ANNAMARIA DE SANTIS

ANALISI MULTIVARIATA E LEARNING ANALYTICS

METODI E APPLICAZIONI

PREFAZIONI DI

PIER CESARE RIVOLTELLA

CORRADO CROCETTA

POSTFAZIONE DI

PIETRO LUCISANO

Pearson

ANNAMARIA DE SANTIS

ANALISI MULTIVARIATA E LEARNING ANALYTICS

METODI E APPLICAZIONI

PREFAZIONI DI

PIER CESARE RIVOLTELLA

CORRADO CROCETTA

POSTFAZIONE DI

PIETRO LUCISANO

Pearson

Grafica di copertina: da StoryBlocks *(www.storyblocks.com)*

Prima Edizione: Settembre 2022

ISBN: 9788891932419

Alla mia meravigliosa famiglia vagabonda che da sempre e per sempre, al di là dello spazio e del tempo, è la mia essenza e la mia forza.

INDICE

Capitolo 6 : CLUSTER ANALYSIS

Capitolo 7 : MULTIDIMENSIONAL SCALING

Capitolo 8 : ITEM ANALYSIS

PREFAZIONE

Pier Cesare Rivoltella

Alex Pentland (2015), in un libro importante di qualche anno fa, gettava le basi di una neoscienza, la fisica sociale. L'oggetto di questa scienza sono i dati, i Big Data, che essa studia grazie al contributo dell'informatica. L'importanza degli analytics si riconosce qui: essa consiste nel rendere possibile analizzare enormi basi di dati producendo in questo modo una discontinuità forte, per non dire un cambio di paradigma, all'interno della ricerca quantitativa. Infatti, quel che diviene possibile è prendere in considerazione l'intero universo di riferimento, senza più bisogno di procedere alla stratificazione campionaria. È la potenza del calcolo a renderlo possibile, con impatti sulla ricerca che solo in un prossimo futuro si potranno apprezzare.

Il libro di Annamaria De Santis riflette proprio su questo tema: il rapporto tra analytics e ricerca quantitativa nel caso specifico della ricerca educativa. Si tratta di una scelta importante e quanto mai opportuna. Infatti, il campo della ricerca educativa, soprattutto nel nostro Paese, soffre una drastica polarizzazione nel dibattito e nella pratica specialistici.

La prima polarizzazione è tra ricerca qualitativa e quantitativa, con una sensibile centratura soprattutto sulla dimensione qualitativa. Non sempre si tratta di una scelta metodologica. Spesso è il risultato della difficoltà a gestire il tecnicismo richiesto dalla ricerca quantitativa, una difficoltà che si spiega con l'assenza di una robusta formazione quantitativa nella maggior parte delle scuole di dottorato. Ma poi anche sull'impostazione metodologica dell'approccio qualitativo vi sarebbe da discutere, poiché anch'esso richiederebbe rigore e la messa in campo di strumentalità e attenzioni non scontate.

La seconda polarizzazione è tra ricerca Evidence Based e non Evidence Based. Questa polarizzazione nasce dall'importazione nel nostro Paese

dell'approccio e dei lavori di Hattie, con una lettura spesso radicale che non sempre rispecchia l'esatta posizione del ricercatore neozelandese. Con l'approccio EBE si è fatta la conoscenza degli studi di meta-analisi attribuendo a essi un valore predittivo enorme in relazione alla valutazione dei processi volta a volta oggetto di studio. Ancora una volta sul secondo estremo dell'asse di oscillazione si trovano gli studi non improntati a una logica EBE, nella maggior parte dei casi qualitativi.

Il risultato di questa doppia polarizzazione ha finito già per produrre molte allergie nei confronti della ricerca quantitativa e altre rischia di produrne in futuro. Per questo lo studio di Annamaria De Santis riveste un ruolo di particolare importanza. Esso offre al giovane studioso, ma anche al ricercatore affermato non quantitativista, l'opportunità di accostarsi a un campo di analisi oggettivamente complicato con la possibilità di ricavarne un'adeguata in-formazione, avviando al contempo la riduzione delle polarizzazione cui facevamo cenno.

L'attenzione va nello specifico all'analisi multivariata che viene presentata con competenza e chiarezza espositiva, grazie anche a una preziosa galleria di esemplificazioni e casi. Il risultato è un libro serio, rigorosissimo dal punto di vista metodologico, informato e aggiornato, corredato da un'ottima bibliografia. Una pista di lavoro che si auspica inauguri una serie di studi in grado, come questo, di rilanciare l'importanza della ricerca quantitativa in educazione.

Milano, luglio 2022

Pier Cesare Rivoltella
Università Cattolica del S. Cuore
Presidente della SIREM
Società Italiana di Ricerca sull'Educazione Mediale

Riferimenti bibliografici

Pentland, *Fisica sociale. Come si propagano le idee*. Roma: LUISS 2015.

PREFAZIONE

Corrado Crocetta

I metodi statistici hanno avuto un enorme sviluppo grazie alla crescente potenza di calcolo dei computer. Per assumere decisioni in condizioni di incertezza è necessario disporre di dati affidabili e di modelli adeguati alla complessità dei fenomeni analizzati.

La crescente disponibilità di dati rende possibile ottenere informazioni molto dettagliate su fenomeni in rapida evoluzione, ampliando il livello di conoscenza delle dinamiche sottostanti. Questo incremento esponenziale dei dati e degli strumenti di analisi non è privo di pericoli poiché la disponibilità di software sempre più user-friendly se da un lato ha facilitato l'accesso a questi strumenti a persone con background culturali molto diversi, dall'altro ha elevato il rischio che chi effettua le analisi non sia in grado valutare gli effetti dell'utilizzo di dati di bassa qualità o di ottenere risultati distorti.

Per questo motivo un manuale di statistica multivariata dedicato al tema specifico del learning analytics è particolarmente utile. In questo volume Annamaria De Santis fornisce gli elementi essenziali per esplorare vari tipi di dati che provengono da ambienti educativi per capire meglio gli studenti e i contesti in cui imparano. In questo libro, scritto in modo chiaro e comprensibile anche ai non addetti ai lavori, si mostra come utilizzando i dati in modo intelligente si possono ottenere modelli di analisi per scoprire informazioni e connessioni sociali, per predire e dare consigli sull'apprendimento.

Gli otto capitoli che compongono questo volume forniscono una panoramica dei metodi di statistica multivariata utili per chi intende utilizzare le tecniche di learning analytics cogliendo le connessioni più significative presenti in grandi banche dati correlate all'apprendimento o per chi vuole analizzare i diversi aspetti educativi per ottimizzare le opportunità di apprendimento degli allievi.

Bari, agosto 2022

Corrado Crocetta
Università degli Studi di Bari Aldo Moro
Presidente della SIS
Società Italiana di Statistica

INTRODUZIONE

Negli ultimi decenni lo sviluppo dell'utilizzo di ambienti online per la formazione ha fatto emergere la disponibilità di grandi quantità di dati in relazione ai processi educativi. Tale disponibilità ha immediatamente stimolato l'utilizzo dei dati per descrivere, conoscere e, in definitiva, poter migliorare l'apprendimento, l'insegnamento, la pianificazione e l'organizzazione della didattica.

Si sono affermate nuove aree disciplinari, per esempio quella dei *Learning Analytics* o dell'*Educational Data Mining* o dell'*Educational Big Data Analysis*. È stato un processo del tutto analogo a quello che ha fatto radicare una disciplina come la *Psicometria*.

Ma cosa hanno in comune questi "nuovi" ambiti disciplinari? La risposta è quanto mai semplice: l'applicazione di adeguati – a volte innovativi – approcci statistici a un preciso campo di indagine, quello educativo (in senso ampio) o quello psicologico-cognitivo (anche questo in senso ampio).

L'obiettivo di questo volume è quello di fornire un contributo per superare le "*separazioni*" disciplinari e mettere sullo stesso piano sia il contributo metodologico quantitativo sia quello applicativo in ambito educativo sollecitando l'avvicinamento di due comunità, quella degli statistici e quella dei pedagogisti, e quindi di due "*culture*" scientifiche e metodologiche. Non una via di mezzo ma un arricchimento reciproco.

L'invito è agli statistici a considerare l'ambito educativo come un ambito di interesse (dopo quello demografico, fisico, economico, finanziario, sociologico, psicologico, sanitario, biologico ecc.) vista sia la grande quantità di dati attualmente disponibili sia il valore e la complessità delle problematiche di un *com-*

plex social ground come quello dei processi di apprendimento, insegnamento, formazione, organizzazione delle istituzioni educative e così via.

L'invito è ai pedagogisti a rafforzare la cultura quantitativa, e quindi la comprensione e la capacità di utilizzo dei principali approcci di analisi dei dati, come integrazione e complemento della capacità di lettura e di analisi dei processi educativi (sempre in senso molto ampio).

Questa "*integrazione*" è la chiave di lettura di questo volume, in cui vengono introdotti gli assunti della ricerca quantitativa in ambito educativo e i learning analytics per poi passare a una rassegna (sicuramente non completa) di alcune tecniche di analisi dati. Cercando di mantenere sempre un equilibrio tra l'aspetto più "*tecnico*" con quello più "*applicativo*".

Nel primo capitolo viene delineato il legame esistente fra ricerca educativa, statistica e tecnologie nella duplice veste di strumenti di analisi e ambienti online in cui si generano dati. Le considerazioni fornite a partire da un'idea di ricerca che ha lo scopo di descrivere, spiegare, agire per porsi al servizio degli esseri umani sono completate con la sintesi delle principali caratteristiche dell'ambito di studi dei *learning analytics* e la definizione degli *open* e *big data*.

Al centro del secondo capitolo sono la definizione, le classificazioni e le fasi dell'analisi multivariata per identificare strutture latenti e relazioni tra le variabili. Essa rispecchia il tentativo, qualora applicata alla ricerca educativa, di sintesi e semplificazione della complessità dei fenomeni per identificare il filo rosso che li lega, genera senso e individua comportamenti e strategie che rendono efficaci apprendimenti e organizzazioni. Aprono il capitolo due approfondimenti sul *data screening*, la preparazione dei dati per l'analisi, e sulla *data visualization* che sfrutta grafici e visualizzazioni per rendere comprensibili fenomeni rilevati su grandi quantità di dati sia per un vasto pubblico che per gli studiosi.

Dal terzo capitolo in poi si procede con la descrizione delle specifiche tecniche di analisi con vari esempi di applicazione nella ricerca educativa.

Le prime ad essere affrontate sono le tecniche di riduzione della dimensionalità, tecniche che, passando da un elevato numero di variabili a un ridotto nu-

mero di dimensioni con una perdita minima di informazioni, sono in grado di esprimere la variabilità del dataset e definirne la struttura latente. Nello specifico, nel capitolo tre si illustrano l'*analisi delle componenti principali*, l'*analisi fattoriale esplorativa* e l'*analisi delle corrispondenze multiple* la cui selezione è effettuata dall'analista in base alle caratteristiche delle variabili (quantitative o qualitative), agli scopi della ricerca (inferenziali o descrittivi), alla tipologia di risultati attesi (numerici o grafici).

Due tecniche di dipendenza finalizzate a definire un modello statistico ossia una rappresentazione, una relazione fra le variabili sono presentate nel quarto e quinto capitolo. Esse sono la *regressione lineare multipla* (capitolo 4) e la *regressione logistica* (capitolo 5), tecniche nelle quali le variabili sono in un rapporto asimmetrico: il comportamento di una variabile dipendente è stimato/predetto a partire dall'andamento di un numero più alto di variabili indipendenti. La scelta dell'una o dell'altra tecnica dipende principalmente dalla tipologia di variabile dipendente: quantitativa nella prima, binomiale nella seconda.

Descriviamo nel capitolo 6 la *cluster analysis*, tecnica il cui scopo è quello di identificare gruppi omogenei (cluster) tra le unità statistiche. Il capitolo successivo è dedicato a una ulteriore tecnica di riduzione della dimensionalità, il *multidimensional scaling*, la cui presentazione è stata posposta per le sue implicazioni come supporto alle tecniche di classificazione.

L'ultimo capitolo del volume è dedicato all'*item analysis*, usata in ambito educativo prevalentemente per questioni di natura docimologica, volta a esaminare le prove di valutazione e stimare l'abilità degli studenti. Qui ne descriviamo l'approccio classico, Classical Test Theory, e probabilistico, Item Response Theory (capitolo 8).

La descrizione di ciascuna tecnica, presentata con un linguaggio generalmente comprensibile anche da chi non ha uno specifico background analitico e tecnico, è arricchita da esempi realizzati su dati reali raccolti nelle attività di formazione svolte presso il Centro Interateneo Edunova dell'Università degli Studi di Modena e Reggio Emilia.

Ampliano le spiegazioni numerosi casi di studio tratti dalla letteratura internazionale nei quali si mostra, talvolta con l'aiuto di tabelle e grafici, come le tecniche sono state utilizzate nella ricerca sui temi dell'educazione. Si tratta di casi individuati nella letteratura più recente e maggiormente citati su Scopus nell'ultimo decennio in riviste peer-reviewed del settore Education in lingua inglese. Gli esempi coprono uno scenario molto ampio sia in riferimento a questioni legate agli apprendimenti sia per questioni che hanno a che fare con la gestione delle organizzazioni formative.

Questo volume è solo un'introduzione, un percorso di formazione ai metodi dell'analisi multivariata per chi si occupa di social science e un pacchetto informativo sulle applicazioni in ambito educativo per chi conosce già le tecniche. Un educatore a cui capiterà nelle mani lo troverà probabilmente tecnico e spigoloso. Uno statistico lo vedrà semplicistico e impreciso. Esso non aspira ad essere esaustivo, né completo, ma prova a proporre un'opportunità che non dovremmo farci sfuggire e che deriva dall'incontro fra i settori dell'educazione, della statistica e dell'innovazione tecnologica: l'opportunità cioè di conoscere e guardare le questioni educative da un altro punto di vista, usando procedure standardizzate, prendendo in considerazione elementi specifici e misurabili (variabili) su piccoli campioni o intere popolazioni, costruendo modelli che possano stimare l'andamento dei fenomeni tenendo in considerazione l'incertezza e le regolarità che attraversano i dati.

CAPITOLO 1

LA RICERCA QUANTITATIVA IN EDUCAZIONE

Al termine del capitolo, il lettore sarà in grado di:

- *definire le principali caratteristiche della ricerca in educazione;*
- *delineare l'approccio probabilistico della ricerca quantitativa in educazione;*
- *spiegare differenze e complementarietà delle tradizioni di ricerca quantitativa e qualitativa;*
- *descrivere il ruolo delle tecnologie nella ricerca quantitativa nella duplice veste di ambiente di raccolta dei dati e strumento di analisi;*
- *fornire una definizione di open data, big data e data visualization;*
- *fornire una descrizione delle principali caratteristiche del software* `R` *per la programmazione e l'analisi dei dati.*

1.1 - Ricerca quantitativa in educazione

Se cerchiamo una definizione completa di ricerca educativa, quella fornita dall'American Educational Research Association (AERA) che pone l'attenzione sull'oggetto della ricerca, sugli scopi e sui metodi può fare al caso nostro:

> *"Education research is the scientific field of study that examines education and learning processes and the human attributes, interactions, organizations, and institutions that shape educational outcomes. Scholarship in the field seeks to describe, understand, and explain how learning takes place throughout a person's life and how formal and informal contexts of education affect all forms of learning. Education research embraces the full spectrum of rigorous methods appropriate to the questions being asked and also drives the development of new tools and methods."*
> *(https://www.aera.net/About-AERA/What-is-Education-Research)*

La definizione esplicita che il campo di studi della ricerca educativa comprende tanto i processi di apprendimento del singolo individuo in contesti formali e informali, quanto il funzionamento di organizzazioni e istituzioni che si occupano di formazione. Metodi rigorosi e appropriati alle domande di ricerca poste sono gli strumenti da utilizzare per fare in modo che questo tipo di indagine possa DESCRIVERE, COMPRENDERE e SPIEGARE i fenomeni educativi.

Non sempre il binomio "metodi rigorosi" e "ricerca educativa" è dato per scontato sia a livello di policy che di percezione da parte dell'opinione pubblica.

Ci viene facile pensare che un vaccino debba essere studiato in laboratorio attraverso l'uso di procedure rigorose e raffinati strumenti statistici di analisi, nessuno di noi si sottoporrebbe a una terapia medica se non fosse stata verificata in questo modo. Allo stesso modo, ci aspettiamo che la sperimentazione di un nuovo materiale per costruire le nostre case o produrre apparecchi tecnologici o gli studi degli investimenti o delle produzioni partano da rigorose procedure di indagine e analisi logico-matematiche di dati.

Più difficile pensare a procedure standardizzate quando si parla di ricerca nel contesto della formazione. Metodi rigorosi sia legati ad approcci qualitativi che quantitativi nella ricerca empirica sembrano non riuscire a cogliere la complessità dei comportamenti degli individui, dei gruppi e delle istituzioni nei percorsi educativi e formativi. Alcuni elementi supportano questa opinione: gli esperimenti non sono replicabili e generalizzabili se non con dovute accortezze; solo in condizioni ben determinate (talvolta in collaborazioni di ambito neuroscientifico o psicologico) possono essere svolti in laboratorio, più comunemente vengono agiti in ambienti "reali"; non si può escludere, e talvolta "misurare", l'influenza del contesto; non si può prescindere dai cambiamenti peculiari degli esseri umani nei comportamenti e nelle azioni. Donald Ary e colleghi (2010, p. 17) elencano alcuni limiti nell'applicazione di un approccio scientifico nelle scienze sociali:

- *Complexity of Subject Matter*: nella ricerca educativa, l'oggetto/soggetto della ricerca è l'essere umano (da solo o in gruppo) con i suoi comportamenti, abilità, stati emotivi, rapporti con gli altri e con gli ambienti. Nei comportamenti umani intervengono variabili che possono influenzare e modificare i fenomeni che stiamo osservando. Quanto osservato in un gruppo avrà una limitata validità per altri gruppi, con molta cautela si possono elaborare delle generalizzazioni.

- *Difficulties in Observation:* l'osservazione nella ricerca sociale rischia di essere meno obiettiva rispetto a quanto lo è nelle scienze naturali perché nell'osservazione e nell'interpretazione dei fenomeni rientrano fattori che possono far riferimento al mondo valoriale e conoscitivo del ricercatore.

- *Difficulties in Replication:* le esperienze educative possono essere replicate per essere osservate ma ci restituiranno risultati sovrapponibili solo in parte e in determinate condizioni.

- *Interaction of Observer and Subject:* la presenza di un ricercatore nel contesto indagato modifica la situazione, la percezione, le azioni da parte degli osservati.

- *Difficulties in Control:* non è possibile applicare un controllo rigido nello studio dei fenomeni educativi come se fossimo in laboratorio poiché c'è la probabilità che intervengano variabili di cui lo scienziato potrebbe anche non essere consapevole.

- *Problems of Measurement:* gli strumenti di misura non sono così precisi come nelle scienze naturali e i fattori che influenzano gli eventi potrebbero appartenere non solo al presente ma anche al passato. Bisogna provare per quanto possibile a controllare le variabili oggettive intervenienti: tempi, spazi, modalità d'azione, risorse umane e così via.

Conoscere i limiti dell'approccio scientifico alla ricerca educativa non significa rinunciare a perseguirlo. Al contrario, significa partire dalla conoscenza dei metodi di indagine per provare a controllare gli ostacoli: scegliendo teorie di riferimento e strumenti di lavoro ben definiti, monitorando gli ambienti e le condizioni in cui si sviluppa una ricerca, selezionando con accuratezza le variabili che possono essere osservate e misurate, ripetendo le misure e replicando gli studi, lavorando su campioni ampi e rappresentativi per diminuire le possibilità di errore e proporre generalizzazioni.

I ricercatori, a prescindere dal quadro ontologico o epistemologico di appartenenza, concordano sul fatto che la conoscenza sia influenzata dal contesto storico e sociale nel quale si svolge e dal quadro teorico e valoriale del ricercatore, la stessa raccolta dei dati è densa di teoria; le conoscenze e le teorie scientifiche non sono definitive ma possono e devono essere sempre riviste (fino anche falsificate) alla luce di nuovi scenari di ricerca e nuovi esperimenti

da proporre (Trinchero, 2002; Trichero & Robasto, 2018; Buscema & Pieri, 2004).

Le posizioni del realismo critico, che aspirano ad individuare relazioni, tendenze e regolarità nei fenomeni osservati partendo dalla consapevolezza che possiamo conoscere la realtà solo "in modo imperfetto e probabilistico" (Trinchero, 2002, p. 26), asseriscono che attraverso metodi qualitativi e quantitativi si possono raccogliere informazioni su una realtà stratificata, articolata, multipla.

Spiegando il senso della ricerca in ambito educativo sui tre livelli – individuo, scuola, istituzioni – nel rapporto con i decisori, Hans Fisher, William Boone e Knut Neumann (2014) aggiungono un altro tassello alla definizione di questo scenario: "*Researchers should be able to tell future teachers how they can increase the probability of their own teaching being of high-quality*". La frase fa riferimento solo a uno degli aspetti e degli scopi della ricerca educativa legato alle pratiche di insegnamento e definisce i caratteri della necessaria applicabilità delle ricerche nei contesti didattici quotidiani e della relazione fra gli attori che prendono parte all'indagine (per un approfondimento sui costrutti della ricerca didattica, si veda Fabbri, 2012). Nella semplicità con cui è formulata, l'aspetto su cui è bene soffermare l'attenzione è il modo in cui viene introdotta l'idea di probabilità i cui tratti contraddistinguono anche i risultati di ricerche svolte in maniera rigorosa in quanto la natura dell'indagine educativa così come la realtà osservata è probabilistica e transitoria. Per condurre ricerche attendibili, guardando a dati e risultati, dicono subito dopo gli autori, vanno considerati quattro criteri:

1. *Obiettività:* si riferisce alla riduzione delle influenze esterne nelle misure e nelle osservazioni e si realizza nelle ricerche quantitative predisponendo rigorosi step nella standardizzazione delle attività di ricerca, nell'uso di misure psicometriche e calcoli statistici di analisi dei dati. Più difficile valutare l'oggettività degli studi qualitativi dove diminuiscono le distanze fra osservatore e partecipanti. L'oggettività di uno strumento di indagine, aggiungono Pietro Lucisano e Anna Salerni (2002), riguarda anche la *"concordanza nella rilevazione da parte di diversi ricercatori (osservatori, siglatori, valutatori, giudici, correttori ecc.)"* (ivi, p. 151) al fine di evitare giudizi soggettivi.

2. *Affidabilità*: considera l'errore casuale presente in ogni misura e verifica la "costanza dei risultati ottenuti della rilevazione e dell'analisi, quando queste siano compiute da persone diverse, con strumenti diversi, in con-

dizioni diverse, ma a parità degli elementi che costituiscono l'oggetto di rilevazione" (Trinchero, 2002, p. 174). Si calcola solitamente come rapporto fra la varianza del valore vero e la varianza totale delle misure. Per verificare l'affidabilità di uno strumento, si usano la ripetizione della rilevazione delle misure (test-retest), la somministrazione e il confronto con i risultati ottenuti dall'uso di strumenti simili che condividono gli obiettivi (forme parallele), la verifica della correlazione fra i risultati ottenuti dividendo lo strumento in due parti (split-half) e il calcolo della coerenza interna come ad es. l'α (Alpha) di Cronbach (Lucisano & Salerni, 2002).

3. *Validità:* siamo soliti esaminare la validità di uno strumento di misura, verificando che esso rilevi ciò per cui è stato realizzato. Si parla quindi di: validità dei *contenuti* come rispondenza dello strumento agli assunti teorici sulla base dei quali è stato costruito; validità di *criterio* come confronto dei risultati ottenuti dallo strumento in esame con altre rilevazioni acquisite con strumenti diversi in maniera sincronica o nel tempo; validità di *costrutto*, come rapporto fra il costrutto teorico di partenza e i risultati misurati dallo strumento. A queste si aggiunge la validità di *aspetto* intesa come approvazione dello strumento da parte dei soggetti che saranno coinvolti nell'indagine. La validità non riguarda soltanto gli strumenti utilizzati ma l'intera ricerca che potrà essere detta valida se ogni sua parte risulterà tale: gli indicatori, i processi, gli strumenti, i risultati. Parliamo, quindi, ad esempio di validità delle conclusioni statistiche e dell'interpretazione dei dati riferendoci alla verifica della validità dell'*analisi dei dati*. Ancora a livello di intera ricerca si distingue fra la validità *interna* che analizza la rispondenza e la coerenza dei risultati ottenuti con gli obiettivi di ricerca e la validità *esterna* che prova la possibilità di generalizzare i risultati anche ad altri contesti se è definito chiaramente il contesto di partenza della ricerca e il campione è individuato in maniera corretta (Lucisano & Salerni, 2002; Trinchero, 2002; Cohen et al., 2007).

4. *Significatività:* si riferisce all'attendibilità dei risultati e trova spiegazione nell'ambito della verifica delle ipotesi, della statistica test e della stima del *p-value* nel contesto della statistica inferenziale. Serve per decidere se l'ipotesi statistica formulata è supportata dall'evidenza empirica e, quindi, può essere attribuita alle caratteristiche e alle relazioni studiate oppure se essa dipende dal caso. La procedura adottata prevede che sia formulata una ipotesi statistica detta *Ipotesi nulla,* H_0, che spesso rappresenta l'ipotesi che afferma che le relazioni o le caratteristiche del fenomeno indagato siano del tutto dovute al caso, e la sua complementare

detta *Ipotesi alternativa, H_1,* che, al contrario, è l'ipotesi che in effetti vorremmo verificare. Fissiamo α, il livello di errore che siamo disposti a tollerare indicando tipicamente come valori soglia 0,01, 0,05, 0,1. Questo corrisponde all'errore di rifiutare H_0 se vera ("errore di prima specie"), di rifiutare cioè che gli eventi siano dovuti al caso se fosse davvero così. Il livello di significatività del test, valore complementare pari a 1 – α, risulterà quindi pari a 90%, 95%, 99%.
Il p-value, probabilità che l'ipotesi nulla sia vera e che quindi i fenomeni siano dovuti puramente al caso, esprime la probabilità che la statistica-test cada nelle regioni critiche, ossia nelle regioni corrispondenti ad aree dell' 1%, 5%, 10% a seconda della nostra scelta del valore α. Calcolare quindi che il p-value sia inferiore al valore soglia considerato, ad es. 0,05, e di conseguenza affermare che il test rileva un livello di significatività del 95% significa dire che l'ipotesi H_0 può essere rifiutata e che quindi le relazioni rilevate non sono solamente frutto del caso (per una più ampia descrizione della significatività in termini didattici si veda Mecatti, 2015).

Questi criteri sono fondamentali per verificare l'attendibilità dei processi di ricerca e ci consentono di attuare scelte; come dice Donald E. Stockes (1997, p. 6), "*Research proceeds by making choices*". In tale processo di scelta che riguarda l'area di studio, i modelli e le teorie di riferimento, le ipotesi, il disegno della ricerca, gli strumenti di misura e di indagine, quelli di raccolta e analisi dei dati, la comunicazione dei risultati ed eventuali follow up, bisogna prendere in considerazione le possibilità, i pregi e i limiti, che ciascuno strumento di lavoro fornisce ponendo come centrale l'obiettivo della ricerca, unico che, come si legge nella definizione dell'AERA riportata all'inizio, può indicarci come disegnare gli step che ci porteranno a verificare le ipotesi di partenza. In un progetto di ricerca dobbiamo fare in modo che tutto concorra a rispondere alla domanda di ricerca formulata a priori. Proprio come quando progettiamo un intervento didattico, a guidare la scelta delle attività e delle metodologie didattiche, delle risorse e degli strumenti da utilizzare sono gli obiettivi formativi che rispondono alla domanda: cosa saprà/saprà fare lo studente alla fine di questo percorso?

Una delle scelte fondamentali in fase di disegno della ricerca è nell'uso di un approccio quantitativo o qualitativo, approcci distinti da metodi e strumenti di rilevazione dei dati e analisi degli stessi. L'uso dei due orientamenti di ricerca è spesso specchio di un differente paradigma di riferimento: i metodi quantitativi rispecchiano l'approccio razionalista-sperimentale e quelli quantitativi il

più recente approccio fenomenologico-costruttivista che, emerso a partire dagli ultimi decenni del Novecento, ha messo in discussione l'uso dei metodi delle scienze dure in educazione favorendo un'indagine dell'esperienza umana a partire dallo studio "dei contenuti della coscienza – quali i desideri, i ricordi, le percezioni e i significati personali" (Vannini, 2009, p. 7).

La Tabella 1.1 mostra le differenze fra i due approcci a proposito di finalità degli studi, disegno della ricerca, strumenti di indagine e di analisi, campione individuato.

Comparison of Quantitative and Qualitative Research		
	Quantitative	**Qualitative**
Purpose	To study relationships, cause and effect	To examine a phenomenon as it is, in rich detail
Design	Developed prior to study	Flexible, evolves during study
Approach	Deductive; tests theory	Inductive; may generate theory
Tools	Uses preselected instruments	The researcher is primary data collection tool
Sample	Uses large samples	Uses small samples
Analysis	Statistical analysis of numeric data	Narrative description and interpretation

Tabella 1.1 - Confronto fra approcci di ricerca quantitativo e qualitativo (Ary et al., 2010, p. 25).

La ricerca quantitativa permette di studiare le relazioni fra i fenomeni identificando interconnessioni o distinguendo caratteristiche che accomunano insiemi di soggetti intesi come singoli individui, gruppi, istituzioni. In quella qualitativa ci proponiamo di esaminare un fenomeno in profondità e interpretarne gli effetti.

Nel filone di ricerca qualitativo si prediligono strumenti di raccolta dei dati come l'osservazione e le interviste non strutturate somministrate a un numero ristretto di soggetti. Nel quantitativo, che è in grado di condurre a risultati più affidabili all'aumentare delle osservazioni raccolte, questionari altamente strutturati o inventory, ricognizioni di dati anagrafici, valutazioni, comportamenti, dati provenienti da strumenti tecnologici o navigazione di ambienti online.

I metodi statistici, utilizzati nell'approccio quantitativo, strumento necessario ed essenziale per numerose altre scienze e discipline − per le scienze naturali e sperimentali come la biologia, la medicina, la fisica e l'ingegneria, così come per quelle sociali come economia e sociologia −, ci danno una lettura ordinata dei fenomeni, ci portano a identificare variabili ben definite, ci mostrano modelli matematici e probabilistici che descrivono regolarità tendenziali nell'andamento delle distribuzioni e nel rapporto fra i fenomeni da indagare. Ci permettono di aggregare indicatori che tengono in considerazione più fattori, come la complessità dei tempi e delle modalità di apprendimento, delle questioni valutative; o i livelli di soddisfazione, le caratteristiche degli studenti e i meccanismi psicologici nell'espressione di un giudizio per attuare scelte consapevoli sulla base di risultati significativi (come in Cafarelli & Crocetta, 2016; Crocetta et al., 2016). Ci consentono di descrivere fenomeni attraverso gli strumenti della statistica descrittiva sia per singole variabili che per due o più; ci mettono in grado di riconoscere dipendenze e interdipendenze fra più variabili (analisi bivariata o multivariata) e di fare inferenze per estendere alla popolazione le osservazioni condotte su un campione rappresentativo. Con grafici e formule "ambiscono, aspirano" a mostrare la complessità dei fenomeni.

Le tecniche statistiche ci danno la possibilità di generare conoscenze aggregando un numero molto elevati di casi da studiare in fenomeni reali e desumendo leggi e regole generali per obiettivi specifici rispondendo alla necessità dell'uomo di razionalizzare comportamenti, elementi della vita culturale e collettiva, risultati mutevoli per poter poi agire e intervenire su essi e migliorare la realtà quotidiana (Piccolo, 2020). La statistica, come metodo scientifico e scienza metodologica, interviene, infatti, "in tutte le situazioni in cui occorre prendere decisioni in condizioni di incertezza" (ivi, p. 15) e usa nello studio di problemi reali tre criteri logico-concettuali: la *sintesi delle informazioni*, la ricerca cioè di un "indicatore riassuntivo di fatti complessi" (ivi, p. 25); la *scoperta del nuovo*, ossia la rilevazione di fatti e connessioni che non erano conosciuti o che vanno verificati; la *dialettica*, il fulcro dell'analisi con metodi statistici che prevede fasi di ricerca interattive e iterative nel ciclo di lettura dei risultati e formulazione di nuove ipotesi, nel rapporto fra dati, modello e teoria (*ibid.*).

Se i metodi quantitativi ci danno questo spaccato sugli eventi, dettagliato e rigoroso, i metodi qualitativi ci fanno cogliere i processi interpretandoli.

Il rischio di usare la ricerca quantitativa è di concentrarsi sui tecnicismi lasciando da parte il senso di realtà. Viceversa nel qualitativo si può enfatizzare l'interpretazione a scapito del rigore e dell'oggettività.

Nello scenario internazionale si assiste attualmente a un superamento dell'opposizione fra i due metodi a favore della valorizzazione dei punti di forza di ciascuno in una relazione di complementarietà: ciò a cui un approccio rigoroso nella ricerca educativa deve tendere è una strutturata analisi qualitativa e una realistica analisi quantitativa. Derivano da questa visione la formalizzazione di strategie di ricerca del *multi methods* e *mixed methods* che consentono l'uso di tecniche di triangolazione di tempi, spazi, teorie, ricercatori e metodi (Cohen et al., 2007, p. 141). I primi, *multi methods*, prevedono l'uso di strumenti e tecniche qualitative e quantitative su uno stesso tema di indagine ma in maniera del tutto indipendente. I secondi, *mixed*, prevedono una connessione stretta fra i due metodi in ogni fase di ricerca sin dalla pianificazione del disegno e dall'individuazione degli obiettivi e del quadro teorico. I metodi qualitativi e quantitativi possono essere utilizzati in maniera sequenziale o parallela come strumento di validazione e confronto nell'elaborazione dei risultati ottenuti dalle due procedure per rispondere a "l'obiettivo conoscitivo del ricercatore, in termini di studio in superficie o in profondità, monoprospettico o multiprospettico" (Trinchero & Robasto, 2018, p. 12).

Fenomeni circolari non vanno pensati solo fra approcci qualitativo e quantitativo all'interno della ricerca sperimentale, ma anche fra ricerca empirica e teorica. Una sorta di ricorsività, un circolo virtuoso, che permette alle conoscenze nei campi dell'apprendimento di divenire in ogni fase più complete e più ampie.

La formulazione di teorie scientifiche è il più grande obiettivo della scienza, scrivono Donald Ary e colleghi (2010). Nell'alternanza dei processi induttivi e deduttivi, di strumenti qualitativi e quantitativi, di ricerche empiriche e teoriche, i fenomeni vengono osservati e descritti, le ipotesi formulate a partire dalle teorie e verificate nelle esperienze pratiche, i risultati analizzati e contestualizzati. Le teorie "*(1) organize empirical findings and explain phenomena, (2) predict phenomena, and (3) stimulate new research*" (ivi, p. 14). Gli autori utilizzano l'esempio degli studi sulla relazione fra la zanzara Anopheles e la malaria negli esseri umani per mostrare l'influenza che gli studi scientifici possono, e anzi devono, avere nella pratica quotidiana: scoperto il ruolo della zanzara nella trasmissione della malattia, gli scienziati hanno potuto spiegare la presenza endemica della malaria in alcune aree, predire come i cambiamenti nell'ambiente avrebbero influenzato la diffusione della malattia, controllare la diffusione della malaria modificando l'ambiente.

DESCRIVERE, COMPRENDERE e SPIEGARE, abbiamo letto nella definizione di ricerca educativa dell'AERA.

SPIEGARE, PREDIRE e CONTROLLARE, scrivono Donald Ary e colleghi (2010, p. 15) parlando di teorie scientifiche e della zanzara Anopheles.

Le due triplette si susseguono, completano il quadro degli obiettivi della ricerca educativa e delineano i tratti costitutivi di un approccio investigativo che si mette a servizio del genere umano. Ci dicono anche perché l'uso della statistica che descrive, spiega e predice è parte di questi processi. Nella definizione del ruolo della statistica come scienza, Domenico Piccolo (2000) usa tre verbi: VEDERE, CAPIRE, AGIRE. Sono azioni semanticamente sovrapponibili a quelle elencate finora che confermano l'idea dell'uso dei metodi statistici come metodo scientifico nei processi di ricerca anche in ambito educativo:

> *"La statistica è quindi scienza che aiuta a vedere il mondo, a fotografarlo a catalogarlo e classificarlo cogliendo, misurando ed esplicitando l'essenziale, il caratterizzante, ciò che fa la differenza. […] capire la realtà all'interno di una impostazione probabilistica nella quale l'esistente è esaminato in rapporto a ciò che poteva accadere e che verosimilmente accadrà. […] agire per raggiungere scopi predefiniti e consentire la fruibilità del mondo, della società, delle risorse e delle opportunità che si offrono." (ivi, p. 27)*

Un approccio quantitativo alla ricerca in ambito educativo valorizza l'interdisciplinarità in primo luogo nella formazione dell'*educational researcher*, in secondo luogo nello svolgimento delle ricerche. L'*educational researcher* ha bisogno di una formazione interdisciplinare, aperta e attenta a tutte le possibili metodologie di ricerca e in grado di comprendere approcci statistici. La comprensione non sempre corrisponde alla capacità di utilizzare tali tecniche in autonomia nelle pratiche di ricerca data la loro complessità. Per questo motivo, le ricerche in ambito educativo devono essere condotte da team di ricercatori con specializzazioni diverse, dalla pedagogia alla statistica, dalla psicologia all'informatica e così via, a seconda degli obiettivi e dei temi delle indagini su cui si lavora. La complessità dei soggetti/oggetti della ricerca in ambito educativo richiede molteplici competenze e punti di vista disciplinari per aspirare alla completezza.

Riprendendo Pier Cesare Rivoltella (2018):

> *"nella complessità in cui viviamo, se anche si provasse a esaurire un problema nel campo dell'educazione con un unico sguardo monodisciplinare, ci si voterebbe probabilmente al fallimento. I temi e*

> *gli oggetti di ricerca sono sfaccettati, indagabili a più livelli, richiedono un punto di vista a sua volta complesso, tanto più proprio in quanto l'oggetto è "ipercomplesso". In questa situazione il caso dello "scienziato poliglotta" è difficile da immaginare: dovrebbe governare troppi saperi in modo specialistico. Così l'idea di una strutturale apertura interdisciplinare dei saperi diviene la più praticabile. Essa comporta il riconoscimento da parte di ogni disciplina di farsi portatrice di una razionalità limitata e la conseguente configurazione del lavoro di ricerca nel senso del lavoro di équipe."*

In questo volume ci concentriamo sull'applicazione dei metodi della statistica multivariata nella ricerca educativa al fine di corroborare l'utilizzo di pratiche di ricerca rigorose e interdisciplinari nel settore educativo.

1.2 - Tecnologia e ricerca educativa

Non è unica la strada che collega la ricerca scientifica e le innovazioni tecnologiche dei processi e dei prodotti. Strade lineari sono quelle che partendo dalla ricerca di base e passando per la ricerca applicata, ci portano a definire nella ricerca industriale nuove tecnologie utilizzabili da tutti. Questa linearità si basa sull'idea che "la ricerca scientifica e l'innovazione tecnologica sono come la lingua parlata e la lingua scritta di ogni cultura. La prima scopre nuove terre, la seconda costruisce strade e ponti perché tutti le possano abitare" (Buscema & Pieri, 2004, p. 19). Siamo consapevoli, tuttavia, che esistono strade alternative come quelle in cui le tecnologie progrediscono e si evolvono in maniera indipendente dai progressi nella ricerca. Così come sono tante le strade a doppio senso nelle quali le conoscenze e i progressi si costruiscono su un doppio binario: "*from scientific discovery to technological innovation*" e viceversa, "*from technology to science*" (Stockes, 1996, p. 20). L'influenza che i due campi di sviluppo hanno l'uno sull'altro ci fa parlare quindi di una tecnologia *science-based* e di una scienza *technology-derived* (*ibid*.).

Il settore dell'educazione, nel corso del tempo, è stato in grado di utilizzare per la formazione – senza contribuire alla loro creazione – tecnologie che non sono nate prettamente per la scuola, pensiamo alla scrittura o ai libri: "*Tools and technology, in their broadest sense, are important drivers of education, though their development is rarely driven by education*" (Laurillard, 2012, p. 2).

Lo sviluppo tecnologico, soprattutto dell'ambito digitale, ha accelerato la sua corsa e il mondo dell'educazione in qualche modo prova a integrare queste tecnologie, come ha fatto altre volte, e personalizzarle in base alle necessità formative, incidendo anche sulle caratteristiche e sulle modalità di funzionamento delle stesse. Pensiamo in questo senso a tutto il filone di ricerca educativa che si occupa di digitale, di formazione e ambienti online, videogame, app, ausili per le disabilità o i disturbi di apprendimento e così via.

Il contributo che lo sviluppo tecnologico può dare nel settore educativo è molto più ampio di quanto possa sembrare, soprattutto se collegato all'ambito della ricerca e in particolare all'uso dei metodi quantitativi di raccolta e analisi dei dati, non solo per i temi di indagine legati al digitale. Possiamo infatti pensare agli ambienti e agli strumenti digitali almeno da due punti di vista:

1. sono ambienti in cui si producono/archiviano i dati;
2. sono ambienti di programmazione e analisi.

Ambienti digitali in cui si producono/archiviano i dati

Negli ambienti online sono disponibili numerose banche dati che contengono dataset che possono essere utilizzati nelle indagini sui sistemi educativi. Alcuni esempi sono i portali che mettono a disposizione dati in modalità *open* che riguardano, sempre riferendoci al settore dell'educazione, l'anagrafica degli studenti, l'organizzazione scolastica e universitaria, le attività di formazione formale, non formale e informale nel tempo ecc. Questi cosiddetti *open data* permettono a chiunque ne abbia le competenze di lavorare direttamente sui dati per stabilire relazioni e trarne informazioni utili. Il Portale Unico dei dati della scuola (https://dati.istruzione.it/opendata/) o il Portale dei dati dell'istruzione superiore (http://dati.ustat.miur.it/) rilasciati dal Ministero dell'Istruzione e dal Ministero dell'Università e della Ricerca sono degli esempi che contengono oltre ai dataset scaricabili anche delle prime rielaborazioni. Su più ampia scala, i portali ISTAT, EUROSTAT, OCSE (per citarne alcuni) forniscono dati utilizzabili in ricerche (per esempio storiche e/o comparative) sulle caratteristiche degli istituti scolastici e universitari, sull'incremento dei numeri degli studenti, sui docenti, su età, genere, titolo di studio, abbandoni, studenti con BES ecc. Troviamo online pubblicati ugualmente in modalità *open* archivi di dataset su molte tematiche come l'European Data Portal (https://www.europeandataportal.eu/) o report di enti nazionali o internazionali sul sistema scolastico come i report annuali "Education at a glance" dell'OCSE.

Il concetto e le pratiche relativi agli open data si sono diffuse a partire dal 2009 in seguito ad iniziative di apertura delle informazioni pubbliche da parte di alcuni governi come gli Stati Uniti d'America, il Regno Unito, il Canada e la Nuova Zelanda. Rendere open i dati semplifica la gestione dei sistemi e genera conoscenza e servizi, dato che potenzialmente chiunque può accedere alle informazioni condivise.

La definizione di open data, che rientra nel più ampio contesto dell'*openess* e della *open definition*, può essere sintetizzata da tre aspetti:

1. *availability and access*: i dati devono essere disponibili preferibilmente mediante la rete in formati modificabili;

2. *re-use and redistribution*: i dati devono poter essere riusati, ridistribuiti e combinati con altri dataset;

3. *universal participation*: i dati devono poter essere riusati da tutti, senza discriminazioni; non è ammessa la clausola del riuso senza fini commerciali.

Queste e altre informazioni sono pubblicate dall'Open Knowledge Foundation (okfn.org) nelle pagine dedicate all'"Open Data Handbook" (opendatahandbook.org). Altro istituto che si occupa di questi temi è l'Open Data Institute (ODI), organizzazione no-profit fondata nel 2012 che ha lo scopo di collaborare con le aziende e i governi "per costruire un ecosistema di dati aperto e affidabile". Fra i fondatori dell'ODI c'è Tim Berners-Lee, conosciuto come inventore del web, il quale ha proposto una classificazione a 5 stelle per gli open data (5stardata.info, www.w3.org/2011/gld/wiki/5_Star_Linked_Data). Ciascuno dei 5 livelli presuppone che il precedente sia stato raggiunto. Si parte da un livello base, ad una stella, in cui è sufficiente che i dati siano disponibili sul web, in qualsiasi formato. Si aggiunge poi una stella per la possibilità di leggere i dati attraverso un software che permetta di rielaborarli. Una stella in più se il software non è proprietario. Il quintetto di stelle è completo quando si introducono gli standard aperti del W3C (WordWideWebConsortium) che prevedono anche l'implementazione di sistemi per collegare dati e database fra loro.

Tuttavia, parlare degli open data nel rapporto fra tecnologie ed educazione è solo un primo step.

Lo sviluppo tecnologico ha aperto altre strade e altri canali per realizzare percorsi di apprendimento o fruire di risorse didattiche conducendo a una ricca riflessione sulle modalità di apprendimento e progettazione dei corsi online

anche nel contesto italiano (Rivoltella, 2022; Piras et al., 2020; Sancassani et al., 2019). È proprio in questi nuovi ambienti digitali per la formazione che abbiamo la possibilità di collezionare una quantità di dati che fino a pochi anni fa non avremmo mai immaginato di raccogliere in un contesto didattico fisico. L'esempio più palese sono i *file log* che registrano ogni evento collegato alla navigazione di pagine nel web, ogni nostro click cioè, registrando la data in cui l'evento si è realizzato, il tipo di evento, l'utente che l'ha attivato.

A partire dall'ambito industriale ed economico si parla di *big data*, sia riferendosi al grande numero di dati registrati, sia alle tecnologie che gestiscono l'immagazzinamento e l'analisi degli stessi. Differentemente dagli *small data* tradizionalmente raccolti in forme strutturate e con lunghi tempi di generazione e analisi, i big data vengono generati continuamente in grandi numeri e in modalità flessibili coprendo intere popolazioni e non solo campioni selezionati (Kitchin & McArdle, 2016).

Olshannikova e colleghi (2016) definiscono tre grandi sfide/difficoltà che riguardano i dati (*data challenge*); l'elaborazione e dunque la raccolta dei dati, l'adattamento a un formato utile all'analisi, l'analisi stessa e la visualizzazione dei risultati secondo modalità più semplici per la comprensione umana (*processing challenge*); la gestione riferita all'archiviazione sicura dei dati nelle fasi di raccolta ed elaborazione (*data management challenge*). Parlando della *data challenge*, l'autore riprende le caratteristiche che Yuri Demchenko (2013a) individua come le *5 V* ossia *Volume*, *Velocity*, *Variety*, *Value* e *Veracity* (Figura 1.1). Il Volume è la caratteristica che naturalmente contraddistingue i big data; ha a che vedere proprio con la quantità di eventi osservati e rilevazioni immagazzinate e richiede che i dati siano "*accessible, searchable, processed and manageable*". La Velocità ci restituisce l'informazione sui tempi di produzione di questi dati: sono generati e vanno processati in maniera molto veloce (*real-time* o *streams*). Con Varietà si fa riferimento alla complessità dei formati dei dati − da strutturati a misti − che aumenta quando osserviamo sistemi biologici, umani e sociali e di conseguenza richiede meccanismi dinamici di archiviazione. Il Valore è quello che i dati aggiungono ai processi e alle attività osservate. Conclude l'elenco la Veridicità che può essere rilevata con l'affidabilità statistica e si basa sull'attendibilità dei dati in base alla loro origine, ai metodi di elaborazione, alle infrastrutture di archiviazione.

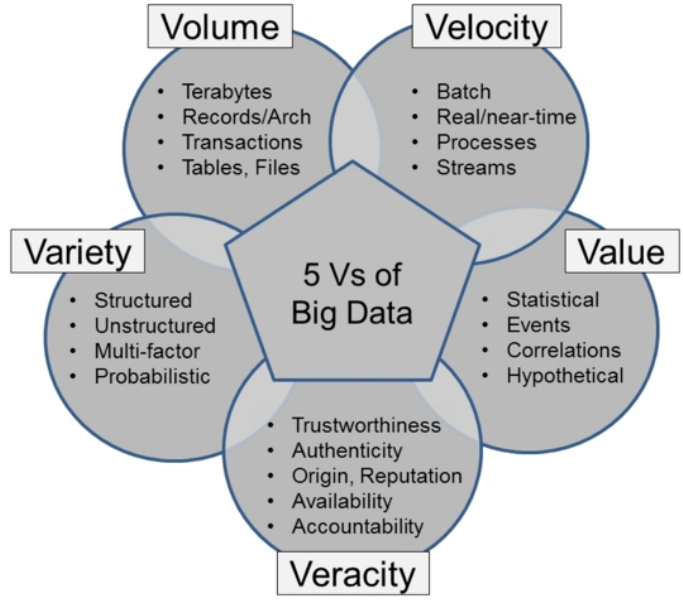


Figura 1.1 - Le 5 V dei big data (Demchenko, 2013a, p. 2/8).

Le parole che iniziano con la *V* sembrano rispecchiare molte delle caratteristiche dei big data: è bene sapere che in uno scritto del 2001 di Doug Laney, le *V* erano solo tre e si limitavano a *Volume*, *Veracity* e *Variety*. Il numero di *V* nel tempo è continuato a crescere e modificarsi: solo per fare qualche esempio a dimostrazione del fatto che siamo in un settore ampio di cui ancora vanno tracciati chiari confini (Kitchin & McArdle, 2016), sono sette per Uddin e Gupta (2014), dieci per Khan e colleghi (2018), 42 nel settore sanitario per Bahri e colleghi (2018), 56 per Hussein (2020). Dato che il tipo di procedure che ruotano attorno ai big data coinvolgono meccanismi di produzione, archiviazione, elaborazione dei dati che si discostano da quelle conosciute fino ad oggi e richiedono la ridefinizione di una intera architettura di gestione delle informazioni, Yuri Demchenko (2013b) ne parla come di un ecosistema.

Ambienti digitali di programmazione e analisi

Fino a tre decenni fa le analisi che prevedevano l'uso di metodi statistici articolati richiedevano tempi molto lunghi. Schede perforate sulla programmazione e sui database venivano inviate a computer spesso collocati anche a km di distanza. Il ricercatore doveva aspettare alcune ore per ricevere i risultati, inter-

pretarli perché non sempre le informazioni erano restituite in modalità lineare e sperare di non aver commesso nessun errore per concludere l'operazione (de Lillo et al., 2007). Oggigiorno ci torna difficile anche solo immaginare quali dovessero i tempi (e l'impegno) necessari a calcolare semplicemente la deviazione standard su un campione di 100 osservazioni prima ancora delle schede perforate. Questo perché per condurre questo tipo di analisi oggi abbiamo a disposizione non solo software ma app gestibili da un dispositivo mobile che teniamo in tasca o nella borsa. In questi sistemi la scrittura di una stringa ci restituisce in tempo reale risultati di calcoli che in passato richiedevano innumerevoli ore di lavoro.

Solo a scopo esemplificativo, nella Figura 1.2 riportiamo quello che utilizzando `R`, uno dei più noti ambienti per l'analisi statistica dei dati, riusciamo a "far dire" a una stringa. Digitando `pairs.panels(meta_exams)` dove `meta_exams` è il dataset che raccoglie 111 osservazioni e 7 variabili, questa funzione ci restituisce un'immagine nella quale sulla diagonale troviamo gli istogrammi che descrivono l'andamento delle singole variabili, sopra la diagonale i valori di correlazione di Pearson, sotto la diagonale diagrammi di dispersione bivariata con media e deviazione standard. Si tratta di un esempio elementare con calcoli e grafici di base che, se da un lato, ci dimostra quanto questi strumenti di analisi e programmazione siano potenti, dall'altro ci permette di sottolineare il processo articolato di azioni e competenze che porta alla scrittura di una sola stringa. Per le azioni: la definizione di un obiettivo di ricerca; la raccolta dei dati partendo dallo strumento di raccolta (magari un questionario o archivi con dati anagrafici e personali); la predisposizione del dataset secondo gli standard che il software possa riconoscere; la selezione della singola funzione da usare fra molte. Per le conoscenze/competenze: di ricerca, per pianificarne le fasi e gli strumenti; disciplinari, dell'ambito della formazione e della didattica; statistiche, per l'analisi dei dati; operazionali, nell'uso dei software.

