è possibile anche una analisi delle intercorrelazioni tra gli esiti delle tre prove.

In aggiunta alle prime prove già individuate con l'appaiamento alla seconda prova, sono state estratte casualmente anche altre prime prove per poter avere almeno 15 elaborati per tipo ed almeno 5 per tipo e per livello scolastico. Per le professionali, che non hanno seconde prove estratte per la correzione, la scelta delle 20 terze prove estratte per completare il disegno complessivo è stata fatta associandole alle prime prove, in modo da consentire analisi di correlazione almeno sulla prima e terza prova di 20 studenti.

Organizzazione del lavoro

Gli elaborati sono stati fotocopiati, dopo un sistematico controllo che ha eliminato correzioni o valutazioni apposte sui fogli dalla commissione vera. Va detto che durante la raccolta delle prove era stato raccomandato alle commissioni di cancellare o nascondere segni o valutazioni eventualmente già riportate dai commissari d'esame, ma ciò non è sempre stato fatto completamente. Le fotocopie degli elaborati estratti sono state raccolte in fascicoli di 5 esemplari l'uno, opportunamente codificati secondo i piani di assegnazione ai correttori individuati dalle tabelle 2 e 3 e da altre analoghe che non sono qui riportate per ovvi motivi di spazio.

Le tabelle 2 e 3 individuano sommariamente anche le varie regioni geografiche in cui occorreva scegliere i docenti correttori. I docenti correttori sono stati individuati attraverso un campionamento casuale di scuole secondo le tipologie necessarie all'esperimento. Condizioni per l'inserimento nella lista dei docenti correttori sono state la disponibilità alla collaborazione e l'avvenuta partecipazione ad almeno una sessione di esami di Stato.

I correttori singoli sono stati raggiunti direttamente per posta, mentre la costituzione delle commissioni è stata proposta direttamente al dirigente scolastico dell'istituto sorteggiato. Nel caso delle commissioni, la proposta di collaborazione inviata al dirigente scolastico conteneva anche tre nominativi dei docenti, sempre casualmente scelti dall'Osservatorio, e si lasciava al dirigente la responsabilità della scelta degli altri tre commissari e del presidente. La modalità è descritta nell'allegato 2. Nei casi di indisponibilità o di rifiuto si procedeva ad un nuovo sorteggio di un docente avente le stesse caratteristiche del docente da sostituire. Da notare che su circa 160 docenti coinvolti nell'esperimento solo 15 hanno rinunciato a collaborare in una attività rischiosa: chi collaborava era l'oggetto dello studio ed accettava di essere messo in discussione, nonostante le ovvie assicurazioni circa l'anonimato degli esiti dello studio.

Molti docenti estratti hanno manifestato la loro sorpresa per essere stati scelti e per aver ricevuto una proposta di collaborazione da un istituto di ricerca poiché si sentivano del tutto fuori dai normali circuiti delle collaborazioni istituzionali. Ciò era il segno che l'intenzione di costituire un campione rappresentativo di docenti 'normali'; gli stessi che si trovano correntemente a correggere le prove durante gli esami di Stato aveva avuto successo. Anche per questo motivo i correttori non hanno ricevuto alcuna forma di istruzione o di addestramento; l'unico elemento per uniformare i comportamenti era costituito da un manuale di istruzioni (vedi allegato B) annesso ai fascicoli che descriveva le finalità della ricerca e, solo per coloro che la usavano, recava la griglia di correzione adottata.

Come si può intuire, la macchina organizzativa necessaria per una rapida ed efficiente distribuzione dei materiali, diretti ad una rete distribuita sul territorio e dispersa in unità singole, ha richiesto una progettazione piuttosto sofisticata, la cui affidabilità dipendeva tutta da un uso sistematico di efficienti data base (archivio dei docenti, archivio dei correttori, archivio delle prove, archivio dei contatti con i correttori e i dirigenti scolastici, archivio amministrativo per i compensi individuali).

La prima lettera inviata per il raggiungimento del campione dei correttori reca la data del 10 gennaio 2001 mentre la chiusura della raccolta è avvenuta alla metà di maggio dello stesso anno. Le correzioni sono quindi avvenute contemporaneamente, attraverso l'invio a ciascuno delle copie necessarie, ed in modo del tutto indipendente poiché nessun correttore individuale conosceva gli altri correttori. I tempi per la restituzione sono stati ovviamente condizionati dalla effettuazione di alcune sostituzioni e dalla necessità di attendere 15 giorni prima dell'invio del fascicolo per la seconda correzione.

Correzione differita

Per quanto riguarda la correzione differita va precisato che questa non era stata chiaramente annunciata nel primo invio. Solo alcuni avevano notato che rispetto alla lettera di incarico che prevedeva 20 correzioni erano state inviati solo 15 elaborati e ne mancavano quindi 5. Per telefono veniva detto che sarebbero stati inviati successivamente ma che nel frattempo dovevano rispedire tutto il materiale e le schede di valutazione compilate. Solo dopo il rinvio del primo lotto di elaborati veniva spedito il secondo con la spiegazione del significato della correzione ripetuta differita nel tempo.

Un correttore di prime prove, rinviando i dati della seconda correzione, ha confessato di aver trattenuto le prime valutazioni, contrariamente a quanto era stato richiesto, ma che durante la seconda correzione aveva evitato di consultarle per procedere in modo indipendente. Tale docente ha così potuto sperimentare l'esistenza della variazione di punteggio ipotizzata nella ricerca ed anche in che senso tali variazioni erano intervenute, confrontando i giudizi analitici che aveva per se stesso redatto a conforto dei punteggi assegnati. Tale esperienza, a detta del correttore, era risultata tutt'altro che frustante ed anzi aveva significativamente contribuito a migliorare la consapevolezza dei criteri impliciti da lui usati nella correzione dei temi.

La raccolta e la registrazione dei dati

Le schede compilate dai correttori, recanti sullo stesso *record* il codice del correttore, il codice della prova e il punteggio assegnato, sono state registrate in un data base in cui le prime due informazioni erano già state preimpostate durante la costruzione del campione ed utilizzate nella distribuzione dei fascicoli delle prove. Ciò ha consentito di evitare errori di imputazione nei due codici che contenevano tutte le informazioni di sfondo utili all'elaborazione. In questo modo, la registrazione dei punteggi assegnati, oltre ad essere più affidabile, è servita ad effettuare una 'quadratura' sistematica di tutto il complicato sistema di fascicoli e schede spedite e ritirate dai numerosi docenti che hanno collaborato.

La prima fase del trattamento dei dati è servita a ristrutturare le informazioni disponibili: il file dei dati inizialmente costituito da un record per ogni correzione è stato riscritto in modo che le correzioni ripetute dello stesso elaborato fossero disposte sullo stesso record, perché l'unità di analisi doveva essere il singolo elaborato; analoghe

riscritture dei file sono state necessarie per effettuare analisi rispetto alla tipologia della prova o al correttore o allo stesso studente.

Prime rappresentazioni dei dati

Per dare una prima idea, facilmente comprensibile ma abbastanza efficace, della situazione presentiamo due grafici che riproducono le distribuzioni dei punteggi assegnati per ogni prova. Grafici sono realizzati con un normale foglio elettronico e la figura consente di percepire direttamente la struttura dei dati di cui stiamo parlando. Nelle fig. 9 e 10 nella prima colonna A appaiono i codici delle prove. Ad esempio nella riga 64 appare il valore L-B4-08201 che identifica una prova di italiano (traccia B4) del Liceo dello studente 01 della scuola 082 del campione. Per ogni prova, sulla stessa riga, appaiono le frequenze dei punteggi assegnati: in questo caso 5 correttori hanno assegnato 11 punti, 1 correttore ha assegnato 12, 4 correttori hanno assegnato 13. La situazione di ogni elaborato è anche illustrata graficamente per marcare visivamente se e come i giudizi si siano concentrati o dispersi lungo la scala dei punteggi assegnabili.

Abbiamo così una prima facile conferma di quanto ipotizzato nello studio ed, anche, alcuni indizi di altrettanti problemi su cui riflettere ed indagare.

Innanzitutto possiamo capire che la situazione illustrata nella figura 9 è migliore della successiva: nella prima tutti i correttori convergono sulla sufficienza piena e differiscono di poco tra loro mentre nella seconda la divergenza è più sostanziale e oppone due gruppi quasi equivalenti: 6 per la sufficienza piena e 4 per una insufficienza grave. Osservando altre situazioni, ad esempio quella della riga 70, possono emergere altri problemi e cioè l'esistenza di singoli valori completamente staccati dal resto dei dati, come se un correttore si differenziasse significativamente dal resto degli altri correttori. Ricordiamo ancora che i correttori hanno lavorato isolatamente e non hanno avuto modo di interagire con gli altri del gruppo poiché non disponevano della lista dei nominativi.

Ci possiamo ora chiedere quale sia il punteggio da assegnare correttamente a ciascuna prova. Quale correttore ha ragione ed ha individuato il punteggio vero? Assumendo la tecnica di assegnazione del punteggio prevista dagli esami di Stato, il *voto* che dovrebbe essere assegnato è quello *votato* dalla maggioranza (solo la prova della riga 68 ha almeno sei concordanze sul punteggio 10) oppure è la media aritmetica di tutte le proposte formulate (12 nel caso illustrato nella fig. 9 e 8 nel secondo caso. Tra le 20 correzioni esaminate dei due esempi, solo una è 'corretta' in quanto ha individuato il voto che alla fine sarà assegnato.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1																
2	PUNTEGGIO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
57	L-B2-25303						3	2	2			1				
58	L-B3-12803								1				4	1	1	3
59	L-B3-21401						1	3			4		2			
60	L-B3-22702							1	1		3	3	1			
61	L-B3-23401						2		1	3	1			2	1	
62	L-B3-33403						1	1	2	1	4			1		
63	L-B4-06803								3		4	2		1		
64	L-B4-08201											5	1	4		
65	L-B4-12101												3	5	2	
66	L-B4-19503									1	1	1	3		2	2
67	L-B4-33401						1		3	3	3					
68	L-C0-20802							1		1	6	2				
69	L-C0-22103			1		2	1				5		1			
70	L-C0-23103				1				1	1	3	2	1	1		
71	L-C0-24603							1		2	1		4		2	

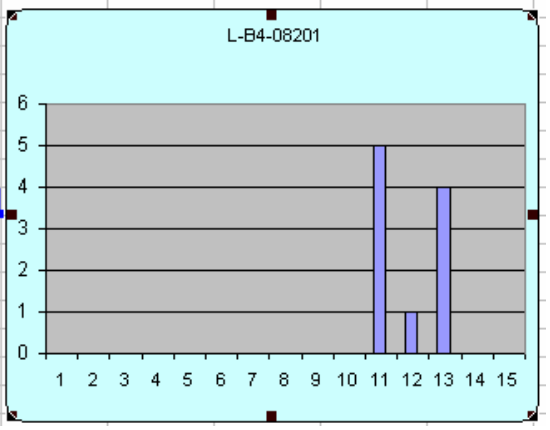


Fig.9 Distribuzione dei punteggi assegnati ad alcune prove

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1																
2	PUNTEGGIO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
57	L-B2-25303						3	2	2			1				
58	L-B3-12803								1				4	1	1	3
59	L-B3-21401						1	3			4		2			
60	L-B3-22702							1	1		3	3	1			
61	L-B3-23401						2		1	3	1			2	1	
62	L-B3-33403						1	1	2	1	4			1		
63	L-B4-06803								3		4	2		1		
64	L-B4-08201											5	1	4		
65	L-B4-12101												3	5	2	
66	L-B4-19503									1	1	1	3		2	2
67	L-B4-33401						1		3	3	3					
68	L-C0-20802							1		1	6	2				
69	L-C0-22103			1		2	1				5		1			
70	L-C0-23103				1				1	1	3	2	1	1		
71	L-C0-24603							1		2	1		4		2	

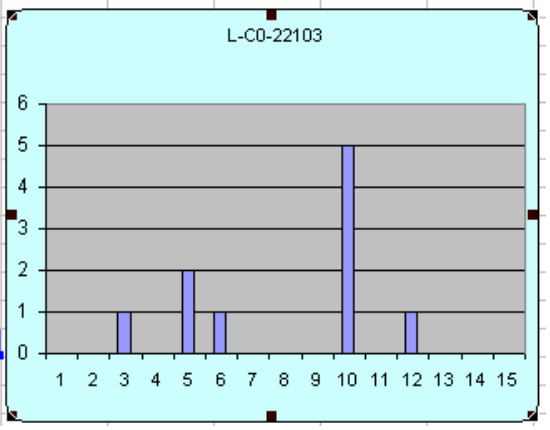


Fig.10 Distribuzione dei punteggi assegnati ad alcune prove

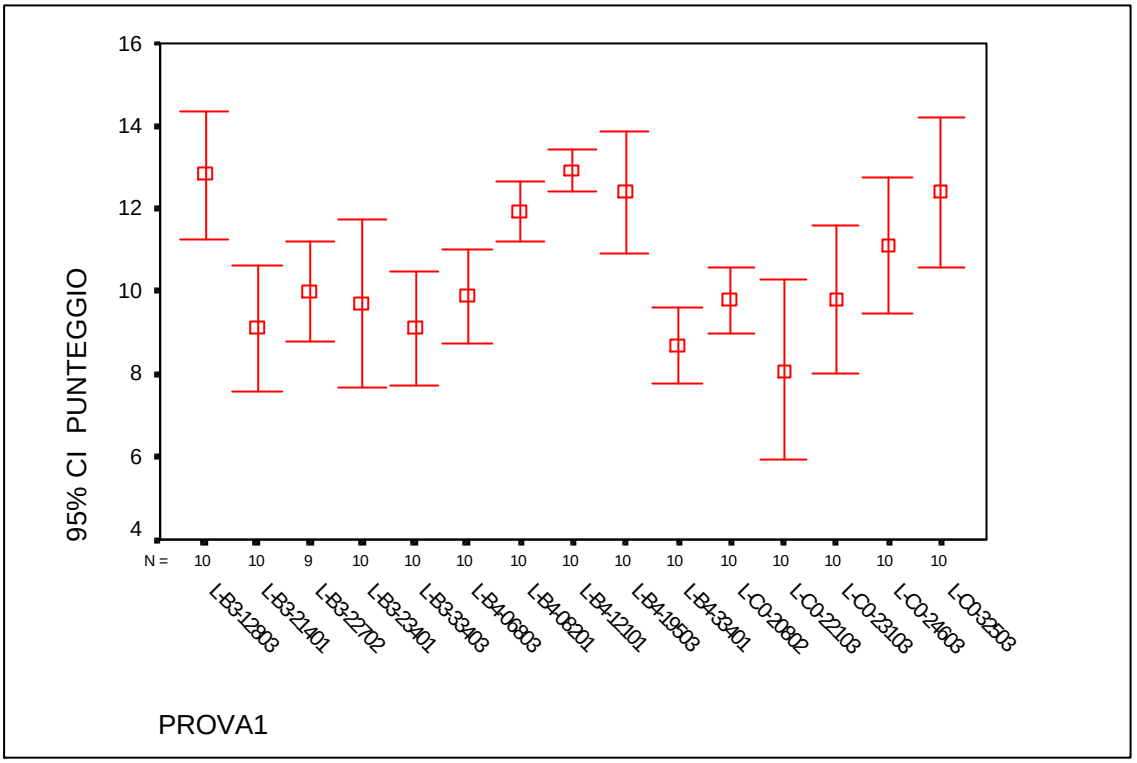


Fig.11 Medie ed intervalli di confidenze dei punteggi assegnati a singoli elaborati

Se però torniamo a considerare questi dati come delle ‘misure’ di un valore che vogliamo correttamente stimare, nemmeno la media aritmetica dei punteggi espressi è il valore ‘vero’ ma solo una stima puntuale di un valore ‘vero’ che con molta probabilità dovrebbe trovarsi in un intorno relativamente piccolo della media aritmetica.

Il grafico della fig. 11 mostra per le stesse prove della figura precedente a quali conclusioni potremmo ragionevolmente arrivare se trattassimo i punteggi come delle misure affette solo da errori casuali, ovvero come campioni casuali dell’insieme delle infinite correzioni, che sono teoricamente possibili, dello stesso elaborato. In ascissa sono riportate le prove e il numero di correzioni della stessa prova, mentre in ordinata sono indicati un punto ed un intervallo di valori. Per ogni prova è identificato un intervallo di confidenza al 95% ovvero l’intervallo in cui, con una probabilità del 95%, dovrebbe trovarsi questo ‘misterioso’ valore vero che con dieci misure abbiamo cercato di individuare. Come è facile osservare, l’ampiezza degli intervalli è molto varia: per alcune prove l’incertezza si restringe intorno a pochi valori interi (L-B4-08201 al 95% dovrebbe avere un valore che si trova tra 11,29 e 12,51 mentre L-C0-22103 dovrebbe avere una valore compreso tra 6,23 e 9,97). Vale la pena di ricordare che il *valore* dell’elaborato, e cioè il punteggio correttamente assegnabile a ciascun elaborato, è una grandezza continua che, seppur in via del tutto teorica, potrebbe essere stimata con una precisione grande quanto si vuole.

Come si può osservare dal grafico di fig. 11 le prove sono tutte del liceo e corrispondono a 3 tracce la B3, la B4 e la C0. Ciascuna traccia è raggruppata nello

stesso fascicolo che è stato corretto dalla stesso gruppo di dieci correttori. I dieci correttori che hanno corretto le prove C0 sembrano più imprecisi dei dieci che hanno corretto la prova B4 poiché gli intervalli sono più ampi, ma potremmo anche supporre che l'accordo dei correttori possa dipendere dalla caratteristica della prova (il tipo di traccia) o dal particolare elaborato da valutare (sui casi eccellenti è più facile convergere mentre ci sarebbe maggior dispersione nei punteggi degli elaborati di valore mediano).

D'altra parte le distribuzioni delle figure 9 e 10 ci mostrano anche l'esistenza di valori anomali, valori che da soli si discostano eccessivamente dal resto dei punteggi. Se questi valori fossero eliminati potremmo ridurre l'ampiezza dell'intervallo fiduciario in cui si trova quasi certamente il valore vero.

Abbiamo sin qui limitato le nostre osservazioni all'esemplificazione di pochi casi. Nelle tabelle seguenti sono complessivamente illustrati tutti i dati relativi agli elaborati corretti nell'esperimento attraverso dei grafici a scatola. Per ogni elaborato, di cui sulle ascisse è riportato il codice, viene rappresentata la distribuzione dei punteggi assegnati. I bordi superiore e inferiore della scatola rappresentano i quartili superiori ed inferiori e quindi contengono il 50% dei punteggi centrali assegnati. La linea all'interno della scatola identifica la mediana del gruppo. Più lunga è la scatola, più grande è la variabilità dei punteggi assegnati dai correttori. Le linee che partono da ciascuna scatola si estendono fino ai punteggi più piccoli e più grandi di uno stesso elaborato e che sono distanti meno di un intervallo interquartile dagli estremi della scatola. I punti al di fuori di questo intervallo, ma con una distanza inferiore a 1.5 volte quella interquartilica dal bordo della scatola, sono stati etichettati dalla procedura di analisi come anomali (O); i punti con distanza superiore a 1.5 volte la distanza interquartilica dal bordo della scatola sono stati etichettati come estremi (E).

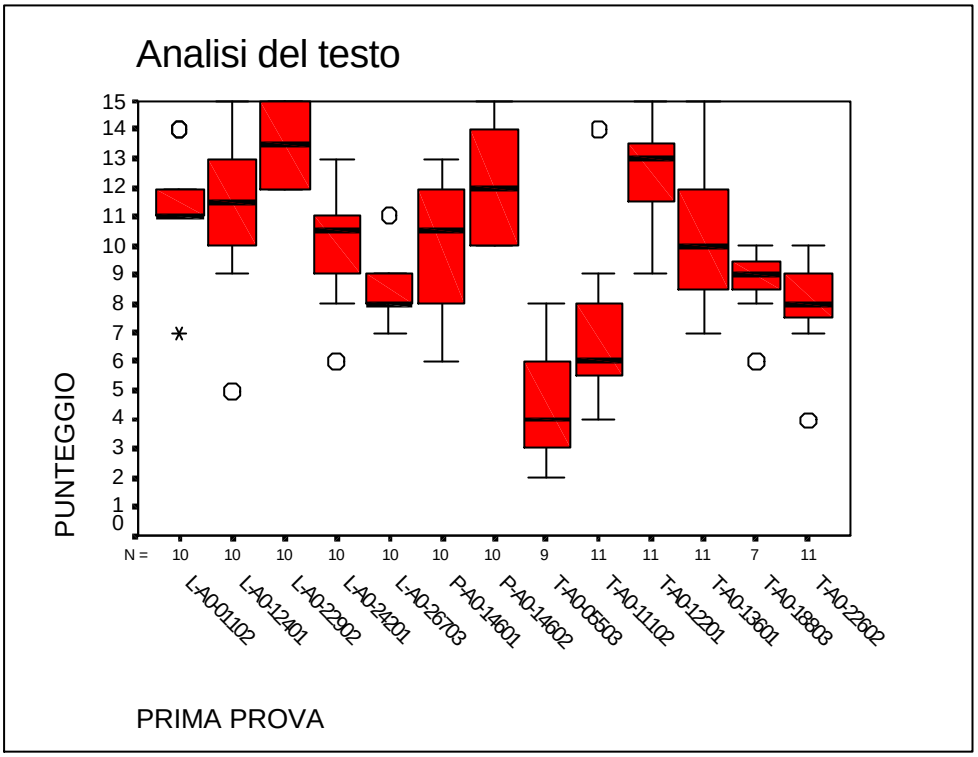


Fig. 12 Grafico a scatola dei punteggi della prima traccia del tema

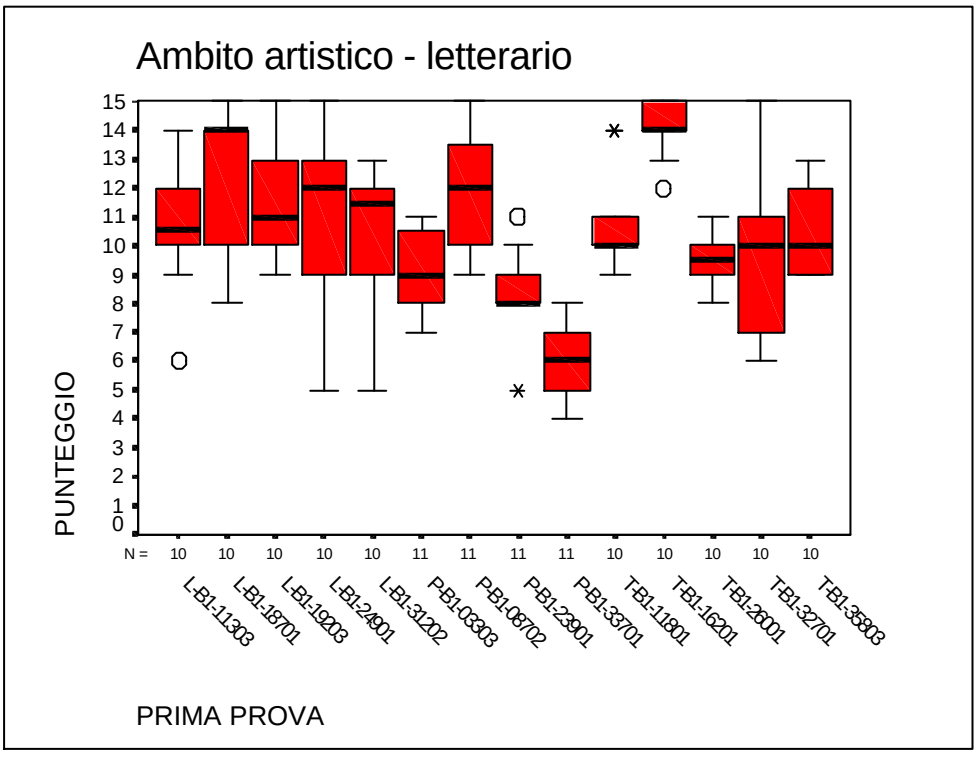


Fig. 13 Grafico a scatola dei punteggi della seconda traccia del tema

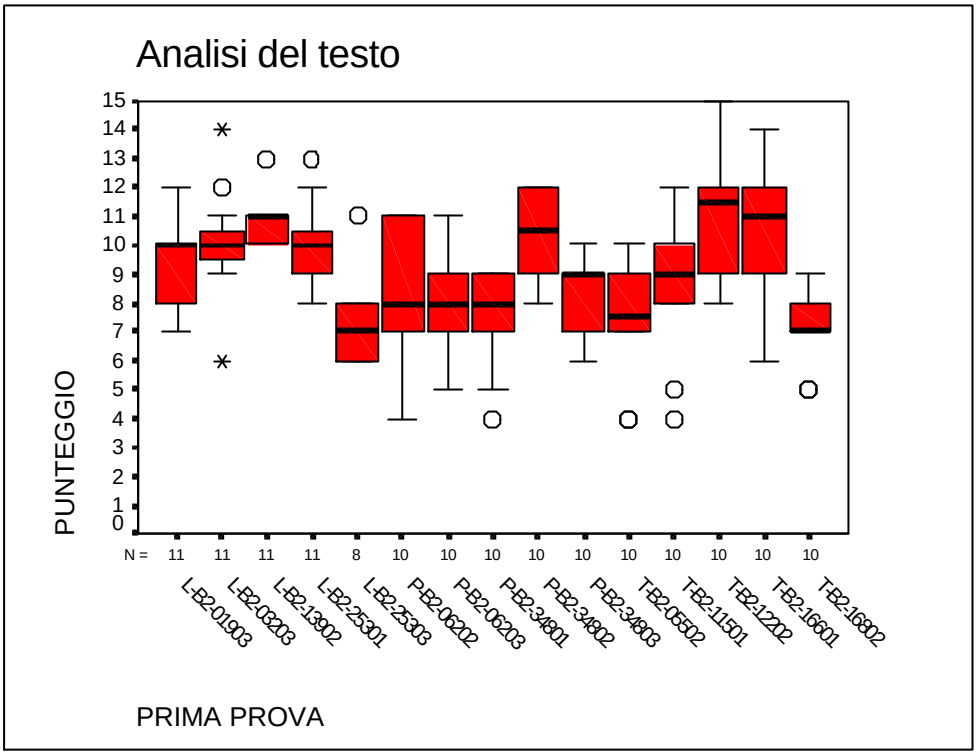


Fig. 14 Grafico a scatola dei punteggi della terza traccia del tema

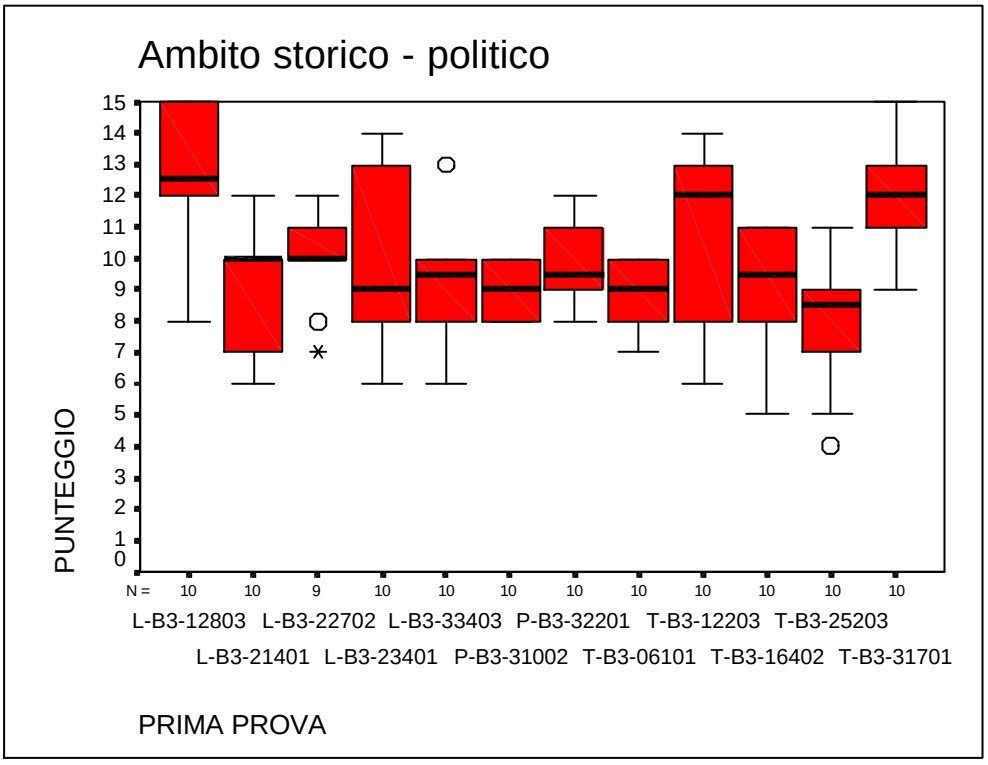


Fig. 15 Grafico a scatola dei punteggi della quarta traccia del tema

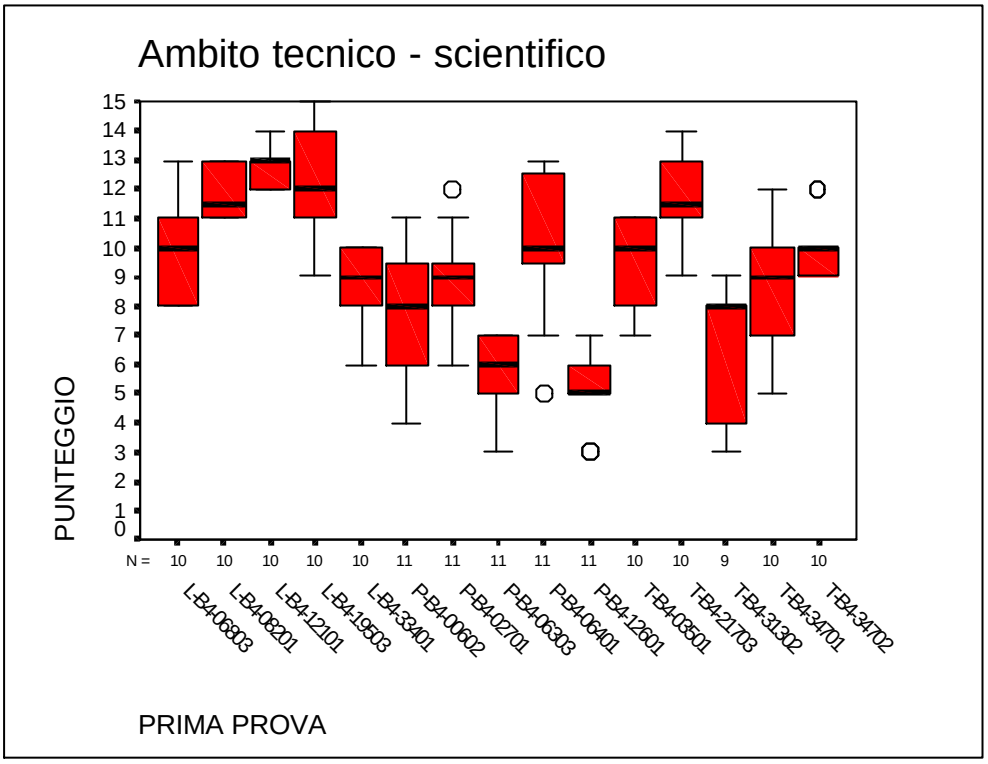