`R` (con la sua interfaccia grafica `R Studio`) è un software open source che può essere installato sul proprio pc o usato via browser per creare e archiviare i propri documenti di lavoro. Sono integrati in `R`, `R Markdown` e `Knitr` che permettono di produrre direttamente dal sistema documenti e presentazioni delle ricerche. Essendo un programma modulare può essere integrato con pacchetti che contengono funzioni specifiche per determinati ambiti di ricerca. Alcuni esempi sono `psych` per la ricerca psicometrica o `ltm` per le analisi nell'ambito dell'item analysis. `R` è un prodotto open e gratuito, ma potremmo aggiungerne molti altri proprietari come `SAS`, `SPSS` o `MatLab`. Per tutti esisto-

no guide free, tutorial, esperti che rispondono nelle community/forum online che si raccolgono attorno a questi strumenti.

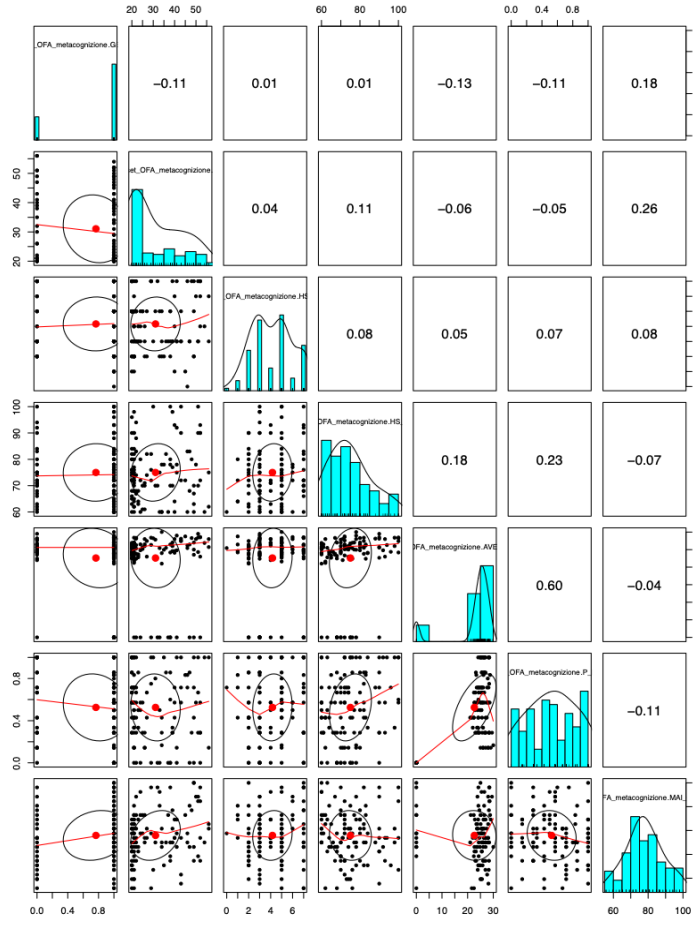


Figura 1.2 - Un esempio della rapidità e complessità con cui lavorano i programmi di analisi e rielaborazione dati. L'immagine è stata generata in R con l'uso della funzione `pairs.panels` (libreria: `psych`) per un dataset composto da 7 variabili e 111 osservazioni. Gli istogrammi sulla diagonale descrivono l'andamento delle singole variabili, sopra la diagonale visualizziamo i valori di correlazione di Pearson fra coppie di variabili, sotto la diagonale sono presenti i diagrammi di dispersione bivariata con media e deviazione standard.

L'esempio in Figura 1.2 ci offre l'occasione per aggiungere un altro tassello al quadro generale di sviluppo nel settore dell'analisi dei dati attraverso le tecnologie digitali. Riguarda il processo di costruzione di grafici e mappe che, pur

avendo radici molto remote, è reso possibile oggi da strumenti operativi che permettono, come abbiamo visto, con pochi click di creare rappresentazioni grafiche di grandi quantità di dati.

Nell'ambito della statistica computazionale, che mette insieme informatica e statistica, ha preso piede negli ultimi 30 anni l'area multidisciplinare di studio e interesse dedicata alla *data visualization* che ripropone in forma grafica (grandi quantità di) dati con due scopi: quello di presentarli in maniera chiara e dettagliata rendendoli più rapidamente comprensibili a un vasto pubblico (*graphics for presentation*) e quello per gli esperti di esplorare i dati per cercare risultati e trarre conclusioni (*graphics for exploration*) (Unwin, Chen & Härdle, 2008). Scaling images e interactive filtering sono alcune delle tecniche per lavorare sull'interattività delle visualizzazioni, nelle quali riuscire rispettivamente a visualizzare in maniera dettagliata solo una parte dell'immagine oppure selezionare solo un sottocampione rispetto al quale visualizzare i risultati (Olshannikova et al., 2016).

Continueremo a parlare di data visualization più approfonditamente nel capitolo 2.

1.3 – Learning Analytics

Nello scenario internazionale, l'ambito di ricerca che ha messo insieme dati e apprendimento è riconosciuto nel termine *Learning Analytics* (LA). Siamo dinnanzi allo sviluppo di un campo di studio che accosta alla ricerca educativa sui temi dell'apprendimento, della didattica, della valutazione, dell'uso delle tecnologie, gli *analytics* con procedure di analisi finalizzate alla definizione di modelli predittivi e lo *human-centered design* attento alle istanze dell'usabilità e della partecipazione (www.solaresearch.org/about/what-is-learning-analytics). Ma anche la conoscenza del web, la statistica avanzata, la data visualization, la psicologia, la linguistica, la filosofia e molto altro ancora.

Da circa un decennio è stato costituito un network interdisciplinare, una community dal nome Society for Learning Analytics Research (SoLAR, www.solaresearch.org) che raccoglie studiosi, professionisti ed enti nell'analisi dei dati su insegnamento, apprendimento e ambienti educativi. Il network organizza la conferenza internazionale LAK, Learning Analytics and Knowledge, e la summer school LASI, Learning Analytics Summer Institute, e gestisce la rivista "Journal of Learning Analytics" (learning-analytics.info).

La definizione più frequentemente usata per parlare di Learning Analytics, adottata da SoLAR nella call per LAK11 (Learning Analytics and Knowledge Conference 2011), la prima conferenza internazionale sul tema, recita:

> *"Con Learning Analytics ci si riferisce alla misurazione, alla raccolta, all'analisi e alla presentazione dei dati sugli studenti e sui loro contesti, ai fini della comprensione e dell'ottimizzazione dell'apprendimento e degli ambienti in cui ha luogo." (Ferguson, 2014, p. 139).*

In essa troviamo le quattro azioni distintive dei LA (misurare, raccogliere, analizzare, presentare), la tipologia di dati (sugli studenti e sugli ambienti), gli scopi (comprendere e ottimizzare).

Esistono molte altre definizioni, fra le quali una più recente propone i LA come il campo in cui "*big data in education meets conventional quantitative approaches*" (Srinivasa & Kurni, 2021, p. V). I LA rispondono alla sfida *educativa* legata ai big data: "come possiamo ottimizzare le opportunità di apprendimento?" (Ferguson, 2014, p. 144). Nonostante le inevitabili sovrapposizioni, settori di studio affini come l'*Educational Data Management* e l'*Academic Analytics* si riferiscono piuttosto rispettivamente alla sfida *tecnica* e alla sfida *politica*: "Come possiamo estrarre valore da questi grandi insiemi di dati correlati all'apprendimento? [...] Come possiamo migliorare sensibilmente le opportunità di apprendimento e i risultati scolastici a livello nazionale o internazionale?" (*ibid.*).

Nel capitolo introduttivo della seconda edizione dell'"Handbook of Learning Analytics" realizzato nell'ambito delle iniziative di SoLAR, per rispondere alla domanda "What is Learning Analytics?", Charles Lang e colleghi (2022) invece di usare una definizione, scelgono di spiegare attraverso quattro dimensioni il binomio LA poiché esso è aperto a innumerevoli interpretazioni come lo sono i due termini che lo compongono.

I LA sono:

1. *una preoccupazione*, di dare senso nel rispetto dell'etica, della privacy, dell'equità, dell'usabilità, alla grande quantità di dati che sono generati nei contesti dell'insegnamento e dell'apprendimento;

2. *un'opportunità*, di usare i dati prodotti in grandi quantità, ad esempio all'interno di Learning Management Systems, per aumentare la conoscenza sull'apprendimento, la comprensione dei docenti sui comporta-

menti degli studenti e la consapevolezza degli studenti stessi sui propri percorsi;

3. *un campo di indagine*, che nonostante abbia ancora confini sfumati per temi e metodologie, indaga tecnologie, dati ed educazione;

4. *una comunità*, o forse un insieme di comunità il cui interesse risiede nei dati e nell'apprendimento e di cui SoLAR è probabilmente la più nota.

I framework che sin dai primi anni di diffusione del campo di studi sono emersi (Greller & Drachsler, 2012; Chatti et al., 2012; Khalil & Ebner, 2015) si focalizzano su un pool di elementi:

- il tipo di dati raccolti: tracce lasciate dagli studenti nella partecipazione ad attività formative, nelle interazioni sociali, i dati personali e le informazioni sul percorso accademico;

- le tecnologie e gli strumenti di analisi (statistica, data visualization, social network analysis e così via);

- gli obiettivi degli studi: riflettere, analizzare, monitorare, predire, intervenire, personalizzare e adattare, gestire tutorato e valutazione;

- gli stakeholder coinvolti quali studenti, docenti e tutor, ricercatori, istituzioni, sia come produttori di dati che come coloro che possono leggerli per valutare le proprie prestazioni (nell'apprendimento da studenti attivando processi metacognitivi e nell'insegnamento da docenti), fornire/ottenere feedback in tempo reale, mettere in atto processi decisionali *data-driven*;

- le competenze necessarie da parte di ricercatori e analisti all'attuazione di tali pratiche di analisi e da parte degli stakeholder alla lettura dei risultati;

- le teorie educative che in maniera implicita ed esplicita fanno da sfondo alla realizzazione dei percorsi formativi nei quali i dati sono ottenuti e raccolti e ai processi di analisi;

- i vincoli legati alla gestione dei dati in termini etici, di privacy e di proprietà che richiedono per la trasparenza procedure standardizzate di comunicazione (informative) con i soggetti coinvolti e affidabili sistemi di archiviazione dei dati (Drachsler & Greller, 2016; Bellini et al., 2019).

Strumento particolarmente usato per restituire agli stakeholders i dati raccolti e analizzati sono le *learning dashboard*:

> *"A learning dashboard is a single display that aggregates different indicators about learner (s), learning process(es) and/or learning context(s) into one or multiple visualizations."*
> *(Schwendimann et al., 2017, p. 37)*

Tali visualizzazioni pensate innanzitutto per docenti e studenti, per il monitoraggio da parte dei primi e l'autoregolazione nei propri percorsi formativi dei secondi, possono raccogliere da una o più piattaforme dati riferibili ai log, alle attività e questionari svolti, a database istituzionali restituendo grafici, mappe, icone e tabelle.

CAPITOLO 2

INTRODUZIONE ALL'ANALISI MULTIVARIATA

Al termine del capitolo, il lettore sarà in grado di:

- *descrivere le fonti di dati in ambito educativo;*
- *descrivere le principali procedure di data screening;*
- *descrivere principi e finalità della data visualization;*
- *definire l'analisi multivariata;*
- *elencare le fasi di realizzazione di un'analisi multivariata;*
- *classificare le tecniche di analisi multivariata.*

2.1 - Dati, variabili, tecniche di rilevazione e campionamento

I dati su cui si può lavorare utilizzando metodi statistici sono i più diversi. O meglio i contesti e i temi di indagine sono diversi ma i dati hanno sempre le stesse caratteristiche. Nei procedimenti di calcolo lavoriamo su numeri e valori da descrivere e mettere in relazione. Ciò che "adegua" l'uso dei metodi alla disciplina (in questo caso il settore educativo) è la scelta dei fenomeni da osservare e, a monte, delle ipotesi e degli obiettivi di ricerca.

In ricerche tese a descrivere e confrontare enti formali e non formali che si occupano di formazione, collezioneremo dati che fanno riferimento alle strutture e alla loro organizzazione, raccogliendo informazioni che hanno a che vedere con i numeri di membri, le caratteristiche anagrafiche degli operatori intesi come educatori e insegnanti, le statistiche legate alle disabilità e così via.

Riferendoci invece ai processi di apprendimento, possiamo studiare comportamenti e rilevare dati personali, conoscenze, abilità, opinioni, stili e modi di fare dei soggetti.

Come già detto nel primo capitolo, negli ambienti digitali siamo in grado di rilevare le azioni dei soggetti collezionando big data di cui sono un esempio i log di navigazione in una piattaforma didattica (Figura 2.1). In essi è registrato

ogni click dell'utente insieme agli altri utenti coinvolti, all'ora e all'ambiente in cui l'attività si è svolta.

Data/Ora	Nome completo dell'utente	Utente coinvolto	Contesto dell'evento	Componente	Evento	Descrizione	Origine	Indirizzo IP
13 luglio 2022, 15:25	Uno Edunova	-	Quiz: Quiz 3	Quiz	Visualizzato modulo corso	The user with id '6' viewed the 'quiz' activity with course module id '42'.	web	155.185.208.62
13 luglio 2022, 15:25	Uno Edunova	Uno Edunova	Quiz: Quiz 3	Quiz	Inviato tentativo quiz	The user with id '6' has submitted the attempt with id '213' for the quiz with course module id '42'.	web	155.185.208.62
13 luglio 2022, 15:25	Uno Edunova	Uno Edunova	Quiz: Quiz 3	Quiz	Visualizzato riepilogo del quiz	The user with id '6' has viewed the summary for the attempt with id '213' belonging to the user with id '6' for the quiz with course module id '42'.	web	155.185.208.62
13 luglio 2022, 15:24	Uno Edunova	Uno Edunova	Quiz: Quiz 3	Quiz	Visualizzato tentativo quiz	The user with id '6' has viewed the attempt with id '213' belonging to the user with id '6' for the quiz with course module id '42'.	web	155.185.208.62

Figura 2.1 - Esempio di Log in una piattaforma Moodle.

Una tra le forme di raccolta dati più note e più diffuse è il questionario che, insieme a colloqui e interviste, permette di raccogliere opinioni e percezioni da parte degli intervistati, ma anche esaminare l'acquisizione di conoscenze o le performance. Distinguiamo questionari di valutazione sommativa predisposti nei corsi nelle forme di esami; i test standardizzati somministrati a livello nazionale e internazionale soprattutto con finalità comparative (es. PISA); test o scale, inventory validate nei contesti di ricerca per misurare, ad esempio, il possesso di abilità metacognitive (es. MAI, Metacognitive Awareness Inventory – Schraw & Dennison, 1994), i livelli di self-regulation nello studio (es. OSLQ, Online Self-regulated Learning Questionnaire – Barnard et al., 2009), le reazioni in presenza di un nuovo strumento tecnologico da utilizzare (es. CES, Computer Emotion Scale – Kay & Loverock, 2008) e così via. Quando pensiamo a questa tipologia di strumenti di ricognizione dei dati, immaginiamo che i questionari, al di là degli argomenti che affrontano, siano costituti da domande a risposta chiusa, con due o più opzioni di risposta, tali da permetterci di abbinare a ciascuna domanda soltanto un valore. È bene sapere che esistono software specifici o pacchetti nei programmi di elaborazione dati finalizzati all'analisi dei testi scritti e dei contenuti che ci restituiscono statistiche e rappresentazioni grafiche

su porzioni di testi o comunicazioni fra più soggetti (come quelle nei social network, ad esempio).

Introduciamo nell'elenco dei tipi di dati rilevabili anche quelli provenienti da sistemi come l'*eye-tracking* che, partendo dallo studio dei movimenti oculari, permette di misurare tempi e frequenze con i quali lo sguardo di un soggetto si sofferma su una parte di una immagine, di un testo o una pagina web. Usati come strumenti di scrittura in caso di gravi disabilità, queste tecnologie sono impiegate nell'ambito della ricerca per studiare in fasce di età diversificate i processi legati all'attenzione, l'usabilità dei sistemi informatici e sempre più auspicabilmente le competenze linguistiche, le attività cognitive e metacognitive, le emozioni, le modalità di apprendimento anche nei contesti multimediali (Porta & Rastelli, 2013; Alemdag & Cagiltay, 2018).

Anche l'osservazione diretta può essere usata come strumento di raccolta di dati quantitativi: in questo caso l'osservatore rileva i comportamenti dei soggetti attraverso griglie e schede che, rielaborate, possono essere usate come variabili che abbinano a ciascun comportamento osservato un numero pari alla frequenza con cui il comportamento è agito.

Abbiamo usato il termine variabile, fondamentale per avviare e comprendere ogni analisi statistica. Le variabili sono gli elementi operativi e osservabili in cui è scomposto un fenomeno e ai quali possiamo attribuire un valore (una modalità) per ciascun caso/soggetto/unità statistica. Il voto nell'esame di Statistica conseguito da ciascuno degli studenti iscritti al secondo anno di un determinato corso di laurea. Oppure l'altezza degli individui nati nel 1983. O ancora il piatto preferito tra 5 elencati da parte dei rispondenti a un'indagine di marketing. Questi esempi mostrano come nella fase di raccolta dei dati, due sono gli elementi da tenere in considerazione: le variabili, quindi gli eventi da osservare, e i soggetti/oggetti su cui rilevare l'occorrenza dell'evento, ossia campioni e popolazioni di riferimento.

Le variabili

Le classificazioni delle variabili sono determinate dalle caratteristiche proprie delle stesse e dalle modalità in cui vengono utilizzate.

Distinguiamo variabili *qualitative* e *quantitative* dal tipo di fenomeno che descrivono e dalla scala con cui vengono rilevate. Le modalità in cui si esprimono le prime sono caratteristiche nominali, le seconde valori numerici. Il colore preferito dei bambini della scuola dell'infanzia rappresenta quindi una variabile

qualitativa, l'età degli stessi è una variabile quantitativa. Le variabili qualitative si distinguono a loro volta in categoriali e ordinali. Il colore preferito, riprendendo l'esempio, è una variabile qualitativa *categoriale* e si misura con l'uso di una *scala qualitativa sconnessa* in cui tutte le opzioni di risposta alla possibile domanda: "quale è il tuo colore preferito?" hanno uno stesso peso. Non c'è un ordine nei colori, una scala che ci dica che il colore rosso ha un valore più elevato per noi del colore verde. Il classico esempio di variabile qualitativa *ordinale* è quello del livello di istruzione. In questo caso le opzioni di risposta all'ipotetica domanda "quale è il tuo titolo di studi?" seguono un ordine ben preciso che ci permette di indicare un titolo come superiore all'altro. La scala usata in questo caso è una *scala qualitativa ordinale* dove le opzioni di risposta saranno licenza di scuola primaria, scuola secondaria di primo grado, scuola secondaria di secondo grado, laurea triennale, laurea magistrale, corsi post lauream, dottorato di ricerca. Per alcune procedure, associando a queste opzioni di risposta un valore numerico, la variabile qualitativa ordinale può essere trattata – seppur con estrema cautela – come una variabile quantitativa. Queste ultime si suddividono in *discrete* e *continue* a seconda che i valori numerici formulati siano rilevati, contati (ad es., numero di smartphone in una famiglia) o misurati con strumenti di misurazione e unità di misura (es. altezza di un individuo misurata in cm). Vengono rilevate con *scale quantitative rapporto* o *non rapporto* che si distinguono per il modo in cui viene ammesso il valore 0: realistico nelle prime, solo convenzionale nelle seconde.

Un particolare tipo di variabili (qualitative) sono quelle definite *dicotomiche* o *binarie*, quelle che cioè possono assumere soltanto due modalità: bianco/nero, vero/falso, sì/no, uomo/donna e così via. Anche le variabili identificate come *dummy* sono dicotomiche (sono quantitative però): esse sono aggiunte nelle fasi di preparazione dei dati per indicare l'occorrenza di un fenomeno e trasformare una variabile categoriale in quantitativa. Una variabile dummy potrebbe ad esempio rilevare il fenomeno: "preferisco il giallo" dove i rispondenti possono scegliere questo colore fra molti. Per convenzione la modalità assunta sarà 1 in corrispondenza dei casi di soggetti che hanno indicato il colore giallo come il preferito, 0 per tutti gli altri colori.

Le trasformazioni sulle variabili possono condurre anche alla costruzione di variabili *composte*, cioè formate da due o più altre variabili del dataset che risultano prossime o per correlazioni statistiche o per legami concettuali e che, combinate, generano informazioni più esaurienti sui fenomeni (per una descrizione più completa si veda Song et al., 2013).

In altre classificazioni legate piuttosto all'uso che si fa delle variabili nelle analisi, possiamo distinguere le variabili *esogene* da quelle *endogene* per distinguere le variabili considerate interne al modello impiegato da quelle esterne. In alcuni dei metodi che descriveremo in seguito per definire relazioni fra i fenomeni osservati, si parla di variabili *indipendenti* e *dipendenti*: si ipotizza che le prime influenzino il comportamento delle seconde che sono in realtà molto spesso quelle al centro del nostro interesse. Nel linguaggio statistico le variabili indipendenti e dipendenti sono anche note rispettivamente con i nomi di *predittori* e *variabile risposta*.

Una distinzione da fare e che ci riporta a quanto detto nel primo capitolo, e cioè alla difficoltà di conoscere fino in fondo la realtà osservata, è quella fra variabili *latenti* e *manifeste*, *osservate* e *non osservate*. Gli eventi osservati sono sempre composti da due elementi non osservabili: il fenomeno in sé e un errore che nasce dalla misurazione, dal campionamento e così via. Vanno quindi distinte, in fase di analisi e discussione dei risultati, le variabili che possono essere osservate e sono manifeste da quei costrutti della cui esistenza siamo consapevoli ma che non possiamo rilevare perché latenti o non osservabili (de Lillo et al., 2007). Di questo avremo modo di parlare in maniera più approfondita a proposito delle tecniche di analisi (già ad esempio nel terzo capitolo).

Il campionamento (ovvero la statistica descrittiva e inferenziale)

L'insieme di soggetti/unità statistiche sui quali si rilevano le variabili dell'indagine contribuisce a definire metodi e strumenti da usare e il tipo di risultato che otterremo.

Quando, come nei censimenti, siamo in grado di rilevare i dati di cui abbiamo bisogno dall'intera *popolazione* di riferimento U, l'analisi statistica che conduciamo è di tipo descrittivo. Potremmo non avere la possibilità di intervistare/contattare tutti i soggetti coinvolti nello studio, non avere i fondi o preferire un'analisi più dettagliata su un gruppo meno numeroso. In questi casi, quando è solo una parte dell'intera popolazione ad essere considerata (anche se siamo interessati a studiare le caratteristiche dell'intera popolazione) cioè un suo sottoinsieme di numerosità n definito *campione*, si entra nell'ambito della statistica inferenziale e si introducono i concetti di probabilità, di stima dei parametri, di errore campionario (poiché da una stessa popolazione possono essere estratti molti campioni diversi). Lo scopo di un'indagine realizzata su un campione è di generalizzare i risultati all'intera popolazione di riferimento. Affinché la generalizzazione sia significativa, il campione deve essere rappre-

sentativo ovvero si rende necessario attivare delle tecniche di "correzione" o di valutazione dell'errore campionario. Spesso ci si affida al caso per assicurarsi che lo sia. Quando l'estrazione dei soggetti che faranno parte del campione avviene con reinserimento del soggetto estratto, si parla di campione *bernoulliano*. In questo caso ogni soggetto ha la stessa probabilità di essere estratto anche dopo l'estrazione. Nel caso in cui non ci sia reinserimento, si parla di campione *casuale semplice*. I due campioni tendono a sovrapporsi quando la loro numerosità è molto elevata.

Fra i campioni *probabilistici*, oltre a quelli già elencati, troviamo anche il campione *sistematico* nel quale sono estratte unità da osservare ogni ***k*** soggetti; *stratificato* nel quale la popolazione viene divisa in strati omogenei e vengono estratti ***n*** soggetti da ogni strato; *a grappoli*, nel quale contrariamente al precedente la popolazione viene divisa in gruppi eterogenei al loro interno in grado di rappresentare la varietà della popolazione. Rientrano fra i campioni *non probabilistici:* il campione *accidentale*, composto dai soggetti più prossimi al ricercatore; il campione *per quote*, dove i soggetti sono scelti fra strati omogenei al loro interno in maniera non casuale; il campione *a valanga*, dove ai primi soggetti coinvolti si chiede di segnalare altri soggetti da coinvolgere; il campione *per dimensioni*, che include in base a delle griglie soggetti che hanno caratteristiche diverse da tutti gli altri inclusi nel campione.

Esistono procedure di calcolo (anche impostate su applicazioni e pagine web) che ci consentono di definire la dimensione ottimale del campione, una volta che, in base ai casi, si conosce la numerosità della popolazione o sono stati fissati i livelli di significatività dei risultati e il livello di errore che siamo disposti a tollerare, considerando quanto i dati raccolti si scostano dai valori reali di ***U***.

2.2 - Preparazione dei dati

La prima operazione per proseguire nell'analisi statistica dopo la raccolta dei dati è quella di preparare il dataset (in pratica una o più tabelle) che contiene tutte le variabili e le unità statistiche osservate. L'obiettivo è spesso quello di avere per ogni riga un'unità statistica e per ogni colonna una variabile (o viceversa). Questa indispensabile operazione serve a mettere insieme ad esempio variabili ottenute da rilevazioni condotte in maniera diversa su uno stesso campione e produrre i file di base sui quali poter lavorare attraverso specifici software. Quando parliamo di preparazione dei dati, non ci riferiamo solo a questa

procedura ma a un più ampio processo di *data screening* che, talvolta, può richiedere anche più impegno e tempo dello stesso processo di analisi e studio dei dati.

Barbara Tabachnick e Linda Fidell (2013) hanno proposto una generica checklist per il controllo dei dati (Tabella 2.1) da riadattare poi ai singoli studi. Essa presenta fasi di analisi sequenziali che si possono comprendere alla luce di quattro elementi che riproponiamo nello stesso ordine e a partire dalle considerazioni delle autrici. Si tratta dell'accuratezza del file dei dati (1), la presenza di *missing data* (2) e *outlier* (3), il controllo di normalità, linearità e omoschedasticità (4).

Checklist for Screening Data

1. Inspect univariate descriptive statistics for accuracy of input
   a. Out-of-range values
   b. Plausible means and standard deviations
   c. Univariate outliers
2. Evaluate amount and distribution of missing data; deal with problem
3. Check pairwise plots for nonlinearity and heteroscedasticity
4. Identify and deal with nonnormal variables and univariate outliers
   a. Check skewness and kurtosis, probability plots
   b. Transform variables (if desirable)
   c. Check results of transformation
5. Identify and deal with multivariate outliers
   a. Variables causing multivariate outliers
   b. Description of multivariate outliers
6. Evaluate variables for multicollinearity and singularity

Tabella 2.1 - Checklist per la preparazione dei dati (Tabachnick & Fidell, 2013, p. 91).

Accuratezza. Un primo livello di verifica dell'accuratezza si ottiene confrontando la corrispondenza dei dati nel dataset con quelli originali della rilevazione. All'aumentare della numerosità del campione saranno necessarie procedure più complete di controllo. Per ciascuna variabile vanno calcolate distribuzioni di frequenze e (per le quantitative) indici di posizione e variabilità; così come vanno disegnati i grafici sull'andamento delle distribuzioni delle variabili coinvolte nello studio. Questo ci permette di essere certi che la media o il range di distribuzione delle variabili risulti accettabile e ragionevole. Fra le coppie di va-

riabili si verifica anche la correlazione e si studiano i casi in cui essa risulta accentuata (inflated) – principalmente quando si considerano variabili composte costituite da stesse variabili – o indebolita (defleated) – quando il range delle risposte di una o più variabile è ristretto. Fenomeni di multicollinearità e singolarità possono emergere rispettivamente quando due variabili sono altamente correlate perché misurano lo stesso fenomeno e quando alcune variabili risultano ridondanti perché una è ottenuta dalla combinazione delle altre. Qualora siano rilevate variabili che presentano valori di correlazione molto alti, si può considerare l'esclusione di quelle che ci restituiscono informazioni simili senza aggiungere spiegazioni ulteriori alla conoscenza del fenomeno studiato.

Durante la fase di verifica dell'accuratezza, infine, si rileva la presenza di outlier e missing data per una o più variabili.

Dati mancanti (missing data). Come facilmente si può dedurre, affrontiamo il tema del controllo dei casi in cui nei dataset risultino informazioni mancanti in corrispondenza di una unità statistica per una certa variabile. È una delle questioni più annose nell'analisi dei dati e per la quale non esistono linee guida troppo rigide e predeterminate. Il ricercatore può eliminare le unità statistiche per cui si rilevano missing data in una o più variabili se non eccessivamente numerose oppure nei casi più semplici può utilizzare le sue conoscenze precedenti del fenomeno per inserire un valore che ritiene verosimile. Potrebbe inoltre decidere di sostituire i missing data con la media o la mediana (con delle ripercussioni sulla variabilità della varianza e sulla correlazione con le altre variabili) o con delle stime come il risultato di un processo di regressione (capitolo 4) reiterato più volte, nel quale sono le altre variabili a predire il valore mancante.

La scelta dell'azione da mettere in campo una volta individuati i valori mancanti dipende dalla numerosità del campione e dalla numerosità e casualità degli stessi missing value.

Se i valori mancanti sono in un piccolo numero rispetto al campione e siamo certi della loro casualità, una delle scelte più frequenti è quella di eliminare le unità statistiche coinvolte dalla rilevazione.

Se il numero dei valori mancanti è molto alto rispetto alla numerosità del campione, bisogna indagare le motivazioni che si nascondono dietro tali valori. Aver collezionato un alto numero di missing data è di per sé un'informazione rilevante ai fini dell'indagine. Quando tali motivazioni sembrano particolarmente rilevanti per lo studio, si può scegliere di sostituire i missing data con la me-

dia della variabile e introdurre una variabile di appoggio *(dummy)* dove si indica con 0 il dato completo e con 1 il dato mancante.

Se le variabili per le quali abbiamo tanti missing data non sono fondamentali o le informazioni che ci restituiscono sono comunque ben rappresentate da altri indicatori, è plausibile non tenerle in considerazione.

Nell'eliminazione delle unità statistiche, soprattutto se l'assenza di valori non è casuale, bisogna verificare che i dati non siano mancanti in una particolare fascia di soggetti. Eliminare i dati in questo caso significherebbe compromettere la rappresentatività del campione. Talvolta in queste situazioni i missing data vengono sostituiti con la media calcolata in una specifica categoria di soggetti partecipanti allo studio invece della media dell'intero campione.

Una buona pratica è quella di ripetere l'analisi con e senza le unità per cui rileviamo missing value per testare le differenze.

Outlier. Si tratta di valori estremi che si discostano molto dalle modalità assunte dalle variabili, incidendo in maniera più rilevante sull'analisi. Si possono far rientrare in questa categoria anche modalità che raccolgono il 90% delle frequenze in variabili dicotomiche.

Distinguiamo gli outlier fra univariati (riferiti a una sola variabile) e multivariati, quando derivano dalla combinazione fra le modalità attribuite a più variabili. Nei grafici gli outlier appaiono come valori lontani dal resto della distribuzione presunta e sono pertanto riconoscibili con un colpo d'occhio. Nella Figura 2.2, ad esempio, vediamo rappresentati i boxplot relativi alla distribuzione dei numeri di iscritti ai Mooc (Massive Open Online Courses) della piattaforma EduOpen (learn.eduopen.org) relativi a 6 categorie di corsi. Sia la categoria in giallo indicata come AHU (Arts and Humanties), così come quella arancione contrassegnata dalla sigla SSC (Social Sciences), presentano numerosi, potenziali outlier, che oltrepassano il limite massimo calcolato dal software nella distribuzione (i "*baffi*" del boxplot). La categoria SCI (Sciences) presenta un unico potenziale outlier che si discosta in maniera estremamente rilevante dal resto dei dati. Questo non implica automaticamente che tali punti "estremi" siano effettivamente degli outlier ma ci indica di approfondirne la natura.

I metodi per determinare i multivariate outlier, come il calcolo della Mahalanobis distance, la discrepancy o l'influence, prendono in considerazione la distanza che esiste fra l'outlier e lo sciame dei dati, e l'influenza che di conseguenza tale distanza può comportare nell'andamento e nell'analisi delle variabili.

Si considera anche per gli outlier la possibilità di cancellare le unità statistiche a cui sono riferiti o di trasformare i valori rendendo la distribuzione più simile a quella ipotizzata (per esempio una normale).

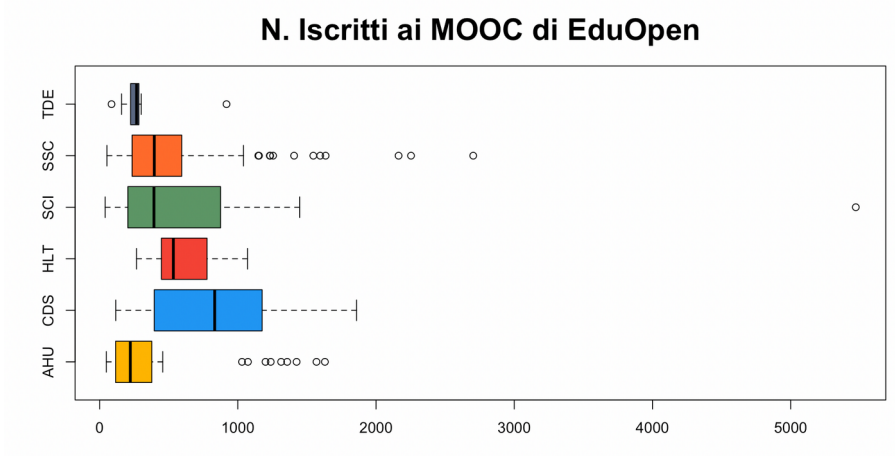


Figura 2.2 - Boxplot degli iscritti ai MOOC della piattaforma EduOpen suddivisi per categoria dei corsi. Vanno indagati i punti che superano il valore massimo come possibili outlier delle distribuzioni.

Normalità, linearità e omoschedasticità. Per applicare alcuni dei metodi di analisi multivariata, bisogna assumere che i dati si distribuiscano secondo una specifica distribuzione, molto spesso una normale multivariata, cioè che ogni variabile e ogni combinazione lineare fra le stesse variabili siano normali, obiettivo complesso da verificare quando il numero di variabili è particolarmente elevato. Tuttavia un buon punto di partenza è provare normalità, linearità e omoschedasticità per le singole variabili. Diamo rapidi cenni di seguito su come procedere:

- ci sono test e grafici che testano l'ipotesi di normalità univariata come i test di Shapiro-Wilk, Anderson-Darling e Lilliefors (Kolmogorov-Smirnov) o il metodo grafico denominato Q-Q plot (plot quantile della distribuzione osservata *vs* quantile della distribuzione di riferimento). Parametri da calcolare utili a testare la normalità sono la *skewness*, che restituisce la simmetria della curva, e la *kurtosis*, che ne verifica il livello di appiattimento e lo spessore delle code. Entrambi devono avvicinarsi allo 0 se le distribuzioni sono normali ma rappresentano soltanto delle condizioni necessarie e non sufficienti.

- l'esistenza di una correlazione lineare fra distribuzioni normali si visualizza in uno scatterplot bivariato quando i punti sono disposti in maniera curvilinea. Talvolta bisogna accettare che la relazione fra due variabili non assuma una forma lineare ma di altro tipo. E, spesso, è il caso più interessante.

- l'omoschedasticità è la proprietà di cui gode un gruppo di variabili che possiedono una variabilità (varianza) simile. Se esiste una normalità multivariata allora le variabili sono omoschedastiche.

Quando sono presenti outlier o quando, pur se necessario per i nostri calcoli, le variabili non rispettano i criteri di normalità, linearità e omoschedasticità, si preferisce applicare processi di *data trasformation* che vanno a modificare le singole modalità di ciascuna variabile e comportano, come conseguenza, modifiche alle scale e alle unità di misura con le quali i dati sono stati raccolti. Lo scopo di tali trasformazioni è quello di migliorare la normalità delle distribuzioni e di rendere confrontabili variabili misurate con scale diverse (ad esempio, "perdendo" le unità di misura con le quali i dati sono stati misurati). Molto noti sono i processi di normalizzazione e standardizzazione. La prima operazione ci consente di modificare i dati in modo che, come percentuali, risultino compresi in un range fra 0 e 1, dove 0 coincide con il valore minimo assunto dalla variabile e 1 il valore massimo. La seconda trasforma i valori di partenza generando una distribuzione che ha il valore 0 come media e 1 come deviazione standard. A queste, inoltre, solo per citarne alcune, si possono aggiungere l'inversa e le trasformazioni monotoniche come quelle logaritmiche, di elevamento a potenza e radice quadrata (McCune & Grace, 2002).

2.3 - Data visualization

Nel TED talk "The beauty of data visualization", (www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization), David McCandless, giornalista esperto di *data visualization* & *information design*, propone molti esempi di come la visualizzazione dei grafici aiuti a comprendere meglio quello che i dati dicono: gli investimenti di milioni di dollari, le paure del mondo, gli aggiornamenti nei social sulle rotture dei fidanzamenti, i budget militari e le forze armate, l'efficacia degli integratori alimentari. Invita a comportarsi da detective fra le informazioni, a giocare con i dati, a puntare su quel qualcosa di magico, dice, che c'è nella rappresentazione grafica ed è in grado di mostrare informazioni che in altri modi non comprenderemmo. Come dimostra la *visualization* creata

dal fisico danese Tor Norretranders che McCandless descrive: YOUR SENSE OF SIGHT IS THE FASTEST. Nel grafico presente nel video, che confronta l'ampiezza di banda dei sensi, emerge che il gusto ha la velocità di una calcolatrice, l'udito e l'odorato quella di un hard disk, il tatto quella di una chiavetta USB e che la vista è potente come la rete di un computer. Ancora più potente, aggiunge, se mescolata con le potenzialità del linguaggio della mente:

> *"if you combine the language of the eye with the language of the mind, which is about words and numbers and concepts, you start speaking two languages simultaneously, each enhancing the other. So, you have the eye, and then you drop in the concepts. And that whole thing – it's two languages both working at the same time."*

La data visualization, configuratasi come disciplina a sé con lo sviluppo di grandi quantità di dati nei sistemi informatici, è un campo al confine fra matematica, informatica, scienze cognitive, ingegneria, statistica, computer graphics, tutte discipline che prevedono l'uso di grafici. In essa convergono i campi della *scientific visualization*, *information visualization*, e il più recente settore di *visual analytics*. La *scientific visualization*, altrimenti detta *spatial data visualization*, ha a che vedere con i processi di rappresentazione dei fenomeni scientifici caratterizzati da una collocazione geografica o in uno spazio bi- o tri- dimensionale. Una visualizzazione appartenente a questo settore potrebbe rappresentare il flusso colorato di un liquido all'interno di un macchinario in base a come lo stesso liquido cambia la temperatura. Fanno parte dell'*information visualization* i processi di produzione di grafici ad albero, reti, tabelle, documenti, serie temporali e così via, che richiedono una rielaborazione visuale pur non essendo necessariamente collocati in uno spazio definito. La collocazione nello spazio (ad es. gli assi cartesiani), in questo caso, è una scelta che viene fatta per la rappresentazione e che è utile per mostrare le relazioni fra gli eventi. Il campo di *visual analytics* deriva dai precedenti due settori e permette la combinazione di data analysis e visualization tools con lo scopo di produrre metodi e tecniche di visualizzazione dei dati. L'obiettivo è di supportare ragionamenti analitici usando tecniche e strumenti che generino interfacce visuali interattive in un processo che deve generare senso a partire da dataset molto numerosi e complessi. Il focus della *visual analytics* è il processo reiterato messo in atto nella costruzione di visualizzazioni che siano sempre più adatte a rappresentare i dati e generare informazioni (Telea, 2014).

La sfida della data visualization è di presentare ciò che è complesso in una maniera completa, comprensibile, dinamica e interattiva. L'obiettivo di questa forma grafica di analisi dei dati, a seconda dei casi, può essere quello di tradurre per un lettore comune dati e relazioni espressi altrimenti da articolate formule matematiche o di mostrare a un ricercatore attraverso il canale visivo uno scenario indagabile anche con numeri e simboli ma con tempi lunghi e operazioni poco intuitive.

Sebbene tutti i package di analisi statistica includano degli strumenti di graficazione dei risultati, per un approccio efficace alla data visualization è necessario utilizzare degli strumenti specifici.

Ci sono nel web decine di video e brevi articoli che classificano i tool per la data visualization. Ne troviamo alcuni che sono vere proprie applicazioni da scaricare come Tableu, altre che si presentano come piattaforme online, ad es. Flourish. Si distinguono per il tipo di grafici che permettono di produrre e per la possibilità di scaricare o usufruire di una versione gratuita. Tool proprietari sono Looker e Knime. La distinzione fra le versioni free e i profili business spesso ricade sulle modalità di condivisione di dati e grafici prodotti, che in base alle soluzioni di acquisto scelte diventano automaticamente pubblici o eventualmente privati. Servono competenze statistiche e informatiche per utilizzare i tool? Questo è un altro elemento da considerare quando ne scegliamo uno: Google Chart ad esempio dà la possibilità di integrare grafici nelle pagine web provvedendo codici in html da modificare che possono non essere noti a tutti. In alcuni casi accanto alla costruzione di grafici, ci sono funzionalità per la creazione di infografiche come per Infogram, piattaforma online, dove si trova un buon numero di tipologie diverse di visualizzazioni da utilizzare all'interno di infografiche, post per i social, presentazioni e report. Un esempio estremamente semplice è in Figura 2.3 dove, a partire da uno dei modelli preimpostati di Infogram sono state riassunte alcune caratteristiche degli studenti iscritti al primo anno del CdL in Digital Education presso l'Università di Modena e Reggio Emilia nell'a.a. 2019/20.

Frequentemente questi tool, danno la possibilità di costruire dashboard e presentazioni che talvolta assumono il nome di storie. I dati e le loro visualizzazioni sono in grado di raccontare come è fatta una società, un gruppo, cosa cambia nel tempo, come due fenomeni sono in relazione fra di loro. Nel formato grafico, riescono a farlo utilizzando un design accattivante con la peculiarità di portare l'attenzione dei "visualizzatori" su specifiche questioni e allo stesso tempo di conservarsi più a lungo e con maggiore enfasi nella memoria e fra le conoscenze.

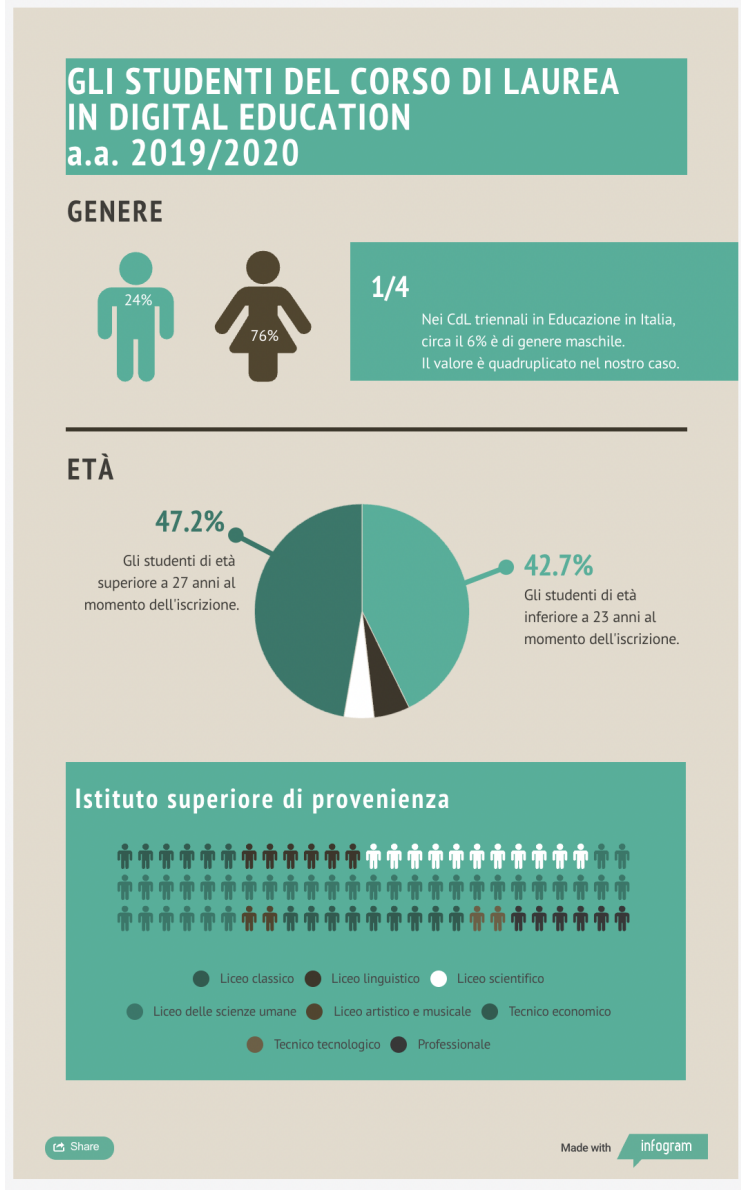


Figura 2.3 - Infografica relativa alla composizione del gruppo del primo anno degli studenti del CdL in Digital Education, Unimore, a.a. 2019/20, realizzata con Infogram https://infogram.com/digital-education-student-1hmr6g78jo7pz6n?live

Usiamo una visualizzazione realizzata sulla piattaforma Flourish per proporre un esempio che, semplificando il meccanismo di funzionamento dei sistemi di data visualization, prova a spiegarli e renderli evidenti. Flourish è una piattafor-

ma online user-friendly per la data visualization che non richiede competenze elevate da sviluppatore o da statistico per creare visualizzazioni dinamiche, efficaci e accattivanti. Oltre a piani di utilizzo business, dispone di un piano di utilizzo pubblico che permette di creare grafici interattivi e storie/presentazioni a partire dai dati e dalle informazioni che essi generano. Nella versione free, anche i dati sono pubblici. Nel Visual vocabulary di Flourish sono comprese oltre 50 tipologie di visualizzazioni che si focalizzano sulla rappresentazione di distribuzioni, variazioni, correlazioni, ranking, serie, suddivisioni, mappe e flussi. Tante sono le opzioni di personalizzazione della grafica con colori, forme, titoli e così via.

Le Figure da 2.4 a 2.7 sono un esempio di una visualizzazione pubblica su questa piattaforma.

Il processo di creazione di un grafico, come processo di analisi dei dati, richiede di individuare ipotesi e modelli concettuali di partenza e di predisporre correttamente un dataset.

I dati caricati (ad es. da un file Excel) possono essere modificati come in un foglio di calcolo e ciascuna variabile (ossia ciascuna colonna) va indentificata con un ruolo nella rappresentazione. Nel caso nelle figure, gli strumenti di Flourish ci hanno chiesto di distinguere le colonne contenenti variabili categoriali da usare per le selezioni riferite a "Group by", "Shade by", "Compare" e variabili continue per l'elenco "Size by".

La visualizzazione (in due formati, circolare e rettangolare) è riferita ai dati raccolti in un'indagine somministrata agli utenti della già citata piattaforma EduOpen per definirne i profili. Il nostro obiettivo è quello di descrivere il campione con una visualization. Ciascuno dei rispondenti corrisponde a uno dei punti del cerchio. Passandoci sopra con il mouse compaiono tutte le informazioni relative al singolo rispondente (Figura 2.4).

Le figure successive mostrano cosa succede quando scegliamo una variabile dai menu nel box di scelta. La Figura 2.5 ci mostra i dati divisi per genere e colorati per livello di competenze digitali. Nella Figura 2.6 viene aggiunta un'altra informazione, quella relativa al numero di corsi completati visualizzata attraverso le dimensioni dei punti del grafico (ricorderete che le variabili quantitative sono quelle relative all'elenco "Size by" e che come tali possono determinare le dimensioni). L'ultima Figura 2.7 contiene anche la distinzione in categorie in base all'età.

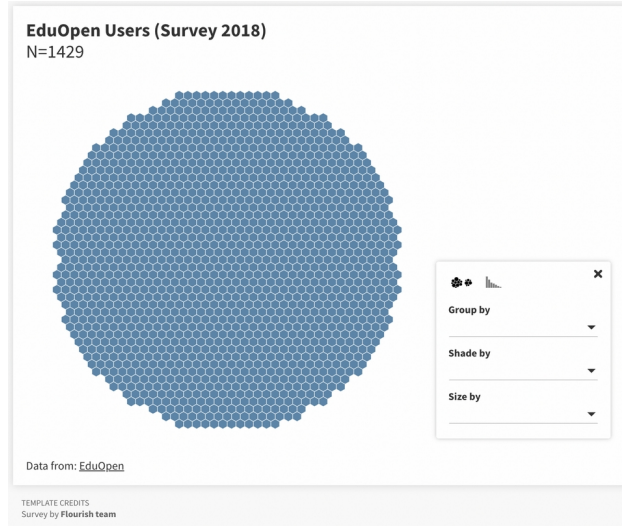


Figura 2.4 - Risultati di un'indagine fra gli utenti di EduOpen visualizzati su Flourish, visualizzazione generale, https://public.flourish.studio/visualisation/5221716/

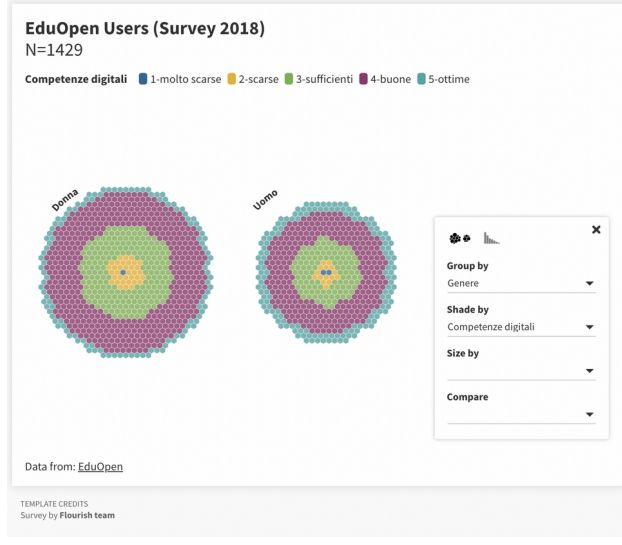


Figura 2.5 - Risultati di un'indagine fra gli utenti di EduOpen visualizzati su Flourish, visualizzazione per genere e competenze digitali, https://public.flourish.studio/visualisation/5221716/

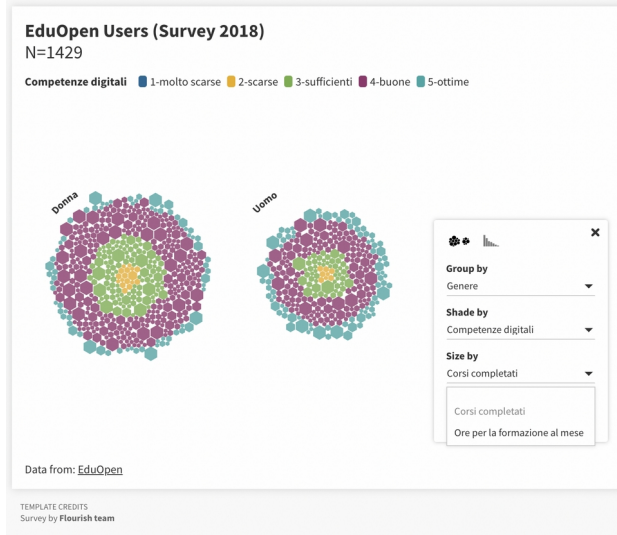


Figura 2.6 - Risultati di un'indagine fra gli utenti di EduOpen visualizzati su Flourish, visualizzazione per genere, competenze digitali e numero di corsi completati, https://public.flourish.studio/visualisation/5221716/

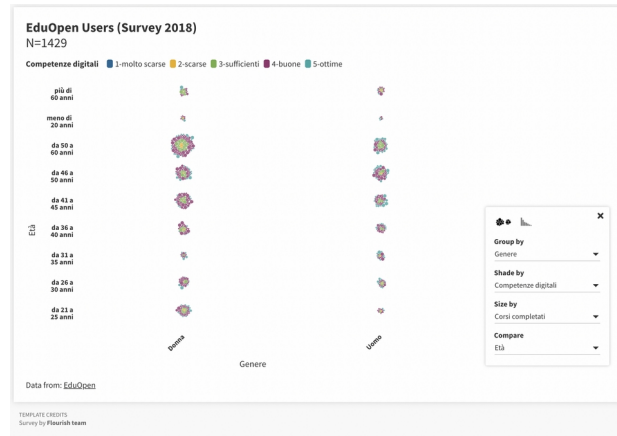


Figura 2.7 - Risultati di un'indagine fra gli utenti di EduOpen visualizzati su Flourish, visualizzazione per genere, competenze digitali, numero di corso ed età, https://public.flourish.studio/visualisation/5221716/

Utilizzando quindi testi, suddivisione nello spazio, colori e dimensioni, il grafico in maniera interattiva riesce a farci visualizzare una quantità di informazioni elevata, sia per il numero di unità che fanno parte del campione che per il nu-

mero di variabili che possono essere combinate fra di loro con molti abbinamenti diversi (il nostro è un esempio semplice, la quantità di dati, di variabili e di relazioni fra le stesse in analisi statistiche e computazionali può essere notevolmente molto più numerosa e più complessa!).