Fig. 16 Grafico a scatola dei punteggi della quinta traccia del tema

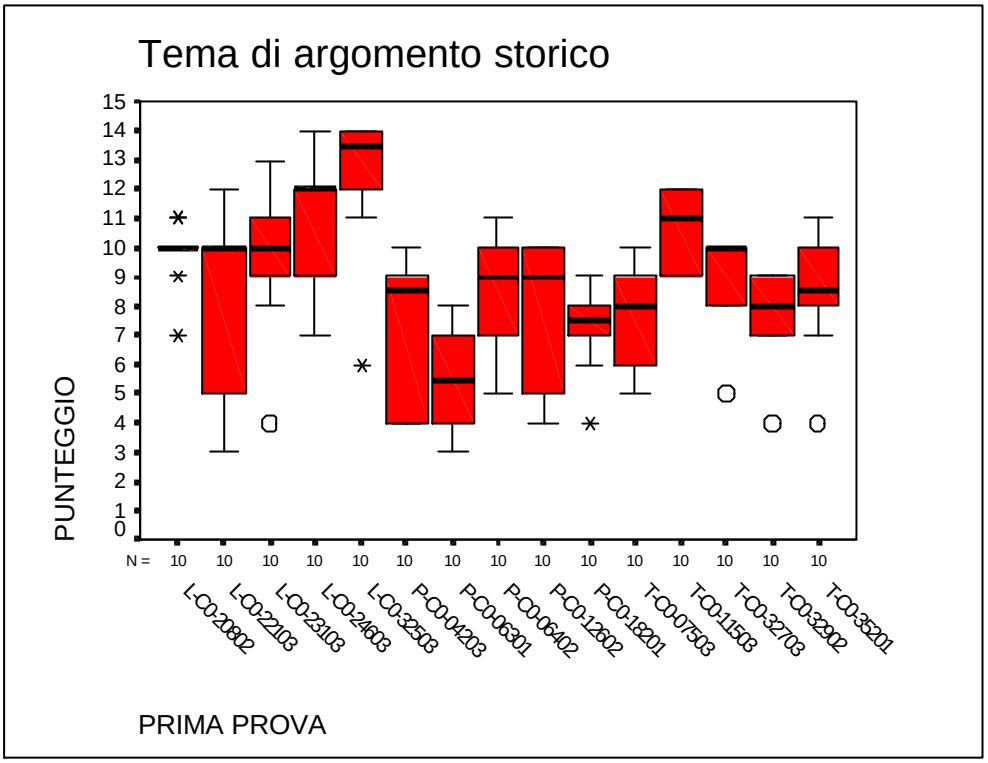


Fig. 17 Grafico a scatola dei punteggi della sesta traccia del tema

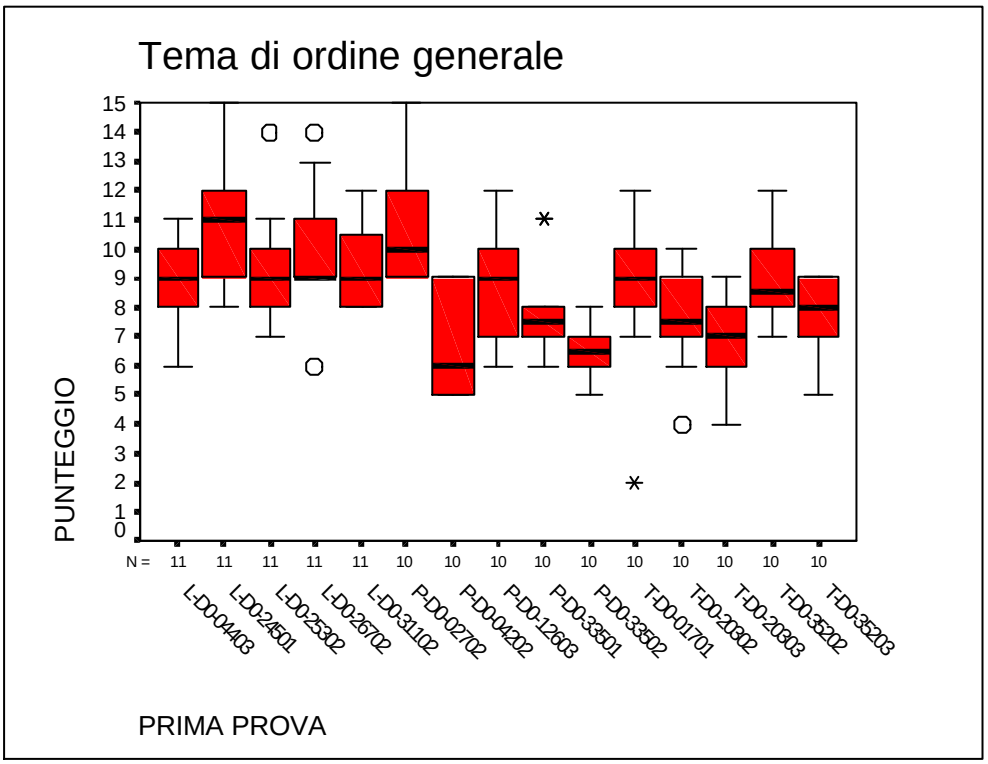


Fig. 18 Grafico a scatola dei punteggi della settima traccia del tema

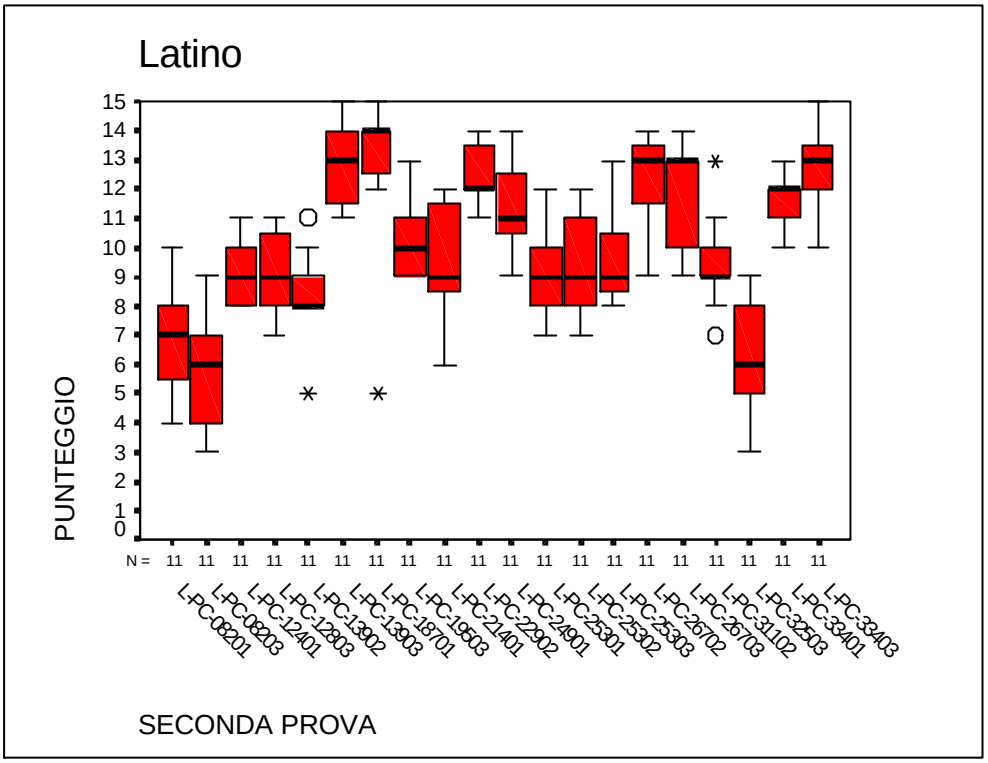


Fig. 19 Grafico a scatola dei punteggi della seconda prova di latino

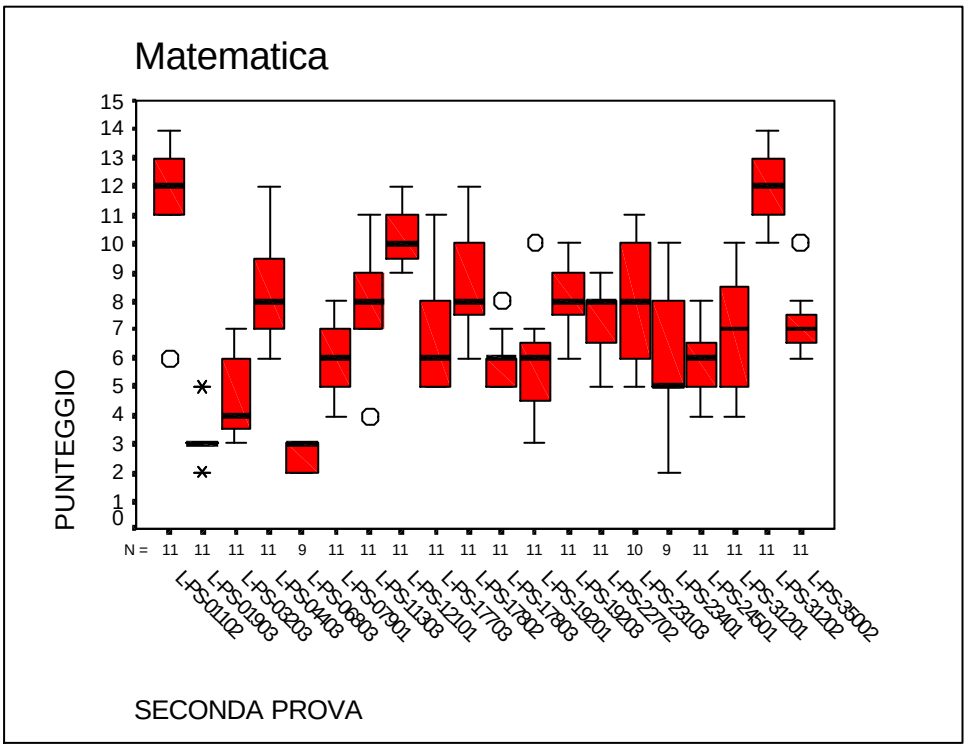


Fig. 20 Grafico a scatola dei punteggi della seconda prova di matematica

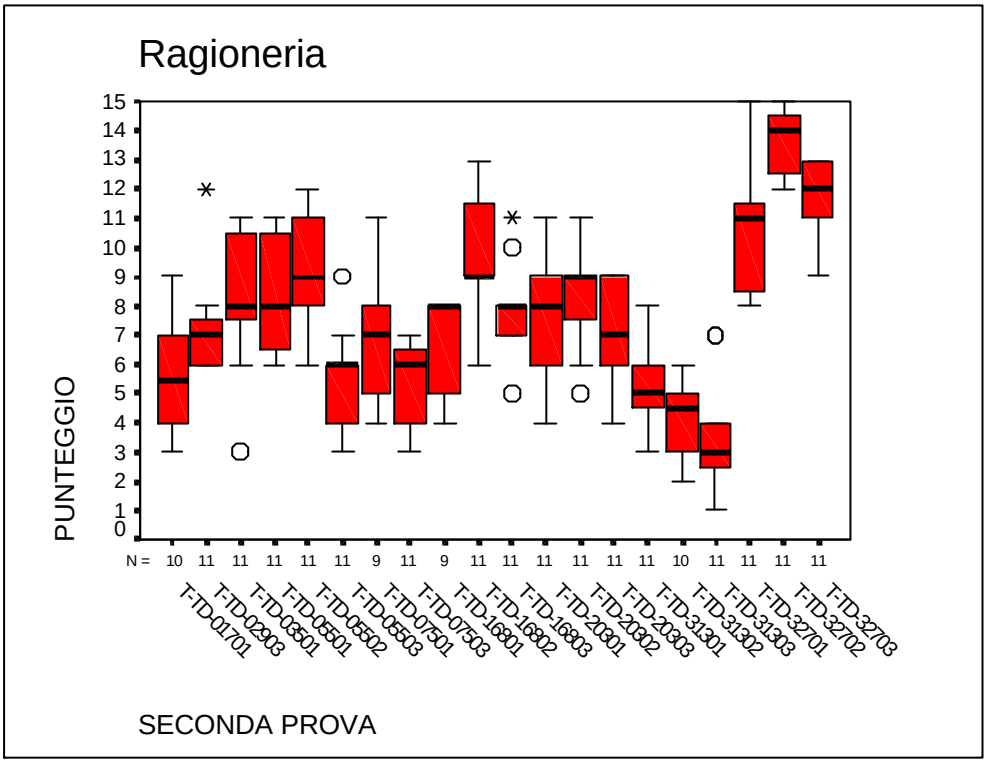


Fig. 21 Grafico a scatola dei punteggi della seconda prova di ragioneria

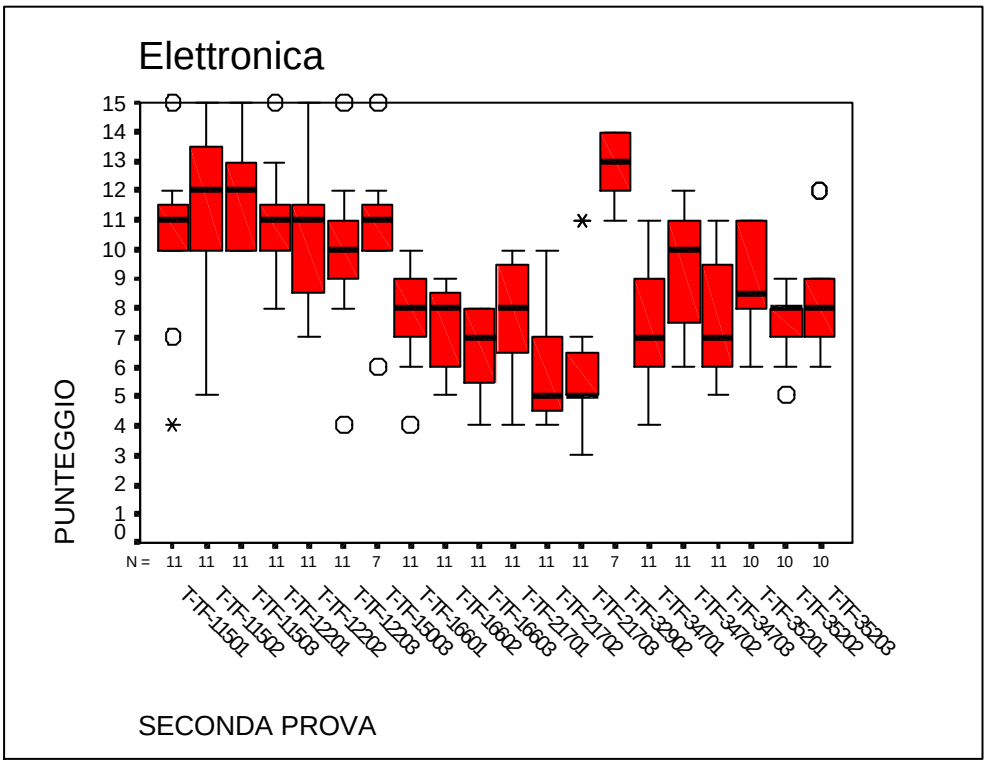


Fig. 22 Grafico a scatola dei punteggi della seconda prova di elettronica

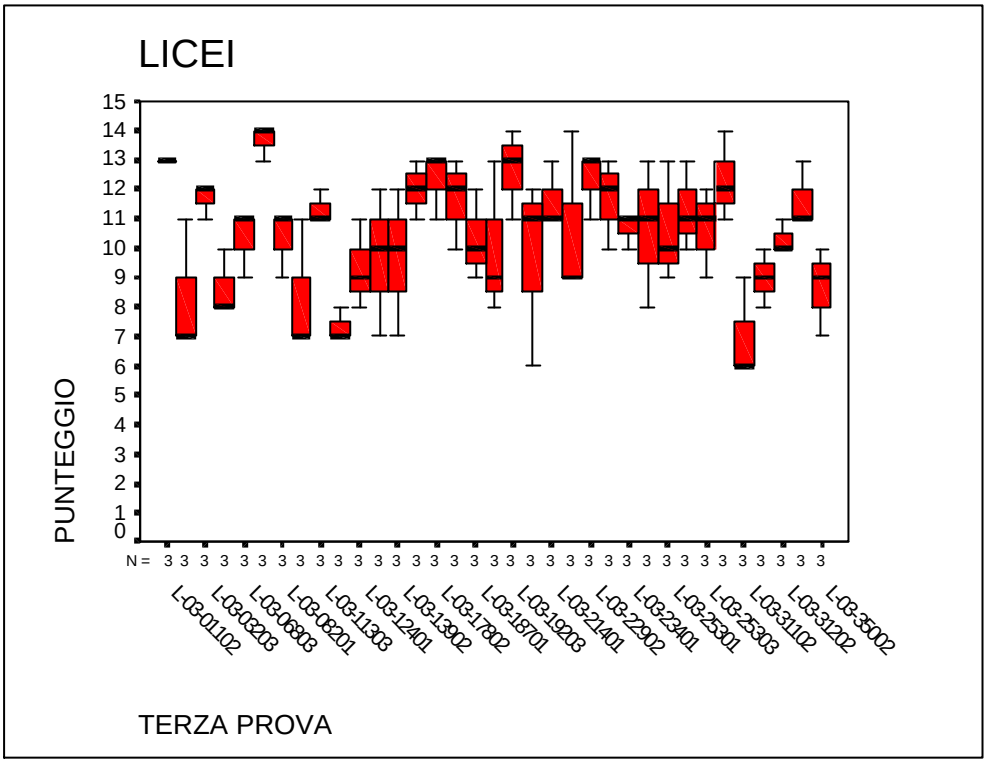


Fig. 23 Grafico a scatola dei punteggi della terza prova dei licei

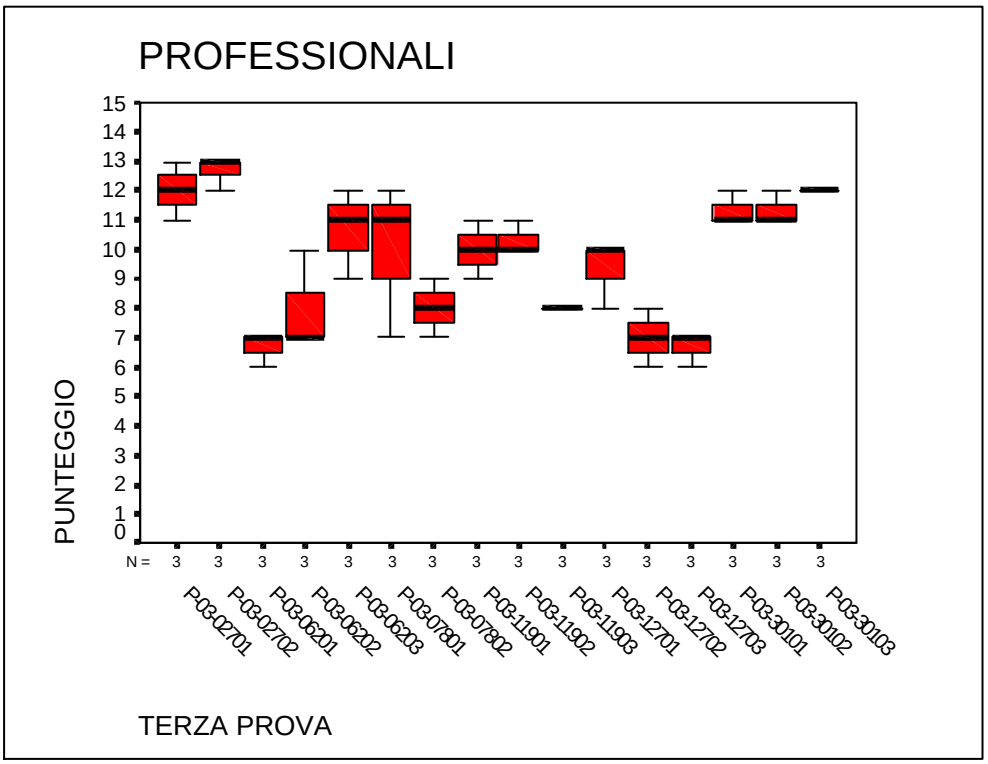


Fig. 24 Grafico a scatola dei punteggi della terza prova dei professionali

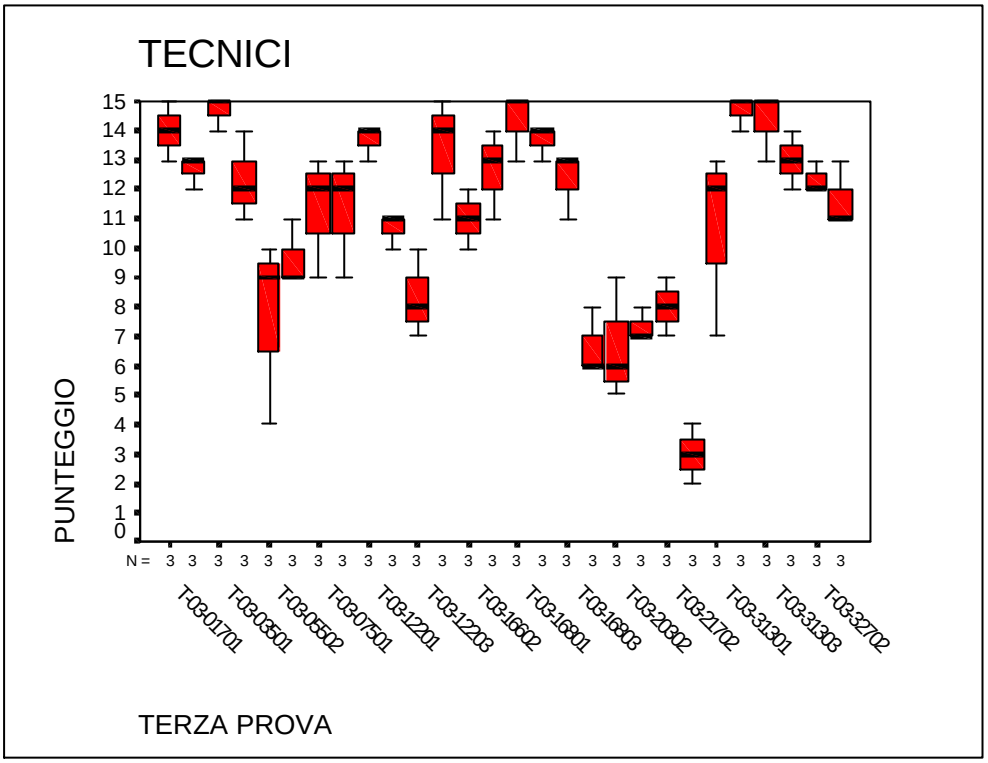


Fig. 25 Grafico a scatola dei punteggi della terza prova dei tecnici

Il calcolo del '*valore vero*'

I grafici a scatola evidenziano la presenza di valori anomali o di valori estremi che, discostandosi troppo dal gruppo degli altri, potrebbero pesare eccessivamente sul valore della media aritmetica rendendo meno precisa la stima del valore vero. Per questo motivo è stato assunto come stima puntuale del valore 'vero' la media aritmetica di tutti i punteggi dello stesso elaborato depurati da due punteggi estremi, un minimo e un massimo. Rispetto a tale valore decimale, assunto **convenzionalmente** come 'voto vero', sono stati calcolati gli errori di ciascuna misura.

La fig. 26 riporta l'istogramma della variabile "errore di misura" e consente di constatare che tale distribuzione riflette le caratteristiche tipiche degli errori casuali di misura e cioè si dispone normalmente seppure con una leggera asimmetria dovuta ad una maggiore frequenza degli scarti positivi rispetto ai corrispondenti scarti negativi poiché il calcolo del valore vero rispetto a cui sono calcolati gli errori ha escluso dal computo i valori estremi, spesso disposti asimmetricamente rispetto al resto dei dati.

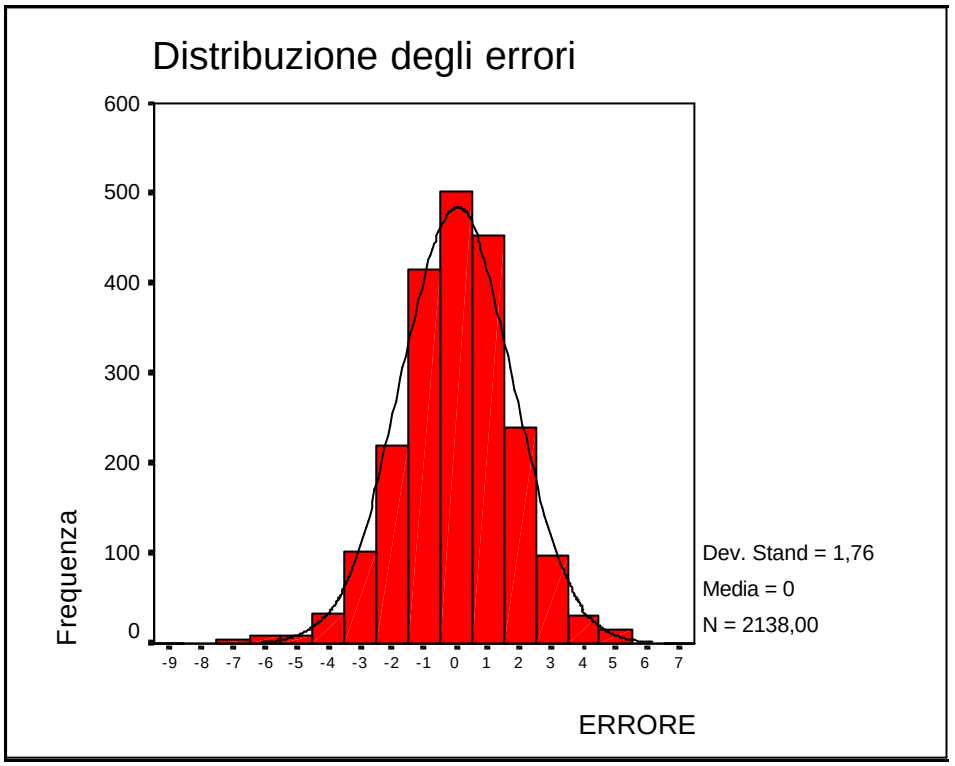


Fig. 26 Istogramma degli errori di misura

Analoga situazione si riscontra se gli errori vengono calcolati calcolando le differenze tra i punteggi assegnati e il valore intero più vicino al valore "vero": in effetti questa è la situazione più realistica in quanto i punteggi utilizzati nella valutazione degli esami di Stato sono solo numeri interi. Trattando il *valore vero intero* è anche possibile contare i casi in cui la determinazione del punteggio è stata esatta: solo il 22,4 % dei punteggi sono 'esatti' il 40% si discosta di un punto, il 20% di 2 punti.

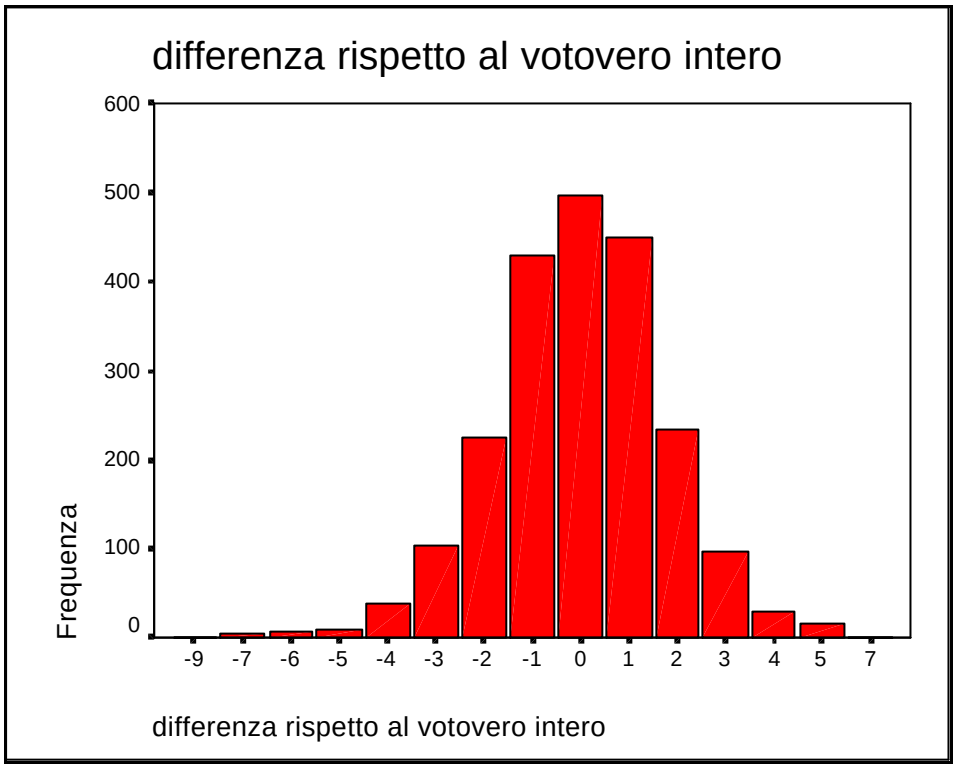


Fig. 27 Distribuzione degli errori calcolati rispetto all'approssimazione intera del 'votovero'

Entrambe le rappresentazioni di figura 26 e 27 illustrano efficacemente l'intensità del fenomeno che stiamo studiando, ma un altro modo per apprezzare le implicazioni pratiche di tale situazione consiste nel calcolare l'ampiezza della gamma dei punteggi espressi per ciascun elaborato. Come si può facilmente osservare dalla tabella 28, nel 90% dei casi la differenza tra il punteggio massimo e quello minimo è maggiore o uguale a quattro punti, nel 30% dei casi tale gamma è maggiore di 6 punti. In media la gamma è di 5,65 punti, che rappresenta un terzo della variabilità totale dell'intera scala del punteggio in quindicesimi.

Gamma dei punteggi assegnati

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	1	1	,6	,6	,6
	2	3	1,7	1,7	2,2
	3	16	8,9	8,9	11,2
	4	34	19,0	19,0	30,2
	5	30	16,8	16,8	46,9
	6	44	24,6	24,6	71,5
	7	23	12,8	12,8	84,4
	8	16	8,9	8,9	93,3
	9	4	2,2	2,2	95,5
	10	6	3,4	3,4	98,9
	11	2	1,1	1,1	100,0
	Totale	179	100,0	100,0	

Tab. 28 Distribuzione della gamma dei punteggi assegnati in ogni elaborato

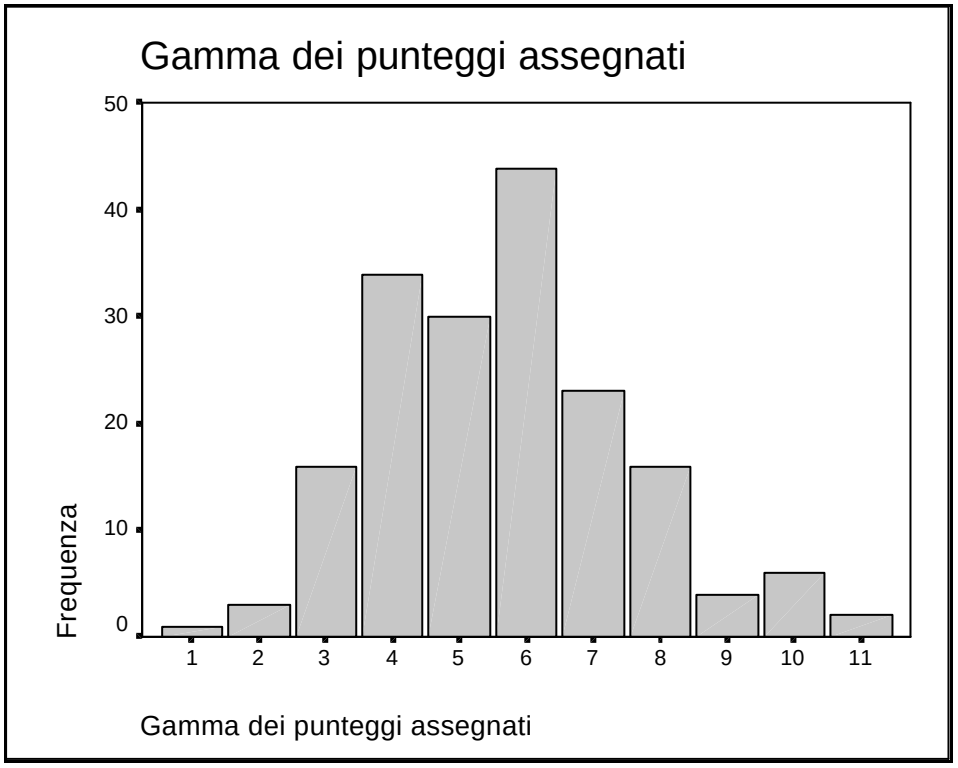


Fig. 29 Grafico della gamma dei punteggi assegnati in ogni elaborato

Per poter ulteriormente visualizzare la natura dei dati raccolti possiamo considerare l'errore assoluto che ci consente di calcolare la media dello scostamento rispetto al valore vero di ciascun punteggio Nell'istogramma della fig. 30 è visibile l'intensità media che ammonta, nel caso della prima e seconda prova, a 1,42 punti mentre per la terza prova l'istogramma successivo, pur presentando una maggiore irregolarità, dovuta al minor numero di casi analizzati, presenta un valore medio pari a 0,80 punti.

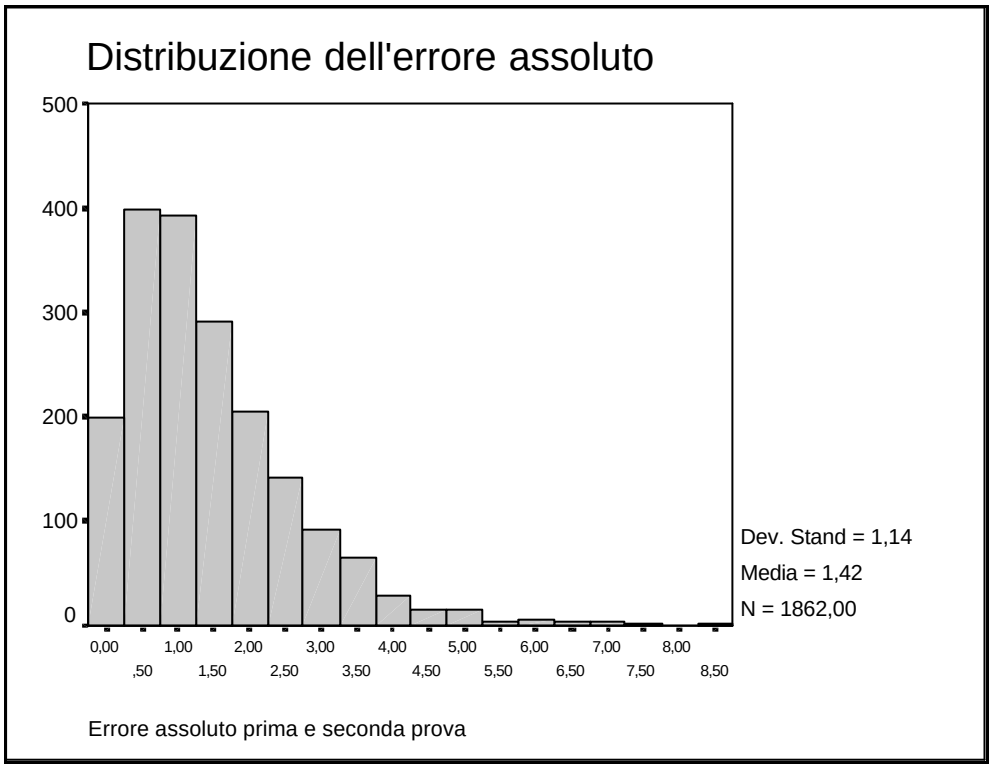


Fig. 30 Istogramma dell'errore assoluto per elaborati di prima e seconda prova

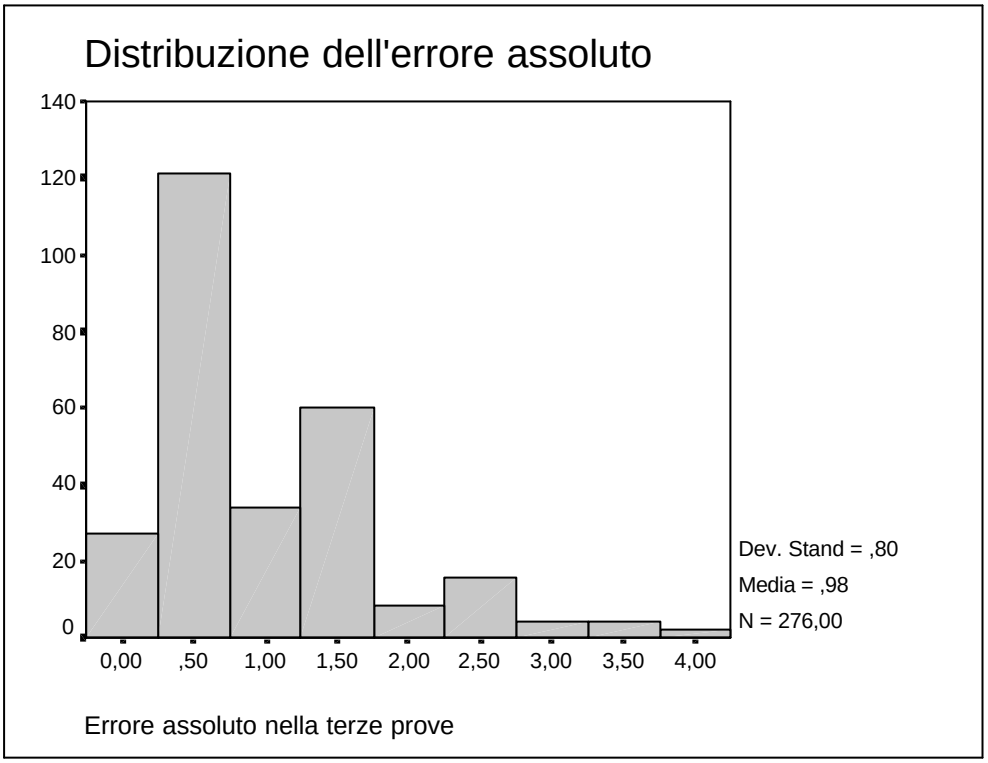


Fig. 31 Istogramma dell'errore assoluto per elaborati di terza prova

Quanto emerge dai grafici è una informazione di particolare interesse poiché marca una prima fondamentale differenza tra la terza prova e le altre e conferma quanto già emerso nei grafici a scatola secondo cui la correzione della terza prova risulta più precisa, anche se effettuata da correttori che non l'hanno pianificata.

Precisione delle correzioni rispetto al tipo di prova

Ci possiamo quindi chiedere se la precisione dipenda dal tipo di prova o addirittura dalla traccia.

Come si può facilmente verificare dalla tabella 32 la precisione dei punteggi della terza prova è più alta anche rispetto alle varie tracce della prima e seconda prova e ciò risulta vero anche se il confronto riguarda solo i punteggi formulati dalle commissioni (v. tab. 33). Le tabelle 32 e 33 riportano anche gli errori relativi cioè il rapporto tra l'errore assoluto e il valore "vero". Le due serie di valori medi dell'errore assoluto e dell'errore relativo, non hanno esattamente lo stesso andamento: ad esempio la correzione della matematica risulta più 'precisa' della correzione del latino se raffrontiamo le medie dei valori assoluti mentre si invertono le cose se confrontiamo gli errori relativi. Ciò può accadere se i punteggi assegnati in latino sono mediamente più grandi dei punteggi assegnati in matematica.

La tabella 32 conferma comunque che le correzioni delle prove di italiano sono più imprecise delle correzioni delle seconde prove che riguardano elaborati più prevedibili e

più facilmente classificabili. In effetti si nota una differenza che però è inferiore a quanto ci si poteva attendere sulla base dei pregiudizi più diffusi: anche le seconde prove si prestano a correzioni imprecise, quasi come accade per i temi di italiano.

Errori rispetto al tipo di prova

	Errore assoluto		Errore relativo
	N	Media	Media
Terza prova	276	,98	,11
Analisi del testo	130	1,51	,17
Ambito artistico - letterario	144	1,44	,14
Ambito socio - economico	152	1,38	,16
Ambito storico - politico	119	1,53	,16
Ambito tecnico - scientifico	154	1,34	,16
Tema di argomento storico	150	1,52	,18
Tema di ordine generale	155	1,34	,16
Latino	220	1,31	,14
Matematica	215	1,22	,18
Ragioneria	214	1,45	,21
Elettronica	209	1,62	,19

Tab. 32 Errori assoluti e relativi rispetto al tipo di prova

Nel caso dell'elettronica va detto comunque che alcuni correttori avevano lamentato la difficoltà di valutare alcuni elaborati a causa della cattiva qualità delle copie disponibili: la lettura di tutti gli elementi di cui era composto l'elaborato non era agevole e ciò può aver influito sulla maggiore variabilità dei punteggi assegnati allo stesso elaborato da correttori diversi.

Va inoltre osservato che tra le tracce della prima prova quella che richiede l'analisi del testo e il saggio breve di ambito artistico letterario presenta errori assoluti e relativi più alti delle altre tracce mentre risulta più precisa la correzione del tema di ordine generale: in base a ciò si potrebbe avanzare l'ipotesi che sulle tracce di tema più consolidate ci sia maggiore precisione rispetto alle tracce e alle forme espositive più innovative in cui manca una diffusa pratica valutativa.