Con un solo colpo d'occhio riusciamo a dire che c'è una prevalenza di donne nel campione, che i livelli di competenza digitale prevalenti sono "sufficiente" e "buono" e anche che fra gli uomini c'è una presenza maggiore di utenti con competenze digitali di livello ottimo, che gran parte del campione ha un'età compresa fra 41 e 60 anni, in particolare fra i rispondenti c'è un gruppo nutrito di donne di età compresa fra 50 e 60 anni con competenze digitali prevalentemente di livello sufficiente-buono.

In numeri avremmo dovuto dire che il 60% del campione è costituito da donne, che il 55% dei rispondenti ha competenze digitali "buone" e il 23% "sufficienti", che il 26% degli uomini ha competenze digitali ottime e che fra le donne questo livello di competenza è raggiunto solo dal 14%, che il 56% degli utenti dell'indagine ha un'età compresa fra 41 e 60 anni, che le donne di età compresa fra 50 e 60 anni sono il 17% del campione totale.

La visualizzazione, in conclusione, ci ha restituito numerose informazioni sulle singole variabili e su gruppi di variabili, che ricorderemo, che potremo rimescolare dagli strumenti di selezione, che sono già di per sé dei risultati e che saranno un punto di partenza rilevante per analisi di approfondimento successive.

2.4 - Linee guida per l'analisi multivariata

Quando avviamo ricerche su fenomeni educativi, che siano su larga scala o in contesti più ristretti, siamo consapevoli dei tanti fattori che intervengono nella determinazione degli eventi. L'uso dell'analisi multivariata risponde alla necessità – che emerge anche nel settore educativo – di maneggiare più variabili in un unico contesto e di sintetizzarle per ottenere informazioni su scenari ampi, ricchi di dati e complessi per via delle tante relazioni che in essi sono stabilite. È duplice il modo in cui opera l'analisi multivariata: da un lato riassume grandi quantità di informazioni in un numero inferiore di indicatori più comprensibili; dall'altro mostra quelle tendenze, quelle regolarità, quelle strutture latenti nelle relazioni fra i dati al fine di costruire modelli predittivi che non solo aderiscono (*to fit* è il verbo inglese usato per esprimere questo concetto) alla struttura del campione osservato ma dell'intera popolazione di riferimento.

Siamo abituati inconsapevolmente a queste operazioni di *summarisation* proprie dell'analisi multivariata. Non solo nel contesto statistico dove, selezionando pochi indicatori, descriviamo distribuzioni a partire da una media e una deviazione standard. Ma anche nella quotidianità, dove ad esempio siamo in grado di percepire dettagli tridimensionali da una foto bidimensionale (Bartholomew et al., 2008).

La modalità attraverso cui si adoperano le tecniche di analisi multivariata porta al pensiero il paradigma della semplessità dell'ingegnere e neurofisiologo francese Alain Berthoz (2011) applicato anche in contesti didattici (Ferrari, 2013; Rivoltella & Rossi, 2012). Con questo termine lo studioso spiega il modo in cui in situazioni molto complesse, per spirito di sopravvivenza nei processi evolutivi, il cervello sia in grado di elaborare soluzioni semplici ma allo stesso tempo "eleganti ed efficaci" che derivano da una riduzione della complessità, si basano sulle esperienze già compiute e anticipano il futuro.

Esse "non snaturano la complessità del reale: non sono né caricature, né scorciatoie, né riassunti" (Berthoz, 2011). Allo stesso modo, nel tipo di analisi oggetto della nostra discussione, attraverso l'uso di tecniche predefinite e sempre incrementabili, molti dati vengono sintetizzati in fattori e poi modelli che senza perdere la complessità del contesto di partenza mostrano soluzioni organiche che contengono nuove conoscenze sul funzionamento dei sistemi, nel nostro caso educativi.

Nella definizione più stringente dell'insieme di tecniche che vanno sotto il nome di analisi multivariata, si sottolinea come in queste procedure le variabili debbano essere casuali e connesse fra loro in modo che le variazioni possano essere osservate solo sull'insieme dei dati e non in maniera isolata sulle singole variabili. Gli effetti registrati devono essere il risultato di eventi e variazioni che si manifestano simultaneamente. Tuttavia, nella pratica si è soliti utilizzare queste tecniche non solo per analisi multivariate ma anche in casi di analisi multivariabili, riferite cioè a più di due variabili, dove non vige il principio di simultaneità.

Per dirla con le parole di Hair e colleghi (2014):

> *"Multivariate analysis techniques are popular because they enable organizations to create knowledge and thereby improve their decision making. Multivariate analysis refers to all statistical techniques that simultaneously analyze multiple measurements on individuals or objects under investigation. […] To be considered truly multivariate, however, all*

> *the variables must be random and interrelated in such ways that their different effects cannot meaningfully be interpreted separately. Some authors state that the purpose of multivariate analysis is to measure, explain, and predict the degree of relationship among variates (weighted combinations of variables)." (ivi, p. 4)*

Prima di passare alla classificazione delle tecniche di analisi multivariata, prendiamo in considerazione due nuclei di informazioni che, riprese da Hair e colleghi (2014), ci danno un quadro completo dei processi di analisi che fanno riferimento alle tecniche multivariate. Gli autori forniscono:

1. un processo fatto di 6 fasi nel quale si svolge un'analisi multivariata in un approccio mirato alla costruzione di un modello di lettura dei fenomeni;

2. linee guida (quasi una filosofia, affermano) per l'analisi multivariata e l'interpretazione dei risultati.

Le Tabelle 2.2 e 2.3 mettono insieme i due elementi attribuendo a ciascuna fase di lavoro (colonna a sinistra) uno dei principi delle linee guida (colonna a destra). Le fasi di lavoro sono divise fra fasi di preparazione (Tabella 2.2) e fasi di costruzione/validazione del modello (Tabella 2.3); fra queste ultime, la quarta e la quinta hanno in comune il principio che valorizza l'errore, inteso come assenza di aderenza del modello ai dati. I principi focalizzano l'attenzione del ricercatore su scelte importanti come la definizione delle dimensioni del campione e la verifica della sua rappresentatività; valorizzano lo studio preparatorio di data screening ritenendolo fondamentale per interpretare i risultati e ripetere le analisi in caso di errore durante l'analisi; inducono a pensare ai modelli ottenuti come strettamente collegati da un lato alle teorie e ipotesi di partenza, dall'altro al contesto pratico di applicazione dei risultati.

Le varie fasi della ricerca legate in un percorso ordinato e lineare e i principi delle linee guida sono schemi generali che vanno adattati alle diverse tecniche; proprio per la loro linearità ci danno la possibilità di guardare ai fenomeni in maniera strutturata, ripercorrere le analisi alla luce dei risultati e anche trovare, sorprendentemente in uno scenario così regolato, soluzioni innovative e creative.

Fasi di preparazione	
FASI DI LAVORO	**LINEE GUIDA**
Stage 1: Define the Research Problem, Objectives, and Multivariate Technique to Be Used Ogni processo di analisi comincia con la definizione degli obiettivi di ricerca e di un modello concettuale da studiare per cercare relazioni di dipendenza o similarità fra variabili. Definito il quadro teorico, gli obiettivi e il tipo di variabili, si hanno gli elementi sufficienti per selezionare la tecnica multivariata da usare.	*Strive for Model Parsimony* Il modello concettuale sul quale si poggia l'indagine deve essere ben chiaro ai ricercatori, rimane la struttura portante dell'intero processo di analisi. Fissare i tratti rilevanti del modello teorico di partenza permette di non trascurare variabili rilevanti, di introdurre solo quelle non superflue e di evitare/spiegare fenomeni di multicollinearità.
Stage 2: Develop the Analysis Plan Come definire il campione? Come gestire le variabili? (e si può aggiungere: come raccogliere i dati? Con quali strumenti?) Questi i quesiti a cui rispondere nella fase di sviluppo del piano di analisi.	*Recognize That Sample Size Affects All Results* Campioni troppo piccoli possono impedire di ottenere risultati accettabili, talvolta anche perché la rispondenza fra dati e relazioni potrebbe risultare eccessivamente positiva. Lo stesso accade per campioni troppo numerosi. Di pari numerosità devono essere gruppi di controllo e sperimentali negli studi sperimentali.
Stage 3: Evaluate the Assumptions Underlying the Multivariate Technique La fase ha a che vedere con l'analisi univariata e bivariata delle variabili, lo studio di outlier e missing data, le correlazioni, l'assunzione di normalità, e altre procedure che ci permettono di capire se i dati raccolti sono in grado di rappresentare relazioni multivariate.	*Know Your Data* La conoscenza dei dati nel dettaglio, a partire dai processi di data screening, è un'operazione centrale per poter interpretare i risultati anche secondo schemi inaspettati ma coerenti con gli eventi osservati.

Tabella 2.2 - Fasi di preparazione e linee guida in un approccio model-building di analisi multivariata (Hair et al., 2014, pp. 21-24, nostra rielaborazione).

Fasi di costruzione e validazione	
FASI DI LAVORO	**LINEE GUIDA**
Stage 4: Estimate the Multivariate Model and Assess Overall Model Fit Stimato il modello, si verifica dunque che esso risponda ai criteri di significatività statistica e pratica e alle relazioni proposte. Se ci si imbatte in soluzioni non adeguate in questa fase, è necessario ricominciare l'analisi: si tratta di un processo da reiterare fino a quando non si definisce un modello che si adegui ai dati in maniera ottimale.	*Establish Practical Significance as Well as Statistical Significance* La verifica della significatività statistica va accompagnata dal controllo dell'aderenza dei risultati di analisi e dei modelli predittivi con gli aspetti concreti dei fenomeni osservati. Oltre a chiedersi: "Sono effetti legati al caso?", bisogna domandarsi: "Sono utili?" (si veda per questa fase anche il principio successivo)
Stage 5: Interpret the Variate(s) Il ricercatore interpreta le combinazione lineari delle variabili in base ai pesi attribuiti. Anche in questo processo potrebbero presentarsi formulazioni non adeguate tanto da richiedere un riadattamento del modello individuato in un ciclo iterattivo.	*Look at Your Errors* In un'analisi multivariata quasi sempre è necessario ripetere le procedure di analisi per arrivare a risultati convincenti, a modelli predittivi funzionali. Le procedure messe in atto, seppur non definitive e fallaci, sono centrali per identificare gli errori e condurre a formulazioni di modelli più coerenti e validi.
Stage 6: Validate the Multivariate Model Tecniche di validazione per verificare il livello di generalizzabilità del modello, come il bootstrapping, vanno messe in campo prima della definitiva accettazione dei risultati.	*Validate Your Results* Poiché nelle analisi statistiche ci si pone nella maggior parte dei casi lo scopo di generalizzare i risultati ottenuti da un campione all'intera popolazione, nelle fasi di lavoro vanno inclusi quei meccanismi di validazione che ci permettono di dire se il campione è rappresentativo della popolazione e se i risultati sono accettabili al di là dall'errore campionario.

Tabella 2.3 - Fasi di costruzione/validazione e linee guida in un approccio model-building di analisi multivariata (Hair et al., 2014, pp. 21-24, nostra rielaborazione).

2.5 - Breve introduzione ai metodi

Le tecniche di analisi multivariata, in continuo aumento, comprendono procedure di calcolo ma anche restituzioni grafiche di figure e tabelle. La tecnica della regressione multivariata, ad esempio, ci restituisce formule e indicatori

che ci permettono di validare il modello che stiamo studiando. Al contrario, nel multidimensional scaling si producono grafici anche a più dimensioni.

Mentre cominciamo a familiarizzare con i nomi delle tecniche che descriviamo nel volume, proponiamo due classificazioni pur partendo dalla certezza che le distinzioni non sono così radicali e che adeguate trasformazioni di variabili, campionamenti ben fatti, verifiche di ipotesi e significatività possono permetterci di utilizzare schemi alternativi di analisi.

Nella prima classificazione (Bartholomew et al., 2008) distinguiamo metodi *descrittivi* e *inferenziali* riproponendo la già segnalata distinzione fra statistica descrittiva e inferenziale.

I metodi che appartengono al primo gruppo hanno lo scopo di descrivere ed esplorare i dati sintetizzando le variabili osservate per rendere la descrizione del dataset più agevole. Appartengono a questa categoria, ad es. le tecniche della cluster analysis o del multidimensional scaling. Se siamo in grado di determinare il processo latente che ha generato i dati e di conseguenza siamo in grado di generalizzare i risultati ottenuti da un campione alla popolazione, staremo lavorando con tecniche inferenziali come la regressione lineare o l'analisi fattoriale.

La seconda classificazione (Hair et al., 2014) differenzia tecniche di *dipendenza* e *interdipendenza*.

Le prime sono quelle in cui possiamo identificare variabili dipendenti e indipendenti. Sono le tecniche in cui cerchiamo di identificare quelle relazioni fra le variabili che ci permettono di predire il comportamento di una variabile indipendente y, noti i predittori x_i. Nel secondo gruppo di tecniche, le variabili sono trattate insieme per definire la struttura latente sottesa all'intero gruppo di eventi osservati.

Per scegliere la tecnica da usare fra quelle di dipendenza bisogna guardare al numero e al tipo di variabili del dataset. Se, ad es., abbiamo un'unica variabile dipendente quantitativa useremo la tecnica della regressione multivariata. Se la variabile è di natura qualitativa, sceglieremo la regressione logistica.

Una presentazione sintetica delle tecniche descritte nel volume è nelle Tabelle 2.4 e 2.5 dove per ciascuno dei metodi, insieme a una descrizione e a una specificazione della tipologia di variabili principalmente utilizzata, viene riportato un esempio di indagini da realizzare attraverso la tecnica in questione.

Ciascuna tecnica (esclusa l'analisi della varianza multivariata) verrà ripresa e spiegata nei prossimi capitoli con descrizioni dei metodi e casi di studio.

Tecniche di interdipendenza		
Nome	**Descrizione**	**Esempio**
Analisi Fattoriale Esplorativa (inferenziale; variabili quantitative)	Consente di concentrare le variabili del dataset in un numero minore di fattori analizzabili con una perdita minima di informazioni per identificare la struttura latente ai dati	Riunire gli item di una scala sulla self-regulation (variabili) in gruppi in modo che le relazioni fra i gruppi possano restituire informazioni di più ampio respiro sul fenomeno indagato
Analisi delle Corrispondenze (descrittiva; variabili qualitative)	È una tecnica di analisi grafica che a partire dalle tabelle di contingenza mostra in una mappa bi- o tri- dimensionale l'interdipendenza fra gli oggetti analizzati per riassumere le osservazioni in pochi indici numerici rappresentativi	Visualizzare graficamente la relazione fra scelta di un servizio in un centro diurno e caratteristiche degli utenti del campione
Cluster analysis (descrittiva; variabili quantitative)	Ha lo scopo di raggruppare le unità statistiche in gruppi simili a partire dalle caratteristiche che le accomunano	Creare dei gruppi omogenei di studenti di un istituto per reddito, età e voti
Multidimensional Scaling (descrittiva; variabili quantitative e qualitative)	Restituisce una rappresentazione grafica in uno spazio a *n* dimensioni delle distanze fra gli oggetti dell'analisi	Visualizzare graficamente la distribuzione degli utenti di un centro diurno a partire dalle *n* caratteristiche dei servizi offerti

Tabella 2.4 - Principali caratteristiche delle tecniche di interdipendenza di analisi multivariata.

Esula dall'elenco nelle tabelle, l'Item Analysis con i due approcci della Classical Test Theory e dell'Item Response Theory (caso specifico di analisi fattoriale per variabili binomiali). Usata in ambito educativo prevalentemente per questioni di natura docimologica verrà presentata nel capitolo 8.

Per approfondire le procedure e le caratteristiche di ciascuna tecnica sono stati centrali i tre manuali di analisi multivariata di David J. Bartholomew e colleghi (2008), Joseph F. Hair e colleghi (2014), Antonio de Lillo e colleghi (2007).

Gli esempi applicativi presenti in ciascun capitolo sono stati realizzati usando il software open source di analisi dati `R` e analizzano dataset reali e raccolti nei progetti e nelle attività di formazione realizzate all'interno del Centro Interateneo Edunova dell'Università degli studi di Modena e Reggio Emilia.

Tecniche di dipendenza		
Nome	**Descrizione**	**Esempio**
Regressione lineare multivariata (inferenziale; una variabile dipendente quantitativa, variabili indipendenti qualitative e quantitative)	Viene utilizzata per predire il comportamento della variabile dipendente una volta che sia nota la relazione lineare che lega questa alle variabili indipendenti	Stimare il voto in un esame degli studenti in un corso essendo noti: età, genere, percentuale di partecipazione alle lezioni, motivazione, media delle prove di valutazione intermedia
Regressione Logistica (inferenziale; una variabile dipendente qualitativa, variabili indipendenti qualitative e quantitative)	Analoga alla precedente, differisce per la tipologia di variabile dipendente	Stimare il successo/ insuccesso in un esame degli studenti di un corso essendo noti: età, genere, percentuale di partecipazione alle lezioni, motivazione, media delle prove di valutazione intermedia
Analisi della varianza multivariata (inferenziale; più variabili indipendenti qualitative e due o più variabili dipendenti quantitative)	È una tecnica che permette, in studi sperimentali, di determinare la significatività delle variazioni in una variabile a partire dal confronto con le varianze delle altre variabili del dataset	Creati un gruppo sperimentale e uno di controllo, verificare che la differenza nei punteggi ottenuti in un test sia determinata dalla modalità di erogazione del corso (presenza/blended) prese in considerazione altre variabili come il genere o il livello di gradimento della disciplina

Tabella 2.5 - Principali caratteristiche delle tecniche di dipendenza di analisi multivariata.

Per ciascuna tecnica, casi ed esempi sui temi più diversi nell'ambito della ricerca educativa (formazione dei docenti, disabilità, qualità degli atenei, scale di valutazione dell'apprendimento, cittadinanza digitale, NEET generation, formazione permanente per adulti e anziani e così via) sono stati individuati nella letteratura più recente e maggiormente citata su Scopus nell'ultimo decennio in riviste peer-reviewed del settore Education in lingua inglese.

CAPITOLO 3

TECNICHE DI RIDUZIONE DELLA DIMENSIONALITÀ: ANALISI DELLE COMPONENTI PRINCIPALI, ANALISI FATTORIALE ESPLORATIVA E ANALISI DELLE CORRISPONDENZE

Al termine del capitolo, il lettore sarà in grado di:

- *definire le finalità d'uso dei metodi di riduzione della dimensionalità;*
- *descrivere le caratteristiche di alcune delle tecniche di riduzione dati: analisi fattoriale esplorativa, analisi delle componenti principali, analisi delle corrispondenze;*
- *illustrare esempi di ricerca educativa nei quali sono state usate tecniche di riduzione della dimensionalità.*

3.1 - Tecniche di riduzione dei dati

Avviamo in questo capitolo la descrizione delle tecniche di analisi multivariata il cui scopo è quello di combinare le variabili misurate in un numero ridotto di fattori o componenti che rappresentano in maniera più immediata, a volte anche meno completa, i dataset da cui sono stati tratti.

Queste tecniche semplificano le procedure di analisi e di interpretazione dei dati riducendo e riassumendo le variabili con una perdita minima di informazioni. Esse conducono alla definizione di modelli che ci restituiscono la struttura latente alla base delle osservazioni reali effettuate e misurate da variabili manifeste. Questi modelli sono più apprezzati quando sono costruiti partendo da un numero limitato di variabili, rispettando il principio di parsimonia o del rasoio di Occam. Più "eleganti", ed anche più efficaci, in ambito statistico sono

quelle procedure che riescono a individuare strutture e relazioni utilizzando un minor numero di variabili.

Le tecniche di riduzione dei dati lavorano sulle dimensioni dello spazio geometrico in cui sono collocate le osservazioni. Nell'ambito geometrico, le unità statistiche sono visualizzabili come punti in uno spazio multidimensionale, spazio che senza una profonda conoscenza matematica difficilmente può essere percepito e visualizzato da chi, come noi, si ferma a una realtà in tre dimensioni. Ridurre le variabili in uno studio significa ridurre le dimensioni dello spazio multidimensionale. Riuscire dall'analisi a riassumere in due o tre fattori la variabilità dell'intero dataset significa semplificare in due/tre dimensioni i sistemi con cui abbiamo a che fare rendendoli visibili in uno spazio bi- o tri- dimensionale di più facile lettura. È questa una risposta a quella che viene definita la maledizione della dimensionalità, ossia la situazione in cui la presenza di troppe variabili, e cioè troppe caratteristiche delle osservazioni che abbiamo in analisi, rende ingestibile applicazioni di metodi, procedure di calcolo e interpretazione dei fenomeni.

Si tratta certamente di operazioni complesse da un punto di vista del calcolo. Per i più esperti i calcoli algebrici sono di grande aiuto nelle procedure; provvidenziale per tutti è il contributo dei software di analisi dati.

Scegliamo di partire nella descrizione delle tecniche proprio con questo gruppo di metodi perché la selezione delle variabili è una annosa questione con cui ci si confronta immediatamente quando si lavora nell'ambito dell'analisi multivariata. Molto spesso la riduzione della dimensionalità non è il punto d'arrivo delle analisi ma una fase propedeutica all'uso di altre tecniche multivariate. Potrebbe essere infatti necessario applicare tecniche di *data summarization* su un dataset con un alto numero di variabili prima di processarlo ad esempio attraverso tecniche di regressione o cluster analysis, metodi che affronteremo nei prossimi capitoli.

Procedure di riduzione della dimensionalità investono anche la ricerca educativa. Un caso particolarmente frequente in cui vengono usate è quello della costruzione e analisi psicometrica di scale e questionari nei quali ciascun item è considerato una variabile e l'analisi dei dati raccolti richiede una riduzione a poche dimensioni per una lettura più rapida dei risultati oppure per identificare quelle relazioni fra fenomeni indagati attraverso le domande che una restituzione meno approfondita (ad es. fatta di sole percentuali in riferimento alle opzioni di risposta) non ci avrebbe permesso di ottenere.

Descriviamo in questo capitolo tre tecniche: l'analisi delle componenti principali (PCA), l'analisi fattoriale esplorativa (EFA) e l'analisi delle corrispondenze (semplici, CORA, e multiple, MCA). Si preferisce inserire la tecnica del multidimensional scaling, seppur classificabile come tecnica di riduzione dimensionale, dopo aver introdotto la cluster analysis per le sue implicazioni come supporto alle tecniche di classificazione.

Per scegliere fra le tecniche di riduzione quella più adatta alle nostre necessità, dobbiamo rispondere nella fase iniziale di ogni studio ad alcune domande: le variabili del dataset sono metriche o categoriali? La ricerca serve prioritariamente per identificare e generalizzare strutture latenti nei dataset oppure ha solo scopo descrittivo? Sarebbe più opportuno visualizzare i risultati in forma grafica?

L'analisi delle componenti principali è una tecnica di estrazione delle variabili che viene usata anche come parte dell'analisi fattoriale. In essa un numero minimo di variabili, dette *componenti*, è selezionato per rappresentare la massima porzione della varianza del dataset. Si tratta di una tecnica descrittiva nella quale le componenti principali sono combinazioni lineari delle altre variabili (metriche).

L'EFA è una tecnica applicata principalmente a variabili quantitative che si fonda sull'ipotesi che le variabili originali possano essere modellate come combinazione lineare di un insieme ridotto di variabili non osservabili chiamate *fattori* e identificati come dimensioni di strutture latenti dei fenomeni osservati. Si tratta di una tecnica inferenziale nella quale sono presenti assunti, test di bontà del fit, significatività statistica, precisione delle stime.

L'analisi delle corrispondenze (semplici e multiple) è una tecnica di riduzione delle dimensioni per variabili categoriali nella quale le modalità assunte dalle variabili rivestono un ruolo centrale nella definizione di un ristretto numero di dimensioni e nella creazione di una rappresentazione grafica in grado di rilevare le distanze che intercorrono fra le modalità. È una tecnica esplorativa e di carattere descrittivo che utilizza il χ^2 (si legge chi quadro) nel calcolo delle distanze e degli indici principali.

3.2 - Analisi delle componenti principali

L'analisi delle componenti principali (*Principal Component Analysis*, PCA) è, come abbiamo anticipato, una tecnica di estrazione dei dati e riduzione della

dimensionalità che viene a volte utilizzata come unica tecnica di analisi in uno studio, altre all'interno di un'analisi fattoriale per l'estrazione dei fattori, altre ancora come primo passo in una ricerca più complessa dove, dopo la riduzione delle dimensioni, vengono applicati altri metodi specifici per rispondere agli obiettivi dell'indagine.

Differentemente dall'analisi fattoriale, la PCA è una tecnica descrittiva, senza intenti di generalizzazione. Si accomuna invece all'EFA, poiché lavora su variabili quantitative. Useremo la PCA, ad esempio, per sintetizzare con pochi elementi i voti conseguiti in tutti gli esami di un determinato corso di laurea da parte di un gruppo di studenti.

Lo scopo principale della PCA è sostituire ***n*** variabili correlate fra loro (esempio: i voti degli esami) con un numero inferiore di variabili, definite *componenti*, fra loro non correlate. La condizione per effettuare questo scambio è quella di conservare la più alta percentuale possibile di informazioni sull'andamento delle distribuzioni osservate. In pratica informazioni sulla loro variabilità o in maniera ancora più concreta sulla varianza delle variabili, misura che assume un ruolo centrale in queste tecniche e ci dice come sono distribuite le osservazioni e come variano rispetto alla media. Per raggiungere questo obiettivo si lavora su una trasformazione dello spazio dimensionale per ridurre la varianza di alcune variabili estraendo dall'analisi soltanto delle componenti che sono dimensioni reali (non ipotetiche e stimate come nell'EFA) e, nella pratica, combinazioni lineari delle variabili dello studio.

Come si procede nell'analisi concretamente? I seguenti step spiegano le procedure solo accennate finora:

1. verifichiamo che le variabili siano in relazione fra di loro;
2. determiniamo le componenti effettuando una rotazione nello spazio dimensionale;
3. scegliamo un numero ridotto di componenti da conservare nell'analisi;
4. verifichiamo la relazione esistente fra le variabili iniziali dello studio e le componenti ottenute.

La prima azione da compiere per applicare la PCA quindi è verificare la relazione fra le variabili iniziali. L'operazione viene compiuta attraverso la determinazione della matrice di covarianza nella quale la presenza di alti valori indica che le variabili variano insieme e dunque sono in una qualche relazione fra

loro. Tuttavia, affinché la differenza di scale nella covarianza non incida sui risultati è preferibile sostituire, come nella pratica accade quasi sempre, la matrice di covarianza con la matrice di correlazione. Alte correlazioni fra le variabili ci dicono che fra di esse esistono delle relazioni.

Ciò fatto, potremo passare quindi alla seconda fase per determinare le combinazioni lineari fra le variabili che identificano le componenti. Avremo tante componenti quante sono le variabili e potremo osservare che la somma della varianza calcolata per le variabili originali misurate è pari alla somma della varianza calcolata per le componenti. Le informazioni sulla varianza delle variabili, cioè sulla variabilità, non si perdono ma sono ricombinate nella varianza delle componenti. Quando le variabili originali sono standardizzate, la varianza per ogni variabile è pari a 1 e di conseguenza la somma delle varianze corrisponde al numero di variabili dello studio, tutte contribuiscono con uno stesso peso sulla varianza totale.

Ma come si trovano le componenti ossia le combinazioni lineari delle variabili osservate? In estrema sintesi bisogna trovare i valori da attribuire ai coefficienti a_{pp} per ottenere le Y_p dalle seguenti relazioni:

(3.1)

$$
\begin{aligned}
Y_1 &= a_{11}x_1 + a_{21}x_2 + a_{31}x_3 + \ldots + a_{p1}x_p \\
Y_2 &= a_{12}x_1 + a_{22}x_2 + a_{32}x_3 + \ldots + a_{p2}x_p \\
Y_3 &= a_{13}x_1 + a_{23}x_2 + a_{33}x_3 + \ldots + a_{p3}x_p \\
&\cdots \\
Y_p &= a_{1p}x_1 + a_{2p}x_2 + a_{3p}x_3 + \ldots + a_{pp}x_p
\end{aligned}
$$

dove con x_p indichiamo le variabili osservate, Y_p le componenti da definire, a_{pp} i pesi che definiscono quanto ogni singola variabile contribuisce a determinare ciascuna componente.

Queste espressioni sono il risultato della trasformazione nello spazio dimensionale a cui abbiamo accennato finora e che può essere spiegata facilmente in un piano considerando un dataset composto solo due variabili di partenza (restiamo nell'esempio dei voti considerando solo due esami).

Poniamo che le variabili (i risultati dei due esami) abbiano un'alta correlazione fra di loro (0,81 nel caso in Figura 3.1). In tal caso, le osservazioni saranno disposte all'incirca lungo una retta nel piano cartesiano. Per riuscire a lavorare sulle stesse osservazioni, riducendo le dimensioni (e dunque il numero delle

variabili) si possono ruotare gli assi nel piano in modo che la varianza di una delle due variabili sia minimizzata e tutta la varianza da studiare appartenga a un'unica variabile. In Figura 3.1 gli assi rotati sono indicati con la linea tratteggiata e le maiuscole *X* e *Y*. Con *d* indichiamo la distanza dei punti dall'asse *X*, distanza che è inferiore a quella dello stesso punto dall'asse *x*. Nel nuovo sistema di assi *XY*, la distanza rispetto all'asse *X* è minimizzata; più elevata è invece la distanza dei punti dall'asse *Y*. La varianza della dimensione *Y (Exam 1)* è ridotta mentre è aumentata quella della dimensione *X*. Le coordinate in cui sono espressi i punti nel nuovo sistema di riferimento sono combinazioni lineari delle precedenti espresse nel sistema *xy* e la varianza della variabile indicata come *Exam 2* ha un peso maggiore ai fini del calcolo.

Come detto già, nella trasformazione non si perde la varianza delle variabili che resta espressa dalla varianza delle componenti.

Con l'uso di questa tecnica assumiamo che la varianza specifica di ciascuna variabile (ossia la varianza determinata per lo più da errori di misura) non influenzi la variabilità osservata che consideriamo come varianza totale e che corrisponde alla variazione derivante dal legame di ciascuna variabile con la componente latente.

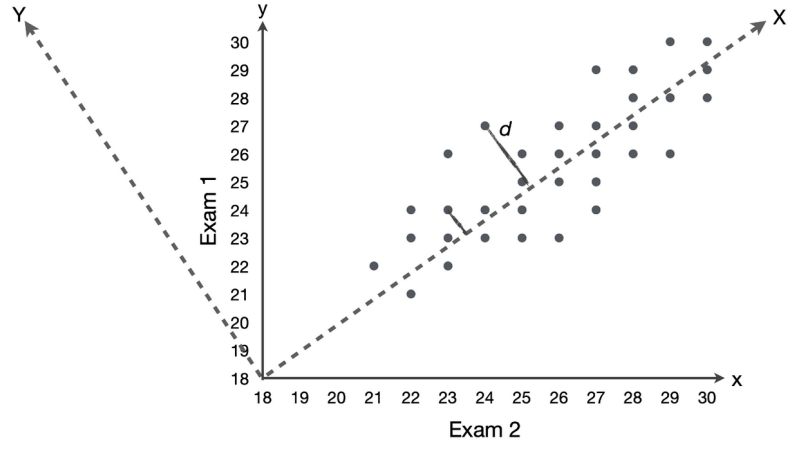


Figura 3.1 - Rotazione degli assi cartesiani e riduzione della varianza nella PCA.

Se la rotazione degli assi risulta semplice in uno spazio a due dimensioni con due sole variabili, trovare il modo in cui ruotare uno spazio a più dimensioni, e di conseguenza trovare i valori delle a_{pp} in base ai quali scrivere la combinazione lineare che ci permette di individuare le componenti principali, può essere difficile e laborioso. È necessario utilizzare l'algebra lineare attraverso cui calcolare autovalori e autovettori della matrice di correlazione o di covarianza (per ripassare gli argomenti relativi ad autovalori e autovettori di una matrice quadrata può essere utile consultare "Algebra Lineare" di Marco Abate, 2000).

Nel calcolo di autovalori e autovettori impostiamo alcune condizioni che permettono di scegliere le a_{pp} ossia:

- la varianza totale delle Y_p è pari alla varianza totale delle x_p;
- la somma delle a_{pp} per ciascuna componente/variabile deve essere pari a 1;
- la prima e la seconda componente sono ortogonali (e quindi non correlate fra loro).

Gli autovalori sono la varianza delle componenti Y_p. La prima componente ha varianza più alta della seconda e così via a cascata. Spesso la prima componente raccoglie la varianza di un numero maggiore di variabili e quindi spiega la maggior parte della varianza del dataset.

I valori delle a_{pp} rappresentano la correlazione fra variabili originali e componenti. Questi valori ci aiutano a capire quante e quali variabili influenzano maggiormente una componente.

Per capire come i meccanismi visti finora si applicano nella pratica, prendiamo in analisi ancora l'esempio del dataset composto dai voti di un gruppo di studenti universitari considerando questa volta sei esami sostenuti. Il nostro scopo è di sintetizzare le informazioni sui sei esami in un numero inferiore di dimensioni. Verificata la correlazione, applichiamo il metodo della PCA, operazione che nei software di analisi statistica si riassume nella scrittura di poche funzioni (per nostra fortuna!). In `R`, ad esempio, è sufficiente la funzione di base `prcomp` (libreria `stats`). I risultati visualizzati con la funzione `summary` appaiono come nella Figura 3.2.

```
> summary(pca)
Importance of components:
                          PC1    PC2     PC3     PC4    PC5     PC6
Standard deviation     1.7650 0.9971 0.77442 0.68932 0.6614 0.61516
Proportion of Variance 0.5192 0.1657 0.09996 0.07919 0.0729 0.06307
Cumulative Proportion  0.5192 0.6849 0.78483 0.86403 0.9369 1.00000
```

Figura 3.2 – Risultati di una PCA realizzata su un dataset con sei variabili (esami). L'analisi dati è stata realizzata in `R` con la funzione di base `prcomp` (libreria `stats`) e visualizzata con la funzione `summary`. Nelle righe troviamo deviazione standard, varianza e varianza cumulata per le sei componenti (colonne).

Osserviamo che 6 variabili ci hanno portato alla definizione di 6 componenti nelle colonne indicata come PC1, PC2, … , PC6. La prima riga ci restituisce la deviazione standard delle componenti. La seconda la proporzione di varianza ossia il rapporto fra la varianza della componente principale (che ricordiamo essere il quadrato della deviazione standard della riga precedente) e la somma delle varianze delle 6 componenti ottenute. In terza riga vediamo poi la proporzione di varianza cumulata ossia la somma della proporzione di varianza per componenti successive.

Usando l'esempio siamo giunti quindi alla terza e alla quarta fase di lavoro nella PCA: (3) scegliere un numero ridotto di componenti e (4) verificare la relazione esistente fra le variabili iniziali dello studio e le componenti ottenute.

Dal `summary` in Figura 3.2, capiamo che la prima componente PC1 spiega il 51,92% della varianza del dataset; la PC2 solo il 16,57%; la terza PC3 il 9,99% e così via. Come detto, il valore della varianza diminuisce andando avanti con le componenti. La proporzione cumulata ci dice che se decidessimo di tenere in considerazione solo le prime due componenti, spiegheremmo il 68,49% (ossia sommando i valori della riga precedente: 51,92% + 16,57%) della varianza delle variabili; aggiungendo una terza, raggiungiamo il 78,48% della varianza spiegata (51,92% + 16,57% + 9,99%).

Considerare la proporzione cumulata di varianza è uno dei metodi per definire il numero di variabili a cui fermarsi. Ovviamente affinché la procedura abbia senso, non possiamo considerare nell'analisi tutte le componenti estratte. Sarebbe sciocco applicare una trasformazione su un dataset di 6 variabili come in questo caso per ottenere 6 componenti. Sarà necessario invece definire un numero di componenti *principali* (da cui il nome della tecnica) da considerare

come definitive e rappresentative dell'intero dataset sacrificando il minor numero di informazioni.

Oltre alla proporzione cumulata della varianza, possono essere usati metodi comuni anche all'analisi fattoriale per definire il numero di componenti a cui fermarsi. Uno di questi prevede che si considerino principali solo le componenti che hanno come autovalore un numero superiore a 1 (*Kaiser-Guttman rule*) affinché spieghi il significato di almeno una variabile originale. Altro metodo è quello della lettura di uno *scree test* (Figura 3.3), nel quale vengono plottati le componenti con gli autovalori. Nella visualizzazione grafica si identifica il punto in cui gli autovalori cominciano ad assumere valori simili fra loro e la retta disegnata (in blu) assume un andamento lineare poiché la varianza singola comincia a dominare la struttura della varianza comune. Le componenti che spesso vengono considerate sono quelle antecedenti al punto di variazione dell'andamento della retta o a quello immediatamente successivo (Es. due o tre in Figura 3.3).

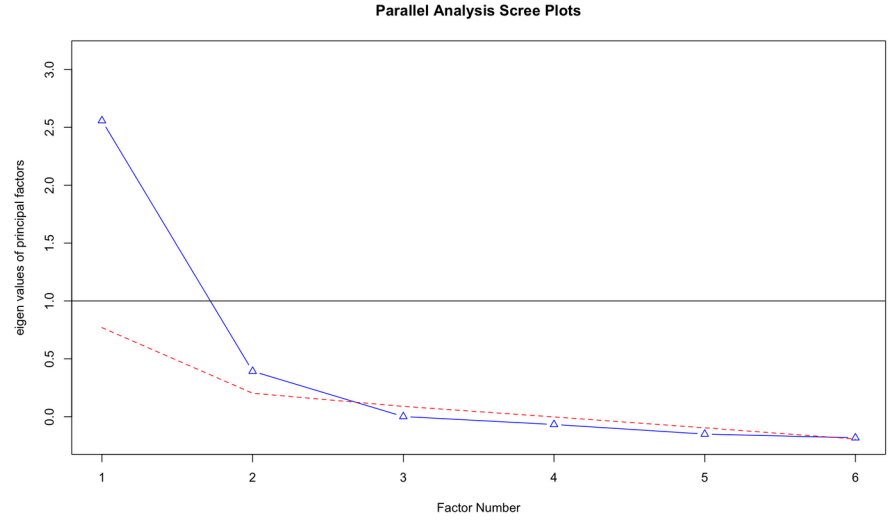


Figura 3.3 - Esempio di scree test. Nella figura in blu è rappresentato lo scree plot reale, in rosso quello stimato. Soltanto un fattore ha autovalore superiore a 1. A partire dal fattore 3, gli autovalori cominciano ad assumere un valore simile. Il grafico è stato realizzato in `R` con la funzione `fa.parallel` (libreria `psych`).

Bisogna decidere in questa fase quindi se "accontentarsi" di rilevare una percentuale di varianza cumulata più bassa e avere un numero inferiore di componenti ma meglio visualizzabili o se introdurre più componenti e aumentare la varianza spiegata, complicando inevitabilmente l'analisi.

Ricordiamo che se si riesce a restare nel numero di due o tre componenti, sarà possibile visualizzare in un piano bi- o tri- dimensionale i dati e sappiamo che la visualizzazione aggiunge sempre molte informazioni in un'analisi.

Nella Figura 3.4, vediamo un plot di sintesi della PCA dove sulle ascisse c'è la PC1, sulle ordinate la PC2 le coordinate sono calcolate applicando le equazioni 3.1, essendo noti le a_{pp} e le x_i). Si tratta della distribuzione delle osservazioni nelle due dimensioni in uno spazio bi-dimensionale. Le frecce rosse che vengono riportate nel piano usando come coordinate rispettivamente la correlazione della variabile x_p con le componenti PC1 e PC2 (Figura 3.5), indicano la quantità e la direzione in cui ciascuna variabile contribuisce a definire le componenti. Il grafico e le coordinate (coefficienti di correlazione ossia le a_{pp}) ci forniscono le informazioni sulla relazione esistente fra variabili originali e le componenti calcolate.

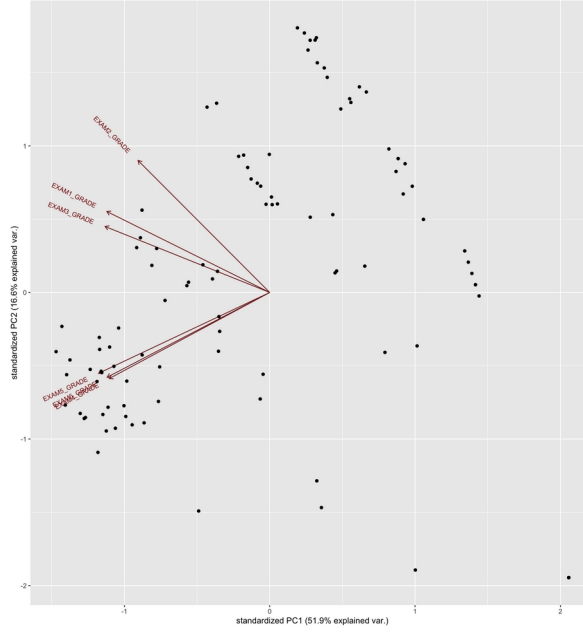


Figura 3.4 - Plot delle componenti PC1 e PC2. La PC1 spiega il 51,92% della varianza del dataset; la PC2 il 16,57%. Le frecce rosse corrispondenti a ciascuna variabile indicano la quantità e la direzione in cui ciascuna variabile contribuisce a definire le componenti.

```
> cor(data, pca$x)
                 PC1        PC2         PC3         PC4         PC5         PC6
EXAM1_GRADE -0.7362153  0.3629897  0.20088321 -0.39073414 -0.36352305 -0.03239170
EXAM2_GRADE -0.5954038  0.5917727 -0.46171321  0.26739338 -0.01080499 -0.10249047
EXAM3_GRADE -0.7442154  0.2957162  0.39760508  0.08925089  0.39347569  0.19446518
EXAM4_GRADE -0.7253141 -0.3859769  0.21139031  0.43256248 -0.30500869 -0.01071538
EXAM5_GRADE -0.7737550 -0.3624765 -0.06157918 -0.15962030  0.23809246 -0.42890006
EXAM6_GRADE -0.7347882 -0.3800229 -0.37367023 -0.17447357  0.02481770  0.38076564
```

Figura 3.5 - Matrice di correlazione fra variabili osservate e componenti.

3.3 - Analisi fattoriale esplorativa

L'analisi fattoriale appartiene al gruppo delle tecniche di interdipendenza che non hanno lo scopo di predire il comportamento di una variabile dipendente ma di spiegare la struttura latente e le relazioni esistenti tra le variabili osservate, massimizzando le informazioni che i dati raccolti possono fornire.

È uno dei *latent variable methods* più consolidati e utilizzati e si fonda sull'idea che p variabili misurate siano una manifestazione di uno o più fenomeni latenti, non osservabili e misurabili direttamente. Ciascuna variabile del dataset è il risultato dell'azione di più fattori. Le modalità e la variabilità delle variabili osservate sono gli elementi da cui partire per individuare i fattori latenti in grado di descrivere i fenomeni in maniera più significativa. L'obiettivo che ci poniamo è quello di individuare un numero di fattori inferiore al numero delle variabili osservate per delineare un modello, cioè una struttura latente generalizzabile anche ad altri campioni.

L'analisi delle correlazioni anticipa l'applicazione della tecnica: affermare che le variabili originali sono correlate fra loro significa dire in questa tecnica che esse sono connesse a uno stesso fattore latente. Quindi la varianza di ogni singola variabile è in parte giustificata dal rapporto che la lega al fattore *comune* (comune non perché generico ma perché correlato a più variabili), in parte alla distribuzione stessa delle osservazioni. Partiamo da un alto numero di variabili correlate fra loro per arrivare, al termine dell'analisi, a un numero ridotto di fattori non correlati: le variabili osservate sono correlate perché collegate a uno stesso fattore; i fattori individuati non sono correlati fra loro perché espressione di dimensioni latenti diverse.

Un esempio molto frequente utilizzato per far comprendere lo scopo dell'analisi fattoriale (e in genere le strutture latenti) è il costrutto dell'intelligenza. L'intelligenza è un costrutto non misurabile e osservabile di per sé ma rilevabile da comportamenti e prestazioni di un individuo. Le domande di un que-

stionario piuttosto che i punteggi in una serie di prove di valutazione possono rivelare alcuni tratti dell'intelligenza. La correlazione fra le variabili utilizzate per misurare e rilevare questi aspetti ci dice che tutte sono legate a un elemento che misurabile *in toto* non è: l'intelligenza appunto.

Possiamo considerare lineare la relazione che esiste fra le variabili manifeste e quelle latenti. Questo approccio propone i meccanismi di predizione (e dunque di regressione come vedremo nel capitolo successivo) fra variabile latente/dipendente e manifeste/indipendenti: possiamo conoscere la variabile latente solo indirettamente e a partire dalle modalità assunte dalle variabili manifeste.

Nell'analisi fattoriale assumiamo che le variabili latenti da cercare siano metriche; esistono altri *latent variable methods* nei quali variano la tipologia delle variabili manifeste e osservate (es. variabili categoriali).

La tipologia di analisi fattoriale che prendiamo in considerazione nella nostra discussione è quella definita *esplorativa* (*Exploratory Factorial Analysis,* EFA) che utilizza meccanismi di data summarization e che, da quanto abbiamo detto finora, si basa su dati reali raccolti per trovare fattori latenti a cui attribuire un significato. L'altra tipologia molto nota di analisi fattoriale è quella *confermativa*: questa tecnica di analisi utilizza i dati raccolti per testare teorie, ricerche e ipotesi sul modo in cui variabili osservate possono essere raccolte per spiegare una certa struttura alla base di un fenomeno in un dataset.

L'EFA più frequentemente usata è la *R-type factor analysis* nella quale, così come discusso finora, il lavoro di riduzione è condotto sulle variabili. Più raramente si usa la *Q-type factor analysis* dove il metodo di riduzione è applicato sulle osservazioni per suddividere queste ultime in gruppi simili. Poiché risulta essere un metodo laborioso, per raggiungere gli stessi scopi se ne preferiscono altri come la cluster analysis (capitolo 6). Diversamente da quest'ultima che per la creazione dei cluster utilizza la distanza fra le osservazioni, la *Q-type factor analysis* lavora sull'intercorrelazione fra le unità statistiche.

Elencate le finalità della tecnica, descriviamo quindi la sua procedura di applicazione.

Innanzitutto serve partire dalla verifica di assunti concettuali e statistici:

- da un punto di vista teorico, il ricercatore deve ipotizzare che esista una struttura latente fra le variabili. I calcoli nell'analisi di per sé non possono garantire che la riduzione o la sintesi delle variabili abbia senso in assoluto;

- le variabili devono riguardare uno stesso nucleo concettuale, elemento che giustifica l'esistenza di fattori latenti in comune fra di esse. Inoltre, il numero di variabili (metriche) deve all'incirca essere il quintuplo dei fattori individuati;

- le deviazioni dalla normalità, dall'omoschedasticità e dalla linearità possono diminuire le correlazioni tra le variabili. Testare la normalità è fondamentale per verificare la significatività dei fattori;

- il campione deve essere omogeneo, non devono esserci differenze fra le osservazioni che possano invalidare o "nascondere" la struttura dei dati osservati. Esso deve essere sufficientemente numeroso: anche se ci sono indicazioni di diversa natura, in genere è suggerito di disporre di un campione fatto da più di 50-100 osservazioni, almeno con 5-20 unità statistiche per ogni variabile.

Poniamo quindi di trovarci nel caso in cui, soddisfacendo gli assunti appena elencati, un analista debba analizzare i risultati di un'indagine composta da oltre 40 item somministrata ai dirigenti delle scuole di primo e secondo grado del vecchio continente per un progetto europeo di comparazione dei sistemi educativi fra i paesi dell'UE. L'indagine ha restituito una grande quantità di dati per la numerosità dei rispondenti e per l'elevato numero di domande somministrate che, come spesso accade, coincidono con il numero di variabili da analizzare. Supponiamo con un certo grado di certezza che ci siano dei fattori comuni alle variabili in grado di descrivere macroaree di comparazione fra i sistemi educativi (ad esempio, l'idea sulla funzione della scuola nel contesto sociale, la percezione del coinvolgimento delle famiglie e del ruolo degli insegnanti, la spinta all'innovazione). Come procedere a questo punto per rilevare tali fattori comuni non osservabili direttamente ma fondamentali per comprendere gli orientamenti dei vari paesi?

Nelle varie fasi di analisi il ricercatore è chiamato a fare delle scelte, scelte che secondo alcune scuole di pensiero potrebbero mettere in discussione l'oggettività della tecnica. Per elencarne alcune: il numero di fattori da estrarre, il metodo da usare per l'estrazione dei fattori, le tecniche di rotazione da applicare.

Per accertarsi di poter usare l'EFA, egli deve verificare la fattoriabilità del dataset e dunque:

- calcolare la matrice di correlazione per verificare le relazioni fra le variabili nel dataset. Se la maggior parte delle correlazioni è inferiore a 0,30, non ha senso usare l'analisi fattoriale.

- applicare il test di sfericità di Bartlett che misura la significatività statistica dei valori nella matrice di correlazione con il χ^2 ; l'ipotesi nulla da invalidare è che la matrice di correlazione sia il risultato di variabili indipendenti.

- calcolare la misura di adeguatezza campionaria (MSA) o di Kaiser-Meyer-Olkin (KMO), un indice fra 0 e 1 dove il valore 1 indica che le correlazioni parziali fra le variabili sono basse, dunque elevata è la correlazione con i fattori comuni. Valori al di sotto di 0,5 non sono accettabili.

Queste verifiche ci permettono di avviare finalmente la procedura di estrazione dei fattori attraverso l'equazione che segue e che descrive il modello generale lineare fattoriale per p variabili osservate e j fattori comuni o variabili latenti:

(3.2)

$$x_p = a_p + b_{p1} f_1 + b_{p2} f_2 + \ldots + b_{pj} f_j + u_p$$

dove x_p sono le variabili osservate, b_{pj} sono i pesi fattoriali (*loading*), f_i i fattori comuni e u_p la varianza che dipende dalla singola variabile e il suo errore.

Ogni variabile è una combinazione lineare di fattori che descrivono elementi latenti. I valori di b_{pj}, pesi fattoriali, ci dicono quanto ciascun fattore contribuisce alla definizione della variabile.

La formula rovescia le posizioni delle x_p e dei fattori f_i (componenti Y_p) rispetto a quanto visto nell'analisi delle componenti principali: qui ci chiediamo quale fattore e con quale intensità contribuisce a descrivere la variabile originale per determinare la struttura latente nei dati. Nella PCA, lo scopo principale è trovare poche dimensioni reali che possano sostituire numerose variabili osservate; di conseguenza gli elementi da rintracciare sono proprio quelle poche dimensioni, ossia come già detto le componenti Y_p.

Nell'espressione (3.2) troviamo anche il parametro u_p assente nella PCA poiché, come ricorderemo, nella PCA si assume che tale valore sia basso o non consistente e che l'unica varianza da tenere in conto è la varianza totale.

Si può dimostrare che i parametri del modello che descrive le relazioni latenti fra le variabili possono essere individuati dalla covarianza fra le variabili manifeste. Sia la determinazione dei pesi fattoriali che il fattore specifico u_p richiedono che sia posta attenzione a come calcoliamo e concepiamo la varianza nell'EFA. L'abbiamo accennato più volte, ma vale la pena esplicitare in maniera più chiara che essa risulta composta da due parti:

- *la varianza comune:* condivisa con le altre variabili dell'analisi, dipende dalla correlazione con i fattori comuni. La indicheremo in seguito con il termine *comunalità* e la calcoleremo come somma della varianza comune spiegata dal fattore o meglio il quadrato dei pesi fattoriali che rappresentano la correlazione/covarianza fra una variabile e i fattori;

- *la varianza unica con il suo errore:* si tratta della varianza specifica della singola variabile che è attribuita alla distribuzione in sé e il cui errore può essere spiegato solo dai processi di raccolta dati e misurazione.

L'analisi delle componenti principali è una delle tecniche di estrazione dei fattori. Poiché in essa la varianza considerata è solo comune, nella matrice di correlazione usata per determinare autovalori e autovettori si conserva la diagonale composta soltanto da 1 (non prendiamo infatti in considerazione la varianza specifica e il suo errore).

Altri metodi, fra cui ad esempio quello dei fattori principali, ammettono invece, come è comune nell'EFA, che la varianza abbia anche una componente unica e quindi che non tutta la varianza sia spiegata nel legame con i fattori comuni. In questi casi, poiché ammettiamo la presenza di errori, nella matrice di correlazione la diagonale sarà composta da valori inferiori ad 1 poiché per via dell'errore la correlazione della variabile con sé stessa non è perfetta; il valore sostituto deriva dalla stima della mutua interazione. Non ci occuperemo di altri metodi possibili come quello dei residui generalizzati, della fattorializzazione immagine o massima verosimiglianza.

Sia che usiamo la PCA, sia che usiamo il metodo dei fattori principali, riportando le equazioni nella forma matriciale, è possibile calcolare i valori dei pesi fattoriali. Non c'è una matrice unica e quindi per stimare gli elementi che compongono la relazione dobbiamo imporre alcuni vincoli come ad esempio che i fattori f_j siano non correlati fra loro e standardizzati (media 0 e deviazione 1); che i fattori unici u_p siano non correlati fra loro e con i fattori comuni; che entrambi seguano una distribuzione normale.

Il numero massimo di fattori estraibili è pari a quello delle variabili originali.

Fra le scelte del ricercatore rientra il numero di fattori a cui fermarsi. A tal proposito una prima valutazione da fare nasce dal confronto tra il numero di fattori che ci aspettiamo di ottenere per motivazioni teoriche e il numero di fattori prodotti dall'analisi statistica e dalle loro comunalità. La somma dei valori delle comunalità di tutti i fattori prodotti nell'analisi è pari a 1, spiega completamente il fenomeno studiato (100%). Tuttavia gli ultimi fattori estratti hanno meno rilevanza nell'analisi (comunalità più basse) e aggiungendoli all'analisi verrebbe meno il vantaggio della riduzione.

Ci sono alcuni criteri per decidere il numero dei fattori da estrarre, criteri che frequentemente vengono usati con modalità iterative in una stessa analisi fino ad arrivare a una selezione di fattori non soggettiva che esprime la struttura delle variabili latenti senza eccessive perdite di informazioni o complicazioni nella lettura dei dati. Il ricercatore potrebbe stabilire prima di avviare l'analisi il numero di fattori a cui fermarsi (*a priori criterion*) o potrebbe valutare di utilizzare nell'analisi soltanto i fattori con autovalore superiore a 1 (*latent root criterion*) o ancora utilizzare lo scree test descritto nel paragrafo precedente.

La matrice dei pesi fattoriali, ossia le correlazioni di ogni variabile con ciascun fattore, ci permette di individuare la struttura latente del dataset. Più sono elevati i valori della correlazione, più i fattori sono rappresentativi della variabile manifesta considerata.

Accade che la matrice dei fattori non risulti di facile lettura: si procede quindi a una rotazione dei fattori per ridurre le ambiguità nell'interpretazione usando gli stessi principi in Figura 3.1. La rotazione è un'invariante rispetto alle distanze e alle composizioni delle variabili. Esistono molte strategie di rotazione che i software statistici includono e applicano automaticamente. Le rotazioni si distinguono in ortogonali (assi perpendicolari) e oblique (assi non perpendicolari). Una delle più note è la variazione ortogonale detta VARIMAX.

L'interpretazione dei fattori (ruotati o non ruotati) ci permette di visualizzare la struttura del dataset. Ciascun fattore diventa a questo punto rappresentativo del gruppo di variabili che ha con esso i valori di correlazione più alti. Il ricercatore quindi potrà attribuire a quel fattore un nome che sia rappresentativo delle variabili con cui è correlato dando più rilevanza nell'interpretazione e nella scelta del significato alle variabili con correlazione più alta.

La Figura 3.6 mostra i pesi fattoriali (*loadings*) e i risultati forniti da `R` in un'analisi fattoriale esplorativa (funzione: `fa`, libreria: `psych`) su un dataset composto da 6 variabili (a scopo esemplificativo abbiamo usato i dati relativi agli esami dell'esempio del paragrafo precedente) con l'uso della rotazione

obliqua OBLIMIN per l'estrazione di due fattori. I due fattori raccolgono la variabilità delle variabili in gruppi di tre. Il riquadro in basso contiene i valori relativi alla somma dei pesi fattoriali al quadrato (*SS loadings*) considerati nell'analisi se superiori a 1 e alla proporzione di varianza per fattore e varianza cumulata come nella PCA. Nell'esempio, la varianza cumulata usando 2 fattori è pari a 0,52, il modello individuato quindi spiega il 52% della varianza del dataset, valore che riteniamo non essere del tutto soddisfacente. Avremmo bisogno di estrarre più fattori per rendere il modello maggiormente rappresentativo della varianza delle variabili di partenza.

```
> efa2$loadings

Loadings:
             MR1    MR2
EXAM1_GRADE          0.753
EXAM2_GRADE          0.669
EXAM3_GRADE          0.658
EXAM4_GRADE  0.668
EXAM5_GRADE  0.812
EXAM6_GRADE  0.737

                 MR1   MR2
SS loadings    1.669 1.449
Proportion Var 0.278 0.242
Cumulative Var 0.278 0.520
```

Figura 3.6 - Pesi fattoriali e risultati di un'analisi fattoriale esplorativa in `R` (funzione: `fa`, libreria `psych`).

Il ricercatore può scegliere se eliminare dal dataset variabili che non sono rappresentate da nessun fattore o variabili che presentano correlazioni alte con più fattori, definite *cross-loading*, poiché viene meno l'indipendenza dei fattori in questo caso legati a più variabili. Altra situazione che può condurre all'eliminazione di una variabile dal dataset è che il suo valore di comunalità sia inferiore rispetto a quello stabilito dal ricercatore. Bassi valori di comunalità corrispondono ad alti valori di unicità/varianza unica: ciò significa che la varianza della variabile è indipendente da quella del fattore, la variabile non è legata al fattore comune. Alternativa per il ricercatore nei casi in cui si presenti una delle situazioni di cui sopra è proseguire nell'analisi dichiarando le incongruenze rispetto ad alcune variabili oppure ripetere l'analisi modificando le variabili e le tecniche di estrazione o rotazione dei fattori.

L'adeguatezza del modello finale ottenuto può essere determinata dalla percentuale di varianza totale spiegata dai fattori come somma delle comunali-

tà delle variabili (varianza cumulata); dal confronto fra la matrice di correlazione originale e quella denominata "riprodotta", calcolata a partire dai pesi fattoriali; attraverso specifici test che provano la bontà del modello verificando l'ipotesi nulla che prevede che la matrice di covarianza fra le variabili osservate abbia la forma definita dal modello stesso.

Lo studio potrebbe interrompersi a questo punto oppure proseguire con l'applicazione di altre tecniche su un nuovo dataset costituito:

- selezionando – anche a partire dalle conoscenze teoriche precedenti – solo una variabile con peso fattoriale più alto per ciascun fattore (la sostituzione del fattore con la variabile fa correre il rischio di perdere informazioni o raccontare soltanto una parte degli aspetti del fenomeno, scartando tutto quello che il fattore rappresenta);

- creando scale (misure composte di alcune delle variabili usando i pesi fattoriali) o punteggi fattoriali (misure composte da tutte le variabili correlate a un fattore) in modo da riuscire a rappresentare tutti gli aspetti di un determinato concetto in una sola misura. L'α di Cronbach è un noto coefficiente di reliability/affidabilità per la verifica della coerenza degli item di una scala fra loro; assume valori compresi fra 0 e 1 e solitamente viene considerato sufficiente con valori al di sopra dello 0,60.

3.4 - Analisi delle corrispondenze

L'analisi delle corrispondenze (*Correspondence Analysis*, CORA) è una tecnica che applica i meccanismi di riduzione della dimensionalità alle variabili nominali ordinali e categoriali. È un metodo descrittivo, al quale possono essere applicati processi che verificano la significatività e la generalizzabilità dei risultati. Si tratta di una tecnica esplorativa di data analysis, non utilizzata per testare ipotesi ma, ancora una volta, per identificare le strutture latenti in un set di dati. Permette di riprodurre graficamente le similarità delle modalità delle variabili calcolate a partire dalle distanze fra oggetti come nella cluster analysis (CA) e nel multidimensional scaling (MDS). Come quest'ultima tecnica, la CORA è un metodo per il *perceptual mapping*. Le tecniche di percentual mapping hanno l'obiettivo di produrre una mappa percettiva ossia una visualizzazione grafica che rappresenta gli oggetti dell'analisi in uno spazio bi- o multi- dimensionale permettendo di cogliere gli aspetti di somiglianza o differenza fra gli stessi. Da un punto di vista geometrico, potremmo dire che utilizzando la CORA ci si

pone l'obiettivo di trovare un piano nel quale collocare tutti i punti corrispondenti agli oggetti dello studio conservando la distanza che esiste fra di essi (Bartholomew et al., 2008).

In sostanza le differenze di questa tecnica dalle altre già viste sono principalmente due:

1. agiamo su variabili non metriche e quindi utilizziamo modalità, frequenze e distanze per lavorare sui dataset;

2. al termine delle procedure di analisi, insieme a un numero di indicatori possiamo ottenere una visualizzazione grafica da interpretare per identificare la distribuzione delle modalità in un numero inferiore di dimensioni che non sempre sono riconducibili a un concetto precisato.

Nell'applicazione dell'analisi delle corrispondenze, non è necessario testare particolari assunti. Ovviamente bisogna verificare di non invalidare procedure di generalizzabilità, completezza e comparabilità fra gli oggetti. Particolare attenzione va dedicata agli outlier che possono generare degli effetti distorsivi sia nell'elaborazione numerica sia nell'interpretazione grafica.

Partiamo dall'osservare il caso dell'analisi delle corrispondenze semplici, quello cioè che prende in considerazione un'analisi condotta solo su due variabili. Riporteremo poi la tecnica nel campo multivariato (analisi delle corrispondenze multiple).

La struttura latente che cerchiamo di identificare attraverso l'analisi delle corrispondenze è esplicitata dalla vicinanza fra le modalità delle variabili. Detto brevemente, il processo che si segue per ottenere e visualizzare tale struttura è quello di definire le frequenze congiunte delle modalità, rilevare le distanze fra modalità diverse, ridistribuire la varianza del dataset fra un numero limitato di dimensioni, rappresentare graficamente le modalità con le loro distanze nel nuovo spazio dimensionale, interpretare visualizzazioni e indicatori ottenuti.

Facciamo un passo alla volta.

La prima operazione da compiere, parlando di variabili non metriche, è quella di calcolare le frequenze congiunte con cui le modalità si manifestano rispetto alla coppia di variabili in esame e costruire quindi la tabella di contingenza risultante come in Tabella 3.1. Non necessariamente nella tecnica le modalità devono essere esclusive. Nell'esempio che usiamo per illustrare il metodo, consideriamo invece un caso in cui lo sono. I dati sono stati rilevati in uno studio sulla percezione dei sistemi di *e-proctoring* (De Santis et al., 2020) attra-

verso un questionario somministrato al termine della prova *proctored* in alcuni corsi di laurea dell'Università degli Studi di Modena e Reggio Emilia di cui riportiamo le risposte di 323 studenti. Parliamo di e-proctoring riferendoci ai sistemi digitali che permettono di monitorare da remoto lo svolgimento delle prove di valutazione attraverso meccanismi di controllo che evitino il plagio o il confronto fra gli studenti e che vengono studiati spesso in relazione a fenomeni legati all'aumento dell'ansia o al peggioramento delle performance (González-González et al., 2020; Kolski & Weible, 2018; Reisenwitz, 2020).

Le variabili riportate nella Tabella 3.1 sono il corso di laurea frequentato (righe) e il livello di gestione dell'ansia nell'uso dei sistemi di e-proctoring dichiarato dagli studenti (colonne).

Corso di laurea	**Gestione dell'ansia**				**Totale**
	L1	**L2**	**L3**	**L4**	
BLD006	14	5	10	11	40
BLD023	27	18	13	10	68
BLD126	7	15	23	30	75
FIM049	17	9	5	4	35
MED054	13	14	17	8	52
MED055	20	13	11	9	53
Totale	98	74	79	72	323

Tabella 3.1 - Tabella di contingenza per le variabili CORSO DI LAUREA (righe) e livello di gestione dell'ANSIA (colonne) in un dataset relativo alla percezione dei sistemi di e-proctoring in ambito universitario.

L'uso della CORA serve a comprendere se esistono associazioni (*corrispondenze*) nel modo in cui negli studenti appartenenti a un determinato corso di laurea l'uso di strumenti di e-proctoring ha influito sulla riduzione dell'ansia in una scala da 1 (in completo disaccordo) a 4 (in completo accordo). La tabella ci fornisce alcune indicazioni: ad esempio, 68 dei 323 studenti che hanno partecipato all'indagine appartengono al corso di laurea indicato con la sigla BLD023; 98 dichiarano che l'utilizzo del sistema di e-proctoring non ha influenzato positivamente la percezione dell'ansia nello svolgimento della prova (livello L1); 17 studenti del corso MED054 ritengono che utilizzare un sistema di controllo durante la prova d'esame ha influenzato abbastanza positivamente la propria gestione dell'ansia (livello L3) e così via.

Per lavorare sulle associazioni fra le modalità, calcoliamo i profili di riga e colonna ossia le percentuali relative e dunque il rapporto fra i valori contenuti in ogni singola cella (frequenza congiunta) e rispettivamente il totale di riga o di colonna (frequenza marginale).

Introduciamo anche altri due indicatori:

- *massa di riga (o colonna)*, ossia il rapporto fra il totale di riga (o colonna) e il totale dei casi;

- *centroidi o medie di profilo di riga (o colonna),* profilo marginale, ossia il rapporto fra il totale di riga (o colonna) e il totale dei casi.

Come è ovvio, masse di riga e centroidi dei profili di colonna coincidono e viceversa, le masse di colonna coincidono i centroidi dei profili di riga.

La Tabella 3.2 rappresenta i profili di riga, la Tabella 3.3 i profili di colonna. È una scelta dell'analista quella di lavorare sulle righe o sulle colonne sia nelle fasi di calcolo che in quelle di interpretazione. Quello che diremo sui profili di riga può essere riportato ai profili di colonna. Come vedremo negli esempi, i software di analisi dati ci permettono con immediatezza di passare dall'una all'altra preferenza.

Profili simili indicano che nell'indagine i partecipanti hanno scelto modalità simili. Più vicini sono i valori, più vicini saranno i punti da disegnare nella visualizzazione finale. Allo stesso modo, più i valori di ogni cella sono vicini a quelli dei centroidi, più i punti saranno vicino all'origine degli assi che rappresenta esattamente il punto medio.

Attraverso quali strumenti di calcolo possiamo "misurare la vicinanza" fra i profili e fra i profili e i centroidi per rilevare i pattern che spiegano le associazioni fra le celle?

Lo strumento usato è il χ^2 sia come test statistico sia come misura di distanza.

Il test del χ^2 della tabella di contingenza ci permette di verificare l'esistenza di una dipendenza significativa fra righe e colonne. Nel nostro caso, possiamo rifiutare l'ipotesi nulla di indipendenza.

La distanza del χ^2 fornisce la distanza fra i profili di riga (o colonna), distanza ponderata rispetto ai profili marginali. Includere nel calcolo i profili marginali evita che questi assumano più rilevanza per modalità con alti totali di riga come accade in un altro tipo di distanza definita euclidea.

Corso di laurea	Gestione dell'ansia				Masse di riga
	L1	L2	L3	L4	
BLD006	0,35	0,12	0,25	0,28	0,12
BLD023	0,40	0,26	0,19	0,15	0,21
BLD126	0,09	0,20	0,31	0,40	0,23
FIM049	0,49	0,26	0,14	0,11	0,11
MED054	0,25	0,27	0,33	0,15	0,16
MED055	0,38	0,25	0,21	0,17	0,16
Media o centroidi dei profili di riga	0,30	0,23	0,25	0,22	1,00

Tabella 3.2 - Profili di riga, masse di riga, media o centroidi dei profili di riga per le variabili CORSO DI LAUREA (righe) e livello di gestione dell'ANSIA (colonne) in un dataset relativo alla percezione dei sistemi di e-proctoring in ambito universitario.

Corso di laurea	Gestione dell'ansia				Media o centroidi dei profili di colonna
	L1	L2	L3	L4	
BLD006	0,14	0,07	0,13	0,14	0,12
BLD023	0,28	0,24	0,16	0,14	0,21
BLD126	0,07	0,20	0,29	0,42	0,23
FIM049	0,17	0,12	0,06	0,06	0,11
MED054	0,13	0,19	0,22	0,11	0,16
MED055	0,20	0,18	0,14	0,12	0,16
Masse di colonna	0,30	0,23	0,25	0,22	1,00

Tabella 3.3 - Profili di colonna, masse di colonna, media o centroidi dei profili di colonna per le variabili CORSO DI LAUREA (righe) e livello di gestione dell'ANSIA (colonne) in un dataset relativo alla percezione dei sistemi di e-proctoring in ambito universitario.

(3.3)

$$\chi^2 = 43{,}159\ , \quad df = 15\ , \quad p = 0{,}0001478$$

Di seguito esprimiamo in formule la distanza euclidea e la distanza del χ^2 per i profili di riga 1 e 2 della Tabella 3.2 e verifichiamo la differenza che intercorre fra le due.

(3.4) DISTANZA EUCLIDEA (profili di riga 1 e 2)

$$\mathrm{d}(1,2) = \sqrt{(0{,}35-0{,}40)^2+(0{,}12-0{,}26)^2+(0{,}25-0{,}19)^2+(0{,}28-0{,}15)^2} =$$

$$= \sqrt{(0{,}0025+0{,}0196+0{,}0036+0{,}0169)} = 0{,}2064$$

(3.5) DISTANZA DEL χ^2 (profili di riga 1 e 2)

$$\mathrm{d}(1,2) = \sqrt{\frac{(0{,}35-0{,}40)^2}{0{,}30}+\frac{(0{,}12-0{,}26)^2}{0{,}23}+\frac{(0{,}25-0{,}19)^2}{0{,}25}+\frac{(0{,}28-0{,}15)^2}{0{,}22}} =$$

$$= \sqrt{0{,}0083+0{,}0852+0{,}0144+0{,}0768} = 0{,}4298$$

Attraverso la stessa formula si può calcolare la distanza fra i profili di riga (o colonna) e i centroidi.

(3.6) DISTANZA DEL χ^2 (profilo di riga 1 e media del profilo riga)

$$\mathrm{d}(1,\mathrm{c}) = \sqrt{\frac{(0{,}35-0{,}30)^2}{0{,}30}+\frac{(0{,}35-0{,}30)^2}{0{,}30}+\frac{(0{,}12-0{,}23)^2}{0{,}23}+\frac{(0{,}25-0{,}25)^2}{0{,}25}+\frac{(0{,}28-0{,}22)^2}{0{,}22}} =$$

$$= \sqrt{0{,}0083+0{,}0526+0+0{,}0164} = 0{,}2780$$

Calcolata la distanza profili-centroidi, possiamo ottenere l'inerzia, indicatore fondamentale nella CORA poiché coincide con la varianza e la sua scomposizione ci permette di definire le dimensioni da includere nei processi di riduzione della dimensionalità.

L'inerzia totale è la somma dei prodotti fra masse e distanze dai centroidi (di riga o colonna) e si può dimostrare che corrisponde al χ^2 normalizzato (diviso cioè per il numero di osservazioni):

(3.7)

$$Inerzia\,totale = \sum (massa\ \ di\ \ riga\ \ i) \cdot d_i^2 = \frac{\chi^2}{N}$$

dove d è la distanza del profilo di riga i dal centroide e N la numerosità del campione.

La somma dell'inerzia calcolata per i profili di riga coincide con quella calcolata per i profili delle colonne. L'inerzia rappresenta la misura della varianza dei profili, il valore dell'inerzia totale viene scomposto negli autovalori che coincidono con la varianza delle dimensioni da estrarre. Possiamo scrivere quindi:

(3.8)

$$Inerzia\,totale = \sum (massa\ \ di\ \ riga\ \ i) \cdot d_i^2 = \frac{\chi^2}{N} = \sum \lambda_k^2$$

dove k è il numero degli autovalori λ in cui viene scomposta l'inerzia ed è pari al numero di dimensioni dell'analisi in questione. Si può dimostrare che k coincide con il valore minimo delle modalità delle variabili osservate meno uno.

Come nelle tecniche precedenti, anche nella CORA i valori degli autovalori sono restituiti in ordine decrescente, dal più grande al più piccolo. Essi indicano quanto la dimensione a cui sono riferiti è rilevante nell'analisi e quindi permettono di selezionare il numero di dimensioni da considerare.

La scelta del numero di dimensioni, come visto nell'EFA e nella PCA, può essere delegata alle rilevazioni da uno scree test (Figura 3.7), al valore della varianza cumulata o alla limitazione a priori del numero di dimensioni da parte del ricercatore. Ricordiamo comunque che questa operazione è un gioco di equilibri: utilizzare più dimensioni aggiunge informazioni nell'analisi ma può renderla più complicata; allo stesso modo ridurre il numero di dimensioni può semplificare l'interpretazione ma ridurre anche le informazioni ottenute.

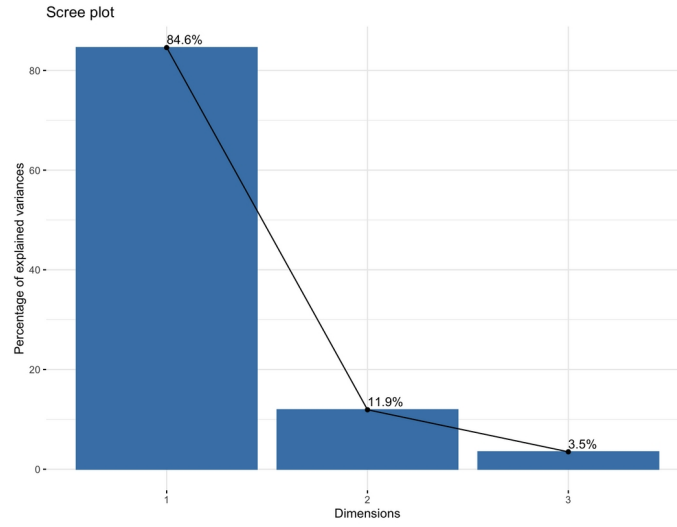


Figura 3.7 - Scree plot per CORA: plot degli autovalori e delle dimensioni (funzione `fviz_screeplot`, libreria `factoextra`).

Dal punto di vista grafico, l'inerzia totale è la somma delle distanze dei punti dall'origine degli assi che, come accennato, si fa coincidere con i profili medi. Una volta calcolata l'inerzia, quindi, si può procedere al calcolo delle coordinate di ciascun punto (ciascuna modalità) come proiezioni dello stesso punto sugli assi.

Nel biplot, l'angolo che si crea fra i segmenti con estremi nell'origine e nel punto che indica la modalità indica il grado di relazione fra due modalità: più l'angolo è acuto, più le due modalità sono similari. Alti livelli di associazione (similarità) fra le modalità saranno visibili nella vicinanza dei punti in una mappa di percezione. Punti più vicini all'origine rappresentano modalità con valori più simili a quelli medi; punti più lontani indicano valori delle modalità che si discostano dai valori attesi. Nell'interpretazione dell'andamento delle modalità si considera anche la loro distribuzione in quadranti e in semipiani positivi e negativi.

Per il modo in cui le distanze sono calcolate (scalate rispetto alle frequenze marginali) nell'analisi delle corrispondenze semplici nei biplot definiti simmetrici si possono confrontare soltanto modalità della stessa variabile. Si è soliti fare generalizzazioni sulle distanze fra le modalità appartenenti a variabili diverse. Il confronto fra modalità appartenenti a più variabili può essere condotto nei biplot asimmetrici nei quali le coordinate dei punti in colonna sono sostituiti con i

valori estremi (considerando il caso estremo in cui tutte le unità statistiche rilevate su una riga cadano in una sola colonna).

Ritorniamo al nostro esempio e proviamo a lavorare sui dati utilizzando `R` e due fra i pacchetti specifici per l'analisi delle corrispondenze: `FactoMineR` e `factoextra` (per indicazioni più complete per questa e le precedenti tecniche si veda Kassambara, 2017).

Con i dati a nostra disposizione, possiamo calcolare il valore dell'inerzia totale che è pari a 0,134.

(3.9)

$$Inerzia\,totale = \frac{\chi^2}{N} = \frac{43{,}159}{323} = 0{,}134$$

La funzione `CA` del pacchetto `FactoMineR` usa come argomento principale la tabella di contingenza senza totali (Tabella 3.1) e restituisce la visualizzazione grafica in Figura 3.8 e gli indici in Figura 3.9.

Otteniamo 3 dimensioni (come ci aspettavamo, il minimo delle modalità per variabile meno uno). La prima dimensione spiega l'84,580% della varianza, la seconda 11,930%, la terza 3,491%. Utilizzando le prime due dimensioni riusciamo a spiegare il 96,509% della varianza del dataset.

Notiamo che sommando l'inerzia per le modalità in riga o in colonna otteniamo 0,134. Sommando gli autovalori (varianza/inerzia) delle 3 dimensioni otteniamo ancora 0,134. Questa uguaglianza non ci meraviglia poiché l'inerzia totale calcolata dalla tabella di contingenza viene ridistribuita come varianza di un numero limitato di dimensioni.

Le coordinate dei punti/modalità, insieme ai contributi (`crt`) e al coseno quadro (`cos2`) vengono restituiti sia per riga che per colonna (Figura 3.9).

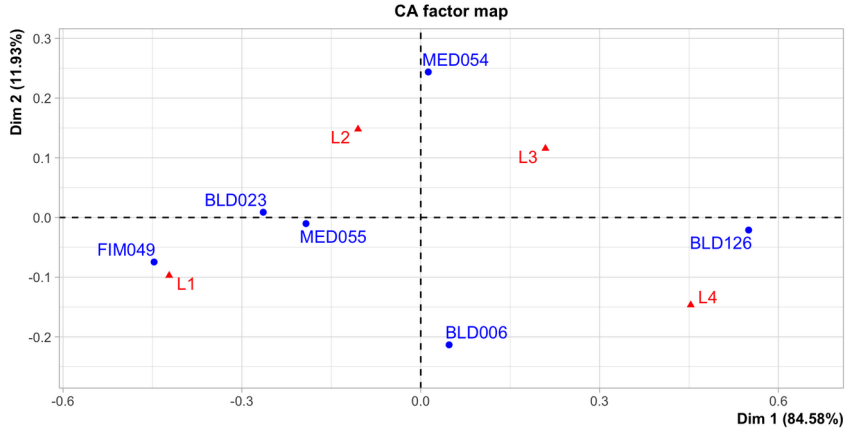


Figura 3.8 - Biplot simmetrico nell'analisi delle corrispondenze semplici riferita alle variabili ANSIA e CORSO DI LAUREA in un dataset relativo alla percezione dei sistemi di e-proctoring in ambito universitario (funzione `CA`, libreria `FactoMineR`).

```
> summary(cora_a2)

Call:
CA(X = t_a)

The chi square of independence between the two variables is equal to 43.15937 (p-value =  0.0001487266 ).

Eigenvalues
                       Dim.1   Dim.2   Dim.3
Variance               0.113   0.016   0.005
% of var.             84.580  11.930   3.491
Cumulative % of var.  84.580  96.509 100.000

Rows
         Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
BLD006 |     8.266 |  0.048  0.249  0.034 | -0.213 35.372  0.682 | -0.138 50.318  0.284 |
BLD023 |    15.140 | -0.264 12.997  0.970 |  0.009  0.101  0.001 |  0.045  9.315  0.029 |
BLD126 |    70.958 |  0.550 62.168  0.990 | -0.021  0.646  0.001 |  0.051 12.780  0.008 |
FIM049 |    22.562 | -0.447 19.184  0.961 | -0.075  3.781  0.027 |  0.051  5.987  0.012 |
MED054 |    10.553 |  0.013  0.023  0.002 |  0.244 59.991  0.906 | -0.077 20.664  0.091 |
MED055 |     6.141 | -0.192  5.380  0.990 | -0.010  0.108  0.003 |  0.016  0.935  0.007 |

Columns
         Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
L1     |    57.087 | -0.422 47.815  0.947 | -0.097 17.993  0.050 | -0.024  3.851  0.003 |
L2     |     9.572 | -0.105  2.242  0.265 |  0.148 31.459  0.524 |  0.094 43.389  0.211 |
L3     |    16.097 |  0.209  9.470  0.665 |  0.116 20.507  0.203 | -0.093 45.565  0.132 |
L4     |    50.865 |  0.453 40.473  0.899 | -0.147 30.041  0.094 |  0.039  7.195  0.007 |
```

Figura 3.9 - Risultati dalla funzione `summary` nelle procedure di analisi delle corrispondenze realizzata in `R` con la funzione `CA`, libreria `FactoMineR`: χ^2, varianza, `ctr` e `cos2` per righe e per colonne.

L'indice `ctr` denota quanto la modalità contribuisce a determinare la dimensione, cioè a determinare la variabilità del dataset. Ovviamente saranno più rilevanti, anche nell'interpretazione dei risultati, le modalità con valori di `ctr` più elevati; da escludere sono le modalità che non influenzano nessuna dimensione. Nel nostro esempio, il valore `ctr` per le modalità BLD006 e MED054 è rispettivamente 62,168 e 59,991. I due item sono quelli che contribuiscono maggiormente alla definizione della variabilità rispettivamente delle dimensioni 1 e 2.

`Cos2`, coseno dell'angolo fra gli assi e la retta passante per l'origine e il punto della modalità, detto anche correlazione al quadrato, indica il livello di qualità della rappresentazione, ossia quanto la modalità è ben rappresentata dalla dimensione. Si tratta di un valore compreso fra 0 e 1. La somma dei `cos2` per ciascuna modalità è pari a 1. Scegliendo un numero inferiore di dimensioni rispetto a quelle possibili, il `cos2` diminuirà. In ogni caso, più tale valore si avvicina a 1, più la modalità è ben rappresentata nel modello costruito. Ancora nel nostro esempio, MED054 ha un `cos2` maggiore nella dimensione 2, questa dimensione rappresenta meglio l'item. Al contrario BLD126 ha un `cos2` di 0,990 per la dimensione 1 ed è questa dimensione che meglio rappresenta le frequenze per il corso di laurea in questione.

Le stesse riflessioni su coordinate, qualità (`cos2`) e contributi (`crt`) sono riportate nella seconda parte del `summary` dedicato alle colonne (Figura 3.9).

La rappresentazione in Figura 3.8 è definita simmetrica, righe e colonne sono rappresentate con le coordinate calcolate in origine. Come abbiamo detto, ciò limita l'interpretazione delle distanza fra righe e colonne. Per confrontare in maniera più affidabile le modalità appartenenti a righe e colonne, dobbiamo lavorare sul biplot asimmetrico (Figura 3.10).

I livelli di ANSIA in entrambi i biplot (Figura 3.8 e 3.10) sono distribuiti nei quattro quadranti. Nel biplot simmetrico i punti della variabile CORSI DI LAUREA si dispongono lungo l'asse della dimensione 1 "mescolandosi" con i livelli d'ansia. Le modalità dei corsi FIM049, BLD023 e MED055 sono sul semiasse opposto al corso BLD126. Così come BLD006 è sul semiasse opposto della dimensione 2 rispetto al corso MED054. In entrambi i casi, si tratta di modalità che, considerate in funzione dei livelli di gestione dell'ansia, si comportano in modi diversi. In breve gli studenti in corsi di laurea collocati su semiassi e semipiani opposti, gestiscono in maniera diversa l'ansia. FIM049 è vicina alla modalità d'ansia L1, così come BLD126 lo è al livello L4. Questa vicinanza lascerebbe ipotizzare che con maggiore probabilità gli studenti del corso BLD126 hanno gestito meglio l'ansia usando i sistemi di e-proctoring rispetto agli studenti del

corso di laurea FIM049. Il biplot asimmetrico mostra le modalità dei corsi di laurea nelle vicinanze dell'origine, i valori sembrano avvicinarsi molto ai valori medi. I livelli di ansia sono distanti dalle modalità dei corsi di laurea. Questa distribuzione spinge verso l'ipotesi che il tipo di relazione che intercorre fra le due variabili è piuttosto debole. Tuttavia, la visualizzazione nella Figura 3.10 è abbastanza frequente nei biplot asimmetrici poiché abbiamo sostituito i valori delle modalità in colonna con quelli marginali e dunque l'ipotesi è tutta da verificare. L'uso dei biplot asimmetrici è consigliato quando l'inerzia (e dunque la distanza dei punti dai centroidi) ha valori elevati che permettono di conservare una giusta proporzione fra le modalità.

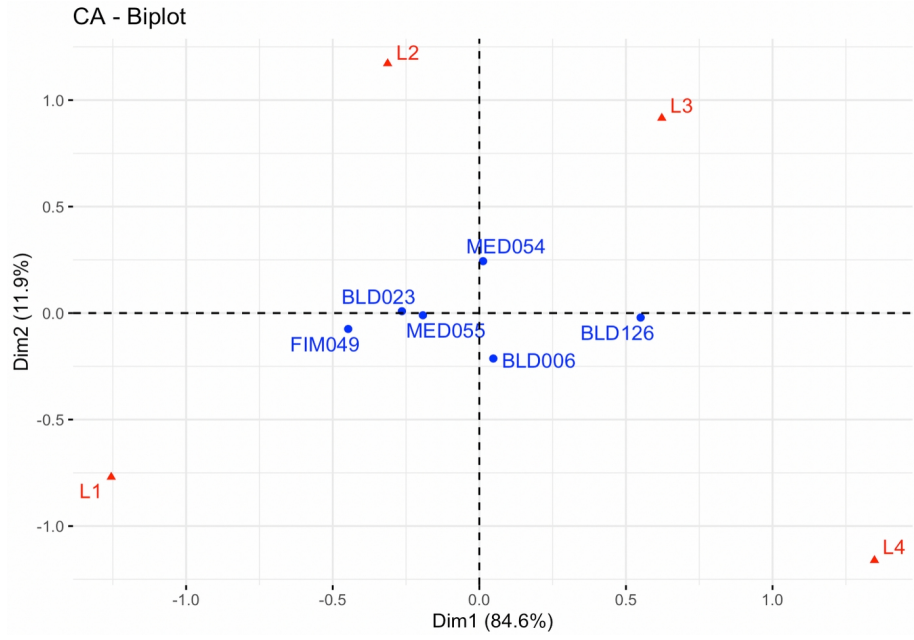


Figura 3.10 - Biplot asimmetrico nell'analisi delle corrispondenze semplici riferita alle variabili ANSIA e CORSO DI LAUREA in un dataset relativo alla percezione dei sistemi di e-proctoring in ambito universitario (libreria `factoextra`, funzione `fviz_ca_biplot`).

I dati usati in questo esempio sono serviti soltanto per generare grafici e calcolare indici, hanno un carattere semplicemente esemplificativo, non aspirano a fornire risposte efficaci a reali domande di ricerca. Utilizzarli ci ha permesso di ripercorrere le procedure di un'analisi delle corrispondenze semplici e comprendere il tipo di risultati a cui è possibile giungere.

Aggiungiamo al nostro esempio la variabile PERFORMANCE in una scala da 1 a 4. Essa rappresenta il livello con cui l'uso dei sistemi di e-proctoring ha influenzato positivamente le prestazioni degli studenti durante le prove d'esame.

Cosa succede quando abbiamo più variabili?

Quando è possibile, il ricercatore prova a riportare l'analisi multivariata a una analisi bivariata. Riporta i dati nella formulazione della Tabella 3.4 dove ciascuna riga rappresenta due modalità (variabili interattive), restando nel nostro esempio, una relativa al corso e una al livello di performance. Gli studenti della riga 1, ad esempio, appartengono al corso di laurea BLD006 e hanno dichiarato che i sistemi di e-proctoring non hanno migliorato la loro performance (livello L1). Di questi 7 hanno un livello L1 di gestione dell'ansia, 0 un livello L2, 1 un livello L3 e 3 un livello L4.

Questa tabella di contingenza può quindi essere usata in una CORA con gli stessi meccanismi visti finora.

Corso di Laurea	**Gestione dell'ansia**			
	L1	**L2**	**L3**	**L4**
BLD006 - PERFORMANCE L1	7	0	1	2
BLD006 - PERFORMANCE L2	5	2	5	4
BLD006 - PERFORMANCE L3	2	3	3	4
BLD006 - PERFORMANCE L4	0	0	1	1
BLD023 - PERFORMANCE L1	15	8	1	4
BLD023 - PERFORMANCE L2	11	4	4	5
...	...	...	...	...
MED055 - PERFORMANCE L3	3	5	4	3
MED055 - PERFORMANCE L4	0	0	0	0

Tabella 3.4 - Tabella di contingenza per le variabili CORSO DI LAUREA, PERFORMANCE e ANSIA in un dataset relativo alla percezione dei sistemi di e-proctoring in ambito universitario. Nella prima colonna sono presenti variabili interattive che rappresentano due modalità combinate da due variabili.

Quando questo non è possibile o è troppo complesso per il grande numero di variabili e modalità, lavoriamo su un'analisi delle corrispondenze multiple (MCA) dove le tabelle di contingenza vengono riproposte nella forma della

matrice degli indicatori (Tabella 3.5) che riporta in un sistema binario il dataset per unità statistiche o della matrice di Burt (Tabella 3.6) che incrocia le frequenze su più modalità.

N.	Corso di laurea						Ansia				Performance			
	BLD 006	BLD 023	BLD 126	FIM 049	MED 054	MED 055	L1	L2	L3	L4	L1	L2	L3	L4
1	1	0	0	0	0	0	1	0	0	0	1	0	0	0
2	0	1	0	0	0	0	1	0	0	0	0	0	0	1
3	0	1	0	0	0	0	1	0	0	0	0	1	0	0
4	0	0	1	0	0	0	0	0	1	0	0	0	1	0
5	0	0	0	0	1	0	0	0	0	1	1	0	0	0
...	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Tabella 3.5 - Matrice degli indicatori. Per ogni unità statistica nelle righe è riportato 1 in corrispondenza della modalità selezionata per ciascuna variabile (colonne) e 0 per le altre modalità.

		Corso di laurea						Ansia				Performance			
		BLD 006	BLD 023	BLD 126	FIM 049	MED 054	MED 055	L1	L2	L3	L4	L1	L2	L3	L4
Corso di laurea	BLD 006	40	0	0	0	0	0	14	5	10	11	10	16	12	2
	BLD 023	0	68	0	0	0	0	27	18	13	10	28	24	12	4
	BLD 126	0	0	75	0	0	0	7	15	23	30	9	16	39	11
	FIM 049	0	0	0	35	0	0	17	9	5	4	7	23	5	0
	MED 054	0	0	0	0	52	0	13	14	17	8	8	23	20	1
	MED 055	0	0	0	0	0	53	20	13	11	9	6	32	15	0
Ansia	L1	14	27	7	17	13	20	98	0	0	0	32	52	9	5
	L2	5	18	15	9	14	13	0	74	0	0	12	32	29	1
	L3	10	13	23	5	17	11	0	0	79	0	8	28	38	5
	L4	11	10	30	4	8	9	0	0	0	72	16	22	27	7
Performance	L1	10	28	9	7	8	6	32	12	8	16	68	0	0	0
	L2	16	24	16	23	23	32	52	32	28	22	0	134	0	0
	L3	12	12	39	5	20	15	9	29	38	27	0	0	103	0
	L4	2	4	11	0	1	0	5	1	5	7	0	0	0	18

Tabella 3.6 - Tabella di Burt. Sulle righe e sulle colonne sono presenti le modalità delle variabili nello studio. Le celle contengono le frequenze con cui si manifestano due modalità.

La Tabella 3.5 anticipa la considerazione che nell'analisi delle corrispondenze multiple possiamo ottenere visualizzazioni grafiche della similarità fra variabili (colonne, Figura 3.11) ma anche fra le unità statistiche (righe, Figura 3.12).

Le funzioni della libreria `factoextra` forniscono delle visualizzazioni dove compaiono insieme i valori di righe e colonne (Figura 3.13) oppure righe con le modalità di una sola variabile (Figura 3.14).

Restano invariati gli indicatori rilevati e la loro interpretazione. In Figura 3.15 vediamo che le prime due dimensioni spiegano il 27,545% della varianza. Le coordinate e i valori `ctr` e `cos2` sono forniti per righe (individui) e colonne (modalità).

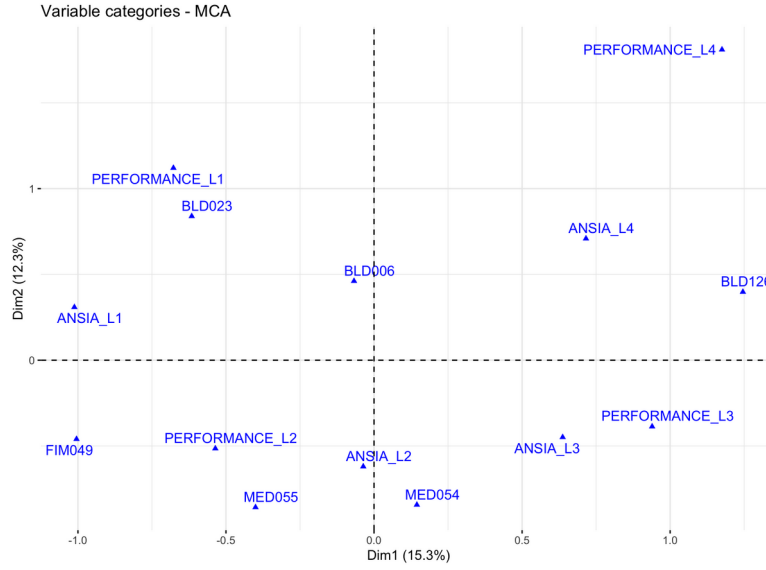


Figura 3.11 - MCA sulle variabili CORSI DI LAUREA, PERFORMANCE e ANSIA in un dataset relativo alla percezione dei sistemi di e-proctoring in ambito universitario (funzione `fviz_mca_var`, pacchetto `factoextra`).

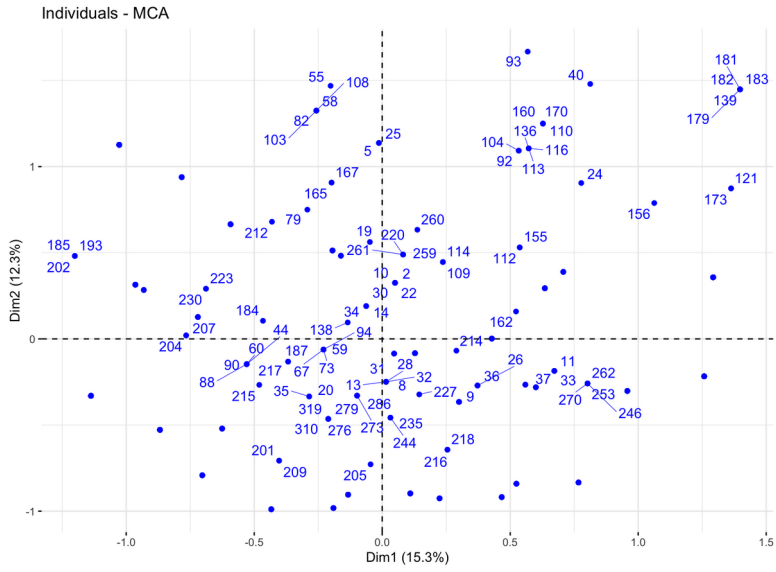


Figura 3.12 - MCA sulle unità statistiche rispetto a CORSI DI LAUREA, PERFORMANCE e ANSIA in un dataset relativo alla percezione dei sistemi di e-proctoring in ambito universitario (funzione `fviz_mca_ind`, pacchetto `factoextra`).

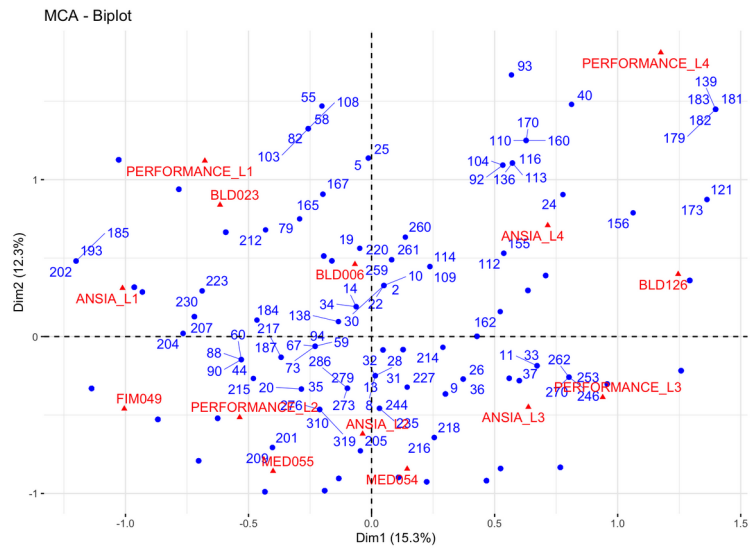


Figura 3.13 - MCA su modalità e unità statistiche rispetto a CORSI DI LAUREA, PERFORMANCE e ANSIA in un dataset relativo alla percezione dei sistemi di e-proctoring in ambito universitario (funzione `fviz_mca_biplot`, pacchetto `factoextra`).

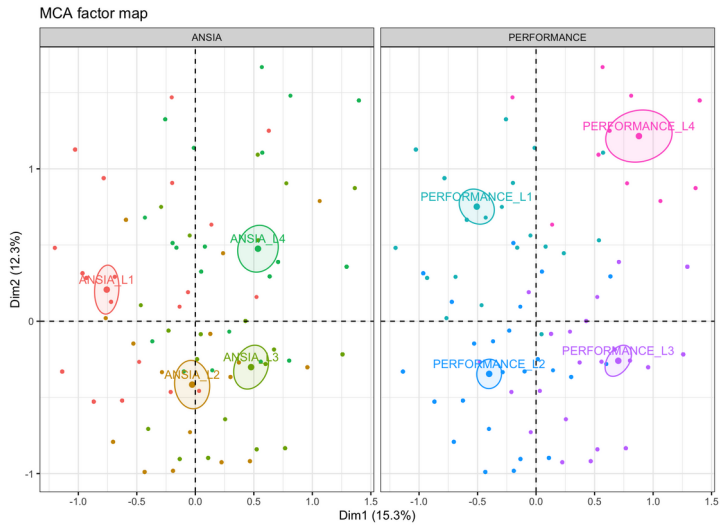


Figura 3.14 - MCA per variabili su modalità e unità statistiche rispetto a CORSI DI LAUREA, PERFORMANCE e ANSIA in un dataset relativo alla percezione dei sistemi di e-proctoring in ambito universitario (funzione `fviz_ellipses`, pacchetto `factoextra`).

```
> summary(mca)

Call:
MCA(X = dm, graph = FALSE)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8   Dim.9  Dim.10  Dim.11
Variance               0.559   0.451   0.384   0.356   0.339   0.331   0.302   0.290   0.246   0.210   0.199
% of var.             15.258  12.287  10.480   9.709   9.232   9.020   8.234   7.917   6.698   5.735   5.429
Cumulative % of var.  15.258  27.545  38.025  47.734  56.967  65.987  74.220  82.138  88.836  94.571 100.000

Individuals (the 10 first)
                  Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
1              | -0.783  0.339  0.140 |  0.939  0.606  0.202 |  0.061  0.003  0.001 |
2              |  0.050  0.001  0.001 |  0.326  0.073  0.027 |  0.883  0.628  0.195 |
3              | -0.783  0.339  0.140 |  0.939  0.606  0.202 |  0.061  0.003  0.001 |
4              | -0.783  0.339  0.140 |  0.939  0.606  0.202 |  0.061  0.003  0.001 |
5              | -0.013  0.000  0.000 |  1.137  0.889  0.271 |  0.194  0.030  0.008 |
6              | -0.720  0.287  0.144 |  0.127  0.011  0.004 |  0.750  0.453  0.157 |
7              | -0.783  0.339  0.140 |  0.939  0.606  0.202 |  0.061  0.003  0.001 |
8              |  0.015  0.000  0.000 | -0.249  0.043  0.016 |  0.470  0.178  0.057 |
9              |  0.372  0.077  0.033 | -0.271  0.050  0.017 | -0.325  0.085  0.025 |
10             |  0.050  0.001  0.001 |  0.326  0.073  0.027 |  0.883  0.628  0.195 |

Categories (the 10 first)
                  Dim.1    ctr   cos2 v.test    Dim.2    ctr   cos2 v.test    Dim.3    ctr   cos2 v.test
BLD006         | -0.068  0.034  0.001 -0.456 |  0.461  1.950  0.030  3.112 |  0.558  3.346  0.044  3.765 |
BLD023         | -0.616  4.755  0.101 -5.705 |  0.839 10.966  0.188  7.775 | -1.014 18.764  0.274 -9.393 |
BLD126         |  1.246 21.470  0.469 12.293 |  0.398  2.720  0.048  3.926 |  0.324  2.113  0.032  3.196 |
FIM049         | -1.005  6.518  0.123 -6.286 | -0.460  1.694  0.026 -2.876 |  0.815  6.236  0.081  5.095 |
MED054         |  0.145  0.201  0.004  1.137 | -0.843  8.462  0.136 -6.625 | -0.790  8.720  0.120 -6.211 |
MED055         | -0.400  1.566  0.031 -3.182 | -0.857  8.921  0.144 -6.815 |  0.658  6.169  0.085  5.234 |
PERFORMANCE_L1 | -0.678  5.758  0.122 -6.278 |  1.120 19.548  0.335 10.381 | -0.766 10.705  0.156 -7.094 |
PERFORMANCE_L2 | -0.536  7.101  0.204 -8.098 | -0.515  8.125  0.188 -7.774 |  0.515  9.551  0.188  7.784 |
PERFORMANCE_L3 |  0.939 16.760  0.413 11.532 | -0.387  3.525  0.070 -4.746 | -0.397  4.364  0.074 -4.877 |
PERFORMANCE_L4 |  1.175  4.585  0.082  5.123 |  1.810 13.507  0.193  7.890 |  1.330  8.551  0.104  5.798 |

Categorical variables (eta2)
               Dim.1 Dim.2 Dim.3
CORSO          | 0.580 0.469 0.523 |
PERFORMANCE    | 0.574 0.604 0.382 |
ANSIA          | 0.525 0.278 0.248 |
```

Figura 3.15 – Risultati dalla funzione `summary` applicati ai risultati di una MCA realizzata in `R` con la funzione `MCA`, libreria `FactoMineR`: varianza, `ctr` e `cos2` per unità statistiche e modalità, `eta2` per le variabili categoriali.