Errori rispetto al tipo di prova (solo commissioni)

	Errore assoluto	Errore relativo
	N	Media	Media
Terza prova	276	,98	,11
Analisi del testo	24	1,77	,19
Ambito artistico - letterario	28	1,82	,17
Ambito socio - economico	29	1,48	,16
Ambito storico - politico	24	1,42	,15
Ambito tecnico - scientifico	29	1,40	,16
Tema di argomento storico	30	1,45	,17
Tema di ordine generale	30	1,35	,17
Latino	40	1,24	,13
Matematica	40	1,50	,22
Ragioneria	40	1,74	,25
Elettronica	35	2,27	,25

Tab. 33 Errori assoluti e relativi rispetto al tipo di prova (solo commissioni)

Precisione delle correzioni rispetto al tipo di correttore

Ci chiediamo ora se la modalità di correzione possa aver influito sulla precisione. Confrontiamo le medie degli errori assoluti e relativi calcolate rispetto alle cinque tipologie di correttore: dalla tabella 34 risulta che la modalità più precisa è quella del correttore singolo mentre quella più imprecisa è realizzata mediante la griglia. Sembrerebbe quindi che nell'assegnazione di punteggi a saggi complessi l'approccio globale, immediato, del correttore singolo, che può operare senza tener conto di altri vincoli esterni (commissioni o griglie proposte dall'esterno e non sufficientemente interiorizzate), sia quella più precisa. Ovviamente i singoli correttori utilizzati nell'esperimento erano liberi di assumere durante la correzione tutte le procedure a cui erano normalmente abituati, ivi compreso adottare proprie griglie di valutazione.

Errori dei punteggi rispetto alla modalità di correzione

	Errore assoluto		Errore relativo
	N	Media	Media
commissione	194	1,52	,17
coppia	196	1,36	,15
griglia	194	1,85	,21
singolo	395	1,21	,14
decimi	25	1,47	,18

Tab. 34 Errori assoluti e relativi rispetto al tipo di correttore

Proseguendo nella riflessione sui fattori che possono influire sulla precisione dei punteggi, possiamo confrontare gli errori medi dei correttori suddivisi secondo lo strato geografico di appartenenza. Emerge che i correttori del centro sarebbero i più precisi, seguiti da quelli del sud e infine da quelli del nord. In questo caso, anche gli errori relativi ci forniscono analoghe indicazioni ma sottolineano il fatto che le differenze non sono troppo vistose. Ovviamente il campione di correttori è troppo esiguo per poter generalizzare questo risultato: la tabella 35 e le considerazioni che ne sono derivate hanno però il valore di un indizio interessante di una eventuale differenziazione territoriale delle pratiche valutative dei commissari anche per effetto di una evidente localizzazione di strumentazioni specifiche per la correzione delle prove complesse.

Errori dei punteggi rispetto allo strato del correttore

		Errore assoluto		Errore relativo	Punteggio assegnato
		N	Media	Media	Media
strato del correttore	centro	636	1,28	,16	9,14
	nord	763	1,42	,17	9,23
	sud	739	1,37	,16	9,08

Tab. 35 Errori assoluti e relativi rispetto allo strato del correttore

Altre caratteristiche dei correttori potrebbero influire sulla precisione dei punteggi: in base alla tabelle 36 e 37 risulta che i correttori maschi sono stati più precisi della femmine ed anche leggermente più severi sia se si confrontano i soli punteggi della prima prova sia se considerano tutti i punteggi della prima e della seconda prova. Poiché le prove da correggere sono state casualmente assegnate ai correttori si può supporre che la loro qualità media sia stata equamente distribuita tra i tre gruppi (maschi,

femmine e commissioni). Le ultime colonne delle tabelle 36 e 37 possono essere lette come indici della diversa severità con cui sono state giudicate le prove: in questo caso non c'è soltanto un effetto casuale degli errori di misura ma anche un lieve effetto sistematico legato al genere dei correttori.

Errori rispetto al genere dei correttori

		Errore assoluto	Errore relativo	Punteggio assegnato
	N	Media	Media	Media
Commissioni	194	1,52	,17	9,11
Femmine	418	1,49	,17	9,68
Maschi	392	1,32	,15	9,17

solo punteggi della prima prova

Tab. 36 Errori assoluti e relativi rispetto al genere di correttori nella prima prova

Errori rispetto al genere dei correttori

		Errore assoluto	Errore relativo	Punteggio assegnato
	N	Media	Media	Media
Commissioni	349	1,59	,19	8,82
Femmine	746	1,40	,17	9,15
Maschi	767	1,35	,16	8,83

Tab. 37.Errori assoluti e relativi rispetto al genere dei correttori (prima e seconda prova)

Una situazione analoga si può riscontrare confrontando il comportamento dei correttori classificati per età. Anche in questo caso l'esame della media dei punteggi assegnati sembra confermare l'esistenza di criteri di valutazione leggermente diversi in cui i più severi sembrano i più anziani ed i meno severi quelli della fascia che va dal 51-simo al 55-simo anno d'età.

Errori e punteggi rispetto all'età dei correttori

		Errore assoluto		Errore relativo	Punteggio assegnato
		N	Media	Media	Media
Classi d'età	30 - 40	257	1,24	,16	9,02
	41 - 50	425	1,48	,18	9,04
	51 - 55	472	1,32	,16	9,15
	56 - 61	314	1,45	,17	8,88

Tab. 38 Errori assoluti e relativi rispetto all'età del correttore

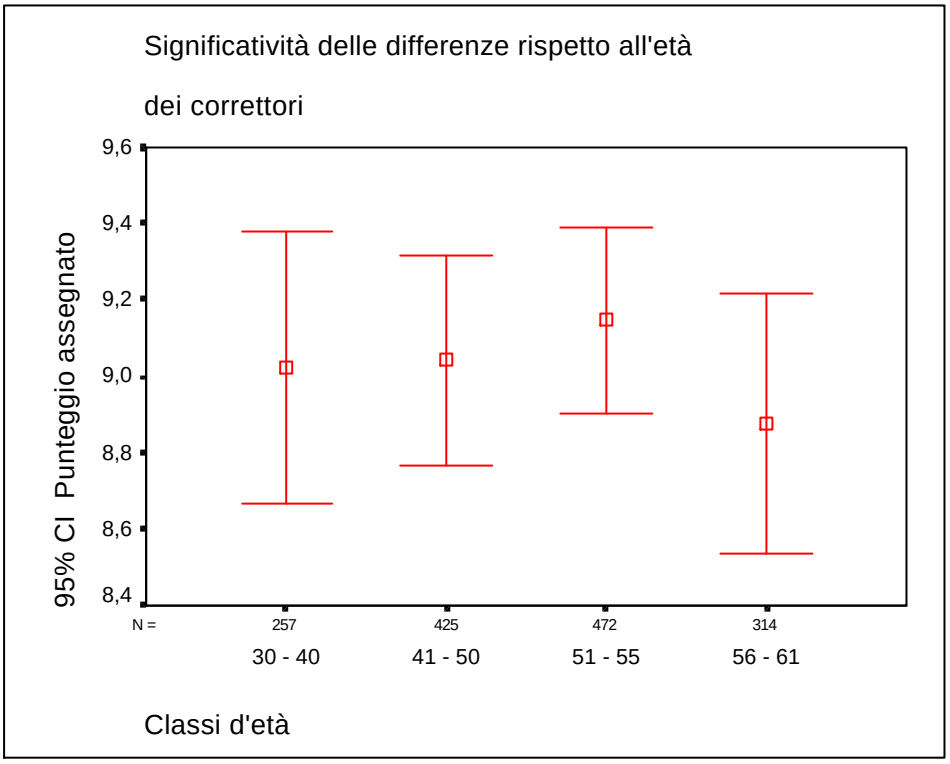


Fig. 39 Stima delle medie dei punteggi assegnati rispetto all'età dei correttori

Il grafico 39 non deve indurci però in facili generalizzazioni poiché le differenze rilevate, seppur interpretabili, sono troppo lievi perché possano essere considerate statisticamente significative con un campione così ridotto di correttori. Anche in questo caso, si può assumere però che gli insiemi di prove assegnate casualmente ai vari gruppi d'età siano equivalenti in media e che le differenze tra le medie dipendano dalla diversità dei criteri di correzione adottati dai vari gruppi di correttori.

Continuando nell'esame delle caratteristiche dei correttori vanno analizzate anche le correzioni ripetute dello stesso correttore. Ricordiamo che ogni prima e seconda prova è stata corretta da uno stesso correttore in tempi diversi, circa 20 giorni dopo la prima correzione. I correttori erano all'inizio dello studio ignari di dover ripetere una correzione già effettuata e quando hanno ricevuto un nuovo fascicolo di cinque elaborati da correggere avevano già riconsegnato le proprie valutazioni e tutto il materiale documentale annesso. Abbiamo raccolto interessanti commenti di correttori che hanno cercato di analizzare le ragioni che li hanno portati ad un cambiamento dei punteggi inizialmente espressi. Tali considerazioni ci hanno rinforzato nella convinzione che la variabilità dei punteggi non è un indizio di scarsa professionalità dei correttori o di poca cura nel lavoro svolto ma è l'inevitabile caratteristica di una procedura di misurazione.

La figura 40 riporta la distribuzione delle differenze assolute tra i punteggi nelle correzioni differite. E' facile notare che in più del 50% dei casi i correttori non confermano il primo punteggio assegnato con variazioni, in qualche caso, di più di due punti.

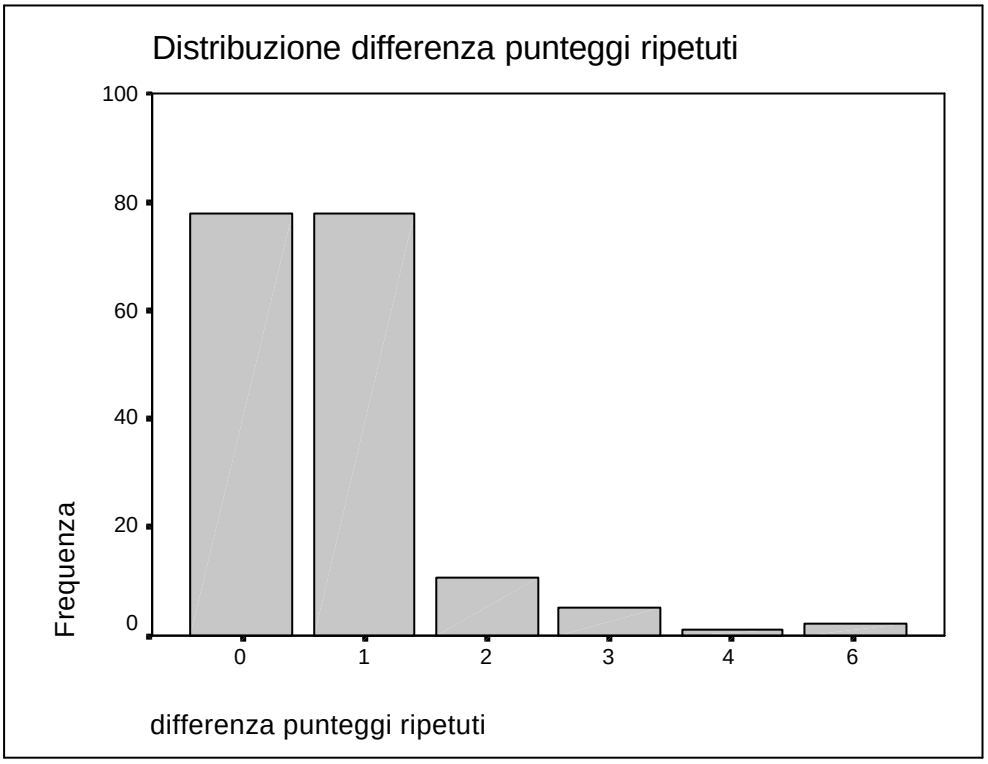


Fig. 40 Stabilità del punteggio in correzioni ripetute dallo stesso correttore

La rilevanza delle divergenze

Senza voler accentuare eccessivamente la gravità del problema, riteniamo utile comunque illustrare alcune implicazioni pratiche connesse alla imprecisione dei punteggi assegnati. Come abbiamo già detto, l'incertezza dei punteggi sembra essere maggiore proprio intorno alla soglia di sufficienza, ma spesso accade che la divergenza tra punteggi diversi sia tale che possano coesistere nel gruppo di valutatori apprezzamenti dello stesso elaborato che corrispondono a livelli qualitativi assai distanti. Per analizzare la rilevanza pratica di queste divergenze abbiamo classificato i punteggi in tre livelli: gli *insufficienti* da 0 a 9, gli *eccellenti* da 13 a 15 e i *medi* tra 10 e 12.

Il grafico di dispersione della figura 41 rappresenta ogni elaborato con un punto sul piano cartesiano: due sono le coordinate, la prima è la percentuale dei punteggi **eccellenti** e la seconda è la percentuale dei punteggi **insufficienti** espressi per quell'elaborato. I punti che rappresentano gli elaborati si disperdono all'interno di un triangolo rettangolo. Gli elaborati che si trovano sui cateti del triangolo sono quelli in

cui una percentuale più o meno alta dei correttori concorda su un solo livello (*insufficienti* o *eccellenti*) ma non coesistono i due livelli estremi di giudizio tra i punteggi assegnati a quell'elaborato. Per tutti gli altri punti del grafico, che non si trovano sui cateti, il gruppo dei correttori si è fortemente diviso ed ha espresso sul medesimo elaborato alcuni punteggi insufficienti ed alcuni altri punteggi eccellenti.

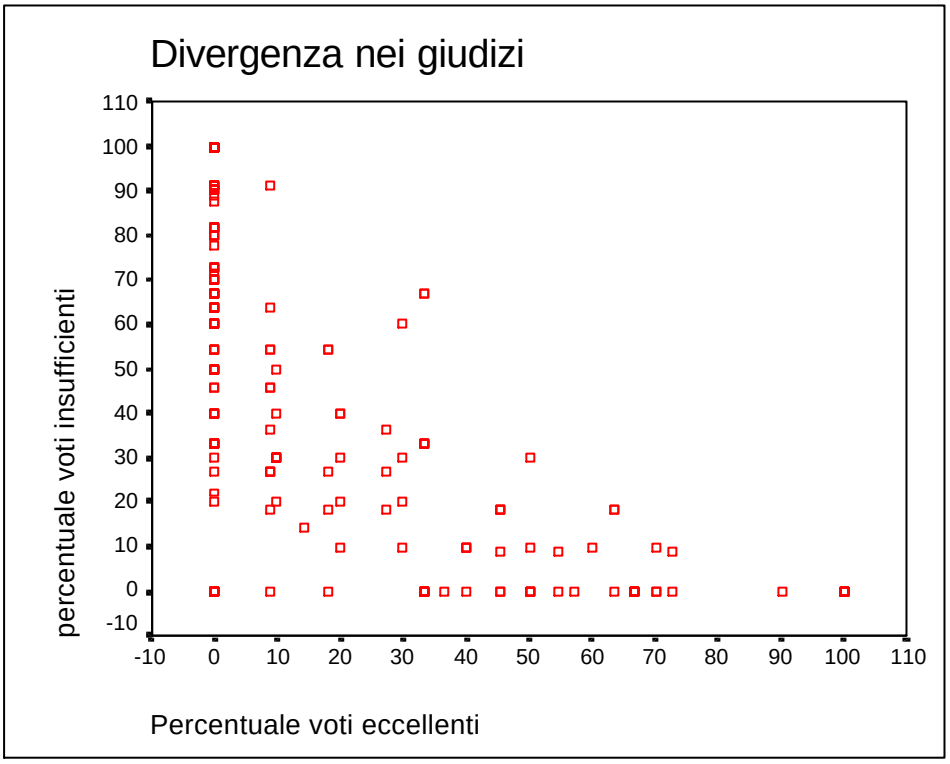


Fig. 41 Grafico di dispersione degli elaborati rispetto alla divergenza dei giudizi

Il grafico della figura 42 rappresenta in tre dimensioni gli stessi dati della figura 41 mostrando come si addensano le frequenze sui vari casi: le tre *torri* che si trovano ai vertici del triangolo corrispondono ai casi di buona concordanza tra i correttori. La più alta corrisponde agli elaborati in cui tutti i correttori concordano sulla insufficienza della prova, quella che si trova all'altro estremo dell'ipotenusa ai casi in cui tutti concordano su punteggi eccellenti e la terza, sul terzo vertice del triangolo, agli elaborati in cui i punteggi si trovano nell'intervallo mediano dei voti sufficienti. Tutti gli altri casi denotano situazioni in cui il gruppo dei correttori si è diviso su livelli qualitativi estremi: ciò fa ipotizzare che non solo le scale numeriche con una estesa gamma di punteggi, come le scale usate negli esami di Stato, ma anche le scale qualitative con pochi livelli (insufficiente, medio, eccellente) possono porre problemi di accordo tra correttori diversi per effetto della variabilità dovuta agli errori di misurazione.

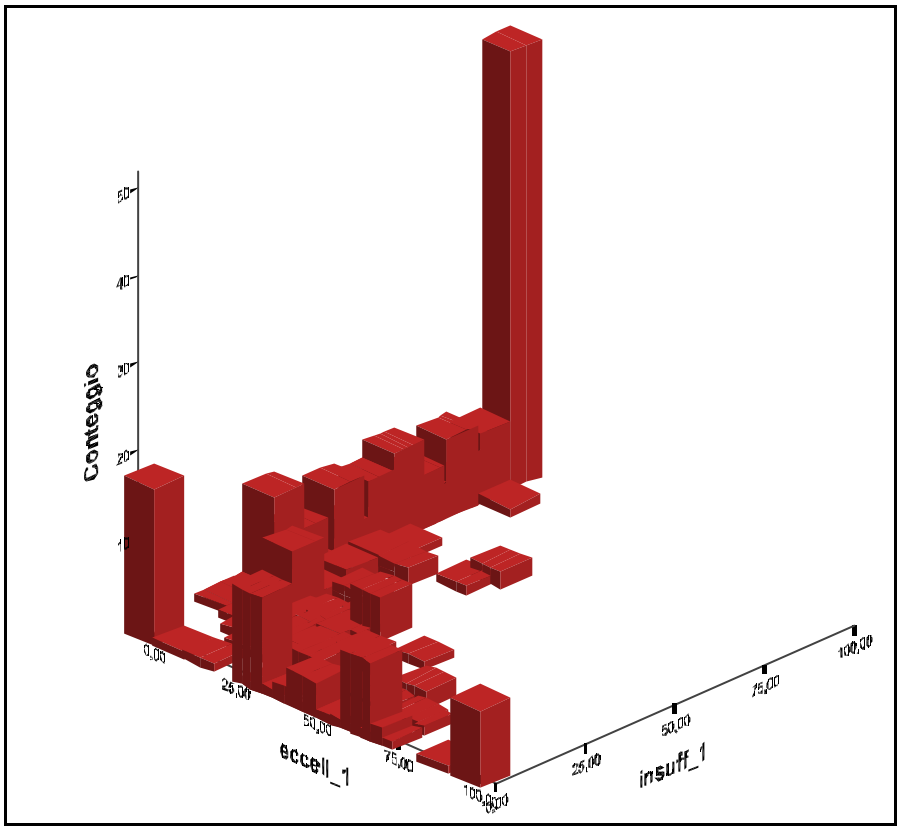


Fig. 42 Rappresentazione tridimensionale degli elaborati rispetto alle divergenze dei giudizi

Confronti dei punteggi assegnati

Sin qui abbiamo riflettuto sugli errori, convenzionalmente calcolati come scarto tra i punteggi assegnati e un "valore vero" convenzionalmente stimato in base alle correzioni ripetute della stesso elaborato.