Nelle MCA, alcune variabili del dataset possono essere considerate supplementari (o illustrative), ciò significa che esse non entrano nell'analisi per estrarre le dimensioni ma i loro valori sono predetti e confrontati con i dati inseriti nel calcolo. Questa possibilità, comune ad esempio alla cluster analysis, permette di usare la MCA come tecnica predittiva.

3.5 - L'uso dell'analisi fattoriale e dell'analisi delle componenti principali nella ricerca educativa

Per validare scale, inventory e questionari nella ricerca educativa, così come in quella psicologica, vengono usate l'analisi fattoriale e l'analisi delle componenti principali, molto spesso in modalità combinata. Una grande fetta di ricerche che usano EFA e PCA riguarda proprio la definizione e la validazione di scale. Gli ambiti di applicazione, come è ovvio, possono essere diversissimi; ci sono scale dedicate agli argomenti più variegati.

Stamatios Papadakis e colleghi (2020), ad esempio, rilevata la diffusione del mercato di produzione e utilizzo di app dedicate ai bambini in età prescolare e fruibili su smartphone o tablet grazie anche alla loro portabilità e attrattività, formulano una scala che permetta a insegnanti, genitori e tutori, di valutare tali app. La scala denominata Evaluation Tool for Educational Apps (E.T.E.A.) è stata redatta dopo aver condotto una review su rubriche, checklist e paper già pubblicati sul tema, utile per definire il set di item da testare e inserire nello strumento di valutazione in costruzione. Essa poi è stata somministrata a 218 futuri insegnanti della scuola dell'infanzia, studenti presso l'Università di Creta a cui è stato chiesto di completare le domande solo dopo aver scaricato tre app educative sui loro dispositivi. Nella scala un numero molto alto di item è stato proposto, numero ridotto poi nell'analisi e sintetizzato in pochi fattori. Il test di Bartlett ($\chi^2 = 1775{,}95,\ p < 0{,}001$) e l'indice di KMO (0,79) hanno dato risultati positivi sulla fattoriabilità delle variabili, cosa che ha permesso di applicare l'analisi fattoriale esplorativa. L'analisi delle componenti principali è stata utilizzata come procedura di estrazione dei fattori. Sono stati selezionati 4 fattori con autovalore superiore a 1 in grado di spiegare il 79,39% della varianza (percentuale di varianza cumulata, vedi Tabella 3.7). I fattori sono stati definiti poi con i nomi di usabilità, efficienza, parental control e sicurezza.

	Factors	Eigenvalues	% of Variance	Cumulative %
1	**Usability**	4.65	35.78	35.78
2	**Efficiency**	2.79	21.45	57.23
3	**Parental Control**	1.80	13.84	71.07
4	**Security**	1.08	8.32	79.39

Tabella 3.7 – Autovalori e percentuale di varianza spiegati dai quattro fattori ottenuti dalla PCA in uno studio sulle app educative (Papadakis et al., 2020, p. 5/10).

Gli autori riportano due tabelle (Tabella 3.8 e 3.9). La prima presenta i valori della struttura fattoriale con i loading (che ricordiamo essere i valori di correlazione fra le variabili osservate e i fattori individuati) calcolati dopo la rotazione VARIMAX. La tabella contiene anche i valori dell'α di Cronbach calcolato per ciascun fattore per indicare il grado di consistenza interna delle sotto-scale. Dei tredici item individuati, tutti i valori dei coefficienti fattoriali sono più alti di 0,651. I primi due fattori raccolgono 4 variabili osservate, il terzo 2 e il quarto 3. I valori dell'α di Cronbach, tutti più alti di 0,73, risultano accettabili. La seconda tabella mostra le correlazioni fra i fattori. Alcune risultano essere significative. Questo dovrebbe meravigliarci perché abbiamo parlato della necessità dei fattori di essere indipendenti fra loro. Tuttavia vediamo (e anche gli autori dello studio lo sottolineano) che i costrutti individuati dai fattori sono ancora legati fra loro: l'efficacia è infatti legata all'usabilità e al parental control.

Negli articoli scientifici che descrivono procedure di ricerca legate alla validazione delle scale, come questo appena presentato, ritroviamo uno schema simile e ricorrente. Gli autori:

- validano la fattoriabilità delle variabili mostrando le correlazioni fra di esse e i risultati dai test di Bartlett e Kaiser-Meyer-Olkin (KMO);
- definiscono il criterio di estrazione dei fattori e di scelta del numero dei fattori (aggiungendo spesso il plot dello scree test);
- definiscono il criterio di rotazione utilizzato;
- forniscono la percentuale di varianza spiegata dal modello aggiungendo frequentemente tabelle sulla varianza cumulata e sui pesi fattoriali;
- riportano il valore del coefficiente α di Cronbach per l'intera indagine o per le singole scale individuate.

	Factors			
	1	**2**	**3**	**4**
Q1 – Suitable instructions	0.898			
Q2 – Button array is consistent	0.880			
Q3 – Buttons facilitate use	0.858			
Q4 – Can be used easily	0.853			
Q5 – Parametrization options		0.859		
Q6 – Suitable multimedia options		0.722		
Q7 – Multiple way for conveying information		0.669		
Q8 – Feedback interaction		0.651		
Q9 – Consulting parents			0.952	
Q10 – Inform parents			0.949	
Q11 – No urging for purchase				0.893
Q12 – No destructive ads				0.852
Q13 – Inform about personal data policy				0.684
Cronbach's alpha of internal consistency	0.91	0.83	0.96	0.73

Tabella 3.8 – Analisi delle componenti principali con la rotazione VARIMAX in uno studio sulle app educative (Papadakis et al., 2020, p. 6/10).

	1	**2**	**3**	**4**
Usability	1	0.590**	0.090	0.055
Efficiency	0.590**	1	0.212**	0.252**
Parental Control	0.090	0.212**	1	0.001
Security	0.055	0.252**	0.001	1
**Correlation is significant at the 0.01 level (two tails).				

Tabella 3.9 – Matrice di correlazione delle quattro dimensioni in uno studio sulle app educative (Papadakis et al., 2020, p. 6/10).

Un esempio nella ricerca italiana nel quale più modelli di analisi fattoriale sono confrontati è lo studio per la validazione dell'University Climate Questionnaire for Distance Education Contexts (UCliQ-DE) di Damiano Felini ed Elisa Zobbi (2022).

Gli autori hanno tenuto il corso di Teorie dell'educazione per le matricole del corso di laurea in Scienze dell'educazione e dei processi formativi presso l'Università di Parma nell'a.a. 2020/21 con studenti che non si erano mai incontrati in presenza e che, per via dell'emergenza sanitaria dovuta al Covid-19, avrebbero svolto totalmente online l'insegnamento tenuto in presenza nei precedenti anni accademici. A partire da un'idea di università vista "not as mere sites of instruction, but as places where people learn and live, build and maintain relationships, develop and grow" (ivi, p. 75), gli studiosi indagano le dinamiche del clima universitario nei contesti di formazione a distanza ponendo attenzione alle relazioni che gli studenti sviluppano con colleghi di corso e docenti, le loro aspettative e percezioni sul clima universitario e la sua qualità. Hanno somministrato un questionario composto da 31 item che prevedevano risposte in una scala Likert a 10 livelli da "assolutamente non vero" a "completamente vero". Cinque fattori sembravano necessari per operazionalizzare il concetto di clima universitario: percezione delle interazioni sociali fra pari, senso di appartenenza alla comunità, aspettative iniziali sul clima che si sarebbe stabilito nel semestre online, percezione delle interazioni sociali fra studenti e docenti, consapevolezza dei limiti nelle interazioni online rispetto a quelle in presenza.

Hanno risposto al questionario 173 studenti.

Dopo l'analisi descrittiva e delle correlazioni fra le variabili (item), gli autori hanno escluso sei item altamente correlati fra loro. Hanno condotto poi tre analisi fattoriali esplorative usando la PCA per l'estrazione dei fattori. Le prime due sono state svolte sul campione completo usando come metodi di rotazione nel primo quello ortagonale VARIMAX e nel secondo il metodo obliquo PROMAX; le due analisi hanno dato risultati sovrapponibili sia nei test di Kaiser-Meyer-Olkin e della sfericità di Bartlett sia nelle soluzioni fattoriali (scree plot e Kaiser-Gutman rule). Raccolgono infatti sei fattori che coprono il 65,2% della varianza; di questi il primo, denominato "percezione delle interazioni fra pari", spiega il 31,3% della varianza. L'ultimo fattore in entrambi i casi ha un solo item.

Per confermare i risultati, adottando un metodo di cross-validation, gli autori sono passati a una terza analisi nella quale hanno diviso il campione in due gruppi e condotto un EFA sul primo sottocampione con metodo VARIMAX, e una sul secondo con il metodo PROMAX. Dal primo gruppo è stata ottenuta

una struttura a sei fattori che spiega il 69,4% della varianza, confermando il primo fattore con il 32,3% e raccogliendo due quesiti nell'ultimo fattore. L'analisi sul secondo gruppo restituisce una soluzione a 8 fattori spiegando il 76,1% della varianza. Il primo fattore da solo ha una varianza del 30,5%. La discordanza fra i due modelli ha portato a fare ulteriori indagini sull'affidabilità: è stato calcolato il coefficiente α di Cronbach per le scale risultanti, in alcuni casi con scarsi risultati. Considerando le tre EFA e le analisi sulla reliability, gli autori arrivano a definire a un modello a cinque fattori con 22 item dove il primo fattore/scala (11 item) raccoglie due aspetti della percezione delle interazioni sociali fra pari, indagata come possibilità e qualità. La percezione del clima universitario è strettamente legata a tale fattore dati gli alti valori della varianza. Seguono gli altri quattro fattori:

1. due aspetti sul senso di appartenenza (3 item): orgoglio di appartenere al gruppo e disponibilità a integrarsi;
2. aspettative iniziali (4 item);
3. percezione delle interazioni sociali fra studenti e docenti (3 item) indagate come possibilità e qualità;
4. consapevolezza dei limiti delle interazioni online (2 item).

Gli autori pubblicano tabelle con i loadings di ciascuno delle 4 EFA e approfondimenti testuali sugli item critici nella consapevolezza che l'utilizzazione del questionario vada verificata su studenti di differenti anni di corso e CdL diversi per essere generalizzata.

Ogni variazione a una scala già formulata, come ad esempio la traduzione in un'altra lingua, l'eliminazione o l'aggiunta di uno o più item, la variazione dei soggetti a cui viene somministrata, richiede una nuova validazione.

Le evoluzioni della scala CLES, elaborata nell'ambito della formazione in ambito sanitario, ce ne dà un chiaro esempio. La scala CLES (Clinical Learning Environment and Supervision), che rappresenta uno dei casi che avremo potuto analizzare fra i tanti, è stata formulata da Mikko Saarikoski e Helena Leino-Kilpi (2002) per valutare l'ambiente e le procedure di supervisione per i futuri infermieri durante i tirocini in reparto. Dopo aver condotto una review della letteratura sul tema, gli autori hanno formulato un'indagine composta da 27 item con risposte su una scala Likert a 5 livelli, organizzati preventivamente in 5 dimensioni. Lo strumento è stato somministrato a oltre 400 studenti in infermieristica del secondo e terzo anno provenienti da scuole di dimensioni diverse col-

locate in differenti aree della Finlandia, paese che ha ospitato lo studio. La validità di contenuto è stata attestata attraverso la review, la validità di facciata/aspetto attraverso il confronto con un gruppo di esperti. Per la validità di costrutto gli autori hanno lavorato usando l'analisi fattoriale esplorativa nella quale i fattori individuati (5, numero che conferma le dimensioni individuate preventivamente dagli autori ma con modifiche nella distribuzione degli item) coprono il 64% della varianza delle variabili; il 40% della stessa deriva dal fattore relativo alla relazione di supervisione fra studenti e infermieri referenti che è quindi l'aspetto a cui prestare maggiore importanza. L'affidabilità della scala è stata calcolata analizzando i valori del coefficiente α di Cronbach per ciascuna sottodimensione.

In un articolo successivo, questa scala viene integrata con una nuova dimensione (Saarikoski et al., 2008) relativa alla presenza di un *nurse-teacher* che si occupi di fornire insegnamenti teorici e pratici agli studenti in formazione. La validazione realizzata sui dati raccolti in un nuovo campione – sempre nel contesto finlandese con le stesse caratteristiche del precedente ma più numeroso (N=549) – porta alla definizione di una nuova scala di 34 item indicata con l'acronimo CLES+T (Clinical Learning Environment, Supervision and Nurse Teacher). In questo caso, gli autori dichiarano di aver lavorato su più modelli di fattorizzazione, sia utilizzando l'EFA che la PCA. Pur riconoscendo che la tecnica più adeguata da utilizzare è l'analisi fattoriale poiché, tra l'altro, è di tipo inferenziale, scelgono la suddivisione in componenti fornita dalla PCA che restituisce 5 dimensioni splittando in due una dimensione risultata particolarmente ricca di item nei risultati dell'EFA con 4 fattori.

Alcune dimensioni della scala CLES+T, quelle relative alle caratteristiche educative del reparto di tirocinio, sono state utilizzate in uno studio più recente (Doyle et al., 2017) per valutare il punto di vista di 150 studenti in infermieristica sugli ambienti deputati al tirocinio clinico nel contesto australiano. In questo caso, la tecnica di analisi per la riduzione delle dimensioni è stata la PCA che ha condotto all'individuazione di due componenti in grado di sintetizzare gli item. Gli autori le definiscono "Happy to help" e "Happy to be here" e rappresentano le caratteristiche che i membri dello staff devono possedere per rendere l'esperienza di tirocinio efficace e soddisfacente per gli studenti.

La scala CLES+T è stata tradotta e validata in più lingue fra cui italiano (Tomietto et al., 2012) e spagnolo (Vizcaya-Moreno et al., 2015).

Attraverso questa scala, abbiamo visto come introduzione e rimozione di nuovi item e traduzioni necessitino di nuove procedure di validazione.

EFA e PCA sono usate nella ricerca educativa anche in altri modi. Presentiamo alcuni casi che si discostano da quelli finora visti e che sono una possibile strada da percorrere nell'analisi quando si lavora su grandi numeri di variabili. I primi due studi riguardano l'uso di analisi fattoriale e delle componenti principali su questionari/raccolte dati che non si intende trasformare in scale ma che, per essere letti e interpretati, vanno necessariamente strutturati secondo uno schema che permetta di focalizzare l'attenzione su un numero ridotto di dimensioni sottese alle misurazioni. Il terzo presenta invece un'azione di riduzione delle variabili propedeutica all'applicazione di altre tecniche di analisi.

Emilio José Delgado-Algarra e colleghi (2019) usano l'analisi fattoriale esplorativa per lavorare su due temi emergenti in ambito formativo. Da un lato ci sono i MOOC (Massive Open Online Courses) che rappresentano un formato di corsi online che dal 2012 ha preso piede nello scenario internazionale e che rappresentano uno strumento open di formazione per chiunque abbia a disposizione un dispositivo e una connessione alla rete (per un approfondimento sui temi della *open education* si veda Nascimbeni, 2020). Dall'altra, troviamo i temi della cittadinanza e della sostenibilità che hanno un posto sempre più rilevante nei contesti educativi formali e sono finalizzati all'aumento della partecipazione ai processi politici e sociali, alla difesa dei diritti universali, all'assunzione di doveri e responsabilità secondo principi etici, allo sviluppo di valori e comportamenti di rispetto e solidarietà a livello mondiale, all'uso corretto delle risorse naturali con risvolti in ambito economico e sociale, alla difesa dell'ambiente. Gli autori raccolgono dati su 161 MOOC dedicati al tema della cittadinanza e della sostenibilità su tre piattaforme internazionali: Coursera, edX e MiriadaX. Rispondono a domande di ricerca sulle caratteristiche tecniche e didattiche di tali MOOC usando gli strumenti della statistica descrittiva. Per comprendere invece quali sono i trend dei MOOC che riguardano i temi della cittadinanza e della sostenibilità e il loro intreccio usano l'EFA. Dopo le verifiche sulla fattoriabilità, usano la PCA per l'estrazione dei fattori e ne selezionano 6 che mescolano 16 indicatori descrittivi dei MOOC di cui 9 sulla categoria della cittadinanza e 7 su quella della sostenibilità. Quattro fattori sono focalizzati su una delle due dimensioni, i restanti due le mescolano. Si tratta dei fattori a cui gli autori attribuiscono questi nomi: "cittadinanza socio-ecologica" (inerente questioni ecologiche e impatto delle attività commerciali su welfare, economia e ambiente) e "cittadinanza austera" (cittadinanza liberale, uso di carbone e acqua). I fattori individuati, senza costituire una scala, sono quindi una lente attraverso la quale analizzare i MOOC che fanno parte del campione.

In maniera simile, James Forrest e colleghi (2015) usano la PCA in un'indagine sul tema del razzismo nel multiculturale contesto australiano dove a scuola bambini e ragazzi mentre sperimentano l'interazione con i pari si confrontano con atteggiamenti di intolleranza che generano malessere e insuccesso nell'apprendimento. Una survey è stata somministrata a livello statale su tali argomenti fra gli insegnanti delle scuole primarie e secondarie e i dati relativi a 1309 docenti sono stati analizzati. L'articolo si sofferma principalmente sulle opinioni degli insegnanti in merito agli obiettivi dell'educazione multiculturale, sull'efficacia delle strategie volte a promuovere l'inclusione, sugli atteggiamenti relativi alla diversità e al multiculturalismo. I docenti sono in grado di andare oltre i loro privilegi essendo molto spesso nel contesto australiano, bianchi, laureati e abbienti? Sono in grado di opporsi al razzismo nella vita quotidiana?

Le alte correlazioni misurate fra le variabili hanno suggerito l'ipotesi che potessero esserci delle dimensioni latenti fra gruppi di domande proposte nell'indagine. È stata così realizzata una PCA con la rotazione VARIMAX attraverso la quale 26 domande (variabili) sono state sintetizzate in 6 dimensioni che hanno permesso di dare una lettura maggiormente consapevole dei risultati della survey.

I dati ottenuti nell'indagine sui docenti sono stati poi comparati con quelli generali della popolazione per comprendere il ruolo che la scuola può esercitare nella comunità per lo sviluppo di pratiche antirazziste e la valorizzazione delle diversità culturali.

L'ultima ricerca che presentiamo propone, come anticipato, l'uso di tecniche di riduzione in concomitanza con altre tecniche di analisi. Nello studio "A study of depression and anxiety, general health, and academic performance in three cohorts of veterinary medical students across the first three semesters of veterinary school", Allison M.J. Reisbig e colleghi (2012) affrontano il tema dei livelli di ansia negli studenti in veterinaria. Usano un alto numero di variabili fra cui, oltre a quelle demografiche e personali (genere, scuola, religione), i risultati da due scale: CES-D (Center for Epidemiologic Studies Depression Scale), sui livelli di depressione dei soggetti patologici e nella popolazione generale e MHI-A (Mental Health Inventory) relativa ai livelli d'ansia. Aggiungono misure su soddisfazione della propria vita, salute generale, successo accademico e 16 item su possibili fattori di stress per gli studenti di veterinaria. L'analisi fattoriale esplorativa è stata utilizzata per determinare la presenza di sotto-scale (e quindi tratti latenti) sui fattori di stress. Ne sono state rilevate quattro con autovalore superiore a 1 (stress universitario, stress di transizione, stress per la famiglia/salute, stress dovuto alle relazioni) che sono state utilizzate in successive analisi

condotte attraverso la MANOVA (analisi multivariata della varianza di cui non ci occupiamo in questo volume) e la regressione lineare (cap. 4) per comprendere quale relazione esiste fra essi e genere, ansia, depressione. Questa ricerca mostra come l'analisi fattoriale esplorativa possa essere usata per riassumere le dimensionalità (in questo caso sui fattori di stress) e ottenere un numero di fattori ridotto da usare in analisi successive attraverso l'utilizzo di altre tecniche di analisi multivariata.

Suggeriamo di consultare fra le ricerche di ambito bibliometrico, ossia di applicazione delle procedure matematiche e statistiche alle review e ai metadati delle pubblicazioni (Donthu et al., 2021; Zupic & Cater, 2015), lo studio di Carlos Rogério Montenegro de Lima e colleghi (2020) che usano l'analisi fattoriale e il multidimensional scaling (tecnica che permette di lavorare anche sulle visualizzazioni delle relazioni, vedi capitolo 7) per condurre una literature-based review sul finanziamento della sostenibilità nell'istruzione superiore con lo scopo di mettere in luce le connessioni fra teorie e autori che si sono occupati di questo tema nei precedenti tre decenni.

Attraverso le due tecniche sono state indagate le relazioni fra citazioni e co-citazioni di un campione di 745 articoli fra i 1880 pubblicati tra il 1994 e il 2018 su riviste internazionali ed estratte dal database Web of Science sul tema. Gli articoli contenevano 19.916 citazioni di cui sono state analizzate soltanto le più rilevanti.

Sono state sviluppate alcune applicazioni in tale ambito, VOSviewer ne è un esempio. Qui è usato BibExcel per estrarre le citazioni e creare la matrice delle distanze che contiene sulle righe e sulle colonne gli stessi documenti/citazioni e nelle celle corrispondenti la frequenza con cui ciascuna coppia di articoli è stata citata insieme in un terzo paper (le co-citazioni sono appunto le coppie di articoli citate in uno stesso articolo). Questa matrice è stata usata per condurre l'EFA con una rotazione VARIMAX. L'EFA ha permesso di raggruppare gli articoli in base a una struttura latente che rischia di sfuggire nelle literature review classiche escludendo effetti dovuti alla soggettività del ricercatore. I fattori individuati, in questo caso cinque, corrispondono alle cinque aree di studio che contraddistinguono il campo di ricerca del finanziamento della sostenibilità nell'istruzione superiore dalle quali si può partire per svolgere ulteriori ricerche, individuare i buchi di sapere e riassumere gli sviluppi in un determinato periodo storico. I cinque fattori spiegano il 64,6% della varianza totale del modello. In questo caso sono gli articoli ad essere le variabili aggregate in fattori (Tabella 3.10).

	Components				
	1	**2**	**3**	**4**	**5**
Lozano et al. (2013a, 2013b)	0.805				
Verhulst and Lambrechts (2015)	0.795				
Barth and Rieckmann (2012)	0.794				
...					
Haigh (2005)		0.831			
Ryan et al (2010)		0.743			
...					
Trencher et al. (2013)			0.880		
Stephens and Graham (2010)			0.816		
...					
Rittel and Webber (1973)				0.794	
Komiyama and Takeuchi (2006)				0.734	
…					
Lidgren et al.(2006)					0.883
Alshuwaikhat and Abubakar (2008)					0.691
Cronbach's alpha	*0.920*	*0.884*	*0.870*	*0.602*	*0.633*
Accumulated variance (%)	*22.045*	*39.130*	*51.750*	*58.638*	*64.648*

Tabella 3.10 - Alcune righe dei risultati dell'analisi fattoriale in una ricerca di ambito bibliometrico (de Lima et al., 2020, p. 450).

3.6 - L'uso dell'analisi delle corrispondenze nella ricerca educativa

Anna Parola e Lucia Donsì (2018) utilizzano l'analisi delle corrispondenze per indagare le dinamiche legate ai NEET (Not in Education Employment or Training), giovani di età compresa fra i 15 e 34 anni che non sono impegnati in attività di formazione né in attività lavorative. Rientrano in tale categoria gruppi con caratteristiche diverse: disoccupati a breve o lungo termine, soggetti inattivi per questioni familiari o di salute, lavoratori scoraggiati o che difficilmente

riescono a rientrare nel mondo del lavoro e così via. Il fenomeno che interessa paesi europei ed extra-europei (Giappone, Nuova Zelanda, Cina ecc.) è strettamente legato alle politiche di occupazione giovanile e porta l'attenzione sulle pratiche sociali ed educative dei contesti nei quali i soggetti coinvolti vivono. Le ricadute sullo sviluppo dell'identità, sulla gestione del tempo, sull'acquisizione di comportamenti scorretti (come fumo, abuso di alcol ecc.) o sulla comparsa di sintomi somatici o psicologici sono alcuni degli effetti collegati alle abitudini di vita dei soggetti identificati nella categoria NEET sui quali si concentrano le ricerche attuali. Lo studio qui presentato e realizzato nel contesto italiano utilizza dati ISTAT nell'indagine multiscopo "Aspetti della vita quotidiana" (2016); le autrici selezionano quelli relativi al territorio campano. I dati fanno riferimento a 501 soggetti di età compresa fra i 20 e i 34 anni e raccolgono opinioni e preferenze su tre macro aree: "modi di vivere il tempo libero, fiducia e partecipazione politica e sociale, salute e benessere percepito" (Parola & Donsì, 2018, p. 35). L'obiettivo della ricerca è quello di confrontare le abitudini e le caratteristiche dei soggetti che rientrano nella definizione NEET con quelle dei soggetti studenti o lavoratori di pari età e provenienza geografica. A tale scopo, nell'articolo viene illustrata la visualizzazione grafica dell'analisi delle corrispondenze multiple realizzata (Figura 3.16). Le 26 variabili con le 82 modalità ad esse collegate vengono rappresentate in un biplot nel quale le due dimensioni spiegano l'80% della varianza (inerzia totale). Alla prima dimensione (*inerzia = 62%*) contribuiscono maggiormente le modalità legate all'uso del tempo: si passa dall'inattività (semiasse positivo) all'attività (semiasse negativo). Sul secondo asse (*inerzia = 18%*) rileviamo le modalità legate alla soddisfazione negativa e positiva sui semiassi corrispondenti. Le tre categorie (NEET, studenti, lavoratori) si collocano in tre quadranti diversi.

Questo permette abbastanza intuitivamente di associare abitudini e comportamenti a un gruppo piuttosto che a un altro. I NEET si trovano nel quadrante caratterizzato da inattività e insoddisfazione: non sono coinvolti in attività sportive, culturali e sociali, definiscono media/scarsa la loro salute, sono sfiduciati nei rapporti con gli altri e nella percezione del futuro; gli studenti sono nel quadrante attività e soddisfazione: sono coinvolti in attività culturali e politiche, definiscono buona la loro salute e credono che la loro situazione futura migliorerà; gli occupati sono nel quadrante attività e insoddisfazione: sono attivi nella vita sociale, fiduciosi negli altri, poco soddisfatti però del loro tempo libero. Dallo studio deriva una definizione dei giovani adulti NEET che, proprio dal titolo dello stesso contributo, appaiono come "sospesi nel tempo", persi cioè nell'eccessivo tempo a loro disposizione e incapaci di progettare un futuro possibile.

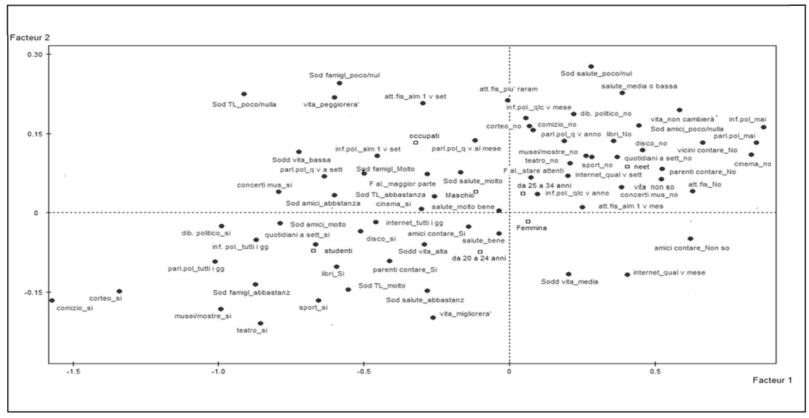


Figura 3.16 – Biplot dell'analisi delle corrispondenze multiple in uno studio sui NEET (Parola & Donsì, 2018, p. 42).

Simile nella tipologia di risultati forniti è lo studio di Emilia Kmiotek-Meier e colleghi (2019) che lavorano sul tema della mobilità per motivi di studio o lavoro in sei paesi europei occupandosi in particolare degli ostacoli che si possono incontrare prima e durante l'esperienza all'estero (risorse economiche, incompatibilità dei regolamenti delle istituzioni, conoscenza della lingua, relazioni sociali e così via). La mobilità è motivo di arricchimento delle conoscenze linguistiche e non solo, dei percorsi formativi, delle esperienze anche interculturali: una ricchezza per l'individuo e le comunità. Gli autori forniscono la mappa percettiva dei risultati raccolti attraverso una survey a cui hanno risposto 1682 soggetti divisi fra lavoratori e studenti di scuola, università e percorsi professionali. Nella tradizione dei multi methods, gli autori hanno integrato tali risultati con 140 interviste attraverso le quali è stato possibile abbinare gli ostacoli a una fase della mobilità (prima o durante).

Restando però nel contesto italiano, interessante per conoscere un'altra modalità di applicazione della MCA è lo studio di Marco Centoni e Antonello Maruotti (2021) sul tema della valutazione dei corsi universitari. In questo caso, poniamo l'attenzione sul modo in cui l'analisi delle corrispondenze multiple è stata usata in combinazione con la cluster analysis gerarchica, tecnica che approfondiremo nel capitolo 6, utile per creare gruppi fra unità statistiche simili all'interno di un campione.

Gli autori si focalizzano sull'importanza che in vari contesti internazionali assume la ricerca sulla valutazione delle attività didattiche da parte degli studenti, valutazione finalizzata a migliorare i percorsi formativi, rilevare punti di debolezza, riorganizzare i materiali, comprendere fenomeni di dropout e così via. Nei processi di assicurazione della qualità nel contesto italiano, ben noti in ambito accademico, gli studenti sono coinvolti in piccoli numeri nelle attività di autovalutazione, valutazione periodica e accreditamento (AVA). Tutti gli immatricolati nelle università, invece, sono interessati nella rilevazione della soddisfazione e della valutazione sui corsi attraverso la compilazione obbligatoria di questionari per ciascun insegnamento con un set di domande definite dall'ANVUR (Agenzia nazionale di valutazione del sistema universitario e della ricerca). I questionari, uguali per l'intero territorio nazionale, differiscono da altri strumenti valutativi in altri sistemi internazionali. Proprio tali questionari sono parte del dataset utilizzato nello studio insieme a variabili relative a caratteristiche personali e di partecipazione ai corsi di laurea rilevate in maniera del tutto anonima per gli studenti coinvolti nell'analisi e iscritti presso l'ateneo LUMSA (Libera Università Maria Ss. Assunta), sede dell'indagine. Gli autori individuano più domande di ricerca che si concentrano sulle relazioni che intercorrono fra la soddisfazione generale degli studenti e la percezione della qualità dell'insegnamento, la struttura dei corsi, il genere e l'interesse ai temi delle discipline da parte degli studenti. La MCA, utile a rilevare le strutture latenti che formalizzano tali relazioni, è stata condotta prima sull'intero campione e in seguito sugli otto corsi di laurea dedicati alle scienze sociali alla LUMSA. La Tabella 3.11 definisce il significato delle prime due dimensioni identificate per ciascuna MCA.

Per ogni analisi delle corrispondenze realizzata, gli autori riportano gli istogrammi che rappresentano il contributo di ciascuna modalità alle prime due dimensioni, il biplot nel quale l'intensità del contributo di ciascuna modalità (`ctr`) è contrassegnata da un colore e, nei casi dei singoli corsi di laurea, anche una rappresentazione dei cluster calcolati a partire dalle coordinate dei punti che rappresentano ciascun studente nella MCA condotta sulle righe. La Figura 3.17 è relativa al corso di laurea in economia. Le prime due dimensioni, nominate "organizzazione del corso" e "qualità del docente", spiegano il 74,3% della varianza. I tre cluster identificati distinguono gli studenti in base alle loro valutazioni: completamente positive, parzialmente positive, negative sia per la soddisfazione riferita al corso che per quella riferita al docente.

Lo studio consente quindi di identificare alcuni elementi particolarmente rilevanti nelle valutazioni degli studenti e, attraverso la cluster analysis, ne valorizza l'eterogeneità.

Degree	Definition of the First Dimension	Definition of the Second Dimension
All	Course organization	Lecturer behaviour
Economic	Course organization	Lecturer quality
Marketing	Course organization	Lecturer behaviour
Psychology	Course expectation	Lecturer performance
Law (Rome)	Lecturer behaviour	Course organization
Law (Palermo)	Lecturer quality	Lecturer performance
Social Work	Course organization	Lecturer performance
Education	Course organization	Lecturer quality

Tabella 3.11 – Definizione delle componenti principali sulle valutazioni da parte degli studenti dei corsi di laurea della LUMSA (Centoni & Maruotti, 2021, p. 7).

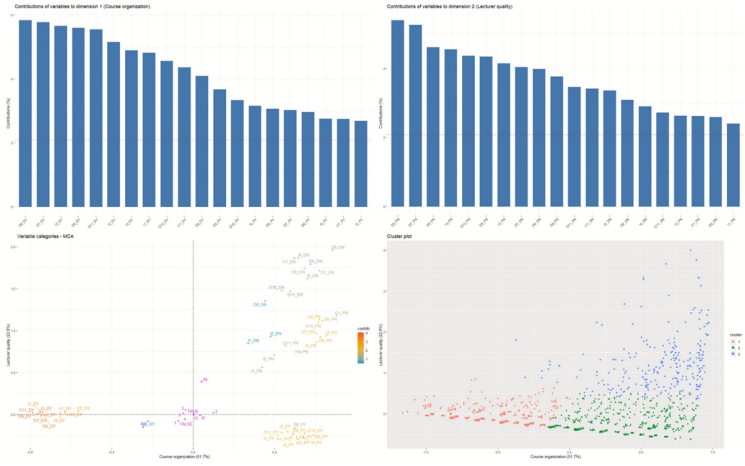


Figura 3.17 - Rappresentazioni grafiche relative al corso di laurea in economia presso la LUMSA. Area superiore: contributi alle dimensioni. Area inferiore, grafico a sinistra: rappresentazione delle dimensioni nel piano; grafico a destra: cluster gerarchico in base alle preferenze degli studenti (Centoni e Maruotti, 2021, p. 8).

Proseguiamo con un caso che, pur restando nel settore dell'alta formazione, si distingue da quelli trattati finora poiché utilizza come unità statistiche le discipline invece degli individui e delle loro opinioni/caratteristiche. Si tratta dello studio "The surprising persistence of Biglan's classification scheme" di Adrian Simpson (2015) dedicato alla classificazione delle discipline dell'alta formazione. Obiettivo dell'analisi è identificare eventuali similarità fra le discipline distribuite fra le istituzioni che le propongono nel Regno Unito. Una precedente classificazione a cui si fa riferimento è lo schema di Biglan, sviluppato sulla base delle osservazioni di accademici statunitensi (contesto diverso a quello dello studio di Simpson). Essa identifica tre dimensioni che distinguono le discipline in *hard/soft* per il livello di condivisione delle teorie di base, *pure/applied* per il rapporto con la risoluzione di questione pratiche, *life/nonlife* per differenziare oggetti di studio che hanno a che fare con sistemi biologici e sistemi inanimati o astratti.

Lo studio analizza più di 23mila corsi di laurea in 113 istituzioni con 82 discipline. Di queste solo 51 sono presenti nello schema di Biglan e classificate in base alle tre dimensioni. L'analisi delle corrispondenze semplici è stata effettuata quindi su una tabella di contingenza di 82 per 113. La Tabella 3.12 mostra le dimensioni per le quali si è registrato un valore dell'inerzia superiore al 2%, le prime due dimensioni coprono il 25,7% della varianza del dataset e dal test del χ^2 sembrano coincidere con le dimensioni hard/soft, pure/applied nel confronto con lo schema di Biglan.

I logistic plot ai lati del biplot in Figura 3.18, che comprende le 51 discipline già classificate, mostrano che il segno delle coordinate dei punti differenzia le discipline in base alla classificazione già nota (corrispondenza dell'89% delle discipline nella dimensione 1, 94% nella dimensione 2).

L'analisi delle corrispondenze è stata ripetuta sull'intero campione e le discipline non classificate sono state inserite in un quadrante che rispecchia posizioni convincenti. La geologia ad esempio è nel quadrante hard/pure, così come la teologia nel soft/pure (Figura 3.19). L'analisi ha una validità predittiva per inserire le discipline non classificate nello schema oggetto dello studio.

Concludiamo questa breve rassegna, suggerendo uno studio di matrice sociologica sul tema dell'uso sociale di Internet fra gli adolescenti in due paesi, Germania e Norvegia, di Tomasz Drabowicz (2017) come esempio di utilizzo delle variabili supplementari in una analisi delle corrispondenze.

Dimension	Value	Inertia (%)	Cumulative inertia (%)
1	0.321318	17.6	17.6
2	0.146521	8.0	25.7
3	0.091230	5.0	30.7
4	0.069890	3.8	34.5
5	0.067433	3.7	38.2
6	0.057522	3.2	41.3
7	0.052371	2.9	44.2
8	0.048724	2.7	46.9
9	0.047405	2.6	49.5
10	0.045890	2.5	52.0
11	0.043948	2.4	54.4
12	0.042319	2.3	56.7
13	0.038560	2.1	58.8
14	0.036828	2.0	60.9

Tabella 3.12 - Dimensioni che contribuiscono per più del 2% all'inerzia del modello di analisi delle corrispondenze per la classificazione delle discipline dell'alta formazione (Simpson, 2015, p. 6).

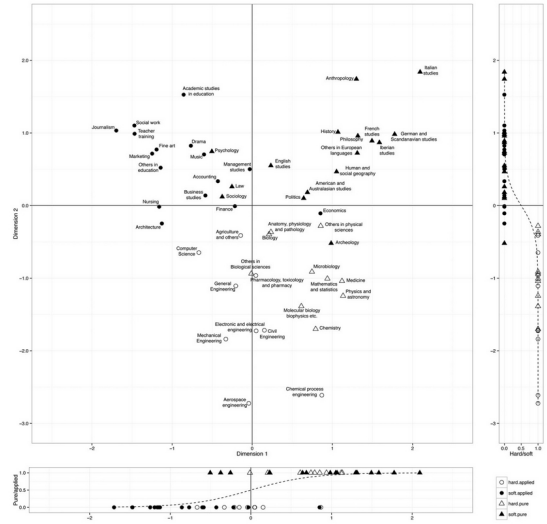


Figura 3.18 - Biplot delle analisi delle corrispondenze sulle prime due dimensioni, con i logistic plot delle discipline universitarie nella classificazione esistente (Simpson, 2015, p. 7).

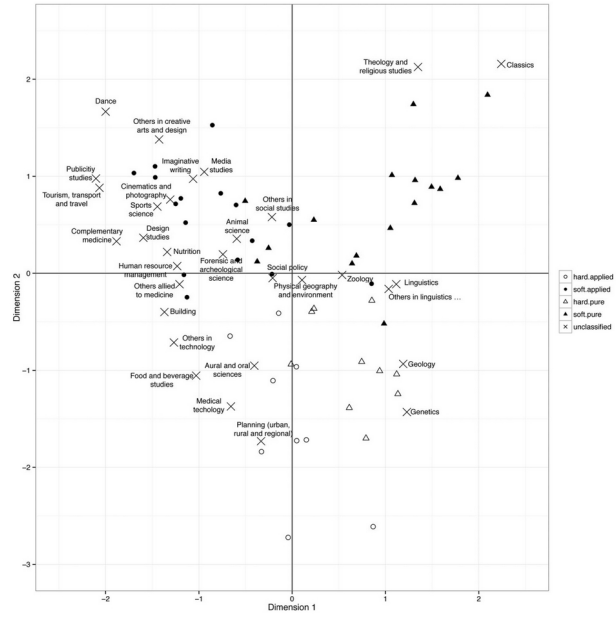


Figura 3.19 - Biplot delle analisi delle corrispondenze sulle prime dimensioni delle discipline precedentemente non classificate in uno studio per la classificazione delle discipline dell'alta formazione (Simpson, 2015, p. 8).

CAPITOLO 4

REGRESSIONE LINEARE

Al termine del capitolo, il lettore sarà in grado di:

- *descrivere le procedure alla base della regressione lineare semplice e multivariata;*
- *spiegare i valori relativi alla significatività e bontà dei modelli di regressione lineare;*
- *elencare esempi della ricerca educativa nei quali è stata utilizzata la regressione lineare multivariata.*

4.1 - Regressione lineare

La regressione lineare rappresenta forse il metodo di analisi bivariata e multivariata più noto nelle scienze sociali. È una tecnica molto utilizzata anche in altri ambiti scientifici (ingegneria, matematica, fisica ecc.) ma con "nomi" diversi: approssimazione, interpolazione, fitting.

Si tratta di una tecnica di dipendenza definita *asimmetrica* poiché alle variabili impiegate non viene attribuito lo stesso ruolo: quelle definite indipendenti vengono utilizzate per stimare (o approssimare) il comportamento di un'unica variabile definita dipendente. Conosciamo già i termini "dipendente" e "indipendente" e il loro significato (vedi capitolo 2). Nel metodo della regressione le variabili indipendenti, x_i, sono anche definite regressori, predittori, variabili esplicative e possono essere di tipo quantitativo o qualitativo (in quest'ultimo caso sono trattate come dummy o opportunamente trasformate); la variabile dipendente, y, detta anche variabile risposta, è una variabile necessariamente quantitativa (se y è una variabile categoriale o ordinale, si parla di regressione logistica, tecnica descritta nel capitolo successivo).

Predire il comportamento della variabile dipendente è lo scopo principale dei metodi di regressione nella formulazione e risoluzione dei problemi di ricerca. La regressione è usata però anche per spiegare gli effetti delle variabili indipendenti sulla variabile dipendente a partire dalla rilevazione della signifi-

cattività e dell'intensità degli stessi effetti, della forza cioè, con cui si manifestano.

Nella pratica, in maniera estremamente sintetica, l'uso di questo metodo ci porta a definire la relazione che lega le x_i alla y e che viene espressa attraverso una funzione matematica in base alla quale possiamo:

- calcolare i valori della y stimati a partire da quelli delle x_i non osservati;
- attribuire un'interpretazione ai parametri che compongono la funzione e definiscono la relazione fra le variabili;
- confrontare modelli costruiti utilizzando variabili e tecniche diverse.

Alla luce di quanto detto, possiamo perciò definire la regressione come un metodo *model-based* ossia finalizzato all'individuazione di un modello statistico e, dunque, di una funzione matematica (definita da una formula e, se possibile, rappresentata attraverso una curva geometrica) che descriva l'andamento della distribuzione dei dati raccolti empiricamente, dando una forma regolare alla relazione che esiste fra la variabile dipendente e la/le variabili indipendenti.

Marcello Galli e Tommaso Minerva (1999) affermano che:

> *"Un modello statistico è una rappresentazione semplificata, analogica e necessaria della realtà derivata da osservazioni sperimentali oltre che da deduzioni logiche. L'aspetto dialettico nella costruzione di un modello statistico deriva dalle opposte esigenze di semplificare la struttura senza perdere in fedeltà, e tale conflitto è ineliminabile. Infatti, tutti i modelli sono intrinsecamente sbagliati: essi sono parzialmente e provvisoriamente utili, e sono destinati a essere sostituiti con l'avanzare del progresso scientifico e l'affinamento della conoscenza. Ciò che realmente conta non è la validità ontologica delle relazioni accertate ma l'efficacia comparata in rapporto agli obiettivi. È l'obiettivo, infatti, che rende utile, efficace e temporaneamente valido il modello." (ivi, p. 5/48)*

Specificare un modello significa: definire la funzione che lega y con le x_i; definire le x_i incluse nella relazione; definire l'insieme dei parametri legati alle x_i. Come abbiamo appena letto, non esistono modelli perfetti e assoluti; esistono modelli che in maniera più completa di altri riescono a esprimere relazioni fra

variabili in relazione agli obiettivo di studio. Pertanto da un lato è necessario fare delle scelte nella fase di analisi anche basandosi sulla conoscenza del fenomeno che si sta studiando, dall'altro è indispensabile capire i limiti di validità dei modelli ottenuti.

La formulazione generica per descrivere la funzione, la relazione cioè fra y e le x_i è

(4.1)

$$\hat{Y} = f(x_i) + \varepsilon$$

dove $\hat{Y}$ è la variabile dipendente stimata, f la funzione matematica/il tipo di relazione, x_i le variabili indipendenti/esplicative/regressori e ε l'errore casuale.

Nella regressione lineare, come dice il nome stesso, la funzione che esprime la relazione fra le variabili è una retta. Non necessariamente la relazione fra due variabili assume la "forma" di una retta: ci sono curve come parabole, esponenziali, logaritmiche, funzioni periodiche, iperboli e altre. Certamente la retta è la curva più semplice da studiare e pertanto la si preferisce alle forme non lineari anche per via di una maggiore solidità del modello statistico sottostante. Molto spesso nel caso in cui le distribuzioni non siano interpolate da una retta, vengono applicate trasformazioni ai dati per poter far rientrare il caso studiato in quello lineare. Non sempre questa operazione è possibile e di conseguenza altri strumenti e altri metodi sono usati per analisi di tipo non lineare.

Distinguiamo la regressione lineare in semplice o multivariata in base al numero di variabili indipendenti considerate. La regressione semplice rientra fra le tecniche di analisi statistica bivariata e prende in considerazione un'unica variabile indipendente come predittore della y. La regressione multivariata utilizza più variabili indipendenti per definire la y.

Partiamo nella nostra discussione parlando di regressione lineare semplice per poter comprendere il processo alla base del metodo multivariato.

Regressione lineare semplice

Come anticipato, nella regressione lineare semplice abbiamo due variabili, x e y, e n osservazioni raccolte empiricamente.

La x potrebbe essere il voto agli esami di maturità per gli n studenti di una scuola secondaria di secondo grado di una certa città e la y la media dei pun-

teggi conseguiti negli esami sostenuti durante il primo anno all'università. Oppure la *x* il numero di libri letti in un anno da *n* bambini della quarta primaria in estate e la *y* la velocità di lettura degli stessi al rientro dalle vacanze. O ancora la *x* la durata di *n* videolezioni inserite in un corso online e la *y* il numero di visualizzazioni da parte degli utenti iscritti.

Il nostro scopo è verificare se esiste una relazione fra le variabili *x* e *y* e capire se questa relazione può essere sintetizzata nell'andamento di una retta. Conoscere la formula matematica che descrive la relazione ci permette di dire con quanta forza il fenomeno rappresentato dalla variabile *x* incide sul fenomeno determinato dalla *y* e ci dà la possibilità di stimare il valore della *y* per nuovi valori della *x* non osservati.

Riprendendo gli esempi che torneranno utili anche nel seguito della discussione: quanto il voto in uscita di uno studente delle scuole secondarie condiziona i punteggi dei risultati negli esami nel primo anno di università? Possiamo stimare la media degli esami delle matricole per un nuovo gruppo di studenti una volta che sono noti i voti degli esami di maturità?

O ancora: che tipo di relazione lega la durata dei video al numero delle corrispettive visualizzazioni? Riusciamo a verificare l'ipotesi che più brevi sono i video e più numerose sono le visualizzazioni da parte degli utenti in un corso online?

Come possiamo scrivere questi fenomeni in un'equazione matematica?

Ipotizziamo che la relazione fra i fenomeni descritti abbia un andamento lineare, rappresentiamo i dati raccolti empiricamente in uno scatterplot e scegliamo la retta più vicina ai punti riportati di cui calcoliamo l'equazione.

L'equazione di una retta generica che lega *y* e *x* in una relazione lineare semplice è:

(4.2)

$$y = a + bx$$

dove

- *a* è detta *intercetta* ed è il valore di *y* in corrispondenza di *x* uguale a 0. Graficamente *a* è il valore di *y* nel punto in cui la retta tracciata interseca l'asse delle ordinate. Negli esempi, *a* è la velocità di lettura di un bambino che abbia letto 0 libri durante l'estate;

- *b* è il coefficiente angolare della retta, ossia l'inclinazione della retta nel piano. In ambito statistico *b* è definito *coefficiente di regressione* e rappresenta il valore aggiunto alla *y* aumentando la *x* di una unità. Negli esempi, *b* è la quantità che aggiungiamo al calcolo della velocità di lettura per ogni libro in più che un bambino legge durante le vacanze estive. *b* può assumere valori positivi e negativi che indicano una dipendenza positiva o negativa fra le due variabili. Assume il valore 0 quando la retta disegnata è parallela all'asse delle ascisse e di conseguenza le variabili sono completamente indipendenti dato che la *y* resta costante per qualunque valore assunto dalla *x* . Negli esempi: *b = 0* se la velocità di lettura resta la stessa qualunque sia il numero di libri letto dai bambini del campione analizzato. Di conseguenza i due fenomeni, lettura estiva dei libri e velocità di lettura, risultano essere completamente indipendenti.

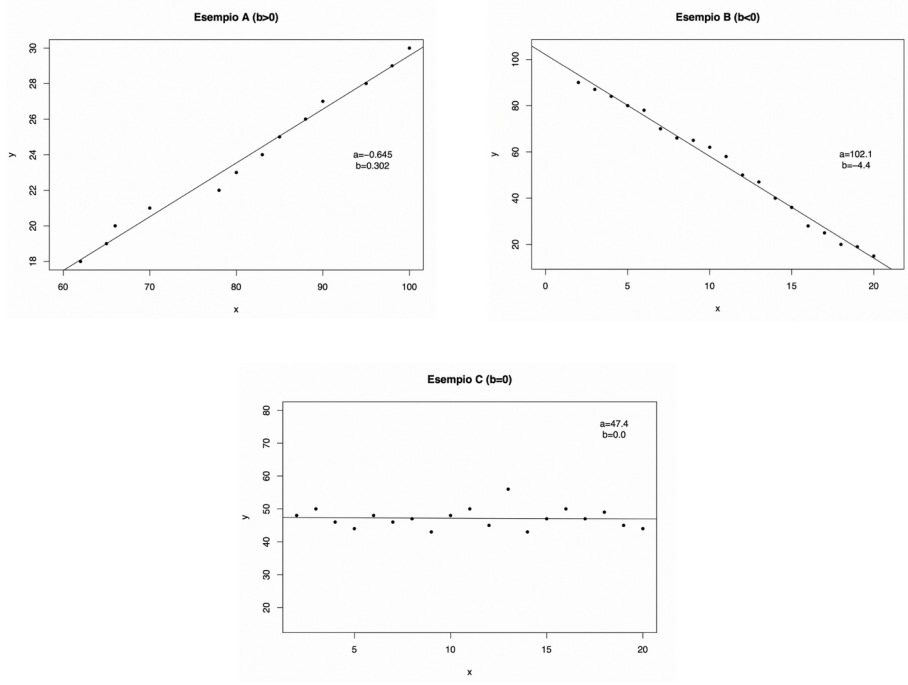


Figura 4.1 - Tre esempi di rette di regressione con $b > 0$, $b < 0$, $b = 0$.

Nella Figura 4.1 sono presenti tre rappresentazioni grafiche di rette di regressione lineare.

Nell'esempio A all'aumentare del valore delle *x*, aumentano anche i valori delle *y*, come nel caso in cui all'aumentare del voto degli esami di maturità, au-

menta anche la media dei voti degli esami del primo anno conseguiti dalle matricole. Il valore di b è in questo caso positivo.

L'esempio B con b negativo potrebbe invece essere la rappresentazione grafica dell'evento in cui il numero di visualizzazioni diminuisce all'aumentare della lunghezza dei video.

Negli esempi A e B osserviamo una correlazione fra le variabili poiché esse variano insieme; nell'esempio C vediamo una perfetta indipendenza di x e y: per qualunque quantità aumenti il valore di x , y resta costante. b in questo caso è pari a 0.

La formulazione statistica dell'equazione della retta di regressione prevede che nel calcolo del valore della y sia aggiunta anche la quantità ε, ossia una variabile casuale che definisce l'errore, la variabilità non spiegata che intercorre fra i valori reali e quelli stimati della y, espressa nella formula (4.3) e visualizzata nella Figura 4.2 come distanza verticale (non perpendicolare!) fra i punti osservati e la retta di regressione.

(4.3)

$$y = a + bx + \varepsilon \text{ dove } \varepsilon = y - \hat{y}$$

Essendo ε una variabile casuale è necessario definirne, ovvero fare ipotesi, sulla distribuzione. Nel caso della regressione lineare l'assunto è che la ε sia distribuita come una normale standardizzata, ossia con media nulla e varianza costante.

Il processo di stima nella regressione lineare consente quindi di ottenere sia una stima del valore della variabile dipendente sia una stima dei residui, ossia gli stimatori ε_i dell'errore.

L'analisi dei residui è una fase molto importante della regressione lineare perché consente di fornire delle valutazioni sulla bontà del modello di approssimazione adottato. In particolare, la loro distribuzione deve risultare il più vicino a una normale con media 0 e varianza costante per soddisfare l'assunto di normalità della componente casuale.

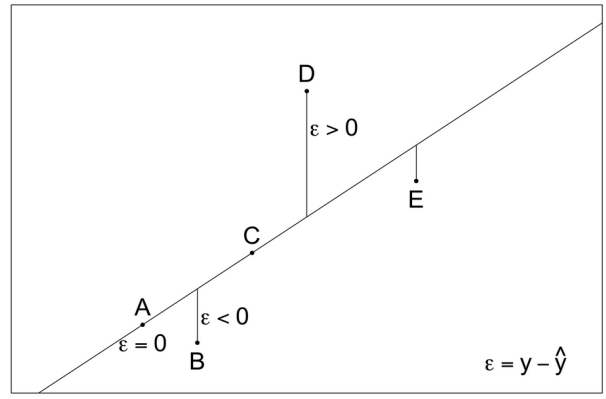


Figura 4.2 – Visualizzazione dei residui. *ε* è pari a 0 se valori reali e stimati coincidono.

Possiamo a questo punto comprendere come calcolare *a* e *b*.

a e *b* vengono calcolati come quelle quantità che rendono minima la somma dei quadrati dei residui, ossia delle distanze fra i valori realmente osservati della *y* e i valori della *y* ottenuti dall'equazione della retta e dunque stimati. Nel calcolo le distanze sono al quadrato; se non lo fossero, la loro somma risulterebbe tendere a 0.

In base a questo metodo, detto dei *minimi quadrati (OLS, Ordinary Least Squares* in inglese*)*, possiamo definire come seguono i valori di *a* e *b* (queste formule non ci devono spaventare perché i principali software di analisi statistica hanno delle funzioni che calcolano in maniera automatica i parametri *a* e *b* delle rette di regressione insieme, ovviamente, come diremo, a molti altri parametri).

(4.4)

$$b = \frac{\sigma_{xy}}{\sigma_x^2}$$

(4.5)

$$a = \bar{y} - b \cdot \bar{x}$$

b è quindi il rapporto fra la covarianza di *x* e *y* e la varianza di *x*.

a la differenza fra la media delle *y* e *b* per la media delle *x* e coincide con il coefficiente di correlazione se *x* e *y* hanno la stessa varianza.

Noti *a* e *b* , siamo in grado di scrivere l'equazione della retta di regressione che interpola la nuvola di punti nello scatterplot e di calcolare i nuovi valori assunti dalla *y* fissando valori per la *x*.

Ad esempio, parlando di velocità di lettura, sapendo che a scopo esemplificativo, *a = 12* e *b = 3*, potremo stimare la velocità di lettura di un bambino della quarta elementare usando la seguente formula:

(4.6)

$$velocità\ di\ lettura = 12 + 3 \times numero\ di\ libri\ letti$$

Per ogni nuovo libro letto, la velocità aumenta di 3 unità (1 libro => velocità 15; 2 libri => velocità 18; 5 libri => velocità 27). Un bambino che non ha letto alcun libro, avrà una velocità di lettura pari a 12 ossia al valore del parametro *a*.

Si tenga presente un aspetto importante: l'ambito statistico è molto diverso da quello analitico. Nell'ambito statistico le variabili sono "casuali", il modello stesso di regressione include una componente "casuale". È indispensabile, pertanto, tenere conto di alcuni aspetti che completano il processo di stima:

- i parametri *a* e *b* sono essi stessi delle variabili casuali e pertanto i valori che si determinano rappresentano delle "stime" puntuali cui sono associati dei livelli di significatività;
- è necessario effettuare una profonda analisi dei residui per valutarne la normalità e l'omoschedasticità (assunti di base del modello) e anche una stima dell'errore che si commette. Una corretta analisi dei residui ci fornisce informazioni importanti sulla validità del modello adottato ovvero sulla necessità di effettuare trasformazioni sulle variabili o addirittura investigare modelli formalmente più complessi.

Regressione lineare multivariata

Molto spesso per spiegare un fenomeno non basta studiare soltanto una tipologia di eventi a esso collegata. Risulta più efficace nell'analisi introdurre più variabili per considerare come ciascuna di esse è collegata al fenomeno che stiamo studiando. Si inseriscono così altre variabili, x_1, x_2, x_3, … , x_n, scelte innanzitutto a partire da un framework teorico e logico che può giustificare il modello statistico che stiamo elaborando. Il passaggio da una variabile indipendente a più variabili indipendenti segna il passaggio dall'analisi bivariata a quella multivariata e di conseguenza da una regressione (lineare) semplice a una multivariata.

In questo caso non possiamo fare affidamento sulla rappresentazione grafica poiché all'aumentare delle variabili, e di conseguenza, delle dimensioni di cui è composto il sistema, non è possibile predisporre una visualizzazione della relazione individuata. Riusciamo a rappresentare graficamente la "forma lineare individuata" solo in modelli con due variabili indipendenti come un piano in un sistema tridimensionale. Tuttavia quello che abbiamo detto (e "visto" in Figura 4.1) per la regressione lineare semplice è sufficiente per comprendere cosa accade quando lavoriamo con più variabili.

Aumentare il numero di variabili nell'analisi aumenta l'accuratezza della predizione nel modello statistico. Determinare i voti degli studenti del primo anno dell'università a partire unicamente dal voto della maturità è molto diverso dal tentare di predirli (statisticamente parlando, meglio stimarli) tenendo in considerazione anche i dati anagrafici, il numero di ore di studio per settimana, il livello di motivazione.

L'equazione del modello statistico con più predittori può quindi essere scritta come segue:

(4.7)

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_n x_n + \varepsilon$$

dove gli x_n sono i predittori e le b_i sono i coefficienti parziali di regressione e indicano quanto la y aumenta a causa di una variabile tenendo costanti tutte le altre.

Cosa significa? Consideriamo il caso già proposto in cui la velocità di lettura sia la nostra y. Oltre al numero di libri letti in un anno (x_1), aggiungiamo in un

modello multivariato altre variabili esplicative come la conoscenza delle regole grammaticali misurata in un questionario (x_2), il livello di piacere nella lettura in una scala con punteggi da 1 a 5 (x_3), la presenza di disturbi dell'apprendimento con una variabile binomiale: diagnosi di DSA, sì/no (x_4). Aumentando di una unità il numero di libri letti in un anno x_1, la y incrementerà il suo valore della quantità b_1 purché i valori di x_2, x_3, x_4 restino invariati. b_1 rappresenta la quantità nel valore della y che distingue uno studente che ha letto 3 libri da un compagno che ne ha letti 4 avendo una stessa conoscenza delle regole grammaticali, uno stesso livello di piacere della lettura e uno stesso quadro in riferimento alla diagnosi di DSA.

Le variabili binomiali o dummy (come ad esempio x_4) non fanno altro che modificare il valore dell'intercetta di una quantità pari al corrispondente coefficiente parziale di regressione quando assumono il valore 1; la loro influenza nel calcolo è nulla quando il loro valore è pari a 0.

Poiché l'individuazione di un modello comporta una semplificazione della realtà osservata, prima di applicare queste procedure, è indispensabile verificare alcuni assunti sulle singole variabili dello studio, sul modello di regressione individuato e sui residui per capire se gli errori verificati sui valori della variabile indipendente sono determinati dal caso o da motivazioni legate ai dati e alla loro rilevazione.

Fra gli assunti da tenere in considerazione:

- *indipendenza dei casi*. Le osservazioni devono essere estratte a caso, non devono essere rilevati errori di misurazione ossia non deve esserci una correlazione fra gli errori di stima della variabile dipendente e i regressori.

- *dimensioni del campione.* Anche se non ci sono regole precise, si assume che dovrebbe esserci almeno un rapporto di 5:1 fra osservazioni e variabili dello studio (Plonsky & Ghanbar, 2018).

- *linearità.* La matrice di correlazione fra y e x_i fornisce informazioni a priori sulla linearità della relazione mentre una adeguata analisi dei residui ne fornisce una robusta valutazione a posteriori. Se l'assunto di linearità non dovesse essere verificato, si possono applicare trasformazioni dei dati in forme lineari (ad es. con i logaritmi), considerare relazioni non lineari (ad es. regressione polinomiale, dove le variabili indipendenti sono elevate a una qualche potenza), o ancora suddividere il campione in sottogruppi dove per ciascuno è possibile identificare un modello lineare.

- *indipendenza dei regressori*. Viene verificata evitando, o limitando, fenomeni di multicollinearità per fare in modo che l'effetto esercitato da ogni regressore sia indipendente da quello degli altri regressori. Se dovessero esserci correlazioni fra i regressori che quindi variano insieme, risulterà difficile valutare gli effetti predittivi di ciascuna variabile perché gli eventi osservati si modificano simultaneamente e si influenzano l'un l'altro.

- *normalità:* la normalità (e quindi la simmetria) della distribuzione dei residui è l'assunto di base per lavorare sulla significatività del modello con intervalli di confidenza e test di ipotesi.

- *omoschedasticità:* la varianza degli errori deve essere costante e non associata a un determinato regressore o alla variabile dipendente. Se la varianza fosse estremamente ampia per alcune osservazioni oppure fosse correlata a un determinato regressore o alla y, nel modello potremmo non aver tenuto in debita considerazione tutte le possibili relazioni fra gli eventi.

Gli effetti di multicollinearità o di bassa correlazione con alcune variabili ci inducono a effettuare una selezione delle variabili (che implica una selezione del modello). Il processo di selezione delle variabili è spesso un processo complesso e non necessariamente univoco.

4.1.1 - Interpretazione e affidabilità dei modelli, selezione delle variabili

Che significato attribuire ai parametri trovati? Quali indicatori ci garantiscono l'affidabilità del modello? E quali indicano il grado di generalizzabilità dei risultati ad altri campioni?

Rispondere a queste domande ci permette sia di interpretare i valori finora calcolati, sia di fare un ulteriore passaggio e muoverci dalla statistica descrittiva a quella inferenziale.

Consideriamo in Tabella 4.1 un esempio dei principali indicatori restituiti dal software di analisi statistica `R` relativi a un modello di regressione lineare multivariata. Lo studio di cui riportiamo i risultati (De Santis et al., 2019) condotto dal team di coordinamento di EduOpen era finalizzato all'individuazione di un approccio per selezionare un modello di regressione lineare multivariata che stimasse la percentuale di completamento dei MOOC della stessa piattaforma da

parte degli utenti (*y*). 24 predittori sono stati raccolti in 4 categorie: profilo degli utenti, partecipazione al corso, attività completate, caratteristiche dei corsi.

Regression model for CRATE
Residual standard error: 0.2165 on 705 DF
Residuals:

Min	1Q	Median	3Q	Max
-1,633	-0,126	-0,0324	0,155	0,547

Multiple R-squared: 0,756
Adjusted R-squared: 0,7504
F-statistic: 136,5 on 16 and 705 DF, p-value: < 2,2e-16

Variable	Coefficient	SE	t-test	p-value
(Intercept) *	-0,247	0,085	-2,897	0,004
GENDER *	-0,049	0,017	-2,804	0,005
DEGREE *	0,013	0,006	2,065	0,039
AGE *	-0,011	0,004	-2,446	0,015
CHILDREN	0,031	0,021	1,514	0,130
SECTOR	-0,003	0,002	-1,526	0,127
EFFORT	0,001	0,000	1,715	0,087
DROPOUT_TOT	-0,013	0,007	-1,865	0,063
DROPOUT_INT	0,012	0,008	1,504	0,133
DROPOUT_LEA	0,014	0,008	1,700	0,090
DROPOUT_NAV	0,012	0,008	1,544	0,123
MOTIVATION *	0,003	0,001	2,043	0,041
CLICKS_TRACKED *	-0,042	0,009	-4,802	0,000
CLICKS_TOTAL *	0,220	0,013	16,919	0,000
CTUTORED *	0,064	0,026	2,517	0,012
CCAT *	0,126	0,029	4,288	0,000
CHOUR *	0,016	0,003	4,899	0,000

LEGEND:
DF = Degree of Freedom
SE = Standard Error
* = variable with p-value < 0,05 at 95% significance level

Tabella 4.1 - Esempio di regressione lineare multivariata realizzata con un processo di selezione delle variabili *stepwise* in cui la variabile dipendente CRATE rappresenta la percentuale di completamento delle attività in un MOOC (ripreso e integrato da De Santis et al., 2019, p. 154). La funzione usata in `R` per le procedure di analisi legate alla regressione lineare è `lm` (libreria: `stats`).

L'analisi ha portato alla conclusione che il modello più convincente fra quelli studiati era quello che attribuiva maggiore potere predittivo alle variabili relative al numero di click degli utenti nei corsi (ossia al loro comportamento in piattaforma).

Nella riga introduttiva della Tabella 4.1 sono elencati i valori che ci restituiscono delle valutazioni complessive sull'intera regressione e cioè:

- *Residual standard error*: è la radice del rapporto fra il quadrato dei residui e i gradi di libertà. Ci dice di quanto i valori della y si discostano mediamente dal valore reale. Più basso sarà tale valore, maggiore sarà la bontà del modello di regressione (se fosse pari a 0, il modello corrisponderebbe perfettamente ai dati osservati). Nel calcolo del RSE vengono utilizzati i gradi di libertà (*df = degree of freedom*) ossia la differenza fra il numero di osservazioni che appartengono al campione o alla popolazione analizzati e il numero dei parametri stimati (pari al numero dei coefficienti parziali di regressione b_i più uno o ancora al numero delle $x_i + y$). Più alto è il numero dei gradi di libertà, più il modello è generalizzabile.

- *Residuals*: le righe riportano i quartili della distribuzione dei residui dai quali possiamo analizzare la simmetria della distribuzione e di conseguenza avere informazioni che contribuiscono a verificare l'assunto di normalità dei residui.

- *Multiple R-squared* (R^2): si tratta del rapporto fra il quadrato delle distanze dei valori stimati della y dalla media e il quadrato delle distanze reali dalla media. Viene definito *coefficiente di determinazione*, indica la quantità di varianza della variabile dipendente spiegata dalle variabili indipendenti. Rappresenta la percentuale di casi spiegati dal modello individuato. Ha pertanto valori compresi fra 0 e 1. In generale, a bassi valori di R^2 corrispondono previsioni meno affidabili. Non è necessariamente vero il contrario. Sicuramente un elevato valore di R^2 rappresenta una indicazione positiva ma, ricordando che in un modello lineare è il coefficiente di correlazione fra i valori osservati (sperimentali) della y e i valori stimati dal modello, dipende fortemente dal numero di variabili incluse nel modello. R^2, infatti, aumenta all'aumentare del numero delle variabili anche se queste possono non essere tutte significative. È un indicatore "debole" di bontà del modello nell'ambito della regressione multivariata.

- *Adjusted R-squared* (*AdjR²*): è un indicatore che completa il precedente poiché tiene in considerazione anche i gradi di libertà (ossia il numero di variabili) del modello nel calcolo della varianza. R^2 aumenta ogni volta che un nuovo predittore (anche non significativo) è aggiunto all'analisi, $AdjR^2$ fornisce valori più stabili. R^2 e $AdjR^2$ sono le misure più dirette di quanto l'equazione matematica individuata interpoli i dati osservati.

- *F-statistic*: è il risultato del test statistico *F* sull'ipotesi che confronta la varianza spiegata dalla regressione con quella non spiegata (residui) tenendo in considerazione il numero di variabili indipendenti coinvolte (nel nostro caso 16) e i gradi di libertà. L'ipotesi nulla verificata dal test è che i coefficienti di regressione siano uguali a 0, cioè che non vi sia una dipendenza fra variabile dipendente e indipendenti e che il modello non abbia una forza predittiva. Se risulta significativo sia per il valore rilevato nelle tabelle di *F*, che per il valore di significatività statistica *p* calcolato, possiamo affermare con sufficiente certezza che il modello può essere predittivo anche per altri campioni della stessa popolazione.

Nella parte centrale della Tabella, le colonne restituiscono informazioni sulle variabili e sulla loro significatività. All'intercetta e a ciascuna variabile considerata nello studio si fanno corrispondere: i valori dei coefficienti di regressione parziale; lo *standard error of the estimate* ossia la deviazione standard dei valori predetti; *t-test* e *p-value*.

I valori sono di per sé variabili casuali che si distribuiscono secondo una distribuzione *t*, su questa calcoliamo il p-value.

I valori dei coefficienti b_i sono affetti da un'incertezza di fondo che rende necessario verificare la significatività di ciascuno di essi, ponendo l'ipotesi H_0 che il valore ottenuto sia solamente frutto del caso. In aggiunta o in alternativa, calcolare lo standard error ci permette di definire l'intervallo di confidenza dei valori assunti dai coefficienti. Più l'intervallo è ristretto, più accurato sarà il modello. L'intervallo di confidenza non deve contenere lo 0 per indicare l'esistenza di una qualche dipendenza, poiché ovviamente assumere che fra i valori che il coefficiente b può assumere ci sia quello nullo, significa accettare che x e y siano indipendenti fra loro (de Lillo et al., 2007).

Dall'esempio della Tabella 4.1 vediamo che la distribuzione dei residui non è simmetrica (le condizioni di normalità dei residui non sono state verificate neppure da altri test usati), che il modello spiega il 75% dei casi (i valori di R^2 e $AdjR^2$ sono quasi sovrapponibili in questo caso) e dalla *F-statistic* risulta signifi-

cativo. Guardando alle variabili nelle righe vediamo che i coefficienti non sono tutti significativi alla luce del p-value e sono molto vicini allo 0. Ad es. EFFORT è pari a 0,001 e di conseguenza indipendente dalla variabile risposta; il $p\text{-}value < 0.087$ non è significativo (le variazioni non sono dovute soltanto al caso). Guardando invece alla variabile denominata CLICK_TOTAL vediamo che il valore del coefficiente di regressione parziale si discosta dallo 0 (è pari a 0,220), lo SE è molto basso e $p < 0.000$ (significativo). Nello studio ci siamo basati su queste osservazioni per arrivare a studiare un modello basato solamente, come dicevamo, sul numero dei click.

Nell'interpretazione dei valori dei coefficienti, i valori delle b_i per ogni predittore vanno letti in base alla scala con cui è espressa la variabile. Per confrontare l'intensità degli effetti dei coefficienti attribuiti alle singole variabili, si lavora con valori di β_i standardizzati che definiscono la variazione osservata sulle unità di variabili standardizzate. Assumono valori da -1 a 1 dove 0 indica assenza di correlazione. Questa formulazione non viene utilizzata per predire il valore della y, ma per rilevare e confrontare l'intensità degli effetti dovuti alle singole variabili.

Un'ultima considerazione va fatta sulla selezione delle variabili da inserire nei modelli di regressione (Paterlini & Minerva, 2010; Galli & Minerva, 1999). Nell'esempio, ritroviamo nelle righe 16 predittori dei 24 considerati nello studio. Non sempre tutte le variabili rilevate vanno inserite nei modelli da definire o perché non aggiungono informazioni rilevanti o perché non sono statisticamente significative. Il teorema fondamentale dell'algebra stabilisce l'esistenza di un polinomio interpolante per qualunque insieme di $n + 1$ punti, al più di grado n. Tradotto in termini più semplici con un numero altissimo di variabili, al più n, possiamo spiegare (interpolare con un polinomio) qualunque fenomeno. Lo scopo delle analisi statistiche invece è quello di usare il minor numero di variabili per rendere i modelli il più possibile esplicativi. Considerando n il numero delle variabili da selezionare, ogni studio potrebbe portarci a $2^n - 1$ modelli possibili. Se n è sufficientemente grande, il numero di modelli possibili può diventare ingestibile anche per i moderni software di calcolo.

La selezione delle variabili indipendenti avviene innanzitutto a partire dal modello teorico e dalle prospettive di ricerca che guidano l'analisi.

Joseph F. Hair e colleghi (2014) sintetizzano tre metodi per la selezione che è preferibile riproporre su set di dati per confermare i risultati.

Nel primo caso, denominato di *Confirmatory specification*, il ricercatore individua le variabili da inserire nel modello in base alle ipotesi e al disegno della

ricerca. Si scelgono le variabili indipendenti che hanno un coefficiente di correlazione più alto con la variabile dipendente poiché sono quelle in una relazione lineare con essa e contemporaneamente hanno un coefficiente di correlazione basso tra di loro.

Il caso opposto, *Combinatorial approach*, elabora tutti i modelli di regressione che possono essere generati da tutti i possibili incroci fra le variabili, considerando regressioni con *2*, *3*, *4*, *… , n* variabili indipendenti con tutte le possibili combinazioni fra le stesse. Verificando la significatività e la bontà dell'interpolazione, si sceglie il modello più adeguato. Si tratta di una soluzione poco usata poiché prevede una forte automatizzazione dei processi e uno scarso coinvolgimento del ricercatore e delle teorie che sono alla base delle attività di ricerca.

L'ultimo caso, molto diffuso, è rappresentato dai *Metodi sequenziali*, che attraverso meccanismi automatici aggiungono e rimuovono le variabili dal set che il ricercatore propone, fino ad ottenere un modello stabile. I metodi sequenziali sono di tre tipi:

- *forward addition*. Si parte nella regressione da un'unica variabile e a questa si aggiungono una alla volta le successive, scelte fra quelle maggiormente significative. Non è possibile in questo metodo eliminare una variabile già introdotta o aggiungere una variabile precedentemente esclusa;

- *backward elimination*. Diversamente rispetto al caso precedente, nella definizione del modello vengono introdotte tutte le variabili dello studio e di volta in volta si eliminano quelle che non incidono sulla bontà della regressione. Come nella forward addition, non è possibile aggiungere o eliminare variabili già considerate;

- *stepwise estimation*. In questo metodo, come nel forward estimation, si parte da una sola variabile, la più significativa, si calcola il modello e la sua bontà. Si aggiunge poi una seconda variabile significativa e si ridefinisce il modello. Se il modello risulta ancora stabile, si prosegue con l'inserimento di una nuova variabile, altrimenti si rimuove l'ultima variabile inserita e si sostituisce con la seguente. In questo caso fino alla fine dell'analisi si può continuare a modificare il gruppo di variabili selezionate aggiungendole o rimuovendole dal set. Questo metodo è quello utilizzato per la regressione in Tabella 4.1.

Numerosi criteri sono stati elaborati per valutare la qualità dei modelli finalizzati a verificare quanto sia bilanciato il rapporto fra il numero di variabili introdotte nel modello e la sua bontà. Fra quelli più utilizzati vi è il criterio di Akaike (*AIC*, *Akaike Information Criterion;* Akaike, 1969; 1978), che fornisce una indicazione quantitativa sulla bontà di adattamento anche in relazione al numero di variabili utilizzate (si ricordi che, generalmente, aumentando il numero di variabili l'adattamento dovrebbe migliorare).

Minore è il valore dell' AIC, migliore è il modello. Confrontando, dunque, più modelli costruiti a partire da uno stesso dataset, il modello più adeguato risulterà essere quello con il valore minimo dell'AIC. E questo aprirà un altro tema – non oggetto di questa pubblicazione – sugli algoritmi utilizzati per determinare il minimo di un indicatore come l'AIC o comunque di una funzione complessa.

4.2 - L'uso della regressione lineare multivariata nella ricerca educativa

La regressione lineare è usata per esplorare le relazioni fra eventi, per predire il comportamento di una variabile una volta che ne sono note altre, per individuare modelli e confrontarli verificandone la bontà o comparando i set di variabili che li definiscono.

Presentiamo tre casi di utilizzo della regressione lineare multivariata in tre diversi studi relativi a: l'uso dei serious games con i ragazzi delle scuole secondarie di primo grado, la definizione dei livelli di cittadinanza digitale degli insegnanti, il successo accademico di studenti universitari in un insegnamento erogato in modalità blended. Solo per un caso i tre articoli si riferiscono al settore del digitale, come è ovvio qualunque tema per il quale siamo in grado di definire variabili e raccogliere dati può essere analizzato attraverso questa metodologia.

Nina Iten e Dominik Petko, nell'articolo "Learning with serious games: Is fun playing the game a predictor of learning success?" (2016), pongono l'attenzione sul ruolo del fun nei serious games a partire dalla considerazione che tanto è stato detto sul ruolo del gioco nell'apprendimento e studi empirici che chiariscano la relazione fra fun e learning nella specifica applicazione dei serious games sono necessari.

Gli autori definiscono 4 ipotesi da verificare applicando la regressione multivariata che possiamo esprimere in maniera semplificata come segue:

- più i bambini si divertono, più saranno disposti a usare i learning games (divertimento e volontà di usare i giochi);
- più i bambini si divertono, più aumenterà il livello di motivazione a impegnarsi negli argomenti di studio del gioco (esperienza di divertimento e autovalutazione della motivazione di apprendere e impegnarsi negli argomenti di studio del gioco);
- più i bambini si divertono, più avranno la percezione di aver acquisito maggiori conoscenze sui temi del gioco (divertimento e autovalutazione dei risultati di apprendimento);
- più i bambini si divertono, più acquisiranno effettivamente conoscenze sui temi di studio (esperienza di divertimento e conoscenze acquisite sugli argomenti del gioco).

Lo studio esplorativo ha coinvolto 74 ragazzi di età compresa fra i 10 e i 13 anni scelti in maniera casuale in 5 classi in una scuola di primo ciclo in Svizzera. Gli studenti hanno usato in tre lezioni un gioco web-based, AWWWARE, il cui scopo è quello di promuovere la media competency fra i bambini, sviluppando le loro abilità critiche nella navigazione fra le pagine di internet. Insieme ai dati anagrafici degli studenti partecipanti all'indagine, sono stati raccolti dati attraverso alcuni test composti da item con risposte su scala Likert a 5 livelli da:

- pre-test e post-test che indagavano le opinioni dei ragazzi su utilità del gioco, semplicità, livelli di divertimento, desiderabilità sociale, abilità personali nell'uso del gioco, paura di usare il gioco e sbagliare, intenzione di usarlo in futuro;
- test sul tema dell'information literacy in cui ai bambini è stato chiesto di valutare la qualità di 9 pagine internet;
- test di autovalutazione del guadagno di apprendimento in termini di motivazione e conoscenze;
- test di valutazione del gioco AWWWARE in termini di chiarezza di obiettivi, approcci strategici, uso delle conoscenze pregresse su Internet, flusso di gioco, assistenza.

Per verificare le ipotesi, gli autori hanno proposto 4 modelli di regressione lineare variando di volta in volta le variabili considerate.

Nel dettaglio:

H1: variabile dipendente = intenzione d'uso; predittori = variabili rilevate nel pre-test sulle opinioni dei bambini

H2: variabile dipendente = autovalutazione del guadagno di apprendimento in termini di motivazione; predittori = variabili rilevate nella valutazione del gioco

H3: variabile dipendente = autovalutazione del guadagno di apprendimento in termini di conoscenze; predittori = variabili rilevate nella valutazione del gioco

H4: variabile dipendente = punteggi nelle prove di valutazione; predittori = variabili rilevate nella valutazione del gioco

Per ciascun modello gli autori riportano i valori su bontà e significatività e le tabelle riassuntive sui valori, lo standard error e la significatività dei coefficienti di regressione parziale e dell'intercetta.

I primi tre modelli sono significativi e presentano un R^2 rispettivamente di 0,56, 0,63, 0,74. L'ultimo dei 4 modelli non è significativo, spiega solo il 20% dei casi [$R^2 = 0{,}20$; $n=74$, $F\,(9{,}64) = 1{,}80$, *n.s.*].

Leggendo i valori dei coefficienti e la loro significatività, nell'articolo si giunge alla conclusione che soltanto la seconda ipotesi può essere accettata. Ad incidere sulle motivazioni d'uso dei serious games risultano piuttosto variabili come le aspettative relative all'utilità e alla semplicità, le conoscenze pregresse sui temi, i feedback.

Al centro del secondo studio che descriviamo è il tema della *digital citizenship* degli insegnanti analizzata nella percezione di sé stessi degli stessi docenti e motivata dalle azioni didattiche che contengono fenomeni di cyberbullismo e spingono gli studenti a coinvolgersi nella vita politica. Lo studio di Moonsun Choi, Dean Cristol e Belinda Gimbert (2018) indaga il modo in cui le caratteristiche personali e psicologiche e l'uso di internet possono influenzare i livelli di digital citizenship dei docenti. Il campione nell'analisi è composto da 348 insegnanti in formazione in un corso blended finalizzato a formare docenti da integrare in distretti scolastici "difficili".

Come strumenti per la raccolta dei dati gli autori utilizzano la Digital Citizenship Scale (DCS) da loro precedentemente formulata e validata. Si tratta di una

scala di autovalutazione composta da 26 item con risposte in una scala Likert a 7 livelli riferita a 5 ambiti: *Technical Skills*, *Local/Global Awareness*, *Networking Agency*, *Internet Political Activism* e *Critical Perspective*.

I docenti hanno risposto, inoltre, a domande su dati generali ed esperienze lavorative, sull'uso di internet a livello personale e nelle attività scolastiche e infine su *self-efficacy* e ansia nell'uso di internet attraverso l'uso di due scale riprese da altri studi (*Internet self-efficacy scale* e *State-Trait Anxiety Inventory*). Questo gruppo di dati ha costituito il set di variabili indipendenti che sono state gradualmente aggiunte nei tre modelli di regressione lineare attraverso i quali gli autori conducono l'analisi sulla relazione fra digital citizenship e caratteristiche dei docenti.

Nel primo modello sono usate come variabili indipendenti soltanto quelle relative ai dati anagrafici e alle esperienze professionali; nel secondo sono aggiunte quelle relative all'uso di internet; il terzo comprende anche i risultati dei test su self-efficacy e livelli d'ansia. In ciascun modello troviamo 5 regressioni: per ciascuna la variabile dipendente è rappresentata da una delle cinque dimensioni valutate dalla DCS.

Confrontare modelli costruiti con variabili indipendenti diverse è una modalità frequente nell'uso della regressione lineare. I modelli gerarchici, come in questo caso, prevedono una introduzione graduale di gruppi di variabili per valutare le differenze nella bontà e significatività dei modelli composti da diversi set di variabili. In questi casi la costruzione dei modelli non ha tanto rilevanza in termini di predizione, quanto di confronto fra soluzioni diverse per trovare quella che meglio riesce a fornire valori più completi e significativi delle relazioni individuate. L'indicatore usato in questo caso per confrontare fra di loro i modelli è R^2 che assume valori più alti per tutte le dimensioni della digital citizenship nel terzo modello ossia in quello che raccoglie tutte le variabili indipendenti rilevate.

Dallo studio risulta che predittori significativi della cittadinanza digitale per gli insegnanti del campione sono gli anni di esperienza lavorativa, l'uso degli social network per l'insegnamento e la percezione di efficacia nell'uso di Internet.

L'ultimo caso che analizziamo è l'articolo "A multivariate approach to predicting student outcomes in web-enabled blended learning courses" di Nick Z. Zacharis (2015) il cui obiettivo è quello di definire un modello per predire gli studenti a rischio di ottenere scarsi risultati nei corsi erogati in modalità blended. L'autore identifica alcuni elementi chiave per l'analisi di cui alcuni stretta-

mente legati al blended, altri al tipo di ricerca condotta. Come afferma, i ricercatori devono scegliere le attività (e di conseguenza le variabili) che possono fornire indicazioni sugli studenti che rischiano il fallimento; molto in queste stime dipende o va interpretato in base al design del corso; i docenti dei corsi blended dovrebbero dedicare una parte delle loro attività didattiche nell'analisi dei dati registrati nelle piattaforme utilizzate (LMS) per le quali nuove competenze sono necessarie. I dati raccolti riguardano 134 matricole in un corso erogato in modalità blended di "Introduction to Computer Science using Java" e sono stati estratti dai log dell'LMS che ospita i corsi (Moodle). Oltre a variabili sul tempo trascorso in piattaforma e sul numero di visualizzazioni delle risorse, di completamento e controllo dei quiz, delle sessioni di connessione, l'autore individua specifiche variabili indipendenti fra cui: *Repo Messages*, come somma di tutte le interazioni con i messaggi prodotti o soltanto letti, importante poiché il corso richiede alti livelli di interattività; *Content Creation Contribution*, che misura la partecipazione degli studenti nella costruzione di wiki e blog.

La variabile dipendente è rappresentata in questo caso dalla performance degli studenti intesa come voto finale del corso.

Gli autori conducono innanzitutto un'analisi sulle singole variabili (29) calcolando per ognuna la correlazione R con la variabile indipendente e il coefficiente di determinazione R^2. In seguito, i 14 predittori significativamente associati con la variabile indipendente sono stati inseriti in una regressione multivariata stepwise dalla quale è risultato, a partire dal valore di R^2, che 4 variabili spiegano il 52% della varianza della y ($AdjR^2 = 0{,}505$). La lettura e la scrittura dei post pare influenzare da sola il 37% della varianza.

Lo studio si conclude con una regressione logistica finalizzata a verificare quanti studenti del campione possono essere considerati a rischio. Della regressione logistica, però, parleremo nel prossimo capitolo.

CAPITOLO 5

REGRESSIONE LOGISTICA

Al termine del capitolo, il lettore sarà in grado di:

- *descrivere le procedure alla base della regressione logistica;*
- *spiegare gli indicatori descrittivi dei modelli di regressione logistica;*
- *elencare esempi della ricerca educativa nei quali è stata utilizzata la regressione logistica.*

5.1 - Regressione logistica

Anche in questo capitolo come nel precedente, parliamo di una forma di regressione destinata alla definizione di un modello nella quale studiamo una funzione, definita "logistica" appunto, che mette in relazione le variabili dipendenti y e indipendenti x_i e calcoliamo i coefficienti b_i corrispondenti ad ogni variabile indipendente x_i. Differentemente dalla regressione lineare, nella regressione logistica la variabile dipendente y è una variabile dicotomica che, come ben sappiamo, può assumere solamente due valori: sì/no; uomo/donna; a rischio/non a rischio; efficace/non efficace, esame superato/non superato. Per via di questa caratteristica, le procedure di analisi della tecnica portano anche a classificare e raggruppare le osservazioni del campione o della popolazione osservata in due gruppi corrispondenti alle due modalità assunte dalla y.

Qualora volessimo verificare la relazione che esiste fra i risultati conseguiti nello svolgimento di alcune prove intermedie in un corso e l'esito finale dello stesso, ad esempio, i due gruppi che potremo distinguere e classificare sono quello degli studenti che superano l'esame e quello degli studenti che non lo passano.

Da un punto di vista grafico, in un'analisi bivariata dove consideriamo una sola variabile indipendente (il voto di una delle prove intermedie), ci aspettiamo che le osservazioni si distribuiscano su due rette parallele corrispondenti ai due gruppi come nella Figura 5.1. Alle modalità assunte dalla variabile y sono attribuiti i valori 1 e 0. Solitamente il valore 1 è attribuito alla modalità che indi-

ca il successo o il manifestarsi di un evento (nel nostro esempio, superamento dell'esame) e il valore 0 l'insuccesso o l'assenza dell'evento (ancora nell'esempio, bocciatura).

Dallo scatterplot possiamo dedurre che chi ha conseguito un punteggio più basso in un'ipotetica prova intermedia si colloca nel gruppo che identifichiamo con la modalità 0 e ha meno probabilità di superare l'esame finale (punti in basso). Al contrario, gli studenti che hanno conseguito punteggi più alti saranno collocati nel gruppo contraddistinto dal numero 1 e in generale avranno più probabilità di successo (punti in alto).

Il superamento degli esami per le osservazioni comprese nella figura fra i voti tra 23 e 25 assume talvolta il valore 1, altre lo 0. Si parla per queste unità di errori di classificazione (*misclassifications*), in quanto non è immediato definire il gruppo al quale esse appartengono.

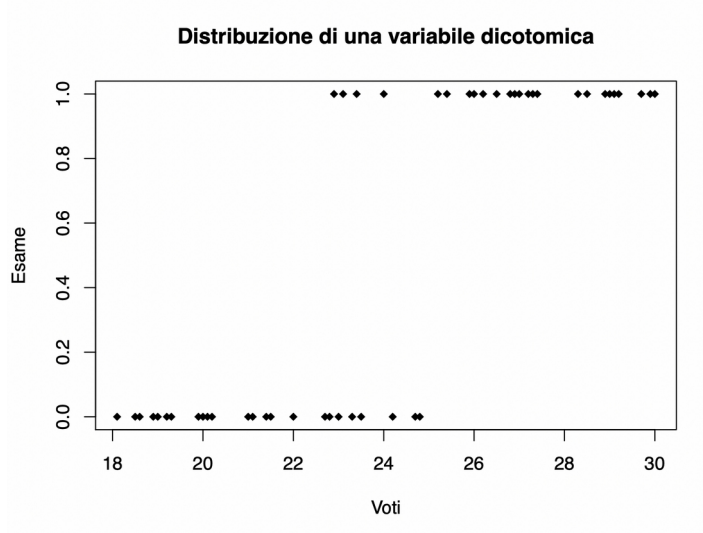


Figura 5.1 – Scatterplot delle variabili VOTI con range da 18 a 30 ed ESAME, variabile dicotomica dove 1 = promozione, 0 = bocciatura.

Quale curva interpola osservazioni così disposte? Osservando la Figura 5.2 possiamo escludere che la retta faccia al caso nostro, la funzione che cerchiamo non può essere lineare.

La curva che meglio interpola questi dati è la curva logistica che ha una caratteristica forma ad S e contiene i valori della *y* fra 0 e 1 senza mai superarli (Figura 5.3).

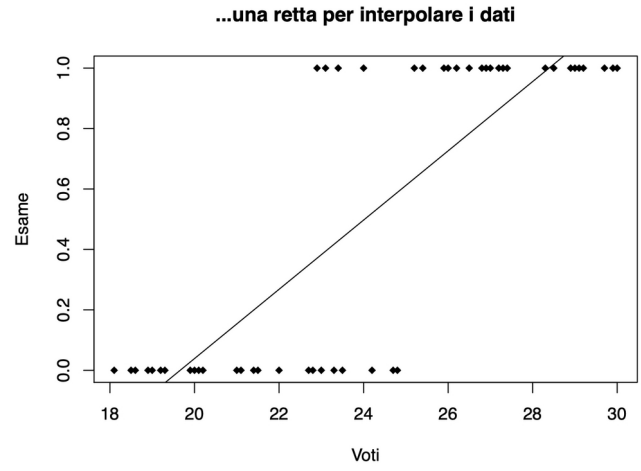


Figura 5.2 - Può una retta interpolare le osservazioni di una variabile binomiale?

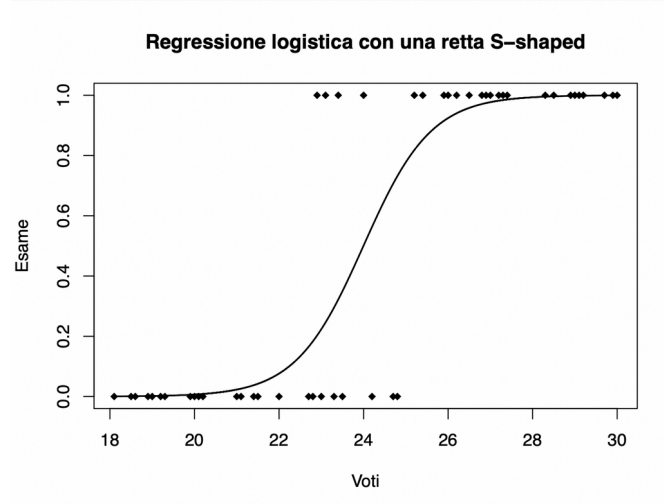


Figura 5.3 - La curva logistica.

Notiamo che:

- attribuiamo alla variabile dipendente *y* i valori di 0 e 1 a seconda che l'evento si verifichi o no;
- la probabilità assume tutti i valori compresi fra 0 e 1;
- i valori della *y* nella curva logistica oscillano fra 0 e 1.

Queste tre condizioni che si sovrappongono ci permettono di lavorare con le variabili binomiali e con le probabilità usando la funzione logistica che ci restituisce la probabilità che un dato evento si manifesti. Per esempio, in Figura 5.4, seguendo l'andamento della curva logistica, vediamo che nel modello costruito gli studenti che conseguono un punteggio pari a 26 avranno il 92% di probabilità di superare l'esame finale del corso ($y = 0,92$).

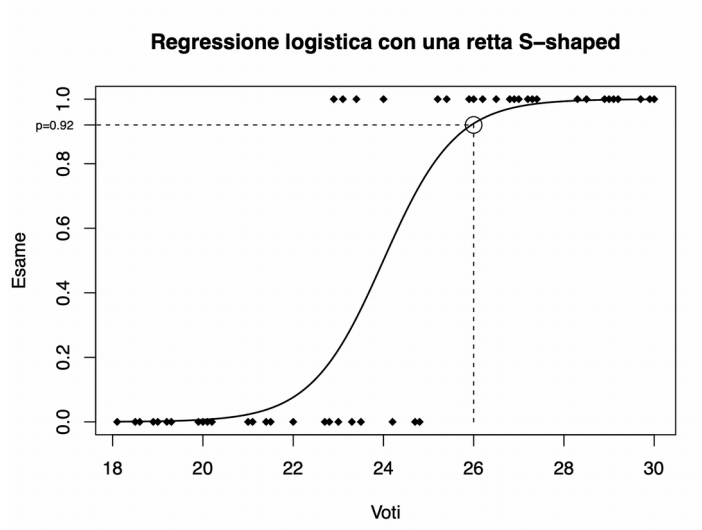


Figura 5.4 - Calcolo della probabilità che si verifichi un evento sulla curva logistica.

L'esempio descritto finora fa riferimento al campo bivariato, aggiungendo variabili indipendenti possiamo riportarlo al caso multivariato.

L'equazione che descrive la curva logistica in ambito multivariato e che possiamo utilizzare per calcolare il valore di p, la probabilità che un dato evento si manifesti (probabilità di successo, $y = 1$) fissati i valori delle x_i, è la seguente

(5.1)

$$p = \frac{e^{(a + bx_i)}}{1 + e^{(a + bx_i)}}$$

Attraverso pochi passaggi algebrici, otteniamo due formulazioni equivalenti dell'equazione sopra riportata.

(5.2)

$$Logit = \ln\left(\frac{p}{1-p}\right) = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_i x_i$$

(5.3)

$$Odds = \frac{p}{1-p} = e^{(a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_i x_i)}$$

Le formulazioni tramite logaritmi (5.2) ed esponenziali (5.3) garantiscono che p sia sempre positiva (l'esponenziale di e è una funzione sempre positiva e l'argomento del logaritmo deve essere positivo affinché possa essere definito).

L'equazione (5.2) è espressione di una funzione con andamento lineare ed è il modello di regressione logistica, detto anche modello *logit*. Il logit non è altro che il logaritmo naturale degli *odds* (*log-odds*) ossia il rapporto fra la probabilità di successo (p) e insuccesso ($1 - p$) di un evento (nel nostro caso, $y = 1$, superamento dell'esame). La trasformazione logit di un parametro dicotomico rappresenta un modello lineare (attenzione: non è la p ad essere considerata come variabile dipendente in un modello lineare definito dall'equazione 5.2 ma l'intero log-odds) ed è il più rappresentativo caso di *Generalized Linear Models* (Agresti, 2013).

Entrambe le formulazioni ci permettono di fare alcune osservazioni utili nel calcolo e nell'interpretazione dei modelli di regressione logistica.

LOGIT. Il logaritmo degli odds, il primo membro dell'equazione (5.2), non ha valori limite superiori o inferiori. Se il logit assume valori positivi, la probabilità che un evento si manifesti ($y = 1 \mid x_i$) è superiore al 50% e dunque è un caso di successo; qualora assuma valori negativi, la probabilità che un dato evento si manifesti ($y = 1 \mid x_i$) è inferiore al 50% (insuccesso).

Risulta infatti che:

$$Logit = 0 \rightarrow p = 0{,}5$$
$$Logit > 0 \rightarrow p > 0{,}5$$
$$Logit < 0 \rightarrow p < 0{,}5$$

ODDS. I valori degli odds, il primo membro dell'equazione (5.3), sono compresi fra 0 (per $p = 0$) e infinito (per $p = 1$). In questo caso, se gli odds assumo-

no valori superiori ad 1, la probabilità che un evento si manifesti ($y = 1 \mid x_i$) è superiore al 50% (successo), qualora assuma valori inferiori a 1, la probabilità che un dato evento si manifesti ($y = 1 \mid x_i$) è inferiore al 50% (insuccesso).

Risulta infatti che:

$$Odds = 1 \rightarrow p = 0{,}5$$
$$Odds > 1 \rightarrow p > 0{,}5$$
$$Odds < 1 \rightarrow p < 0{,}5$$

Nella formulazione degli odds il valore critico è rappresentato da 1, valore che distingue probabilità di successo o insuccesso e permette di classificare in due gruppi le osservazioni con y maggiore o minore di 1. Per la formulazione logit, il valore critico è lo 0: in questo caso valori positivi e negativi distinguono gruppi e probabilità di successo e insuccesso (solo per completezza facciamo notare che i due punti coincidono poiché $ln1 = 0$). Se il logit tende a 0 oppure gli odds a 1, non c'è dipendenza fra le variabili osservate.

Prima di esaminare i risultati ottenuti da una regressione logistica, concludiamo questa introduzione rispondendo a due domande che finora non abbiamo considerato nella discussione.

1 - Quali assunti bisogna verificare sul dataset per applicare la regressione logistica?

La regressione logistica viene spesso scelta fra le altre tecniche perché ha meno assunti da soddisfare. Sappiamo che la variabile risposta non è normale perché la y (e il suo errore) seguono la distribuzione binomiale nella quale la varianza non è costante. Di conseguenza né gli assunti sulla normalità, né quelli sull'omoschedasticità vanno verificati.

Vanno tenuti in debito conto gli assunti relativi all'assenza di collinearità fra le variabili e all'indipendenza dei casi osservati.

Altro fattore da controllare riguarda le dimensioni del campione che deve prevedere non meno di 10 osservazioni per parametro osservato.

2 - Come calcoliamo i valori b_i ?

Per la natura non lineare della relazione e per l'eteroschedasticità della variabile dicotomica, non possiamo utilizzare il metodo dei minimi quadrati come per la regressione lineare. I valori dei coefficienti vengono selezionati in base al metodo della massima verosimiglianza. La funzione di verosimiglianza (*likeli-*

hood) identifica la probabilità che un evento si manifesti con una certa probabilità. Nella selezione dei coefficienti scegliamo quelli che rendono massima la verosimiglianza e cioè che rendono massima la probabilità che in un nuovo campione gli eventi assumano i valori di probabilità già osservati. Non è la probabilità che un evento si verifichi ad essere massima, ma è massima la probabilità che un dato evento assuma una certa probabilità di verificarsi.

Nell'esempio in Figura 5.4, la massima verosimiglianza è rappresentata dalla massima probabilità che nel modello che costruiamo o in altri campioni gli studenti che conseguono il punteggio di 26 in una prova intermedia abbiano la probabilità del 92% di essere promessi nell'esame finale.

L'algoritmo per calcolare la massima verosimiglianza permette di ricalcolare i parametri più volte fino a quando l'errore nel calcolo risulta accettabile.

5.1.1 - Output di una regressione logistica

Come sempre (e per fortuna!), i software di analisi statistica ci aiutano nello svolgere i calcoli e valutare la bontà dei modelli individuati. In Figura 5.5 si osservano i risultati restituiti da `R` utilizzando la funzione `glm` per definire un modello di regressione logistica. I dati riportati fanno riferimento all'esempio che stiamo utilizzando in questa discussione per il quale abbiamo creato un dataset ad hoc e per il quale, adesso che abbiamo familiarizzato con i concetti alla base della regressione logistica, consideriamo tre variabili indipendenti corrispondenti ai risultati degli studenti in tre diverse prove di valutazione intermedia.

Dall'uso della funzione `summary`, le prime indicazioni che otteniamo sono quelle relative alla formula usata nella creazione del modello, con le variabili considerate, il tipo (`binomial`) e il dataset.

Saltano immediatamente all'occhio poi il valore dell'intercetta a – non di particolare interesse nella regressione logistica – e i coefficienti di regressione b_i che, in maniera molto simile alla regressione lineare, rappresentano la quantità in più con cui ciascuna variabile indipendente contribuisce al calcolo della probabilità quando la variabile considerata aumentata di un'unità e tutte le altre restano stabili. Il test solitamente usato per verificare la significatività dei coefficienti non è il t-test come per la regressione lineare ma il test di Wald che parte dal calcolo dello z value.

```
Call:
glm(formula = Esame ~ Voti1 + Voti2 + Voti3, family = binomial,
    data = voti.esame.6)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.57267  -0.15086   0.02495   0.25065   1.96087

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -31.7140    10.2578  -3.092  0.00199 **
Voti1         4.1079     3.2603   1.260  0.20767
Voti2        -2.9577     3.1731  -0.932  0.35127
Voti3         0.1679     0.3611   0.465  0.64191
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 67.908  on 48  degrees of freedom
Residual deviance: 20.132  on 45  degrees of freedom
AIC: 28.132

Number of Fisher Scoring iterations: 7
```

Figura 5.5 - Output di una regressione logistica in R realizzata usando la funzione `glm` (libreria: `stats`).

Noti i coefficienti b_i, possiamo calcolare dalla formula (5.1) la probabilità che un dato evento si realizzi considerati i valori delle variabili utilizzate ($y = 1 \mid x_i$).

Dalla formula che segue (5.4), noti i risultati di uno studente nelle tre prove di valutazione intermedie nel nostro esempio, potremo calcolare la probabilità che, nel modello di regressione logistica costruito, lo stesso studente superi l'esame finale.

(5.4)

$$p_{y=1} = \frac{e^{(-31.7 + 4{,}1 \cdot Voto1 - 3{,}0 \cdot Voto2 + 0{,}2 \cdot Voto3)}}{1 + e^{(-31.7 + 4{,}1 \cdot Voto1 - 3{,}0 \cdot Voto2 + 0{,}2 \cdot Voto3)}}$$

Calcolare le probabilità ci permette anche di verificare quanto il modello rispecchi la realtà osservata: quanti casi di bocciatura predetti dal modello corrispondono a quelli reali? Quanti casi di promozione?