Quali altre considerazioni possiamo fare analizzando direttamente i punteggi raccolti? Quali possono essere le implicazioni se assumessimo gli elaborati come rappresentativi del complesso degli elaborati prodotti negli esami di Stato? Come abbiamo mostrato sopra, le prove analizzate sono un campione casuale a tutti gli effetti ma tale campione è troppo limitato rispetto alla totalità degli elaborati prodotti negli esami. Dobbiamo pertanto assumere i risultati presentati in questa parte del rapporto come un esperimento mentale utile a comprendere e a formulare nuove ipotesi di lavoro evitando però di incorrere in indebite generalizzazioni.

Innanzitutto è possibile confrontare la distribuzione dei punteggi assegnati nell'esperimento con quella generale dei dati della sessione dell'anno 2000 da cui è tratto il campione di prove. La figura 43 consente tale confronto per la prima prova: le barre rappresentano le frequenze relative di tutti i punteggi assegnati nella prima prova dai correttori dell'esperimento mentre la linea spezzata si riferisce alla distribuzione

osservata su tutti i punteggi assegnati nella sessione 2000. La successiva figura 44 confronta invece la distribuzione dei 'voti veri interi' così come li abbiamo convenzionalmente calcolati nel nostro esperimento.

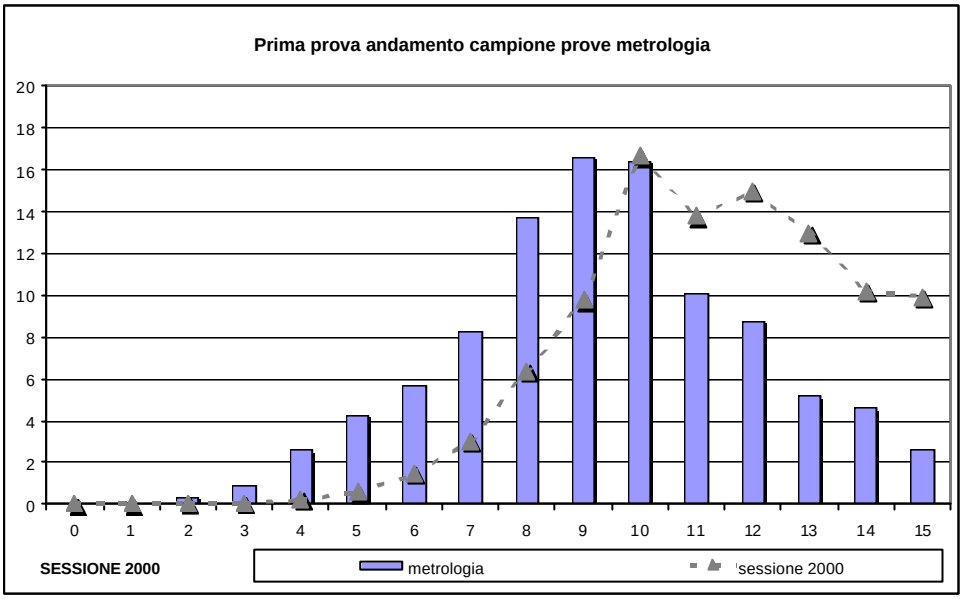


Fig. 43 Confronto distribuzione punteggi assegnati nello studio sperimentale con distribuzione dell'universo (prima prova)

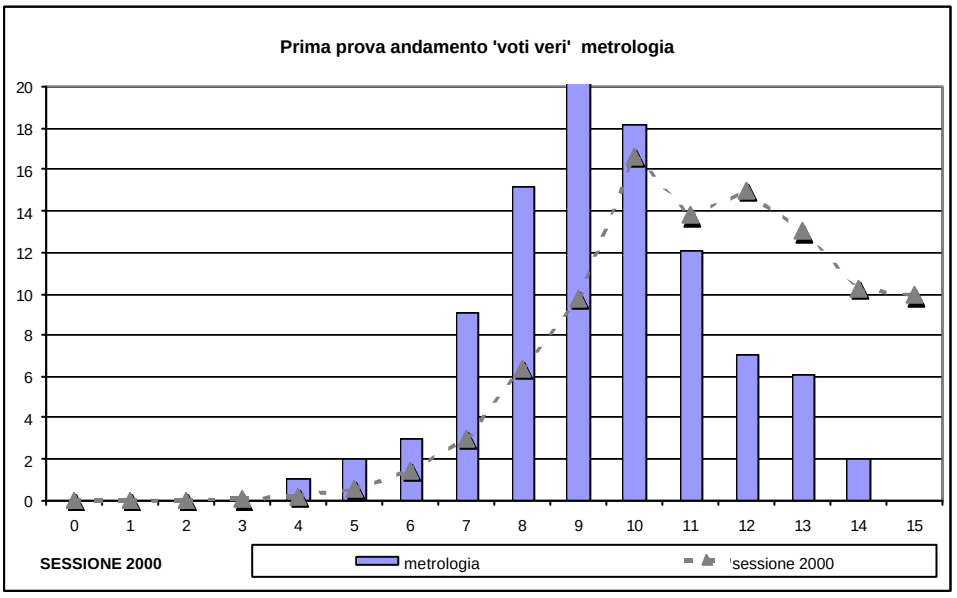


Fig. 44 Confronto distribuzione voti "veri" del campione di metrologia con distribuzione dell'universo (prima prova)

In entrambi i grafici è visibile uno spostamento della distribuzione assegnati nello studio sperimentale verso i valori più bassi. Se si assume che il campione degli elaborati

usati nell'esperimento sia rappresentativo del totale degli elaborati della sessione, emergerebbe da questi due grafici delle figure 43 e 44 che da parte dei correttori dell'esperimento vi sia stato un uso della scala dei punteggi diverso da quello dei commissari d'esame: mentre nei commissari la preoccupazione dell'esito finale porta ad usare prevalentemente la parte superiore della scala, quella che assicura la sufficienza, nelle correzioni dell'esperimento la scala è stata usata in modo più esteso senza saturare il valori più alti. Possiamo chiederci allora: quale distribuzione riflette meglio la situazione reale?

Tralasciamo di applicare questa stessa analisi alla seconda prova poiché abbiamo potuto correggere solo 4 discipline (latino, matematica, ragioneria ed elettronica), mentre una comparazione tra le distribuzioni dei punteggi della terza prova potrebbe essere più proponibile per la maggiore omogeneità dei criteri di formazione del campione degli elaborati.

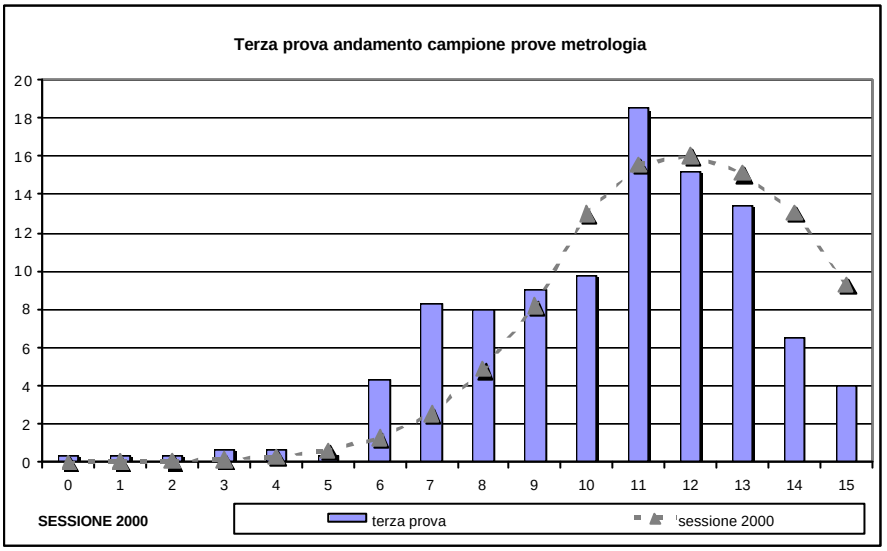


Fig. 45 Confronto distribuzione punteggi assegnati nello studio sperimentale con distribuzione dell'universo (terza prova)

Il grafico 45, che compara la distribuzione dei voti veri interi assegnati nello studio sperimentale con quella dei punteggi effettivi della sessione 2000, presenta delle irregolarità dovute al basso numero di elaborati ma, rispetto alla situazione della prima prova del grafico 44, presenta un migliore adattamento alla distribuzione generale e quindi conferma che la modalità di correzione delle terze prove è rimasta più stabile e simile a quella delle commissioni vere, anche in una situazione artificiale come quella dell'esperimento.

Ciò porterebbe ad ipotizzare che la terza prova non solo è corretta con una maggior precisione ma anche che i criteri e l'uso delle scale siano più stabili e facilmente esportabili tra contesti diversi.

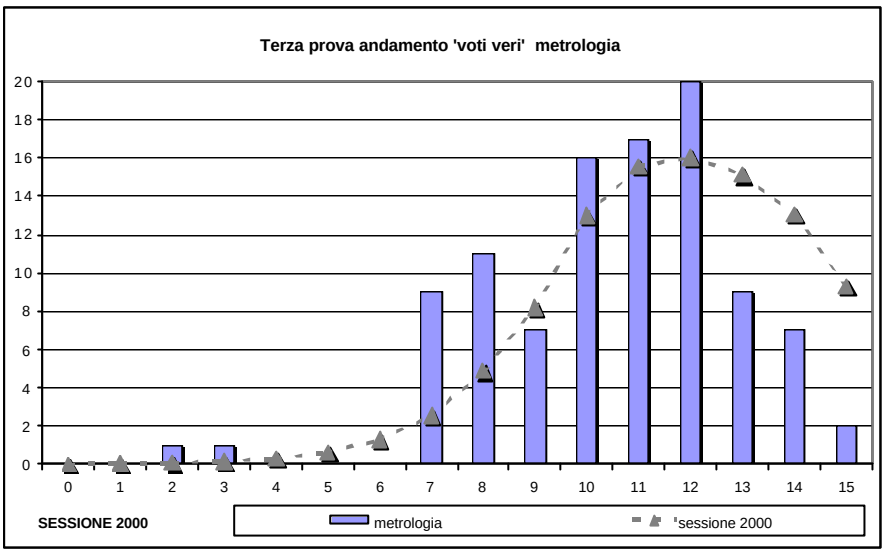


Fig.46 Confronto distribuzione voti "veri" dello studio sperimentale con distribuzione dell'universo (terza prova)

L'analisi del valore "vero" intero ci consente anche di effettuare alcuni confronti dei valori medi rispetto ad alcune variabili strutturali. La tabella 47 mostra le medie dei punteggi "veri" calcolate per ogni tipo di prova per i vari ordini scolastici. Poiché l'assegnazione delle prove ai correttori è stata fatta casualmente rispetto agli ordini scolastici delle prove da correggere, possiamo ritenere che la valutazione 'vera' sia stata fatta con una stessa *metrica* rispetto al tipo di istituto per cui, almeno per gli elaborati che abbiamo esaminato, possiamo dire che la tabella rappresenti le differenze di esito tra i vari ordini scolastici. I confronti possono essere analizzati sia leggendo i dati nella tabella orizzontalmente (a parità di tipo di prova) sia verticalmente (a parità di ordine scolastico). Da notare che le seconde prove analizzate (latino e matematica per i licei e ragioneria ed elettronica per i tecnici) sono state valutate meno positivamente della prima e terza prova.

Valori medi dei punteggi rispetto al tipo di scuola

		ordine scolastico prove						Totale	
		licei		professionali		tecnici		Media	Dev. stand.
		voto vero intero		voto vero intero		voto vero intero			
		Media	Dev. stand.	Media	Dev. stand.	Media	Dev. stand.		
Tipo prova	Prima prova	10,51	1,70	8,39	1,81	9,17	1,93	9,42	2,00
	Seconda prova	8,60	2,85	,	,	8,30	2,44	8,45	2,64
	Terza prova	10,48	1,72	9,65	1,87	10,95	3,09	10,50	2,41

Tab. 47 Punteggi medi per tipo di prova e di ordine scolastico

I confronti tra i punteggi medi dei vari strati territoriali da cui provengono le prove (v. figura 48) sembrano contraddire le differenze riscontrate nelle statistiche ufficiali ma le numerosità delle prove sono troppo basse per poter avere indicazioni statisticamente significative sulle differenze. Anche in questo caso una metrica uniforme nei giudizi delle prove ottenuta casualizzando l'assegnazione dei correttori, consentirebbe di verificare oggettivamente se le differenze che appaiono nelle statistiche ufficiali corrispondono a situazioni di fatto o da diversi criteri di valutazione da parte delle commissioni disperse sul territorio.

Per ottenere stime che apprezzino significativamente differenze di un punto tra i cinque strati territoriali occorrerebbe avere un campione di circa 500 elaborati, per differenze tra gli strati di mezzo punto servono circa 2000 elaborati contro i circa 280 corretti in questo studio.

La situazione è un po' più chiara se ci riferiamo ai dati di una sola prova. Ad esempio per i temi, la figura 50 mostra che gli intervalli di confidenza sono più ristretti e le differenze tra gli strati sono più marcate. In questo caso l'andamento della distribuzione osservata nell'esperimento è simile a quello dell'intera popolazione ma per ottenere intervalli di confidenza minori di un punto occorrerebbero circa 300 elaborati e per saggiare significativamente differenze di mezzo punto occorrerebbero circa 1200 elaborati. Questo studio ne ha corretti esattamente 99.

Le considerazioni precedenti ed in particolare la valutazione della dimensione dei campioni non sono un ozioso esercizio di stile ma costituiscono un valore aggiunto dello studio che abbiamo realizzato: i dati raccolti consentono di pianificare studi comparativi basati sulla correzione ripetuta degli elaborati degli esami da parte di correttori che assicurino una metrica uniforme sul territorio (anche per prove complesse e non solo per prove oggettive). Ovviamente appare chiaro che i costi di una simile procedura sarebbero molto alti comunque molto più alti di quanto si spende usando test oggettivi per le comparazioni di sistema.

Valori medi dei punteggi rispetto allo strato territoriale

		voto vero intero	
		Media	N
strato territoriale della prova	nordovest	9,14	36
	nordest	9,26	35
	centro	9,76	80
	sud	9,24	62
	sudisole	9,88	66

Tab.48 Punteggi medi nella prima prova rispetto allo strato territoriale

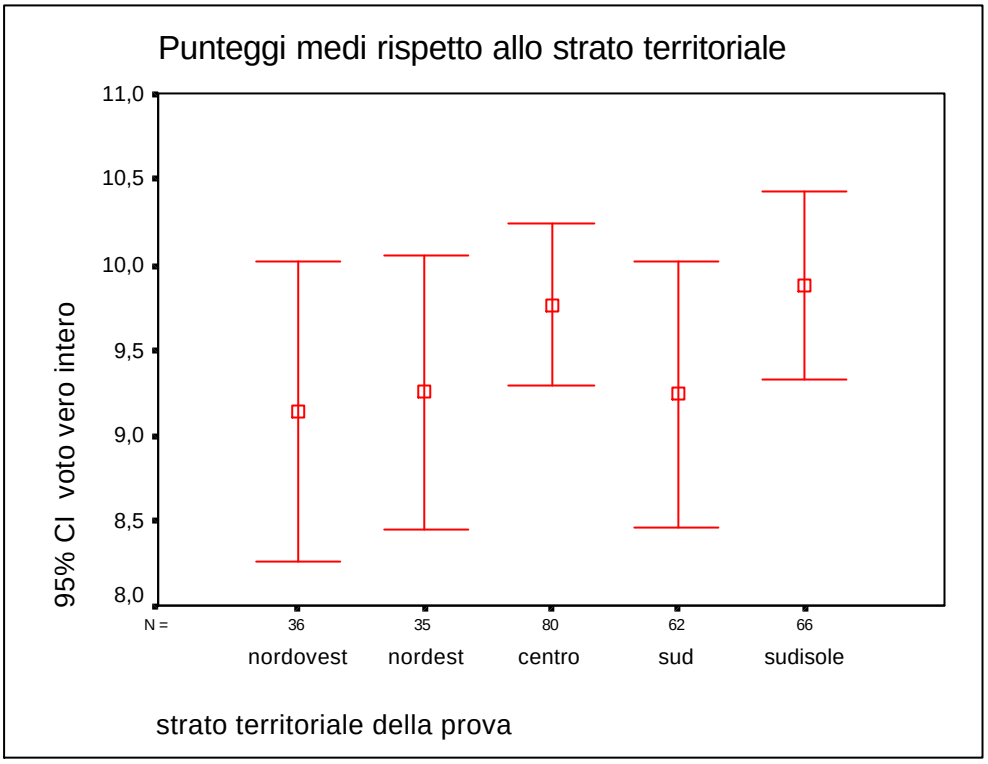


Fig. 49 Stima dei punteggi medi nella prima prova rispetto allo strato territoriale (tutte le prove)

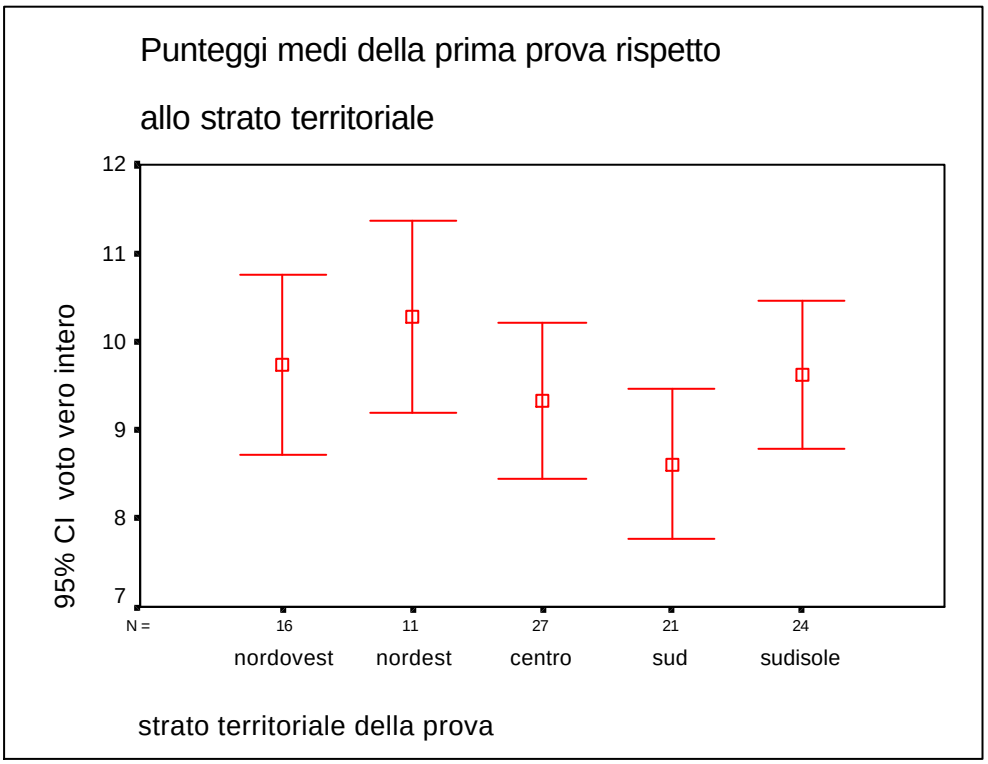


Fig. 50 Stima dei punteggi medi nella prima prova rispetto allo strato territoriale (prima prova)

Alcune implicazioni pratiche

L'esplorazione sin qui condotta ci conduce a formulare nuove ipotesi di lavoro e a trarre alcune prime conclusioni.

Quante correzioni indipendenti servono per ottenere una stima abbastanza precisa del voto "vero"? Abbiamo effettuato il calcolo prova per prova poiché abbiamo verificato che l'errore di misura può variare sensibilmente anche rispetto allo stesso tipo di prova o con la stessa traccia. La tabella 51 riporta il numero di correzioni ripetute indipendenti per ottenere una stima del valore vero di ampiezza un punto: la prima colonna riporta la media aritmetica delle correzioni ripetute calcolate prova per prova mentre la seconda colonna riporta il valore massimo osservato cioè il numero necessario perché la precisione richiesta sia ottenuta per tutte le prove corrette e non solo per alcune. Tale tabella ripropone un'altra implicazione pratica della imprecisione nella assegnazione dei punteggi: per ottenere quantificazioni affidabili paragonabili alle prove oggettive occorrerebbe affrontare costi e sopportare tempi di attesa difficilmente accettabili.

prova	correttori	
	media	necessari
Terza prova	43	165
Analisi del testo	67	121
Ambito artistico letterario	61	145
Ambito socio economico	57	110
Ambito storico politico	63	128
Ambito tecnico scientifico	49	106
Tema di argomento storico	67	147
Tema di ordine generale	53	111
Latino	48	126
Matematica	44	94
Ragioneria	57	100
Elettronica	75	132

Tab. 51 Correttori necessari per avere una stima del voto vero con un intervallo di confidenza inferiore ad uno.

Per una ricostruzione dei risultati veri.

L'analisi dei dati dell'esperimento di metrologia ci ha condotto a riflettere su molte implicazioni pratiche e su alcune possibilità di ulteriori ricerche per rendere maggiormente affidabile l'accertamento dei risultati attraverso saggi scritti e prove strutturate.

La quantificazione dell'errore casuale, compiuto da chi corregge una prova scritta dell'esame di Stato, ci ha spinto ad effettuare un ulteriore esperimento, questa volta sui punteggi effettivamente assegnati nella sessione d'esame 2000, sessione da cui sono tratte le prove scritte usate in questo studio. Tale simulazione/esperimento è possibile poiché disponiamo dei dati analitici ufficiali di quasi tutta la popolazione degli studenti esaminati. Ancora una volta il valore delle considerazioni che seguiranno è soprattutto legato alla possibilità di riflettere, di formulare ipotesi, di capire meglio per avviare eventualmente nuove ricerche empiriche. Serve soprattutto a stimolare negli attori principali del processo (i commissari che valutano) una attenzione critica sugli effetti micro e macro delle loro scelte.

Riprendiamo in considerazione le caratteristiche degli errori di misura dei correttori del nostro esperimento ricordando che si tratta di variabili distribuite normalmente (v. fig.26) con media 0 e deviazione standard dipendente dal tipo di prova. La tabella 52 riporta il valore delle deviazioni standard per i tre tipi di prova.

Errori di misura osservati nell'esperimento

Tipo prova	casi	Media	Dev. std.
Prima prova	1004	,0000	1,8432
Seconda prova	858	,0000	1,7872
Terza prova	276	,0000	1,2695

Tab. 52 Deviazione standard degli errori di misura per tipo di prova

Durante gli esami, gli elaborati scritti ricevono, ovviamente, una sola correzione e quindi a ciascun elaborato viene assegnato un punteggio affetto da un errore casuale che possiamo assumere abbia le caratteristiche osservate nel nostro esperimento. Il valore "vero" della prestazione osservata sarà un valore reale che si discosta dal punteggio assegnato con la stessa distribuzione dei probabilità con cui si distribuiscono gli errori che abbiamo osservato nell'esperimento.

Che succede se ad ogni punteggio ufficiale sommiamo un errore casuale distribuito normalmente, così come sono distribuiti gli errori osservati nell'esperimento? Otterremo dei nuovi punteggi che potremmo considerare altrettanto plausibili, altrettanto 'veri'. Spero che il lettore abbia un leggero sussulto e si senta un po' destabilizzato ma questo ragionamento, che sembra sconvolgere l'ufficialità degli esiti,

è perfettamente equivalente all'affermazione, che ci sembra sempre più evidente, secondo cui i punteggi assegnati sono affetti da errori casuali.

Proseguiamo quindi nella nostra simulazione dopo aver ricalcolato tutti i punteggi sommando errori di diversa deviazione standard come indicato dalla tabella 52 ed approssimando il valore ottenuto all'intero più vicino.

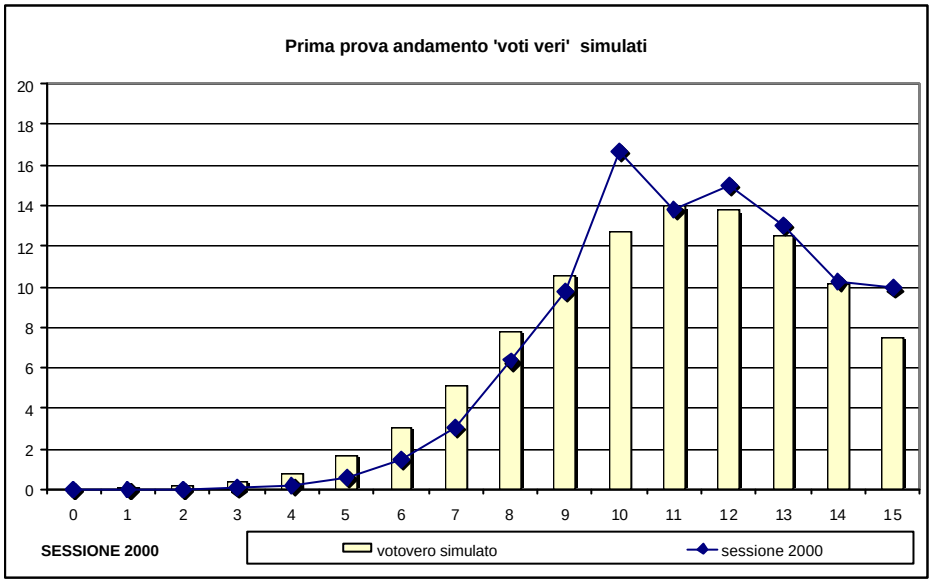


Fig.53 Ricostruzione della distribuzione "vera" della prima prova

Dopo aver escluso i casi in cui il nuovo punteggio usciva dall'intervallo di definizione della scala, abbiamo provato a studiare la distribuzione di frequenza dei punteggi simulati mettendola a confronto con quella osservata nella statistica ufficiale. La figura 53 riporta gli andamenti della prima prova: l'irregolarità del picco modale sulla soglia della sufficienza è scomparsa e la nuova distribuzione simulata assume un andamento certamente più simile alla regolarità con cui un attributo complesso si distribuisce su una popolazione molto vasta. Quale distribuzione è più 'vera'? Saremmo portati a rispondere che sia quella simulata e non quella ufficiale.

Analoghe considerazioni possono essere sviluppate per la seconda e la terza prova riportate nelle figure 54 e 55. Nel caso della terza prova le due distribuzioni, quella statistica e quella simulata si somigliano tra loro molto di più delle prime due confermando le considerazioni già esposte all'inizio di questo rapporto e che cioè la terza prova presenta caratteristiche metrologiche migliori delle prime due.

Cerchiamo di riflettere ora su altre implicazioni pratiche, non più di sistema, ma riferite ai singoli candidati. Se il punteggio simulato, che a livello macro ha caratteristiche più realistiche, dovesse essere adottato come 'vero' e sostituisse quello ufficiale, quale sarebbe l'effetto per i singoli candidati? Ovviamente alcuni vedrebbero il proprio punteggio aumentare mentre altri avrebbero punteggi inferiori in qualche prova scritta. Cosa succede in particolare nell'intorno della soglia di sufficienza? Abbiamo analizzato

i dati della prima prova calcolando la tabella di contingenza determinata dai due punteggi (ufficiale e simulato) ed ottenendo il numero dei casi che, per effetto della perturbazione introdotta dall'errore, scavalcano la soglia della sufficienza verso l'alto o verso il basso.

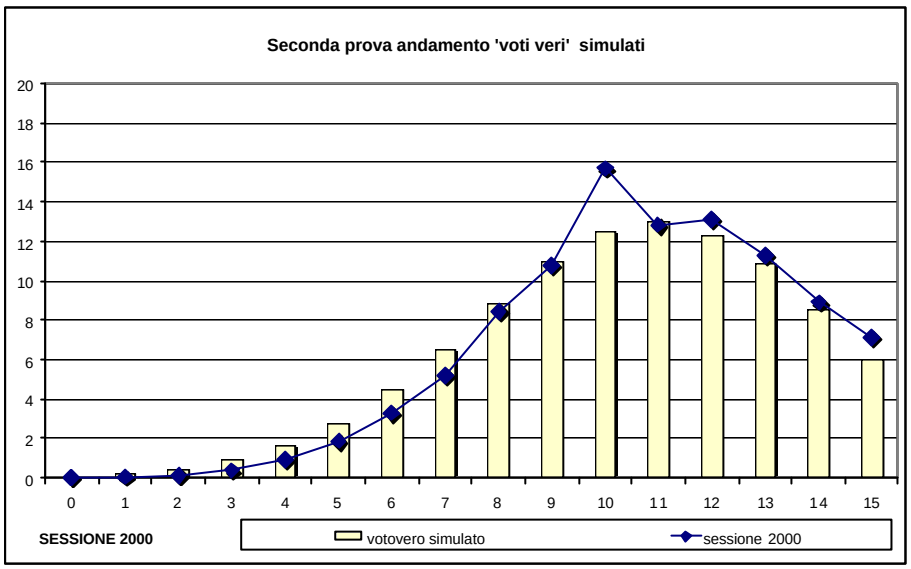


Fig.54 Ricostruzione della distribuzione "vera" della seconda prova

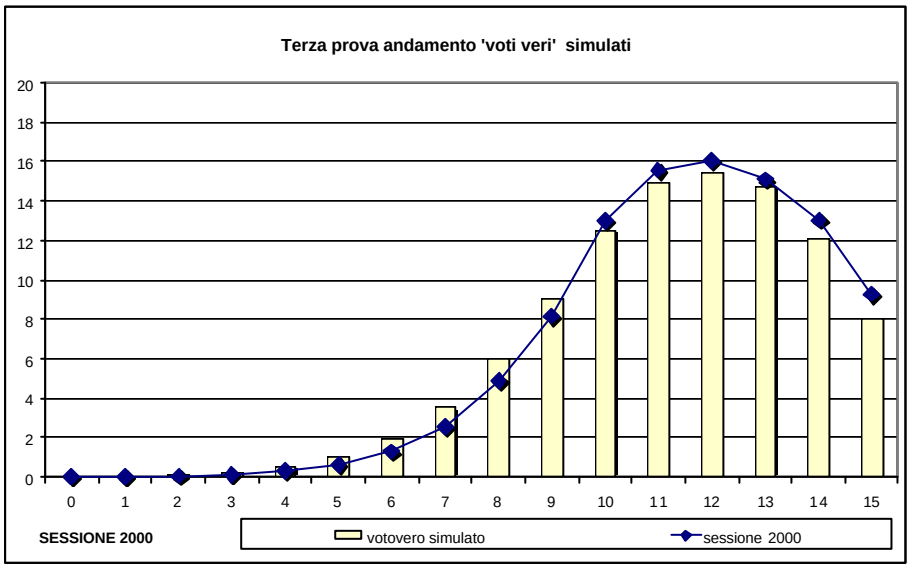


Fig.55 Ricostruzione della distribuzione "vera" della terza prova

Assumiamo come ipotesi di lavoro che i punteggi ricalcolati con la simulazione siano più vicini alla situazione reale e quindi siano i punteggi 'veri'. Su 380.437 casi registrati nel nostro archivio la situazione determinata dalla simulazione è descritta dalla tabella 56: abbiamo chiamato 'aiutati dalla commissione' coloro che hanno ricevuto un

punteggio ufficiale maggiore o uguale a 10 punti ma che con la simulazione (punteggio "vero") hanno ottenuto un nuovo punteggio inferiore a 10; i 'penalizzati dalla commissione' sono coloro che avendo avuto un punteggio ufficiale insufficiente hanno ottenuto nella simulazione un punteggio maggiore o uguale a 10.

Se l'accuratezza dei punteggi assegnati ai temi fosse quella da noi riscontrata nell'esperimento di metrologia, il 5,9% dei candidati sarebbe stato ingiustamente penalizzato dall'errore di misura della commissione contro un 12% che invece ne avrebbe avuto un vantaggio. Si badi bene che qui non stiamo parlando dei casi, per fortuna marginali, di palese ingiustizia o di disfunzioni dovute a qualche commissario incapace, ma stiamo ragionando su andamenti dovuti alla sola variabilità legata ad errori casuali di misura.

	N	%
Aiutati dalla commissione	46.098	12,1
Penalizzati dalla commissione	22.797	5,9

Tab.56 Effetti della simulazione sui valori prossimi alla sufficienza

Assumendo come "vere" le distribuzioni ottenute nella simulazione (sommando ai punteggi assegnati ufficiali un errore casuale), gli elaborati di italiano giudicati insufficienti dovevano essere il 29,4% e non il 21,4% come risulta dai dati ufficiali, nella seconda prova il 31% di insufficienti ufficiali dovrebbe aumentare al 36,7% e nella terza prova dal 18% si passerebbe al 22,3% della simulazione. In sostanza il valore 'vero' ottenuto dalla simulazione metterebbe a nudo una situazione peggiore di quella emergente dai risultati ufficiali.

Quali sono gli effetti della simulazione sul punteggio maturato alla fine delle prove scritte? Anche in questo caso mettiamo a confronto le due distribuzioni, quella ufficiale e quella simulata e notiamo che la correzione introdotta dalla simulazione elimina quella intenzionale irregolarità presente sul 60 e restituisce una distribuzione che più di tutte segue la curva normale propria di una competenza complessa distribuita su una popolazione molto vasta.

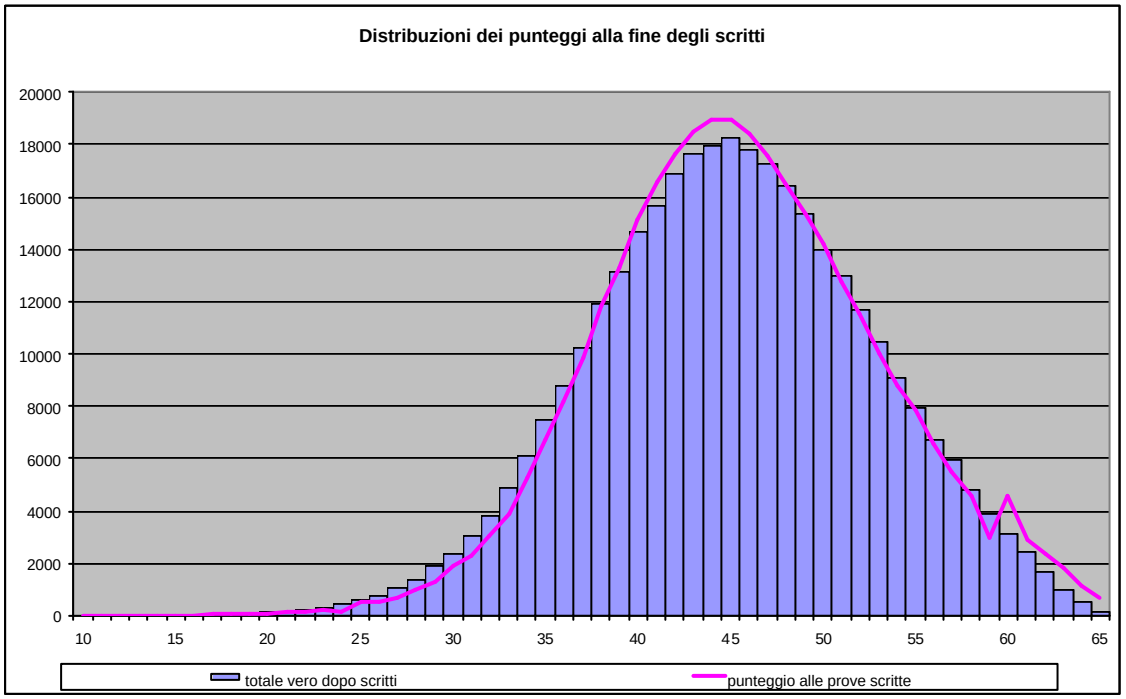
Quale sarebbe stato il risultato degli esami se non si procedesse con la prova orale? Quali sarebbero i risultati se non si aggiustasse intenzionalmente l'esito alla fine degli scritti con un orale che serve a compensarne il valore per assegnare un voto finale stabilito globalmente?