Questo tipo di calcolo può essere formalizzato per validare una regressione logistica (così come per altre tecniche di analisi) con la costruzione di una "ma-

trice di confusione" nella quale si confrontano i valori osservati e i valori predetti per verificare quanto il modello sia accurato e in grado di predire in maniera corretta il verificarsi di un evento.

Nella Tabella 5.1 a doppia entrata, poniamo nelle colonne i valori di superamento e bocciatura all'esame finale predetti tramite la formula (5.4) e nelle righe i valori osservati. Il nostro modello ha classificato correttamente 43 osservazioni del campione costituito da 49 unità (89%). Infatti dei 25 studenti che nel mondo reale hanno superato l'esame, per 21 usando la formula (5.4) abbiamo riportato una probabilità di superare l'esame superiore al 50% (successo). Allo stesso modo, la formula (4) ci ha permesso di individuare in maniera corretta 22 dei 24 studenti che sono stati bocciati.

		Valori predetti		
		Superamento esame	**Bocciatura**	**Totale**
Valori osservati	**Superamento esame**	21	4	25
	Bocciatura	2	22	24
Totale		23	26	49

Tabella 5.1 – Matrice di confusione nella quale confrontiamo valori osservati e stimati.

Ritornando alla Figura 5.5 dopo questa breve regressione sulle matrici di confusione, poniamo attenzione agli altri parametri da valutare:

- i parametri relativi alla devianza che ci forniscono informazioni sull'adeguatezza del modello. Ne troviamo tre: i residui della devianza e i valori che sintetizzano la devianza nel modello nullo e nel modello residuale. I residui ci dicono quanto ogni osservazione contribuisce ad incrementare la devianza e sono calcolati come differenza fra il modello di verosimiglianza e i valori stimati. Come per i residui della regressione lineare, si presume che la distribuzione dei residui sia simmetrica ed è per questo che fra i valori riportati troviamo i quartili della distribuzione (*Deviance Residuals*); *Null Deviance* è l'indicatore relativo al modello nullo che è calcolato usando l'intercetta come unico parametro (si vedano anche i gradi di libertà indicati), rappresenta quindi un modello senza predittori; *Resi-*

dual Deviance è l'indicatore calcolato per un modello con tutte le variabili dello studio.
Se la *Null Deviance* e la *Residual Deviance* assumono valori simili, il modello costruito con tutte le variabili perde di significatività perché l'inserimento delle variabili indipendenti non comporta differenze rispetto al modello senza predittori;

- *AIC* (*Aikake Information Criterion*), indice che abbiamo incontrato anche nella regressione lineare e che usa la funzione di verosimiglianza per stimare la bontà di un modello alla luce dei parametri utilizzati nell'analisi. Si usa per fare comparazioni fra modelli. Il modello da preferire nel confronto è quello che presenta il valore di AIC più basso. Il calcolo dell'AIC penalizza modelli molto complessi che usano tante variabili dipendenti, talvolta non utili;

- il parametro di dispersione per le variabili binomiali, fissato in automatico ad 1. Esso indica la variabilità dei dati e quanto distano dai valori centrali;

- *Fisher scoring iterations*, numero di iterazioni necessarie nel metodo della massima verosimiglianza per interpolare i dati. Quante più iterazioni sono necessarie per definire un modello, tanto più esso risulta debole.

Per verificare la bontà di un modello, inoltre, viene calcolato l'indicatore $PseudoR^2$ di cui esistono più versioni e formulazioni. Per quanto riguarda l'interpretazione, esso è sovrapponibile per grandi linee al valore R^2 della regressione lineare. Ha valori fra 0 e 1 e confronta il modello nullo con quello proposto.

È frequente, comunque – e con questo chiudiamo la rapida carrellata sugli indicatori legati ai modelli di regressione logistica–, che negli articoli scientifici compaiano pochi dei parametri riportati finora. I valori che più spesso sono riportati e che permettono di giungere a conclusioni e verificare ipotesi sono i valori dell'odds ratio.

Con alcuni passaggi algebrici, si può dimostrare che e^b, dove b è uno dei coefficienti di regressione, corrisponde all'odds ratio, il rapporto fra gli odds calcolati sulle probabilità di un evento rispetto a due categorie (nella formula che segue donna/uomo).

(5.5)

$$e^b = \frac{(p\ di\ superare\ l'esame\ per \leq donne)}{(p\ di\ non\ superare\ l'esame\ per \leq donne)} : \frac{(p\ di\ superare\ l'esame\ per\ gli\ uomini)}{(p\ di\ non\ superare\ l'esame\ per\ gli\ uomini)}$$

Dalla formula

(5.6)

$$(e^b - 1) \cdot 100$$

calcoliamo la differenza fra le probabilità di successo fra le due categorie scegliendone una come quella di riferimento, nel nostro caso il genere femminile.

Poniamo che per esempio $e^b = 1{,}25$. Usando la formula, diremo che le donne hanno una propensione a superare gli esami rispetto agli uomini maggiore del 25%.

Gli odds ratio non permettono un confronto di percentuali fra le variabili indipendenti ma ci permettono di fare riflessioni solo sulla relazione del singolo predittore con la variabile risposta.

Insieme agli odds ratio bisognerebbe riportare anche gli intervalli di confidenza al 95% che ci forniscono informazioni sull'incertezza della stima. Se questi comprendono il valore 1, è possibile che le modalità confrontate (uomo/donna) non incidano sulla variabile indipendente e di conseguenza ci sia indipendenza fra variabile dipendente e indipendente.

5.2 - L'uso della regressione logistica multipla nella ricerca educativa

Quali sono le caratteristiche degli studenti che usano *mobile LMS* per partecipare alle attività didattiche nel contesto universitario? Quali quelle di chi non li usa?

Chi sono gli studenti che si iscrivono a corsi universitari online? Chi preferisce quelli tradizionali?

Due studi che descriviamo di seguito rispondono a queste domande rispettivamente nel contesto coreano e statunitense.

L'uso della regressione logistica in entrambi permette di distinguere gli studenti in due gruppi, quelli che usano mobile LMS e quelli che non li usano, così come quelli che si iscrivono ai corsi online e quelli che optano per i corsi tradizionali.

Ricerche di questo tipo permettono di progettare e realizzare interventi didattici più appropriati alle modalità e abitudini di studio degli studenti utilizzando gli strumenti digitali o tradizionali in base alle necessità. Allo stesso modo, procurano informazioni alle istituzioni formative sulle lacune da colmare nei servizi predisposti e sulle opportunità per potenziare i percorsi di apprendimento offerti e renderli maggiormente funzionali rispetto alle necessità degli studenti, allargando il bacino di iscritti nella formazione e scegliendo i canali e le modalità formative adeguate in maniera critica e guidata dai dati e dalla ricerca.

Il diffuso uso dei sistemi mobile e l'impegno degli atenei nel fornire contenuti anche attraverso questi canali sono il punto di partenza della ricerca di Insook Han e Won Sug Shin (2016) finalizzata all'individuazione di una possibile relazione fra background individuale, fattori psicologici ed esterni nell'adozione di mobile LMS da parte degli universitari. A 2000 studenti che hanno frequentato una delle 10 edizioni di un corso di "Introduzione al Cyber Learning" in una università online della Corea del sud è stato somministrato un questionario suddiviso in due parti, la prima dedicata a raccogliere dati sulle caratteristiche personali (età, genere, stato occupazionale, uso del mobile LMS) e la seconda con domande di carattere psicologico (self-efficacy, disponibilità all'innovazione, attitudine nell'uso delle tecnologie, utilità percepita) o di contesto (pressione sociale nell'uso del mobile e livello percepito di accessibilità dei sistemi mobile) in una scala Likert a 5 livelli. Usando la regressione logistica, gli autori hanno costruito tre modelli: il primo comprende variabili che hanno a che fare solo con i dati personali, il secondo aggiunge anche variabili relative alle caratteristiche psicologiche e il terzo comprende tutte le variabili rilevate. Quest'ultimo è il modello che ha un valore di $PseudoR^2$ più alto (0,145). Nelle tabelle descrittive dei dati, gli autori riportano il valore dei coefficienti b_i, dell'errore standard e dell'odds ratio che, come abbiamo detto, è il valore che ci permette di leggere con più semplicità i risultati e confrontare gli effetti di uno stesso regressore.

L'articolo propone anche una seconda domanda di ricerca finalizzata a stabilire la relazione fra uso del mobile e risultati accademici sullo stesso campione per la cui risposta si procede nell'analisi con l'uso della regressione lineare.

I predittori significativi dell'uso di LMS mobile tra gli studenti intervistati risultano essere da questo studio la self-efficacy, l'innovatività, la facilità d'uso e l'utilità percepite tra i fattori psicologici e la pressione sociale tra i fattori esterni. I risultati permettono sia di comprendere le scelte degli studenti nell'uso dei mobile LMS, sia di ipotizzare percorsi di supporto e formazione sui vantaggi di questi strumenti a determinate categorie di studenti.

Spostandoci nel contesto statunitense, lo studio "From the periphery to prominence: an examination of the changing profile of online students in American higher education" (Ortagus, 2016) analizza le caratteristiche e le iscrizioni a corsi online su un ampio campione formato da circa 240.000 studenti registrati dal "National Postsecondary Student Aid Study" (NPSAS) negli anni 2000, 2004, 2008, 2012. La logica utilizzata è sovrapponibile a quella dello studio precedente: quali studenti del campione appartengono al gruppo degli iscritti a corsi universitari online? Quali appartengono al gruppo degli iscritti a corsi tradizionali? Quali sono le caratteristiche degli studenti appartenenti ad entrambi gruppi? È possibile rilevare una qualche relazione fra le caratteristiche degli studenti e la loro scelta di iscriversi o meno ai corsi online?

Oltre a usare strumenti di statistica descrittiva per descrivere le distribuzioni delle variabili identificate, l'autore usa la regressione logistica di cui riporta come risultati gli odds ratio con gli standard error. La variabile dipendente è quindi la partecipazione ai corsi online espressa da due modalità: la prima "alcuni insegnamenti online" (1-99%) e la seconda "tutti gli insegnamenti online" (100%). Le tre ipotesi formulate intendono verificare una correlazione positiva fra iscrizione a corsi interamente online e:

1. vincoli di tempo e spazio. Tradotto in variabili: stato occupazionale e civile, genere, genitorialità, età, servizio prestato nell'esercito;

2. caratteristiche sotto-rappresentate nel tempo nell'istruzione superiore. Tradotto in variabili: appartenere a una minoranza, bassa fascia di reddito, prima generazione di studenti in famiglia;

3. caratteristiche delle istituzioni. Tradotto in variabili: dimensioni e tipologia (pubblica o privata) dell'istituzione, anni di durata del percorso formativo.

Le differenze nella scelta di iscriversi a corsi in modalità totalmente online possono essere spiegate secondo l'autore anche dal concetto della microeconomia di costo opportunità, mostrando la possibilità di iscriversi a corsi online come la migliore soluzione se confrontata in termini di costi per soggetti che in caso contrario non proseguirebbero gli studi perché costretti a lasciare il lavoro, per i numerosi impegni familiari o altro.

Un tema completamente diverso da quelli delle precedenti ricerche è affrontato ugualmente attraverso la tecnica della regressione logistica nell'articolo di Paul L. Morgan e colleghi (2017) che si pone nel dibattito sulla sovra- o sotto- rappresentazione degli studenti appartenenti a particolari gruppi etnici fra quelli certificati per una qualche disabilità nelle scuole statunitensi. Dall'analisi della letteratura che gli autori ci propongono, tanti sono gli studi sul tema di interesse anche governativo, particolarmente delicato a livello sociale e sanitario. Si tratta di un fenomeno che deve tenere in conto aspetti diversi, quali i risultati scolastici e le difficoltà collegate al raggiungimento degli stessi, pregiudizi, disuguaglianze legate alle scuole frequentate (con alta concentrazione di particolari etnie e con scarse risorse) o alle condizioni familiari. L'articolo ci mostra come l'analisi dei dati e la regressione logistica possono essere utili strumenti per leggere problemi di questo tipo. Gli autori utilizzano i numerosissimi dati provenienti dall'indagine "National Assessment of Educational Progress" (NAEP) somministrata dal 1969 dal National Center for Education Statistics, entità federale delegata dal Congresso a raccogliere e analizzare dati relativi ai livelli di istruzione e ai risultati accademici negli Stati Uniti e in altre nazioni. L'indagine NAEP raccoglie dati sull'andamento scolastico degli studenti statunitensi di alcuni gradi (quarto, ottavo e dodicesimo) in molti ambiti fra cui quelli utilizzati nello studio qui presentato, ossia le abilità di lettura e di calcolo. I test somministrati hanno alti livelli di affidabilità e i metodi di campionamento sono predisposti in modo da raccogliere informazioni da ogni giurisdizione.

Le dimensioni del campione analizzato sono considerevoli: 183.570 studenti di quarto grado (anno 2013), 165.540 dell'ottavo (anno 2013) e 48.560 del dodicesimo (anno 2009).

Le variabili riguardano:

- lo stato di disabilità corrispondente a: la predisposizione per lo studente di un programma individualizzato di educazione (Individualized Education Program – IEP) indispensabile nella scuola per ricevere servizi di special education; la disabilità per la quale il bambino è segnalato (disturbi

emotivi, autismo, difficoltà nel linguaggio, ADHD, disabilità cognitive ecc.);

- l'etnia di appartenenza (bianca, nera, ispanica, asiatica, indiani d'America, nativi del Pacifico, *multiple races*);
- i risultati ottenuti nei test sull'abilità di lettura (e calcolo) del NAEP;
- genere, reddito familiare (misurato come accesso gratuito ai servizi per la mensa), condizione di ELL (English Language Learning) in riferimento all'apprendimento dello studente della lingua inglese.

Per ogni grado scolastico, gli autori descrivono 4 modelli di regressione logistica con diversi gruppi di variabili indipendenti: nel primo solo l'etnia, nel secondo genere e risultati NAEP, nel terzo genere e variabili aggiuntive per valutare unicamente le influenze sociali e demografiche; nel quarto tutte le variabili (con un intervento in più per ridurre gli effetti dell'appartenenza a una scuola piuttosto che a un'altra). Calcolano inoltre i modelli per le singole disabilità e ripropongono il secondo modello anche per anni diversi sui dati disponibili.

Ancora una volta, i risultati delle regressioni logistiche che ci vengono restituiti sono espressi nella formulazione degli odds ratio.

Solo a scopo esemplificativo vediamo alcuni dati rilevati da una delle tabelle nello studio (Figura 5.5). Il primo modello evidenzia che i bambini che rientrano nell'etnia degli Indiani d'America hanno maggiori probabilità di rientrare fra i bambini speciali rispetto ai bambini bianchi. Il valore dell'odds ratio è infatti pari a 1,27. Usando la formula (5.7) otteniamo che 27% è il valore di incremento.

(5.7)

$$(e^{b}-1) \cdot 100 = (1{,}27-1) \cdot 100 = 27\,\%$$

I risultati cambiano negli altri modelli nei quali sono inseriti fra le variabili indipendenti i risultati NAEP. Guardando ai coefficienti significativi (Tabella 5.2), notiamo ad esempio che i bambini neri o ispanici nel secondo modello hanno una probabilità rispettivamente inferiore del 57% e 63% rispetto ai bambini bianchi di essere inseriti fra i disabili.

I valori del terzo modello non si discostano dal primo quanto quelli del secondo o del quarto, cosa che può giustificare studi precedenti, ci dicono gli

autori, che non hanno considerato nel fenomeno gli aspetti prettamente scolastici e legati all'apprendimento.

Il quarto modello pare essere il più rigoroso e indica delle disparità fra le etnie ma anche fra gli studenti di sesso femminile (30%), a basso reddito (25%) e ELL (67%) che rientrano fra quelli che hanno meno probabilità di essere identificati come disabili.

	IEP			
	Model 1	**Model 2**	**Model 3**	**Model 4**
Black	1.03	0.43***	0.84***	0.44***
Hispanic	0.91**	0.37***	0.75***	0.51***
Asian	0.47***	0.47***	0.47***	0.64***
American Indian/Alaskan Native	1.27***	0.58***	1.09	0.62***
Native American/Pacific Islander	1.23+	0.42***	1.07	0.46***
Multiple race/ethnicity	0.82**	0.77***	0.77***	0.80***
Reading composite score		0.27***	-	0.23***
Male			1.99***	1.70***
Free or reduced lunch eligible			1.64***	0.75***
ELL			1.06	0.33***
School fixed effects				X
+p < .10. *p < .05. **p < .01. ***p < .001.				

Tabella 5.2 – Odds Ratio nella regressione logistica in uno studio sul riconoscimento delle disabilità per gli studenti del quarto grado di etnie diverse, N = 183.570 (Morgan et al., 2017, p. 310).

Analizzando un campione molto ampio, con un numero ragionevole di variabili, lo studio giunge alla conclusione che nelle scuole statunitensi i bambini delle minoranze hanno meno probabilità di essere identificati come disabili e quindi di ricevere servizi educativi speciali rispetto ai bambini bianchi che hanno le stesse caratteristiche e che uno sforzo collettivo da parte di ricercatori, *policymaker*, professionisti deve essere disegnato e condotto per ridurre questa disparità.

Ultimo tema che affrontiamo in studi che utilizzano la regressione logistica è quello del rapporto fra formazione e benessere psico-fisico.

Il riferimento è a uno studio (Narushima, Liu & Diestelkamp, 2016) il cui scopo è quello di identificare gli effetti di lunghe partecipazioni degli adulti e anziani in attività di formazione permanente sul loro benessere psicologico, prendendo in considerazione stato di salute, vulnerabilità e altre variabili demografiche quali età, genere, stato civile e così via. È condotto su un campione di adulti e anziani del Canada, paese che ha aggiunto il LifeLong Learning fra le politiche di invecchiamento attivo in linea con l'Active Ageing Framework dell'Organizzazione Mondiale della Sanità, che, lavorando sulle idee di salute, partecipazione e sicurezza, ha incoraggiato politiche e pratiche relative all'invecchiamento utili per gli individui e l'intera società, considerando la tarda età come un momento di conoscenza e competenza e non come un momento di declino.

Il campione, depurato dai missing data, è rappresentato da circa 400 adulti e anziani da 60 anni in su che hanno risposto a un questionario avendo partecipato a un programma pubblico di educazione continua non formale organizzato da una scuola insieme ad alcuni enti come community centres, centri di apprendimento per adulti, scuole superiori e case di riposo su temi che vanno dall'arte alla musica, al fitness all'informatica e così via.

Segue una descrizione delle variabili trattate.

- *benessere psicologico*, misurato con lo Psychological General Well-Being Index (PGWBI) in 22 item su stati depressivi, ansia, stress, soddisfazione di vita, vitalità e così via con score da 0 a 5. Il punteggio massimo totalizzabile è pari a 110. La variabile dicotomica creata (variabile dipendente) distingue le due modalità "in difficoltà" / "positivo" in base ai punteggi rispettivamente inferiori a 73 / uguali o superiori a 73.
- *vulnerabilità*, calcolata attraverso un indice definito dagli autori che tiene in considerazione l'età e somma più indicatori: indicatori sulla salute (malattie croniche, esercizio fisico, difficoltà nella vita quotidiana quale camminare, salire le scale, sentire), stato socio-economico (reddito e livello di istruzione, stato civile) e supporto sociale (partecipazione in altre attività sociali, numero di soggetti su cui l'anziano può contare, numero dei co-abitanti).
- *durata della formazione*, suddivisa in tre periodi superiori ai 4 mesi: 4-18 mesi (short), 19-84 mesi (middle), superiore a 49 mesi (long).
- *Self-perceived health* (SRH), rilevata attraverso le modalità "eccellente, molto buona, buona, scarsa" in risposta alla domanda: "Rispetto ad altre persone della tua età, come definiresti la tua salute in generale?".

Gli autori propongono due modelli di regressione logistica, in entrambi la variabile dipendente è lo stato di salute e la variabile indipendente la durata della formazione. Nel primo prendendo in considerazione come covarianti età, genere, livello di vulnerabilità. Nel secondo età, genere e in più il valore SRH.

Insieme ai valori dell'odds ratio nei due modelli ci sono restituiti anche gli intervalli di confidenza al 95% di significatività, mostrati in grafici come in Figura 5.5.

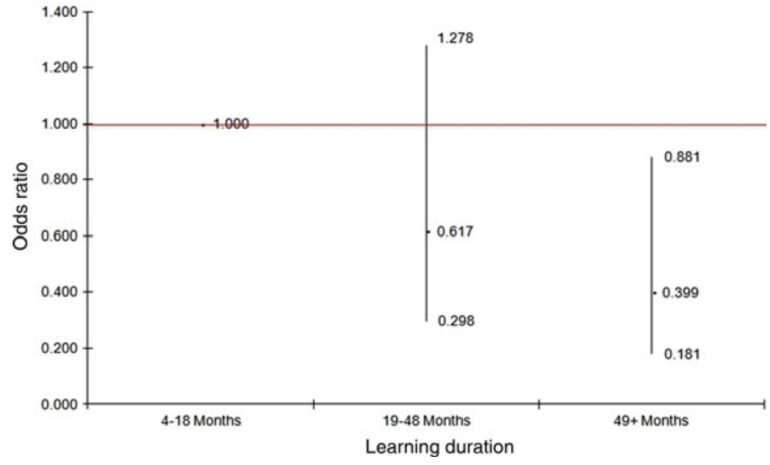


Figura 5.5 – Odds ratio e intervalli di confidenza al 95% per durata dell'apprendimento, genere, età, e gruppo di rischio in uno studio condotto usando la regressione logistica sugli effetti della formazione permanente sul benessere psicologico (Narushima, Liu & Diestelkamp, 2018, p. 666).

Nell'intervallo di confidenza per gli odds ratio della durata di 19-48 mesi è compreso anche il valore 1, cosa che non permette di usare questo range per trarre conclusioni.

In entrambi i modelli, partecipare alla formazione per periodi superiori ai 48 mesi aumenta il benessere psichico del 60%, sia che inseriamo fra i confondenti l'indice di vulnerabilità sia che ci sia il self-rated health, elemento che dimostra i benefici di partecipare alle attività sociali in contesti non formali per un invecchiamento attivo e un benessere in età adulta.

Il titolo di studio non influenza il livello di benessere, probabilmente perché le necessità cambiano nel tempo.

Del rapporto fra titolo di studio e stato di salute si occupano Jamie L. Lynch e Paul T. von Hippel (2016) i quali lavorano su un campione composto da circa 9mila studenti facenti parte della coorte del 1997 che nel periodo fra i 15 e i 31 anni ha risposto a un'indagine sulla Self-Rated Health (SRH) a partire dalla domanda "In generale, come va la tua salute?". Gli autori si chiedono se un migliore stato di salute conduca a raggiungere più successi scolastici, se al contrario più formazione possa migliorare lo stato di salute o se entrambi gli elementi, titoli di studio e stato di salute, siano ugualmente conseguenze di fattori relativi a stili e modalità di vita.

In questo caso viene usato come metodo di analisi la regressione logistica ordinale che rappresenta nelle regressioni multinomiali una generalizzazione della regressione logistica nella quale la variabile dipendente è rappresentata da una variabile ordinale sui livelli attribuibili allo stato di salute. Essa si basa sullo stesso meccanismo d'analisi della regressione logistica e considera di volta in volta soltanto una delle modalità della variabile come quella di riferimento. Non approfondiremo però questa tecnica in questo contesto in cui ci siamo concentrati unicamente sulle variabili binomiali.

CAPITOLO 6

CLUSTER ANALYSIS

Al termine del capitolo, il lettore sarà in grado di:

- *descrivere i principi e le fasi della cluster analysis, il concetto di distanza e gli algoritmi di clustering;*
- *interpretare un dendrogramma;*
- *elencare esempi della ricerca educativa nei quali è stata utilizzata la cluster analysis.*

6.1 – Cluster analysis

Probabilmente anche quest'oggi ci sarà capitato di riorganizzare gli scomparti del frigo per tipologia di alimenti e data di scadenza, di riordinare le banconote nel nostro portafogli per taglio, di ripensare gli scaffali della libreria del salotto riempendo quelli più bassi con i libri per bambini, posizionando al centro quelli scolastici e in alto saggi e romanzi per una lettura piacevole e rilassata. Sono situazioni in cui abbiamo a che fare con la creazione di gruppi fra gli oggetti sulla base di una o più caratteristiche. Al termine di queste attività, probabilmente noteremo che gli oggetti raggruppati possiedono elementi comuni anche per altri fattori che non ci aspettavamo: i cassettoni del frigo pieni di frutta sono più colorati del resto degli scomparti; le banconote di grosso taglio hanno dimensioni maggiori; i libri per bambini molto spesso hanno una copertina cartonata diversamente dai testi scolastici.

Molte sono le circostanze nella vita quotidiana in cui ci ritroviamo a raggruppare gli oggetti di nostro interesse. Lo facciamo perché abbiamo bisogno di creare classificazioni e tassonomie (quante banconote di tagli diversi abbiamo nel portafoglio?), perché cerchiamo caratteristiche che accomunano elementi diversi (quali volumi fra quelli che abbiamo in casa sono destinati ai più piccoli?) o ancora perché desideriamo ridurre a una quantità osservabile le informazioni generate da un numero elevato di casi particolari (nel nostro frigo c'è quanto serve per un'alimentazione completa e varia?).

Queste pratiche della quotidianità sono comuni a numerosi contesti di ricerca, ogni settore scientifico nel quale sia necessario produrre tassonomie, gerarchie e raggruppamenti fra gli elementi studiati. Discipline che fanno uso di processi di raggruppamento, solo per fare qualche esempio, sono l'archeologia, il marketing, la biologia e l'astronomia, la medicina e la linguistica.

Assumere l'esistenza di una naturale struttura in gruppi all'interno di una popolazione o di un campione ci permette di verificare ipotesi sulle caratteristiche degli stessi gruppi e generarne di nuove sul funzionamento dei sistemi analizzati.

La cluster analysis comprende un insieme di metodi che permettono di raccogliere gli oggetti dell'analisi (unità statistiche, raggruppamenti e talvolta anche variabili) in gruppi (*cluster*) nei quali le osservazioni hanno caratteristiche simili all'interno e dissimili dagli altri raggruppamenti. La composizione dei cluster deve garantire che all'interno dei gruppi ci sia un'alta omogeneità fra gli oggetti che li compongono e che i gruppi costituiti all'interno del campione risultino eterogenei fra di loro. La distanza fra le unità di un cluster deve essere perciò minima mentre deve essere elevata la distanza fra i vari gruppi individuati affinché si possano definire i profili caratterizzanti ciascun gruppo e rilevare le differenze che intercorrono fra di essi.

Lo scatterplot in Figura 6.1 mostra la distribuzione di un campione di studenti del CdL in Digital Education dell'Università degli Studi di Modena e Reggio Emilia diviso in quattro cluster in base ai risultati in trentesimi conseguiti in due esami. La collocazione nello spazio bi-dimensionale mostra evidenti distanze fra i gruppi: appartengono al gruppo A gli studenti che hanno sostenuto entrambi gli esami, al gruppo B gli studenti che hanno sostenuto solo l'Esame 2, al gruppo C gli studenti che hanno sostenuto solo l'Esame 1, al gruppo D gli studenti che non hanno sostenuto nessuno dei due esami (0 è il punteggio attribuito agli studenti che non hanno sostenuto l'esame). Notiamo che la distanza fra i punti interni ai gruppi è bassa, mentre la distanza fra i gruppi cerchiati in rosso è più rilevante. È questa una prima soluzione per distinguere i gruppi nel campione ma non è l'unica. Sia nel gruppo A che nel gruppo B appaiono come più distanti gli studenti che hanno conseguito un punteggio appena sufficiente e che quindi, in una soluzione diversa a più cluster, potrebbero essere collocati in un gruppo diverso. Nell'analisi di un campione, possiamo scegliere di determinare molti gruppi che di conseguenza contengono un numero inferiore di unità più simili fra loro oppure un numero inferiore di gruppi di più alta numerosità che risultano meno omogenei all'interno. Aumentando il numero dei gruppi, aumenterà la similarità fra gli oggetti che li compongono e

otterremo profili più simili fra le unità, riducendo anche il numero di informazioni perse nella descrizione generale delle caratteristiche del cluster. Tuttavia, allo stesso tempo, lavorare su un numero più elevato di cluster potrebbe comportare un incremento di complessità nell'analisi.

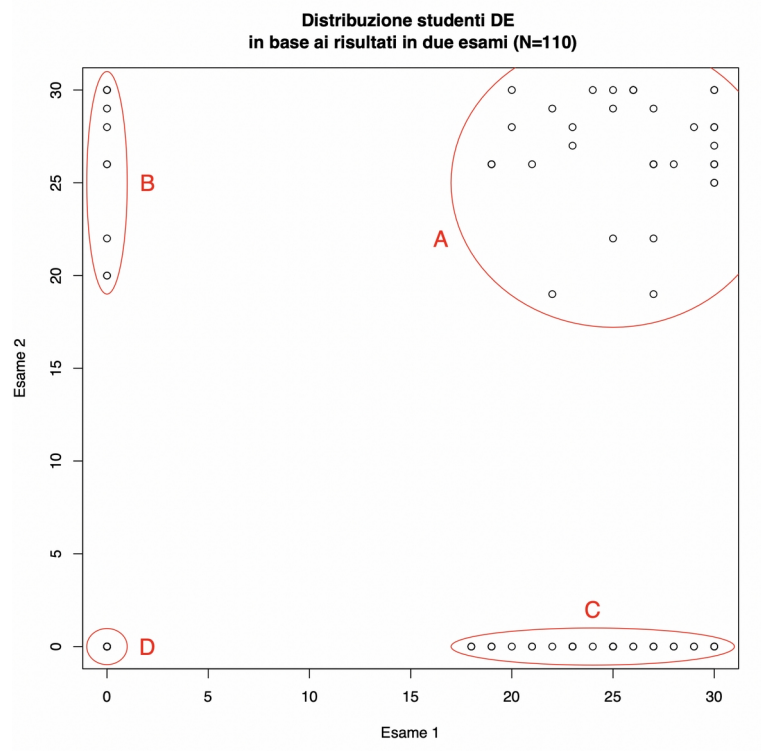


Figura 6.1 - Scatterplot dei voti conseguiti in due esami dagli studenti del CdL di Digital Education. Nella figura si distinguono nei cerchi in rosso 4 cluster con caratteristiche simili.

Lo scopo principale delle tecniche di cluster analysis non è tanto quello di generalizzare i risultati ottenuti quanto di individuare naturali raggruppamenti nei campioni. Si tratta infatti di un metodo descrittivo e non inferenziale per cui molti degli assunti sulla normalità, linearità e omoschedasticità non sono così rilevanti come in altre tecniche.

La cluster analysis è considerata una tecnica esplorativa, che cioè permette di studiare la composizione di un campione per trovare tendenze strutturali nella similarità fra le unità statistiche. Può avere comunque finalità confermative se confrontiamo il dataset osservato con una struttura in gruppi ipotizzata sulla base di precedenti studi e analisi. Nella tecnica si mette in atto un processo di

riduzione che l'analisi fattoriale conduce sulle variabili: nella cluster analysis le osservazioni vengono sintetizzate in un numero ridotto di gruppi con una perdita di informazioni che le procedure di clustering provano a minimizzare.

6.1.1 - Fasi di realizzazione di una cluster analysis

Nel processo di clusterizzazione il ricercatore sceglie le variabili da sottoporre all'analisi e i metodi da utilizzare per misurare distanze e creare raggruppamenti; definisce le regole per fermare il processo (*stopping rules*) di definizione del numero di cluster a cui fermarsi e interpreta i profili interni a ciascun cluster (proprio come accade per noi nella riorganizzazione della libreria: scegliamo la caratteristica in base alla quale dividere i libri fra gli scaffali, valutiamo a quale categoria appartiene ogni testo e cerchiamo similarità fra i gruppi così creati). Gli interventi da parte del ricercatore e le diversità fra le numerose tecniche di clustering che nel tempo sono state prodotte impediscono che i processi di analisi portino a soluzioni univoche e ci fa parlare della cluster analysis come di un processo scientifico e al contempo come di un'arte (Hair et al., 2014, p. 428).

Tre sono gli elementi chiave nella realizzazione di una analisi cluster: il modo in cui misuriamo la similarità − e di conseguenza le differenze, che si traducono in una misura di distanza − fra gli elementi che costituiscono il gruppo, la procedura attraverso cui costruiamo i cluster, l'interpretazione e la validazione dei gruppi costituiti.

Descriviamo di seguito il processo di applicazione della tecnica che è schematizzato nella Tabella 6.1.

La fase iniziale di lavoro prevede la definizione degli obiettivi dello studio e la selezione delle variabili da prendere in considerazione nella costituzione dei cluster. Nello screening del dataset è importante verificare la presenza di:

- fenomeni di multicollinearità fra le variabili che potrebbero modificare il peso delle dimensioni osservate nella costituzione dei gruppi: variabili altamente correlate fra loro che magari sono riferibili a uno stesso set di informazioni assumono maggiore rilevanza nella composizione dei cluster;

- corretta costituzione del campione che deve essere numeroso e rappresentativo della struttura della popolazione: se così non fosse, raggruppa-

menti della popolazione sottorappresentati nel campione potrebbero essere assimilati a outlier;

- omogeneità delle scale con cui sono espresse le variabili: qualora le modalità siano espresse in scale diverse, è necessario standardizzare le variabili.

Concluse le operazioni di preparazione dei dati, si passa alla fase di creazione dei gruppi e l'applicazione di algoritmi per raggruppare i cluster.

L'articolazione dei gruppi parte dalla definizione della similarità delle unità statistiche che li compongono. La similarità rappresenta il grado di corrispondenza fra tutte le variabili usate nell'analisi.

Per le variabili quantitative esistono due metodi per misurare la similarità.

Il primo e anche il meno usato, più focalizzato sugli schemi e le relazioni fra le variabili, è il calcolo del coefficiente di correlazione fra i profili di due unità. Valori più alti di correlazione sono espressione di una maggiore affinità fra le unità che quindi verranno inserite in uno stesso cluster poiché simili.

Il secondo metodo, più che calcolare la similarità, verifica la dissimilarità, in quanto non tiene in considerazione la somiglianza fra due unità ma la distanza che esiste fra le stesse. Il metodo calcola quanto due oggetti sono distanti fra di loro e raggruppa quelli più vicini.

Le distanze possono essere calcolate in più modi. Parlando dell'analisi delle corrispondenze (capitolo 3), abbiamo in parte affrontato questo tema che ora riprendiamo e approfondiamo.

Fra le più note distanze vi è la distanza euclidea, che possiamo visualizzare in analisi bivariate in un piano bidimensionale come la lunghezza del segmento che unisce i due punti di cui stiamo misurando la similarità e che in maniera più generale nelle analisi multivariate corrisponde alla radice quadrata della somma del quadrato delle distanze fra tutte le variabili che definiscono le unità statistiche. Viene anche usata in alcuni casi direttamente senza estrazione della radice quadrata e con l'aggiunta di pesi che possono essere attribuiti alle singole variabili qualora sia necessario assegnare ad alcune un peso maggiore nel processo di clusterizzazione.

Altra distanza è quella di Manhattan, anche detta *city-block*, che somma le distanze assolute fra le variabili. Una sua variazione è conosciuta come distanza di Chebyshev, dove la distanza è il valore massimo fra le differenze dei valori assoluti.

Fasi di preparazione	
Stage 1: *Objectives of Cluster Analysis*	Scegliere l'obiettivo dello studio (è esplorativo o confermativo? È finalizzato a creare una tassonomia, a sintetizzare i dati, a identificare relazioni latenti esistenti?) Selezionare le variabili da includere nello studio
Stage 2: *Research Design in Cluster Analysis*	Verificare che il campione sia sufficientemente numeroso affinché tutti i gruppi siano ben rappresentati Verificare la presenza di outlier che potrebbero compromettere la clusterizzazione o perché fuori dal campione (troppo distanti) o perché espressione di un gruppo di dimensioni troppo piccole sottorappresentato Definire le procedure per calcolare misure di similarità e distanza Se necessario, standardizzare i dati (variabili, distanze, osservazioni) per uniformare le scale
Stage 3: *Assumptions in Cluster Analysis*	Accertarsi che il campione sia rappresentativo della struttura della popolazione Verificare casi di multicollinearità delle variabili che possono incrementare il peso di alcune dimensioni nella definizione dei gruppi
Fasi di costruzione e validazione	
Stage 4: *Deriving Clusters and Assessing overall fit*	Definire quale procedura di partizione usare per formare i cluster Definire la stopping rule, la regola cioè che determina il numero di gruppi a cui fermarsi Se necessario, ripetere le operazioni nei casi in cui vengano rilevati outlier, gruppi di scarsissima numerosità, risultati poco significativi
Stage 5: *Interpretation of the Clusters*	Descrivere i profili risultati nei cluster ottenuti (molto spesso a partire dai centroidi)
Stage 6: *Validation and Profiling of the Clusters*	Validare la soluzione finale di clusterizzazione su campioni diversi o a partire da prederminati criteri di validità

Tabella 6.1 - Sintesi delle fasi di realizzazione di una cluster analysis (Hair et al., 2014, pp. 425-451, nostra rielaborazione).

Semplificando, da un punto di vista geometrico possiamo considerare in un piano bidimensionale la distanza euclidea come la lunghezza del segmento di estremi A e B nella Figura 6.2 e la distanza di Manhattan come la somma delle lunghezze dei segmenti AC e BC.

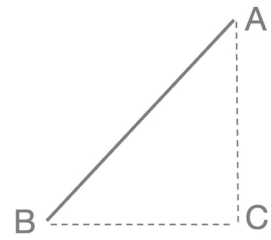


Figura 6.2 - Raffigurazione geometrica in un piano bidimensionale delle distanze euclidea (AB) e di Manhattan (AC + BC).

Sia la distanza euclidea che la distanza di Manhattan sono casi specifici della distanza di Minkowski di ordine p espressa dalla formula dove x_i e y_i sono le coordinate dei punti fra i quali calcolare la distanza (distanza di Manhattan per $p = 1$ e distanza euclidea per $p = 2$).

(6.1)

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

Aggiungiamo all'elenco la distanza Mahalanobis nel cui computo non ha influenza la maggiore o minore variabilità delle variabili.

Le variabili qualitative sono usate più raramente in procedure di cluster analysis e in seguito a processi di dummizzazione. Per misurare le distanze in questi casi si utilizzano misure di associazione che nei casi più semplici verificano per ogni unità statistica la percentuale di concordanza all'interno di un gruppo di variabili. Vengono quindi raggruppate le unità che mostrano una percentuale simile di modalità (1,0) nelle variabili dummy o di concordanza fra le modalità attribuite a più variabili.

Alla scelta del tipo di distanza/misura di similarità che decidiamo di utilizzare, segue la scelta dell'algoritmo da usare per formare i gruppi.

Si distinguono tecniche *gerarchiche* e *non gerarchiche*, applicate di volta in volta in base ai contesti della ricerca.

Le prime organizzano le osservazioni ad albero (si vedano le Figure 6.3 e 6.4).

Si parte dal considerare le osservazioni singolarmente e calcolare le distanze che intercorrono fra tutte. Le due unità più vicine vengono associate in un cluster. Si misurano le distanze di questo primo cluster dalle altre unità e di nuovo le distanze inferiori indicano quali osservazioni/cluster raggruppare. Si ripete l'algoritmo fino ad ottenere un unico gruppo costituito dalle n osservazioni del campione. Questa procedura è detta *agglomerativa* poiché partiamo da tutte le osservazioni e, aggregandole, giungiamo a individuare il numero ottimale di cluster per il nostro studio. Nelle procedure gerarchiche agglomerative la distanza fra i gruppi da formare può essere calcolata con algoritmi diversi: per aggregare due unità/cluster possiamo prendere in considerazione la distanza minima fra le osservazioni che lo compongono (*single-linkage*) o al contrario la distanza massima (*complete-linkage*). O ancora si può calcolare la similarità come media fra tutte le distanze fra ciascuna coppia di oggetti che fanno parte dei due cluster (*average linkage*); si può considerare la distanza minima fra i centroidi dei due cluster dove per centroide indichiamo il punto che ha per coordinate i valori medi delle osservazioni sulle variabili nel cluster; o si può usare il *Ward's method* che raggruppa le unità che portano al minimo incremento possibile della devianza in ogni livello di fusione, apprezzando la perdita del minor numero di informazioni nell'aggregazione di una coppia di oggetti.

Opposte agli algoritmi agglomerativi sono le procedure gerarchiche definite *divisive*, meno utilizzate. In esse si procede in maniera inversa: tutte le osservazioni vengono considerate come un unico gruppo che viene diviso in sottogruppi più piccoli a mano a mano che aumentano le distanze fra le osservazioni. Si distinguono fra gli algoritmi divisivi quelli *monotetici* che usano solo una variabile per dividere i sottogruppi e quelli *politetici* che usano tutte le variabili del dataset.

Una tipologia di grafico usata per rappresentare gli algoritmi gerarchici è il *dendrogramma* che negli esempi delle Figure 6.3 e 6.4 presenta sulle ascisse le distanze fra le osservazioni e sulle ordinate le osservazioni stesse. In grafici di

questa natura, gli outlier si visualizzano come gli oggetti che vengono agganciati per ultimi nei gruppi.

In entrambe le figure, sono rappresentati in ambito multivariato i raggruppamenti generati su un cluster di 20 studenti (il numero estremamente basso è stato utilizzato solo a scopo esemplificativo) a partire dai voti conseguiti nei 7 esami del primo anno nel CdL in Digital Education [il dataset completo e ampliato da altre variabili è stato utilizzato con la cluster analysis in due studi che riguardano i sistemi di tutoring (De Santis et al., 2021a) e i fattori che influenzano il successo accademico (De Santis et al., 2021b)]. Ci chiediamo se possano essere rilevate delle similarità nei profili degli studenti che hanno scelto nel corso del primo anno di sostenere uno o più esami di settori disciplinari diversi. Sono quindi 7 le variabili scelte per operare la clusterizzazione. Nella Figura 6.3 è stata usata la distanza di Manhattan e l'algoritmo agglomerativo del complete-linkage; nella seconda la distanza euclidea e l'algoritmo single-linkage. In `R` le distanze sono state calcolate con la funzione `dist` (libreria: `stats`) e i cluster con la funzione `hclust` (libreria: `stats`). Notiamo che nei due grafici i livelli di fusione e la composizione finale dei cluster differisce.

Accade frequentemente che l'uso di metodi diversi per descrivere le distanze e per aggregare i gruppi restituisca soluzioni con cluster diversi per numero e oggetti che li compongono. Ripetere le operazioni con più metodi e confrontare i risultati aumenta la ragionevolezza della nostra suddivisione e ci porta a identificare fra le classificazioni quella maggiormente sensata e utile. Queste differenze nei processi di clustering ribadiscono quanto sia importante partire da teorie e ricerche precedenti, ipotesi concrete e giustificazioni di ciascuna fase di lavoro per gestire al meglio la scelta delle variabili e dei metodi e l'interpretazione dei risultati.

Le procedure non gerarchiche di raggruppamento dei cluster partono dalla definizione del numero ***k*** di cluster da formare sulla base dei *semi* ossia punti di riferimento rispetto ai quali viene calcolata la distanza di ciascuna osservazione del campione. Talvolta i semi del cluster vengono definiti dal ricercatore stesso a partire da altre ricerche o analisi precedenti, altre volte sono generati attraverso selezioni random. L'algoritmo non gerarchico più noto, definito *k-means*, è una procedura a partizione iterata: i gruppi vengono formati attorno ai semi identificati. In seguito vengono calcolati i centroidi di ciascun cluster e ricalcolate le distanze delle unità dai centroidi. In base ad esse, le unità sono ridistribuite in nuovi gruppi.

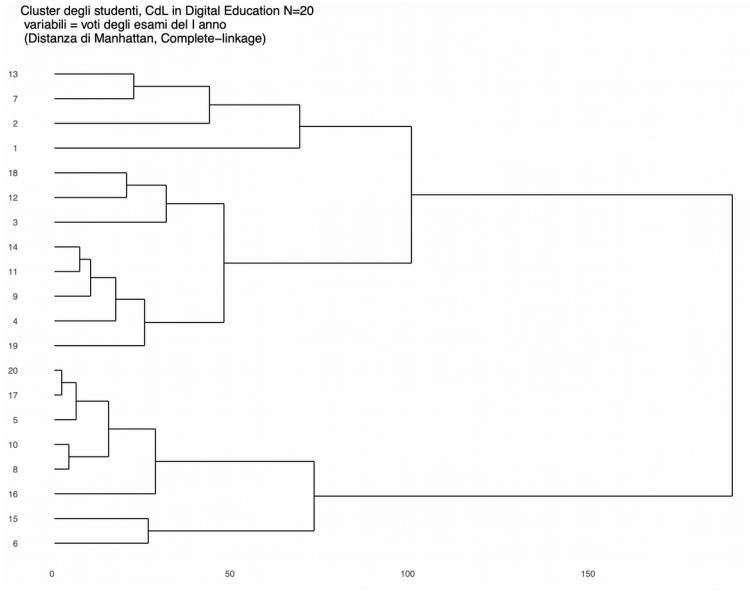


Figura 6.3 - Esempio di dendrogramma (distanza di Manhattan e algoritmo gerarchico agglomerativo del complete-linkage). Il grafico è stato realizzato in `R` con la funzione `ggdendrogram` (libreria: `ggdendro`).

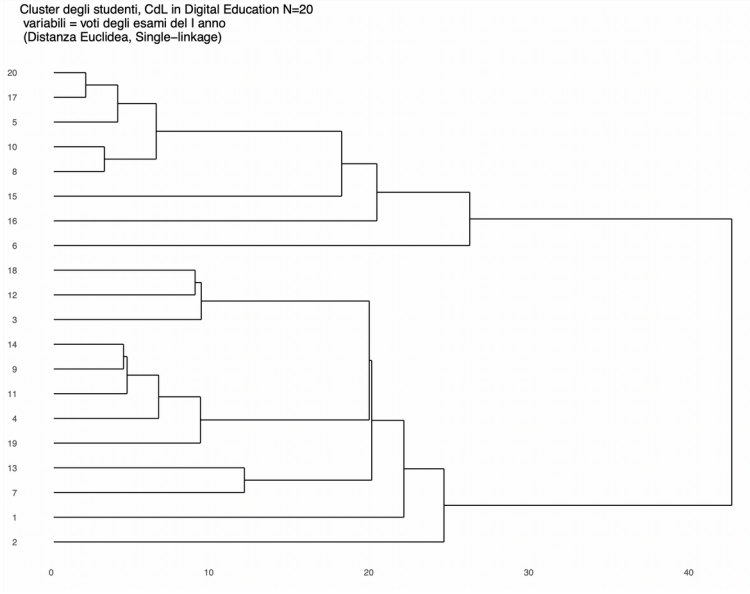


Figura 6.4 - Esempio di dendrogramma (distanza euclidea e algoritmo gerarchico agglomerativo del complete-linkage). Il grafico è stato realizzato in `R` con la funzione `ggdendrogram` (libreria: `ggdendro`).

L'algoritmo viene ripetuto fino a quando, al momento del computo dei centrodi, questi conservano una posizione fissa. Gli algoritmi che funzionano sull'identificazione dei semi in maniera sequenziale o parallela non prevedono lo spostamento di una unità statistica fra gruppi in base al modo in cui si assegnano le osservazioni, cosa che invece è fatta da algoritmi di ottimizzazione più raffinati.

Fra le procedure non gerarchiche si annoverano anche i metodi di ottimizzazione e gli algoritmi genetici (Paterlini & Minerva, 2001; Pattarin, Paterlini & Minerva, 2004).

A quanti gruppi fermarsi? Quanti cluster rappresentano in maniera ottimale il campione?

Anche se non esiste una risposta univoca a tali domande, ci sono buone pratiche e processi di analisi che portano ad identificare soluzioni valide da un punto di vista teorico e pratico.

Nei casi dell'uso di metodi gerarchici il processo di clustering viene interrotto quando si raggiunge un grado di eterogeneità interna non accettabile. L'eterogeneità può essere misurata come la distanza dei livelli di fusione. In genere si è soliti considerare tagli a metà della distanza totale, che vediamo nei dendrogrammi in figura sull'asse delle ascisse. Si guardi a scopo esemplificativo la Tabella 6.2 riferita ai dati della Figura 6.3, nella quale abbiamo riportato i livelli di fusione superiori al decimo per mostrare il comportamento negli ultimi livelli, dato che nei precedenti l'incremento della distanza è sufficientemente costante. La prima colonna indica il livello di fusione, la seconda e la terza i cluster e le unità raggruppate (con il segno -) e la quarta colonna la distanza misurata. Il taglio dei cluster si fissa quando le differenze fra i livelli di fusione sono troppo elevate. In questo caso esemplificativo, il ricercatore valuterà se fissare il taglio dopo il livello 13 o 15, dopo aver analizzato la composizione dei cluster ottenuti. Tale analisi viene condotta per cluster con il computo degli indici di posizione per ciascuna variabile e la visualizzazione grafica della distribuzione dei dati.

Ultime fasi di lavoro in una analisi cluster sono l'interpretazione e la validazione dei risultati. Il ricercatore, a partire dalle suddivisioni fra le osservazioni, determina i profili dei soggetti appartenenti a ciascun cluster. Le caratteristiche dei centroidi possono sintetizzare le proprietà degli elementi dell'intero cluster e, qualora non ben diversificate fra i gruppi, indicare che le operazioni di clusterizzazione vanno ripetute. La significatività pratica della clusterizzazione può essere validata ripetendo l'operazione su altri campioni oppure suddividendo

il campione di partenza in due in modo da verificare i risultati. Criteri di predittività derivanti da presupposti teorici o concettuali possono allo stesso modo permettere di validare le classificazioni ottenute.

Livello di fusione	Elemento 1	Elemento 2	Distanza
...	...	...	...
10	-19	7	25
11	-6	-15	26
12	-16	6	28
13	-3	8	31
14	-2	9	43
15	10	13	47
16	-1	14	68
17	11	12	72
18	15	16	99
19	17	18	188

Tabella 6.2 - Livelli di fusione per il processo di clustering della Figura 6.3.

Un ultimo aspetto da tenere in conto è il fatto che l'attribuzione di una unità statistica a un gruppo ovvero a un altro non sempre è definita (*crispy/hard attribution*) ma, trovandosi in prossimità dei bordi, un'unità potrebbe appartenere a un gruppo oppure a uno vicino (*fuzzy attribution*) senza che questo alteri significativamente la struttura dei cluster. È compito del ricercatore definire queste situazioni limite anche basandosi sulla profonda conoscenza del fenomeno.

6.2 - L'uso della cluster analysis nella ricerca educativa

Presentiamo alcuni studi che hanno utilizzato la cluster analysis come tecnica di analisi, nella consapevolezza di non poter essere esaustivi né nella descrizione delle ricerche, né nella varietà dei casi.

È frequente che questa tecnica venga utilizzata in concomitanza con altre, a volte per classificare, altre per individuare l'organizzazione latente in gruppi che caratterizza dataset di vario genere. Le ricerche, non senza eccezioni e casi particolari, prendono in considerazione come unità statistiche nell'analisi sia in-

dividui che, meno frequentemente, organizzazioni. Questo comporta che il metodo venga utilizzato ad esempio in ricerche che restituiscono informazioni sulle strategie di apprendimento e sull'acquisizione di abilità cognitive e non cognitive e, parimenti, in studi che analizzano le modalità di funzionamento di organizzazioni che a vari livelli si occupano di formazione.