Estratto della distribuzione cumulata negli intorni delle soglie				
	Punteggio	Simulato		Osservato
	35	10,1		8,2
	36	12,7		10,6
	37	15,6		13,4
	38	19,1		16,9
	39	22,9		20,8

Estratto della distribuzione cumulata negli intorni delle soglie				
	Punteggio	Simulato		Osservato
	40	27,2		25,2
	41	31,8		30,0
	42	36,7		35,2
	43	41,9		40,6
	44	47,1		46,1
	45	52,4		51,6
	46	57,6		57,0

Tab. 57 Distribuzione cumulata del punteggio alla fine degli scritti sulla soglia della sufficienza

Se fissiamo la soglia di sufficienza sui due terzi della scala, come accade per scale degli scritti e cioè su 43,3 punti, il 35,2 % sarebbe insufficiente secondo i dati ufficiali contro 36,7 % dei valori simulati. Se invece la soglia viene fissata sui 60 centesimi della scala, come accade per il voto finale, riscontreremmo che il 16,9% non ha raggiunto la sufficienza secondo la distribuzione ufficiale mentre tale percentuale sale al 19,1% se adottiamo come vera la distribuzione simulata. In ogni caso, qualunque sia il punto di vista secondo cui si analizzano i dati, troviamo che la coda di sinistra della distribuzione dei punteggi assegnati alla fine degli scritti, prima della 'sanatoria' degli orali, che raggruppa i candidati che non hanno raggiunto la sufficienza, è ben più consistente del 5% finale dei non diplomati.



Tab. 58 Distribuzione del punteggio alla fine degli scritti

Interessante notare che, anche in questo caso, l'effetto cumulativo della correzione dei valori osservati effettuato nella simulazione elimina quella piccola irregolarità della distribuzione osservata sulla soglia del 60 che corrisponde alla possibilità di assegnare il bonus finale.

Riprendiamo dunque il filo del nostro discorso iniziale ed in particolare cerchiamo di analizzare gli effetti della simulazione sulla distribuzione del voto finale. Sommiamo quindi i punteggi 'veri' degli scritti (quelli perturbati da noi con gli errori di misura) ai valori ufficiali del credito, dell'orale e del bonus.

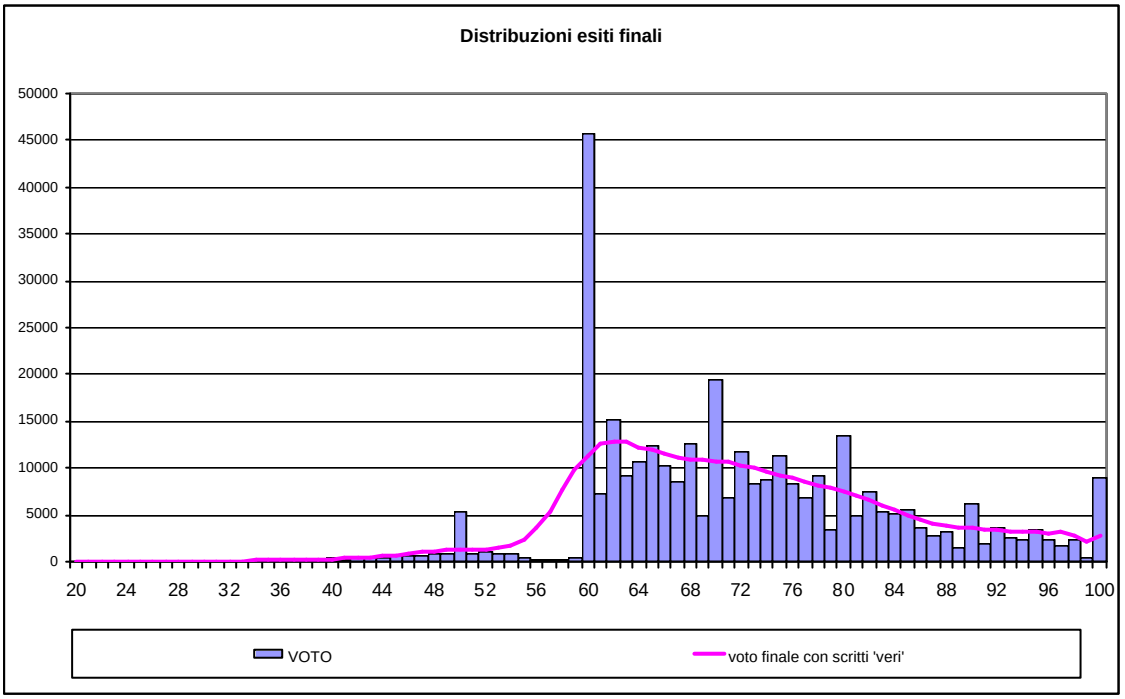


Fig.59 Esito finale ufficiale ed esito "vero" simulato

Il grafico di figura 59 mette a confronto la distribuzione statistica osservata, rappresentata con le colonne con quella ricostruita con la simulazione: l'aver perturbato i punteggi degli scritti con gli errori casuali osservati nel nostro studio ha avuto un effetto piuttosto vistoso ovvero ha eliminato quei picchi sui valori soglia dovuti al fatto che quando si valuta l'orale si conosce esattamente il punteggio assegnato negli scritti. Un effetto analogo si avrebbe se la commissione assegnasse il voto dell'orale avendo dimenticato l'esatto valore dei punteggi assegnati agli scritti: non avremmo quegli arrotondamenti che rendono meno frequenti i 69, i 79, gli 89 e i 99. Ma l'effetto più vistoso sarebbe che una parte di coloro che hanno avuto il 60 sarebbero classificati al di sotto della soglia di sufficienza: nella distribuzione del voto "vero" il 12,7% del totale non otterrebbe il diploma.

La disponibilità dei punteggi effettivi in ciascuna prova per quasi tutti i candidati (punteggi ufficiali) ci consente di andare oltre nella nostra simulazione ed in particolare di esaminare gli effetti di una diversa ripartizione dei punteggi tra le varie prove.

Una prima ipotesi consiste nello scambiare il peso del credito e dell'orale ovvero nell'assumere che attraverso il credito scolastico si possa ottenere fino a 35 punti mentre con il colloquio solo 20 punti. La variabile *votoc1* ottenuta sommando il credito e il colloquio, ricalcolati con i nuovi pesi, e i punteggi degli scritti simulati viene rappresentata nel grafico della figura 60 dalle barre chiare. Ciò che si può facilmente osservare è che l'effetto dell'aggiustamento sulla soglia della sufficienza sparisce completamente e coloro che sono classificati al di sotto della sufficienza salgono al 19% (assegnando un peso maggiore al credito scolastico e diminuendo il peso dell'orale). La trasformazione ha un effetto visibile anche sui punteggi alti aumentando leggermente le frequenze nella coda di destra. Dal punto di vista metrologico la distribuzione di *votoc1* essendo meno concentrata riesce a discriminare meglio su tutta la gamma dei punteggi, dai più alti ai più bassi.

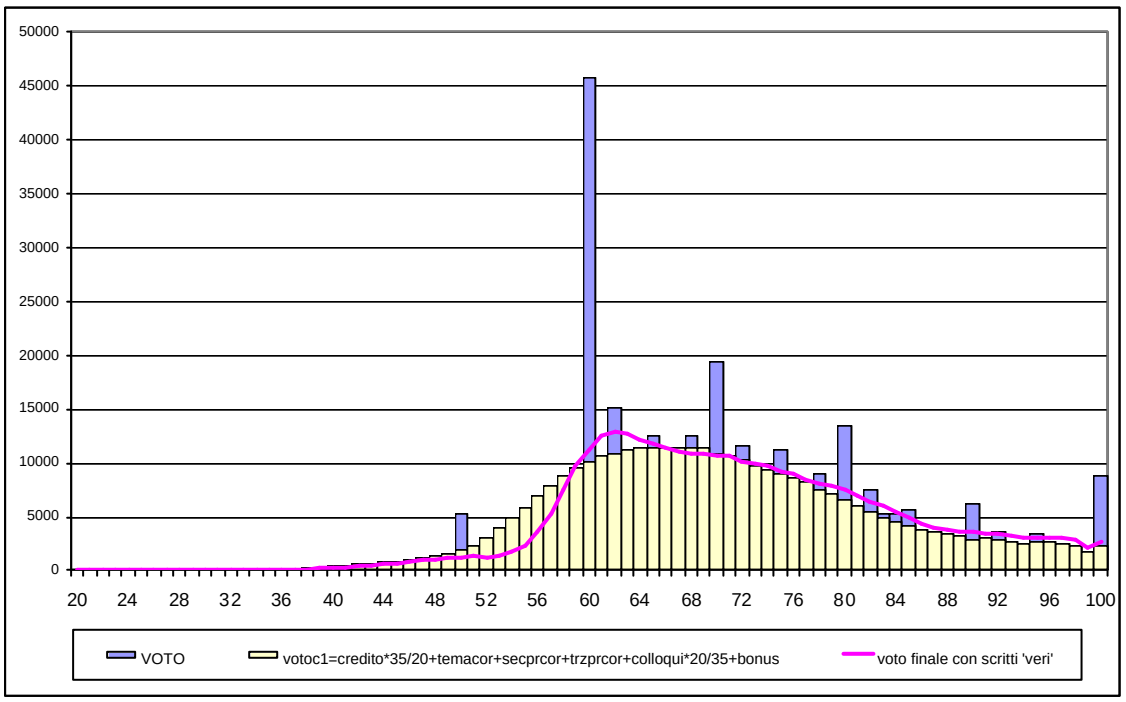


Fig.60 Esito finale ufficiale, esito con diversa escursione dei punteggi ed esito "vero" simulato

Ovviamente si può procedere ulteriormente nella simulazione prospettando altri scenari, ad esempio ipotizzare che l'orale abbia lo stesso peso degli scritti, che il punteggio residuo sia assegnato tramite il credito scolastico e che vi sia l'eliminazione del bonus: la variabile *votoc2* che nel grafico della figura 61 è rappresentata dal tratteggiato, ha una distribuzione molto simile alla variabile *votoc1* ma aumenta ulteriormente la selettività del punteggio incrementandone la frequenza nei punti di flesso della distribuzione. In questa ipotesi gli insufficienti salgono al 21%

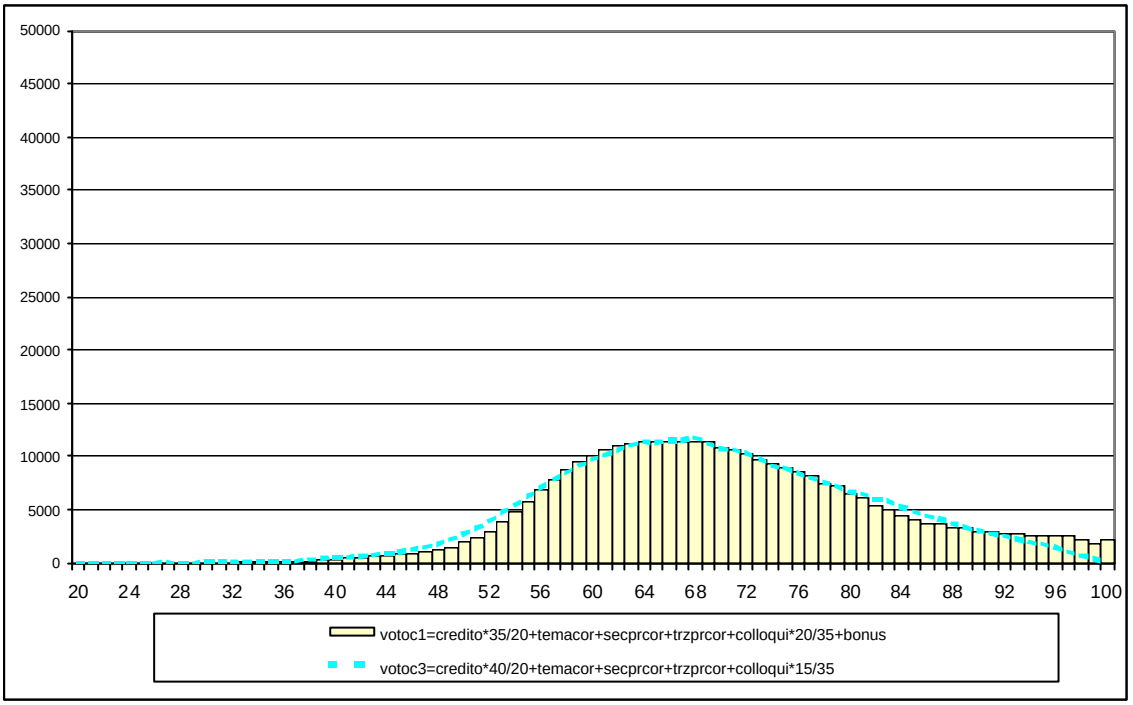


Fig.61 Esito finale ufficiale secondo pesi diversi assegnati ai vari punteggi

Conclusioni

Difficile trarre delle conclusioni univoche da uno studio così delicato. Avremmo reso un pessimo servizio alla scuola se diffondessimo l'idea che siccome ogni misura è affetta da un errore casuale ineliminabile tanto vale rinunciare ad effettuare misure in campo educativo. Tale rischio è sempre presente e potrebbe rinforzare l'idea che con un po' di buon senso tutti i problemi possano essere risolti al meglio.

Ambizione di tale studio era quella di rendere maggiormente consapevoli tutti gli attori della vicenda degli esami di Stato della delicatezza dei problemi legati alla valutazione in cui l'equità del giudizio è fortemente legata alla qualità delle operazioni di accertamento e di analisi delle prestazioni prodotte dai candidati. Vorremmo che il problema della precisione delle 'misure' diventasse una consapevolezza diffusa e che ciò promuovesse due atteggiamenti:

1. maggiore flessibilità e disponibilità al confronto intersoggettivo tutte le volte che si propongono o utilizzano dati e 'misure' che si riferiscono a prestazioni degli studenti;
2. sistematica ricerca di un miglioramento della precisione delle stime attraverso la replica delle 'misure' e attraverso l'affinamento degli strumenti di misura utilizzati.

Quanto abbiamo verificato nell'esperimento non delegittima la forma dell'esame di Stato introdotta dalla riforma del ’97 ma evidenzia dei problemi ineliminabili in qualsiasi forma di accertamento e valutazione. Il nuovo esame di Stato nei suoi presupposti teorici e normativi propone appunto una soluzione al problema identificando una pluralità di accertamenti indipendenti che dovrebbero concorrere al miglioramento della precisione delle stime del punteggio 'vero'. I dati ufficiali relativi agli esiti mostrano però che prevalgono le abitudini più consolidate e che cioè permangono degli approcci di tipo globale che in qualche caso introducono delle autentiche distorsioni sistematiche degli esiti finali. Questo studio ha cercato quindi di promuovere e rinforzare un processo di adattamento del mondo della scuola ad una visione della valutazione finale più 'oggettiva' e più scientifica, più consapevole dei vincoli posti dall'esigenza di migliorare la precisione dei punteggi assegnati.

Lo studio fa inoltre emergere un problema sostanziale su cui occorrerà riflettere collettivamente:

> le prove scritte mostrano delle carenze di rendimento che, secondo il giudizio delle commissioni, riguarderebbero una porzione di candidati che va dal 20 al 30% a seconda del modo in cui vengono effettuati i calcoli. Se il campione delle prove ricorrette nello studio fosse rappresentativo della situazione generale il giudizio dei nostri correttori su tali carenze sarebbe assai più esigente. Il problema non riguarda il meccanismo dell'esame ma il funzionamento e l'efficacia della scuola secondaria superiore. Un frazione così alta di 'insufficienti' è accettabile? Si può far qualcosa per alzare il livello e ridurre le distanze tra i migliori e i peggiori? Occorre forse cambiare i livelli di accettabilità? Siamo convinti che tutti possano e/o debbano raggiungere un sicuro livello di sufficienza alla fine di un percorso formativo ben orchestrato?

Un esame di Stato quale quello previsto dalla riforma del ‘97, potenzialmente selettivo ed esigente, è stato volutamente ‘addomesticato’ nei primi due anni di attuazione per evitare traumi, rifiuti o rigetti e per dare il tempo ai ragazzi e alla scuole di adattare i propri ritmi alla nuova situazione. Una completa attuazione del processo di riforma doveva passare proprio attraverso la soluzione del problema metrologico che abbiamo posto al centro dell'attenzione del nostro studio: l'uso indipendente di una pluralità di accertamenti o 'misure' in cui gli errori accidentali siano il più possibile ridotti. In questo senso la modificazione della composizione della commissione introdotta nella sessione 2002 dal ministro Moratti interrompe tale processo perché difficilmente una commissione interna riuscirà ad apprezzare il valore di singole prove senza tener conto della valutazione globale del candidato, già fortemente consolidata nell’esperienza dei docenti della classe.

Ma aldilà della questione specifica degli esami di Stato, se riusciremo a migliorare l'attendibilità dell’accertamento degli apprendimenti e del controllo formativo e sommativo potremo avere uno strumento in più per ridurre il numero di quella parte di popolazione di giovani che sembra trarre poco vantaggio da molti anni di permanenza nelle aule scolastiche.

Il progetto è stato realizzato nell'anno 2001 sotto la direzione scientifica del prof. *Benedetto Vertecchi*.

Raimondo Bolletta Responsabile, Disegno sperimentale e Pianificazione delle procedure, Campionamento delle prove e dei correttori, Elaborazione dei dati.

Lina Grossi e Silvana Serra Messa a punto della griglia per la correzione della prima prova

Per la complessa realizzazione della codifica degli elaborati per il raggiungimento del campione dei correttori, per la diffusione dei documenti della raccolta e della registrazione dei dati ha operato, seppure in modo non esclusivo, lo staff dell'Osservatorio costituito da *Monica Amici* (coord. segretariale) *Cristina Crialesi, Caterina Ponzo, Emanuela Cuzzucoli Cecilia Carnevale e Maria Teresa Catanese.*