Vitomir Kovanović e colleghi (2019) utilizzano la cluster analysis per identificare le principali strategie di studio degli iscritti a un MOOC sulla programmazione erogato su edX. La ricerca intende verificare l'esistenza di differenze significative nei gruppi di utenti costituiti in base alla tipologia di engagement nel corso e qualificati, in secondo luogo, dal voto finale conseguito e dai livelli percepiti delle tre dimensioni chiave dell'apprendimento in una community of inquire (cognitive presence, social presence, teaching presence). Le modalità di fruizione dell'ambiente didattico, le strategie di apprendimento degli studenti e il loro livello di engagement nel corso sono state identificate a partire dai log e in particolare da 29 variabili calcolate come misure ottenute dal tracciamento in piattaforma di oltre 20mila studenti. Fra le altre: i numeri di accessi al corso e di navigazione nelle pagine, i numeri di valutazioni completate e delle videolezioni visualizzate, i numeri di partecipazione alle discussioni. Agli studenti è stato somministrato un questionario in ingresso centrato su dati anagrafici, motivazioni e aspettative e uno in uscita che, oltre a domande sulla soddisfazione, conteneva item sulle percezioni dello sviluppo di un modello di community of inquire nell'esperienza di apprendimento online.

Nella cluster analysis gli autori hanno utilizzato il metodo agglomerativo di Ward calcolando la distanza euclidea. Le differenze fra i gruppi sono state validate usando la tecnica di analisi MANOVA. Sono state prese in considerazione le distanze fra i livelli di fusione sul dendrogramma per distinguere tre gruppi corrispondenti a tre profili di studente-tipo, descritti poi a partire dai valori del centroide di ciascun cluster (Figura 6.5). Gli autori hanno attribuito a ciascun profilo un aggettivo identificativo, parlano infatti di *limited*, *selective* e *broad users*. Il gruppo più numeroso è quello indicato con l'aggettivo limited che raccoglie studenti poco motivati, che non partecipano alle attività e che scelgono i video fra gli oggetti didattici presenti sulla piattaforma. Il termine selective identifica gli studenti che hanno un coinvolgimento medio e non partecipano alle discussioni; sono quelli focalizzati sull'ottenimento del certificato e che centrano le loro attività sulla visualizzazione dei video e sugli assessment. Ultimo gruppo è quello dei broad users contraddistinti da un alto coinvolgimento, l'uso di tutte le risorse nel corso, un'attenzione speciale a risorse e tool, motivati da fattori professionali e curiosità.

Nelle conclusioni, a partire dalle caratteristiche identificate nei cluster, i ricercatori abbinano a ciascun gruppo un intervento didattico più utile: gli utenti limited, che non mirano a concludere il corso, avrebbero probabilmente bisogno di interventi per essere coinvolti nei processi piuttosto che per lavorare sui contenuti; i selected user, oltre a un coinvolgimento più ampio nelle discussioni, avrebbero bisogno di interventi focalizzati sull'acquisizione di conoscenze e competenze, poiché interessati a questo. I broad users, già molto coinvolti nel corso, potrebbero essere resi moderatori nelle attività per supportare i partecipanti meno impegnati. Individuare l'appartenenza di uno studente a uno di questi gruppi permette dunque di attivare proposte di individualizzazione dei percorsi formativi nei confronti dei singoli utenti anche in una piattaforma con numeri così alti di iscritti.

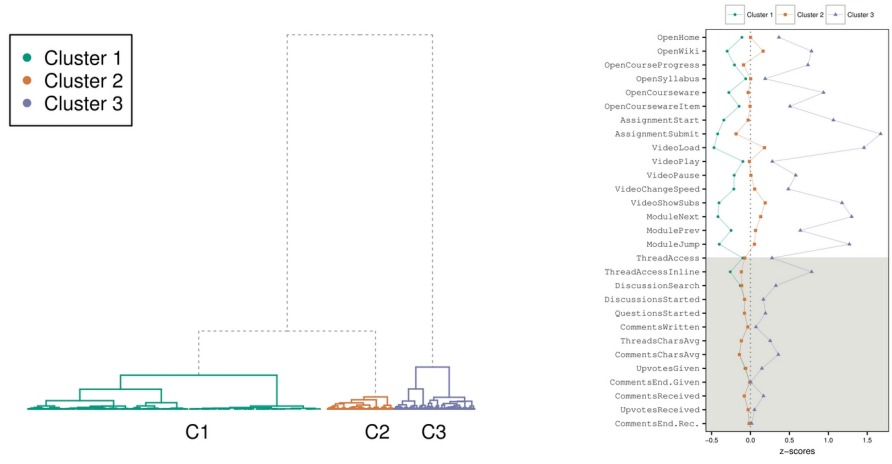


Figura 6.5 - A sinistra, dendrogramma della cluster analysis realizzata in una ricerca sulle strategie di studio in un MOOC nella quale sono stati ottenuti tre cluster fra i partecipanti. A destra, confronto fra le caratteristiche dei tre cluster dello stesso studio, in grigio i risultati relativi alle discussioni online (Kovanović et al., 2019, p. 25-26).

In maniera simile, anche uno studio condotto da Nick A. Stites e colleghi (2019) utilizza la partecipazione in un ambiente di apprendimento, definito attivo e collaborativo, per identificare gli "archetypical patterns" nell'uso delle risorse di un corso blended di Dinamica e, al contempo, le motivazioni e i valori che guidano gli studenti nell'uso di ciascun modello. Per la raccolta dei dati in questo caso non vengono utilizzati i log ma un questionario somministrato al termine del corso a 581 studenti in Ingegneria. Nel questionario è stato chiesto

di indicare con quale frequenza ciascuno dei 9 tipi di risorsa presente nel corso è stata fruita. Le risposte relative alle frequenze di utilizzo delle risorse sono state utilizzate come variabili nell'analisi cluster.

Gli autori hanno identificato 14 modelli di clustering e hanno selezionato quello che meglio si adeguava ai dati raccolti attraverso il criterio di verosimiglianza del Bayesian Information Criterion (BIC). Sono stati creati 9 cluster (si veda la Figura 6.6) nei quali, per ognuno, le frequenze d'uso si concentrano su diverse risorse o attività. La scelta delle risorse da parte di uno studente aiuta a far luce anche sulle sue capacità autoregolative nella ricerca di supporto nello studio.

Tale attività, definita *help-seeking behaviour*, è stata l'argomento al centro delle interviste condotte con 44 studenti appartenenti ai 9 gruppi. Lo studio valorizza la presenza di risorse didattiche di natura diversa nelle piattaforme che ospitano corsi online. La loro varietà può rispondere ai bisogni diversificati degli studenti e i docenti, note queste informazioni, possono avere chiaro il modo in cui ciascuna variazione nelle attività e nelle risorse influisce sulle modalità di studio di ciascuno studente.

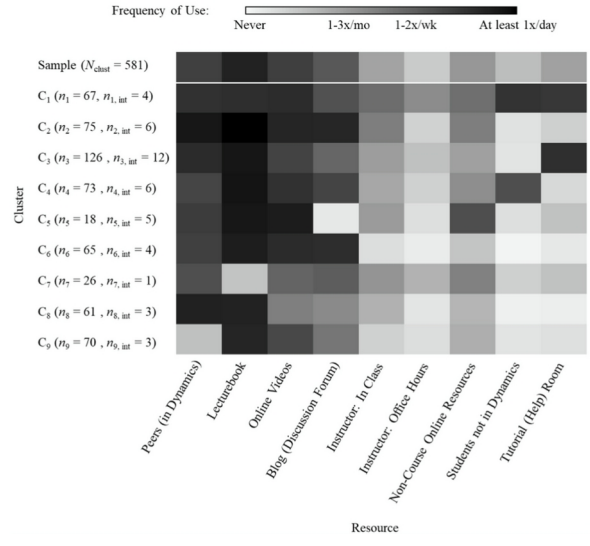


Figura 6.6 – Valori medi delle frequenze relative all'uso delle varie risorse da parte degli studenti suddivisi per cluster. Accanto ai nomi dei cluster, sono riportati la loro dimensione e il numero dei soggetti estratti per l'intervista (Stites et al., 2019, p. 1745).

Po-Yao Chao (2016) conduce una ricerca fra gli studenti universitari in un corso in Informatica nel tentativo di comprendere quali pattern di comportamento gli studenti adoperano per la programmazione visuale in ambienti grafici e se essi sono il risultato di performance diverse nella risoluzione di problemi computazionali. 158 studenti di un'università di Taiwan hanno utilizzato un ambiente di programmazione grafica creato ad hoc in un corso finalizzato ad acquisire i principi del linguaggio di programmazione C++. Nelle otto settimane del corso sono stati previsti quattro incontri della durata di un'ora; ai partecipanti è stato chiesto di usare l'ambiente per risolvere problemi nelle fasi di introduzione, formazione e pratica, al fine di completare gli assignment. Sono stati collezionati i log file e da essi sono stati calcolati 10 indicatori su computational practice, computational design e computational performance. La cluster analysis è stata realizzata prendendo in considerazione come variabili i 5 indicatori della computational practice. Per l'analisi è stato usato prima il metodo di Ward che ha condotto alla definizione del numero dei cluster; si è fatto poi ricorso al metodo k-means per ottimizzare i risultati con la ridistribuzione delle unità statistiche. Sono stati così individuati 4 cluster corrispondenti a differenti approcci alla pratica computazionale. A ciascun approccio sono stati associati anche gli indicatori relativi al computational design e alla computational performance. I risultati ottenuti hanno ricadute nell'ambito della didattica per ridefinire le caratteristiche degli ambienti di apprendimento di programmazione per i giovani programmatori e dei percorsi formativi più efficaci per far acquisire competenze di problem solving e programmazione nel settore informatico.

Fino a questo punto abbiamo mostrato come attraverso la cluster analysis sia possibile usare informazioni di varia natura sugli studenti per raggruppare profili e adeguare le pratiche didattiche in base ad essi. I tre studi presentati, inoltre, hanno in comune il contesto di svolgimento della ricerca, ossia un ambiente di formazione online. Uno studio svolto su attività in presenza in una scuola superiore è quello di Jennifer A. Schmidt, Joshua M. Rosenberg, e Patrick N. Beymer (2018). Lo studio differisce dai precedenti non solo per la differenza di ambiente (online e in presenza) ma anche perché in questa ricerca le unità statistiche non sono gli studenti ma quelli che gli autori definiscono *momentary engagement profiles* (MEP) ossia i profili rilevati in un dato momento rispetto a tre tipi di engagement – behavioral, cognitive e affective – durante le lezioni di scienze. L'obiettivo dell'articolo è quello di raggruppare i MEP in profili simili e allo stesso tempo stabilire se esiste una relazione fra l'engagement, il tipo di attività svolte in quel determinato momento e la presenza di attività di scelta da parte degli studenti rispetto ai compagni con cui lavorare, le attività da svolgere, i materiali da usare e così via.

Seguendo le procedure dell'Experience Sampling Method (Csikszentmihalyi, 2014), in due periodi dell'anno scolastico per 5 giorni, 244 studenti sono stati dotati di un cercapersone che si attivava per due volte in varie fasi di una lezione. Al segnale, gli studenti dovevano rispondere ad alcuni item relativi ai livelli di engagement e alle attività di scelta. Le lezioni sono state registrate e ciascun momento è stato contrassegnato in base a 10 tipi di attività: lezione, laboratorio, test, discussione, lavoro in gruppi e individuale, video, presentazione, attività informali e altro. Sono state così collezionate 4136 risposte che hanno fotografato i livelli di engagement e di scelta nelle lezioni di scienze in un dato istante corrispondente a una particolare attività didattica.

Sui MEP è stata condotta una cluster analysis in 2 step: nel primo una cluster gerarchica agglomerativa del tipo complete-linkage con il calcolo della distanza euclidea al quadrato; in seguito, a partire dai risultati ottenuti nel primo step, è stato organizzato il secondo con il metodo k-means che, come già detto, ha il vantaggio di ottimizzare i risultati. La ricerca ha restituito 6 cluster identificando profili istantanei distinti in base ai livelli di behavioral, cognitive e affective engagement. I profili sono stati definiti come *universally low*, *reluctant*, *pleasurable*, *rational*, *moderately full*, *full engagement*.

Lo studio raccoglie informazioni sul complesso rapporto (multidimensionale e legato al contesto) di coinvolgimento degli studenti nello studio delle scienze (il cluster più ampio è quello dell'universally low) e fornisce indicazioni ai docenti per determinare quelle attività didattiche che nella pratica possono maggiormente coinvolgere gli studenti.

Cosa succede quando le unità statistiche sono programmi o enti di formazione?

Laura Perna e Elaine Leigh (2017) studiano 289 *promise programs* dei college statunitensi (programmi equiparabili a premi e borse di studio che mirano a incentivare l'iscrizione ai college e migliorare il contesto culturale) per comprenderne le caratteristiche, il design, l'impatto sociale e definire un framework di riferimento per organizzarli e categorizzarli.

Dopo aver selezionato i programmi in base ad alcune caratteristiche prestabilite, aver creato un manuale per uniformare i risultati nella selezione da parte di ogni membro del team di ricerca e definito un processo di assicurazione della qualità, gli autori hanno realizzato una cluster analysis su un dataset composto da variabili qualitative usando il metodo gerarchico agglomerativo con algoritmo average-linkage. Hanno costruito tre modelli di clustering modificando il numero e il tipo di variabili coinvolte in base a precedenti studi e a consi-

derazioni sulle caratteristiche dei programmi (come tipologia di sostegno finanziario e richiesta di residenza in un determinato stato/città). Hanno ottenuto per ognuno dei tre modelli soluzioni diverse a 6 gruppi. L'analisi ha permesso di identificare fattori che differenziano i programmi (la sponsorizzazione statale, il tipo di sostegno finanziario e di istituzioni in cui la borsa di studio può essere utilizzata, criteri di merito o di necessità) e ha aperto la strada a ricerche future che a partire da queste informazioni possano individuare modalità di progettazione dei programmi promise (e di conseguenza variabili specifiche) che promuovano la partecipazione all'higher education di gruppi di studenti distinti.

Altro esempio è una ricerca condotta da Rosa Puertas e Luisa Marti (2019) che si occupa del tema della sostenibilità negli atenei. Lo scopo dell'indagine è quello di costruire un indicatore che, a partire dal precedente UI GreenMetrics, possa classificare in maniera più completa gli atenei misurando anche l'efficacia delle misure intraprese. Nello studio la cluster analysis è utilizzata per distinguere gruppi fra 719 università a partire dai 6 indicatori descritti nell'UI GreenMetric World University Ranking: la struttura, l'uso dell'acqua e dell'energia, i programmi di riciclo dei rifiuti, l'organizzazione dei trasporti, la formazione e la ricerca sulla sostenibilità. I 4 gruppi omogenei identificati in base ai livelli di sostenibilità degli atenei (*low*, *medium low*, *medium high*, *high*) hanno rappresentato il punto di partenza sui quali i ricercatori, usando altre metodologie di analisi, hanno potuto formulare un nuovo indicatore (DEA-GreenMetric) e una classificazione che tenga conto degli sforzi compiuti dagli atenei per gestire lo sviluppo sostenibile e controllare le questioni ambientali.

Gli esempi presentati di ricerche condotte in ambito educativo con l'uso della cluster analysis mostrano le finalità esplorative e descrittive del metodo che permette di identificare tendenze latenti nelle modalità di raggruppamento delle unità statistiche, siano soggetti in formazione oppure programmi/enti educativi, che contribuiscono a ripensare le policy e le pratiche didattiche.

CAPITOLO 7

MULTIDIMENSIONAL SCALING

Al termine del capitolo, il lettore sarà in grado di:

- *descrivere le tecniche di multidimensional scaling;*
- *distinguere le caratteristiche del multidimensional scaling metrico e non metrico;*
- *elencare esempi della ricerca educativa nei quali è stato utilizzato il multidimensional scaling.*

7.1 – Multidimensional scaling

Con le tecniche di *Multidimensional Scaling* (MDS) entriamo nell'ambito della statistica computazionale nella quale l'uso di algoritmi nelle procedure di analisi è prioritario rispetto all'applicazione di procedimenti analitici.

Si tratta di tecniche esplorative di interdipendenza nelle quali, riproponendo quanto detto nel capitolo 3, l'interesse del ricercatore è quello di rivelare la struttura latente dei dati frequentemente tramite rappresentazioni grafiche. Già traducendo il termine "multidimensional scaling", comprendiamo che siamo di fronte a tecniche di riduzione della dimensionalità nel contesto multivariato. Anche in questo caso, usando i metodi di MDS possiamo passare da n variabili (dimensioni) a k dimensioni che sono combinazione delle n variabili. La scelta di k frequentemente cade sul numero 2 o sul numero 3, perché così facendo si riesce a rappresentare i dati in un piano bi-dimensionale o tri-dimensionale. Ma questo è solo un limite della capacità di visualizzazione del ricercatore, non esiste alcun valido motivo (vedremo in seguito) per limitarsi a due o tre dimensioni.

Quando l'oggetto di studio sono le variabili, le rappresentazioni grafiche ottenute sono definite mappe di percezione; quando ad essere visualizzate sono le unità statistiche, si parla di mappe di classificazione.

Come nell'analisi delle corrispondenze e nella cluster analysis, il punto di partenza nell'applicazione del MDS è il calcolo delle prossimità, le distanze cioè. Tuttavia, rispetto alle due precedenti tecniche emergono delle differenze:

- l'analisi delle corrispondenze, come il MDS, viene applicata per ridurre le dimensioni e ugualmente restituisce visualizzazioni grafiche molto efficaci ma agisce soltanto su variabili non metriche calcolando le distanze fra le frequenze congiunte. Al contrario le tecniche di MDS possono essere applicate a variabili sia metriche che non metriche con procedure diverse, come vedremo.

- la cluster analysis pur lavorando sulle distanze con variabili metriche e non metriche non è finalizzata alla data reduction ma al raggruppamento di unità statistiche. Vedremo, soprattutto nel paragrafo 7.2, come MDS e cluster analysis siano usate spesso in combinazione per ridurre le dimensioni in un dataset e successivamente creare cluster fra gli stessi dati.

L'esempio che frequentemente si incontra nei manuali di statistica per spiegare il MDS si rifà alle mappe geografiche: applicando il MDS si riesce infatti a passare dalle distanze fra le città alle coordinate che ci permettono di disegnarle nel piano. Negli esempi, la configurazione risultante si avvicina alla distribuzione delle città su una cartina geografica. Talvolta potrebbe tuttavia essere necessario specchiare, traslare o ruotare nel piano le configurazioni grafiche risultanti dall'analisi (e dunque i punti). Questo non ci sorprende poiché gli algoritmi lavorano sulle distanze e non sull'orientamento dei punti. L'interpretazione dei dati quindi può essere fatta a meno di rotazioni, traslazioni o simmetrie (tutte trasformazioni che mantengono le distanze).

L'esempio mostra in maniera semplice e convincente che gli algoritmi computazionali alla base del MDS permettono di passare dal calcolo delle distanze fra le unità statistiche alla determinazione delle coordinate che indicano la posizione dei punti nello spazio multidimensionale (nel piano per $k = 2$) a meno di trasformazioni invarianti rispetto alla distanza (rotazione, traslazione, e così via). Alla base di questo gruppo di tecniche vi è un problema di ottimizzazione: l'obiettivo delle procedure che ci apprestiamo a descrivere è fare in modo che le distanze osservate siano il più possibile simili a quelle calcolate in una soluzione di dimensioni inferiori, vogliamo cioè che la differenza fra le due distanze (osservata e calcolata) sia la minima possibile o, al meglio, nulla.

Prima di addentrarci nella descrizione dell'uso della tecnica, un'osservazione è necessaria: nonostante la rappresentazione grafica sia uno dei risultati più

interessanti delle tecniche di MDS, non sempre queste ultime vengono usate per la visualizzazione; in alcuni casi all'analista interessa soltanto usare il MDS per ridurre la dimensionalità e analizzare in un numero inferiore di variabili la varianza di un insieme di dati.

7.1.1 - Applicazioni delle tecniche di multidimensional scaling

Come sempre, la prima domanda che ci poniamo nel momento in cui ci troviamo ad applicare una tecnica di analisi riguarda gli assunti da verificare.

Nel MDS non abbiamo assunti da esaminare se non le indicazioni relative al numero di casi: esso dovrebbe essere superiore al quadruplo delle dimensioni desiderate. Per una soluzione bi-dimensionale, quindi, dovremmo avere almeno 9 unità statistiche da osservare (Hair et al., 2014, p. 489).

Detto ciò, guardiamo alle fasi di lavoro, ripartendo dal concetto di distanza e da qualche esempio. Immaginiamo di desiderare l'opinione degli studenti su *n* libri di testo o su *n* piattaforme online o su *n* metodologie didattiche.

Possiamo chiedere agli studenti di indicare quanto i differenti libri di testo (così come le piattaforme o le metodologie) sono diversi o simili fra di loro. I dati raccolti in questo caso sono già le distanze e sono denominate *prossimità dirette*. Se dal punto di vista dei calcoli parlare di differenze o di similarità non comporta grandi variazioni, dal punto di vista formale e interpretativo siamo dinnanzi a due casi diversi.

Potremmo altrimenti chiedere di esprimere opinioni su alcune caratteristiche di libri, piattaforme, metodologie. In questo caso non raccoglieremo direttamente le distanze fra gli oggetti ma potremo calcolarle con una matrice di distanze come visto nei capitoli 3 e 6. Si parla in questo caso di *prossimità indirette*. Nel calcolo tipicamente viene utilizzata la distanza euclidea. Sappiamo che usando la distanza euclidea su *n* unità è possibile rappresentarle in *n - 1* dimensioni. Tuttavia, qualora con la distanza euclidea non si ottengano risultati soddisfacenti si può passare alle distanze di Manhattan o altre di cui abbiamo già parlato (cap. 6).

Definite similarità o differenze, prossimità dirette o indirette, l'analista dovrà scegliere l'approccio di analisi dei dati fra due: *classico/metrico* e *ordinale/non* metrico.

Nel primo approccio, gli algoritmi portano a definire le coordinate dei punti sulle *k* dimensioni individuate mantenendo le distanze fra di essi il più possibile

simili alle distanze originali. Nel secondo la definizione delle coordinate (e quindi la disposizione dei punti nello spazio) non avviene sulla base del valore delle distanze ma in base all'ordinamento delle stesse, facendo in modo quindi che se la distanza fra i punti/unità statistiche A e B è numericamente la quarta più alta del dataset, anche la distanza nel grafico fra i punti disegnati sia la quarta più alta senza necessariamente conservarne il valore numerico.

Facciamo un esempio usando un'unica variabile e un'unica dimensione. Immaginiamo di voler rappresentare graficamente la distanza fra i voti all'esame di "Tecniche di analisi dati" di cinque studenti (Tabelle 7.1 e 7.2).

Studenti/esse	A	B	C	D	E
Voto	25	28	26	30	22

Tabella 7.1 - Voti di 5 studenti/esse del corso di "Tecniche di analisi dati" espressi in trentesimi.

	A	B	C	D	E
A	/				
B	3	/			
C	1	2	/		
D	5	2	4	/	
E	3	6	4	8	/

Tabella 7.2 - Matrice delle distanze fra i voti degli studenti in Tabella 7.1 (ricordiamo che la distanza è una funzione definita positiva).

Usando un approccio metrico, nella rappresentazione dei punti (nella scelta delle coordinate in uno spazio a una dimensione) dovremmo rispettare esattamente i valori delle distanze come in Figura 7.1 dove, ad esempio, la distanza AC è pari a 1, la ED a 8 e la AB a 3 così come leggiamo nella matrice delle distanze in Tabella 7.2. Nell'approccio ordinale o non metrico invece distribuiamo i punti corrispondenti agli studenti conservando semplicemente l'ordine delle distanze senza preoccuparci dei valori numerici. In Figura 7.2, infatti la di-

stanza fra i punti (voti degli studenti) E e D è certamente la più ampia ma non è pari a 8, così come la distanza fra A e C è la più piccola ma non è pari a 1. Allo stesso modo le distanze CE e CD sono uguali fra loro, le quarte nell'ordine di grandezza (essendo AC la distanza più piccola pari a 1, BC e BD pari a 2 e quindi seconde nell'ordine, AB e AE pari a 3 e quindi terze nell'ordine) ma non valgono 4 (la distribuzione dei punti nelle Figure 7.1 e 7.2 è puramente esemplificativa e non è il frutto di rigorose procedure di analisi).

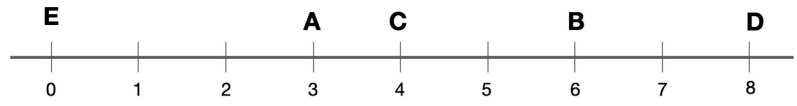


Figura 7.1 - Approccio classico/metrico del MDS: è rispettato il valore numerico delle distanze.

Figura 7.2 - Approccio ordinale/non metrico del MDS: è rispettato l'ordinamento delle distanze.

È facile capire che nel MDS metrico le variabili dovranno necessariamente essere quantitative dato che sono i valori delle distanze ad interessarci; nel MDS ordinale invece, dato che centrale è l'ordinamento, possono essere osservate anche variabili qualitative che introducono nell'analisi valutazioni soggettive come negli esempi precedenti relativi alle opinioni sul confronto di libri di testo o metodologie didattiche.

Nell'approccio classico/metrico, lo scopo è trovare una configurazione su un basso numero di dimensioni in cui le distanze reali e calcolate fra i punti risultino simili e dunque, trovare una nuova matrice distanza con un numero inferiore di dimensioni nella quale gli oggetti conservano la stessa distanza fra loro.

Per tale ragione le coordinate sono calcolate attraverso l'analisi delle componenti principali per le k variabili scelte oppure attraverso il metodo dei minimi quadrati minimizzando in una regressione lineare la differenza fra le distanze reali e quelle teoriche.

Il criterio per misurare la bontà dell'approssimazione è la funzione di *stress* o *fitness* di cui esistono più varianti. La calcoliamo qui come radice quadrata del rapporto fra la sommatoria del quadro delle differenze fra le distanze osservate δ_{ij} e teoriche d_{ij} e la sommatoria del quadrato delle distanze calcolate.

(7.1)

$$Stress = \sqrt{\frac{\sum_{i<j}(d_{ij}-\delta_{ij})^2}{\sum_{i<j}(d_{ij}^2)}}$$

Se le distanze reali e quelle ottenute dalla rappresentazione grafica coincidono, e quindi se la matrice originale delle distanze osservate coincide con quella delle distanze calcolate, il valore dello stress sarà pari a 0.

L'algoritmo di riposizionamento che calcola le coordinate dei punti nel nuovo sistema a k dimensioni lavora per *trial & error* (esistono vari algoritmi ma quasi tutti procedono per tentativi successivi => *trial,* in cui si mantiene di volta in volta la configurazione migliore scartando quelle peggiori => *error*). Semplificando, l'algoritmo imposta i punti, calcola le distanze d_{ij} e la funzione di stress. Se il valore di quest'ultima non è soddisfacente, ripete il ciclo in un processo iterativo: modifica la configurazione dei punti, calcola le distanze d_{ij} e la funzione di stress. La procedura si interrompe quando il valore dello stress raggiunge un valore accettabile oppure si giunge a convergenza di un valore da noi indicato per lo stress o i tentativi oppure quando anche ripetendo le operazioni non si ottengono miglioramenti. Nelle prime configurazioni nelle quali distanze osservate e sperimentali sono molto diverse fra loro, lo stress può tendere a 1. I valori dello stress sono considerati buoni quando sono prossimi allo 0,05; valori attorno a 0,20 indicano un *poor fit*.

Nell'approccio ordinale/non metrico, la configurazione dei punti cercata deve conservare l'ordinamento delle distanze osservate. Calcoliamo quindi, dalle distanze sperimentali d_{ij}, le $\hat{d}_{ij}$ (definite dissimilarità) tali che queste ultime siano nello stesso ordine di rango delle distanze reali (o in quello inverso per le similarità). Per passare dalle d_{ij} alle $\hat{d}_{ij}$, si usa il metodo della regressione monotona dei minimi quadrati (regressione non necessariamente lineare

purché la funzione di regressione sia crescente o decrescente rispettando l'ordinamento delle distanze).

Nella funzione di stress le distanze osservate vengono confrontate con quelle $\hat{d}_{ij}$ ottenute dalla procedure di approssimazione (*fitting*). Si usa in questo caso il *Kruskal's stress, tipo I* (detto semplicemente stress).

(7.2)

$$Kruskal's\ Stress = \sqrt{\frac{\sum_{i<j}(d_{ij}-\hat{d}_{ij})^2}{\sum_{i<j}(d_{ij}^2)}}$$

I valori dello stress possono essere plottati in uno scree plot con il numero di dimensioni per scegliere il numero di dimensioni più adeguate a descrivere il dataset, quando non si può scegliere il numero di dimensioni a priori per un qualche motivo.

Lo stress è uno dei metodi per valutare la bontà di adattamento della configurazione dei punti (e quindi delle coordinate, i valori delle dimensioni) nell'analisi, ma non l'unico.

Altra soluzione è quella di rappresentare in un plot distanze teoriche (ordinate) e distanze osservate (ascisse) dalla distanza minore alla maggiore. Questo grafico è noto come *diagramma di Shepard*. Se il fattore stress è pari a 0, i punti saranno su una retta che passa per l'origine. Una regressione lineare, tra distanze osservate e distanze generate dalla procedura di ottimizzazione, può contribuire a mostrare la bontà di adattamento del posizionamento multidimensionale. Il valore di R^2 ci dice quanto scarto delle distanze viene spiegato in una relazione lineare; se R^2 è prossimo a 1, la configurazione ripropone in maniera soddisfacente le distanze osservate.

Anche nel MDS attribuire nomi agli assi può essere un utile strumento di interpretazione dei risultati. La posizione di un punto in un quadrante positivo/negativo non contraddistingue l'oggetto in base al segno ma all'oscillazione fra le dimensioni individuate sugli assi.

Utilizziamo due dataset che già conosciamo per applicare il MDS: il dataset dei risultati degli esami degli studenti del corso di laurea in Digital Education usato nel capitolo sulla cluster analysis (cap. 6); i risultati a un questionario sull'e-proctoring usato già come esempio nell'analisi delle corrispondenze multiple (cap. 3). In entrambi i casi siamo davanti a prossimità non dirette, dati

raccolti non direttamente come distanze/differenze ma come dati/opinioni in riferimento alle unità statistiche.

Il primo dataset è composto da 6 variabili e 110 osservazioni (ridotte a 101 escludendo gli studenti che non hanno conseguito alcun esame). Usare il MDS in questo caso ci permette di cogliere la struttura che sottende i meccanismi di superamento degli esami a partire dalle distanze fra i voti conseguiti nelle prove. Usiamo in questo caso un MDS classico perché ci troviamo di fronte a variabili metriche e di conseguenza non solo possiamo tenere in considerazione l'ordinamento delle distanze ma anche il valore reale delle stesse. In `R` usiamo `cmdscale`, una delle funzioni della libreria di base `stats`. Per scegliere il numero di dimensioni da inserire nell'analisi, possiamo valutare il valore degli autovalori della matrice di trasformazione. Il numero degli autovalori calcolabili è pari a quello delle unità statistiche del campione. Plottando gli autovalori con i numeri delle dimensioni possibili otteniamo un grafico come in Figura 7.3.

L'analisi di un simile grafico si presta a considerazioni soggettive da parte del ricercatore. Dove ci fermiamo? Se prendiamo un grande numero di dimensioni (per esempio 8 o più) tratteniamo la quasi totalità di informazione ma perdiamo quasi completamente l'obiettivo di ridurre drasticamente le dimensioni del campione. Se prendiamo poche dimensioni (ad es. 2) cogliamo l'obiettivo di una importante riduzione della dimensionalità ma rischiamo di perdere molta informazione utile. Bisogna bilanciare queste due esigenze contrapposte. La regola del pollice è quella di considerare i punti in cui c'è una significativa variazione di pendenza con successiva stabilizzazione o comunque una riduzione meno pronunciata. In questo grafico (Figura 7.3) sembrerebbe un punto tra 3 e 4.

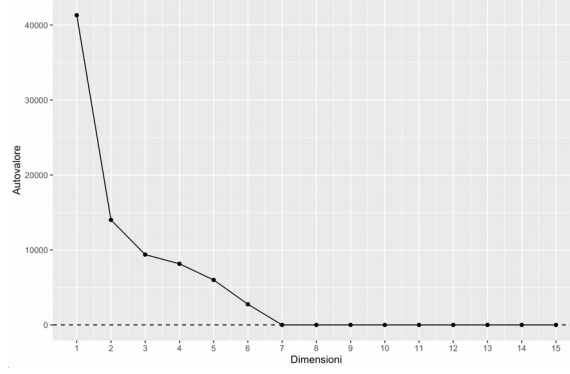


Figura 7.3 - Plot degli autovalori per il dataset relativo ai voti ottenuti a 6 esami da 110 studenti del CdL in Digital Education.

Per semplicità di analisi, come caso esemplificativo, usiamo una rappresentazione a due dimensioni (Figura 7.4). La funzione `cmdscale` a partire dalle distanze calcolate (in questo esempio usiamo distanze euclidee) restituisce le coordinate dei punti o meglio i valori delle due dimensioni in cui si riassumono le 6 variabili di partenza (Tabella 7.3).

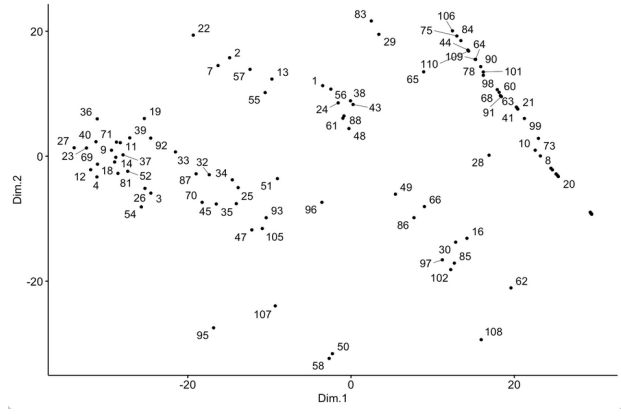


Figura 7.4 - MDS metrico per il dataset degli studenti del CdL in Digital Education: risultati della funzione `cmdscale` (library `stats`) rielaborati in forma grafica attraverso la funzione `ggscatter` (pacchetto `ggpubr`).

	Dim.1	**Dim.2**
1	-3,48	11,29
2	-14,87	15,76
3	-24,53	-5,91
4	-31,14	-3,32
5	24,50	-1,93
...	...	...
100	15,21	15,50
101	14,41	16,83

Tabella 7.3 – Valori delle 2 dimensioni dopo l'applicazione del MDS sul campione degli studenti del CdL in Digital Education con 6 variabili (voti degli esami sostenuti). Nella rappresentazione grafica le colonne Dim.1 e Dim.2 sono le coordinate dei punti nel piano.

L'analisi potrebbe continuare lavorando sul grafico e sulle dimensioni oppure facendo a meno del grafico e lavorando soltanto sui valori delle dimensioni considerate come nuove variabili da utilizzare con una tecnica di analisi diversa.

Un possibile modo per proseguire è quello di applicare una cluster analysis utilizzando come variabili le due dimensioni individuate. In Figura 7.5 è ad esempio graficata la visualizzazione di una cluster analysis non gerarchica, metodo kmeans con $k = 6$. Vediamo già dal grafico che nei gruppi 1 e 3 i punti sono più vicini fra di loro (meno distanti e più simili), che il gruppo 5 è il più ridotto e il primo in rosso quello più numeroso. Altre osservazioni possono essere aggiunte a partire da un'analisi descrittiva dei cluster: scopriremo che al primo gruppo appartengono gli studenti che hanno sostenuto più esami con voti più alti, ai cluster 2 e 4 studenti con un andamento medio e ai restanti tre gruppi studenti che presentano delle difficoltà. Queste informazioni potrebbero essere utili per mettere in atto strategie di tutorato personalizzate per gruppo oppure per collegare i successi accademici a fattori altri come dati anagrafici, numeracy e literacy, livelli di motivazione e così via.

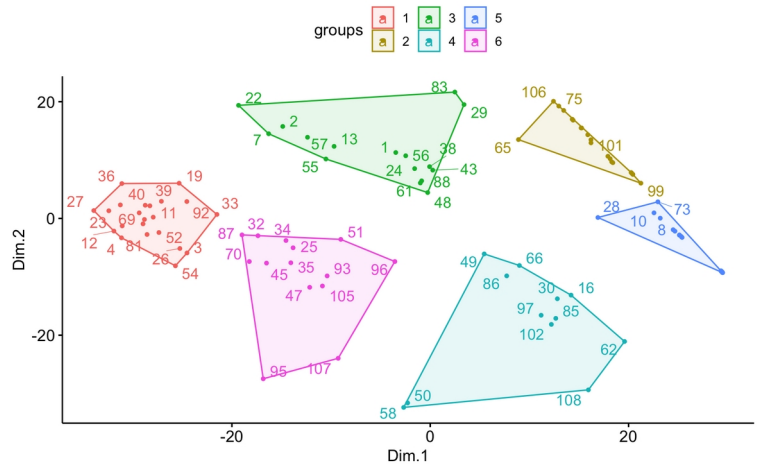


Figura 7.5 - Cluster analysis (metodo kmeans, $k = 6$) successiva al MDS metrico per il dataset degli studenti del CdL in Digital Education.

Per verificare la bontà dei risultati utilizziamo il diagramma di Shepard, nel quale ordiniamo e grafichiamo le distanze teoriche rispetto a quelle reali (Figura 7.6). In questo caso notiamo che i punti si addensano sulla retta bisettrice che passa per lo 0, anche se vi è una certa dispersione al di sotto della retta.

Ossia le distanze calcolate tendono a essere minori di quelle reali. Questo è soltanto un effetto dell'algoritmo utilizzato per minimizzare la differenza presente nell'equazione (7.1). Del resto trattandosi di un quadrato di differenze tra due distanze si otterrebbe lo stesso risultato per valori simmetrici rispetto alla bisettrice.

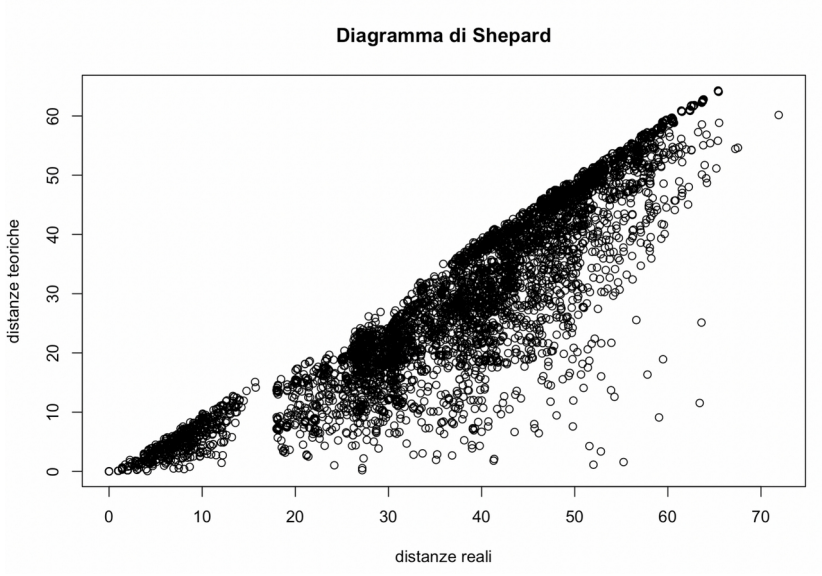


Figura 7.6 - Diagramma di Shepard per MDS metrico (i valori nel plot sono stati calcolati usando la funzione `Shepard`, libraria: `MASS`).

Su questi stessi dati è possibile applicare anche un MDS non metrico. Poiché l'approccio classico è un sottogruppo dell'approccio ordinale, otterremmo una configurazione simile. Nel nostro caso specifico le due configurazioni (Figure 7.4 e 7.7) tendono a convergere a meno di una trasformazione di simmetria/mirroring.

Proponiamo uno scaling multidimensionale ordinale sul secondo dataset citato, composto da variabili ordinali e riferito alle opinioni (espresse in una scala a 4 livelli) degli studenti sugli elementi migliorati dall'uso dei sistemi di e-proctoring nello svolgimento degli esami, in particolare: concentrazione, attenzione, gestione del tempo, ansia, comprensione, motivazione. Il MDS applicato in questo caso utilizza un numero inferiore di dimensioni rispetto alle 6 variabili per descrivere le opinioni degli studenti sui sistemi di e-proctoring. Le dimensioni identificate contengono le opinioni sui sei elementi e le riuniscono.

Esistono più funzioni per svolgere un MDS ordinale. Molto utilizzata è `isoMDS` della libreria `MASS`. Questa, tuttavia, funziona solo in caso di distanze non nulle e positive. Di conseguenza esclude la possibilità di conservare in uno stesso campione unità statistiche con comportamenti identici. Come alternativa abbiamo usato nell'esempio precedente (Figura 7.7) e useremo anche in questo che segue la funzione `metaMDS` del pacchetto `vegan`. Mentre il software calcola la soluzione che meglio ottimizza i dati nel campione nel numero di dimensioni indicate fra gli argomenti, il sistema ci mostra il valore di stress per ciascun tentativo di ottimizzazione dei dati. Viene definita alla fine una configurazione con un univoco valore di stress.

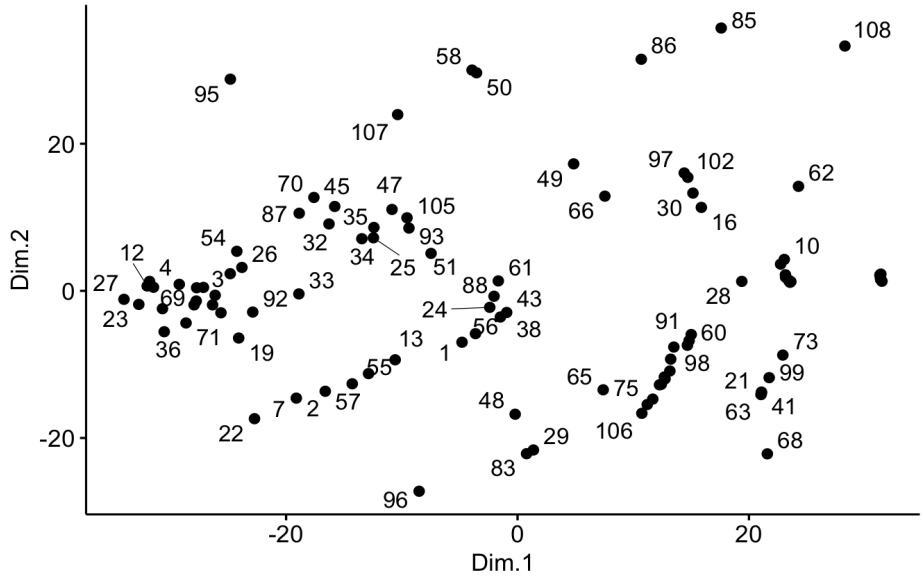


Figura 7.7 - MDS non metrico per il dataset degli studenti del CdL in Digital Education: risultati della funzione `metaMDS` (libreria: `vegan`) rielaborati in forma grafica attraverso la funzione `ggscatter` (libreria: `ggpubr`). La configurazione tende a convergere con quella in Figura 4.4 a meno di una simmetria/mirroring.

In Figura 7.8 vediamo i valori dello stress calcolati su un massimo di 8 dimensioni (dato che, come abbiamo detto, plottare i valori dello stress ottenuti da più soluzioni con numeri diversi di dimensioni è una tecnica per scegliere il numero di dimensioni da utilizzare nell'analisi).

Questo è un caso di difficile interpretazione. Qual è il numero ottimale di dimensioni da selezionare? Non esiste un punto in cui sia evidente una variazione di tendenza. Si propenderebbe tra 3 e 4 bilanciando la necessità di ridurre il numero di dimensioni con quella di trattenere la maggiore quantità di informazione. Se consideriamo le prime 4 coordinate, queste saranno quelle su cui applicare, ad esempio, una cluster analysis.

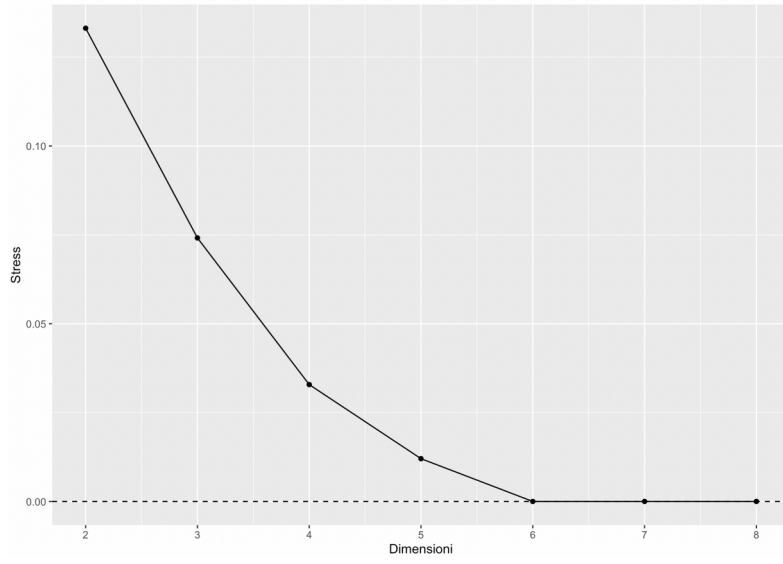


Figura 7.8 - Plot dimensioni/stress per individuare il numero di dimensioni da utilizzare nell'analisi.

Tuttavia per una rappresentazione grafica visualizziamo i risultati della soluzione a 2 dimensioni (Figura 7.9). Calcoliamo poi i cluster con kmeans e $k = 3$ (Figura 7.10).

Lo stress calcolato per la soluzione a 2 dimensioni è pari a 0,13, un valore che comincia ad essere meno rassicurante rispetto alle indicazioni che ci siamo dati (*good = 0.05*) ma accettabile. Per ottenere configurazioni che meglio riescano a descrivere il dataset, oltre a usare soluzioni a più dimensioni, potremmo fare ulteriori tentativi di analisi modificando la tipologia di distanza che abbiamo calcolato (passare dalla distanza euclidea a quella di Minkowski o Manhattan). Il diagramma di Shepard (Figura 7.11) e i valori di R^2, tuttavia, indicano che il modello fitta bene con i risultati ($R^2 = 0{,}982$).

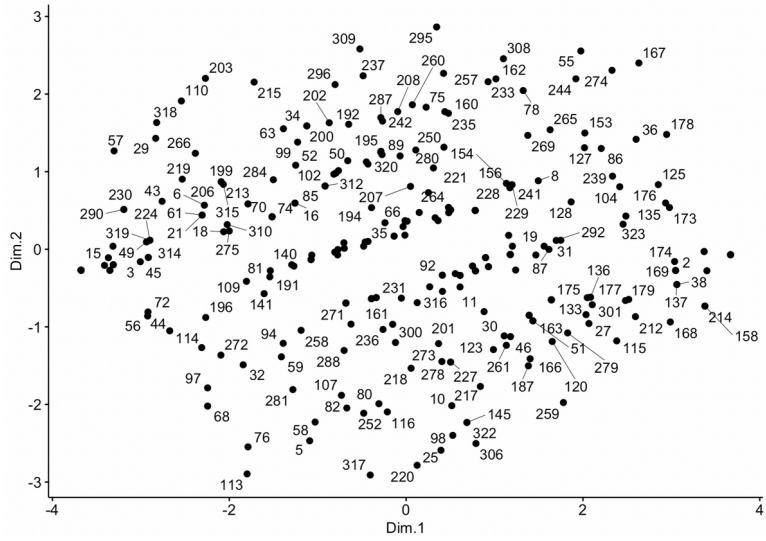


Figura 7.9 - MDS non metrico per il dataset sulle opinioni relative ai sistemi di e-proctoring.

Figura 7.10 - Cluster analysis (kmeans con *k = 3*) successiva al MDS non metrico del dataset sui sistemi di e-proctoring.

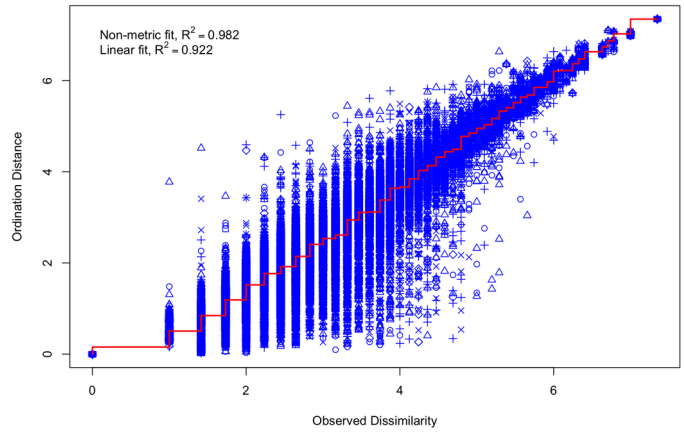


Figura 7.11 - Diagramma di Shepard per MDS non metrico (funzione `stressplot`, libreria: `vegan`).

7.2 - L'uso del multidimensional scaling nella ricerca educativa

Un primo esempio dell'uso del Multidimensional Scaling in ambito educativo viene fornito nell'articolo "Building the role of ICT coordinators in primary schools: A typology based on task prioritisation" di José C. León-Jariego e colleghi (2020). Gli autori riflettono sulla figura dell'*ICT coordinator* definita nelle scuole primarie dell'Andalusia dal 2003, figura che per alcune caratteristiche ci riporta a quella italiana dell'animatore digitale individuata dal Piano Nazionale Scuola Digitale (MIUR, 2015). Il MDS viene usato sia per comprendere la percezione delle funzioni di questa figura da parte degli stessi insegnanti che per identificare gruppi di ICT coordinator con caratteristiche simili. Agli ICT coordinator andalusi è stato somministrato un questionario a cui ha risposto all'incirca la metà degli intervistati (101). Nella seconda sezione del questionario è stato chiesto di valutare l'importanza di 8 funzioni relative al ruolo indagato definite in base a precedenti framework scelti dagli autori.

In una prima fase quindi la tecnica è stata usata per visualizzare la distribuzione delle funzioni nel piano cartesiano (algoritmo ALSCAL). La Figura 7.12 riporta i risultati di una soluzione con stress value pari a 0,17 e $R^2 = 0{,}998$. Nella rappresentazione grafica le due dimensioni degli assi riflettono:

- sulle ascisse: il contesto, dalle funzioni relative alla classe a quelle relative alla comunità scolastica;
- sulle ordinate: la complessità, da azioni di supporto tecnico alla diffusione di esperienze innovative.

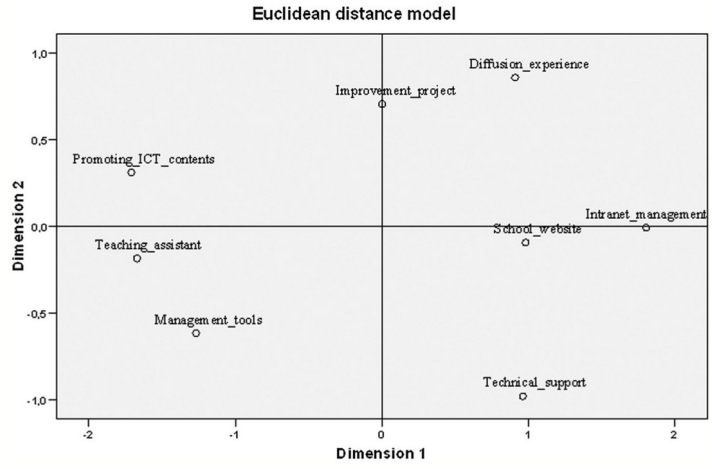


Figura 7.12 - Mappa percettiva delle funzioni dell'ICT coordinator (León-Jariego et al., 2020, p. 7).

In un secondo momento la mappa è stata realizzata raffigurando come punti i rispondenti (algoritmo INDSCAL); la cluster analysis gerarchica (metodo di Wards) condotta sulle coordinate ottenute nel MDS ha identificato una soluzione con tre gruppi di coordinatori con simili punti di vista sulle proprie funzioni (Figura 7.13). Considerando l'interpretazione data agli assi, vediamo che:

- il cluster 1 (*n = 13*) è nel quadrante attività d'aula/compiti complessi. La funzione percepita dai coordinatori in questo quadrante è quella di promuovere l'uso delle tecnologie in classe;
- il cluster 2 (*n = 49*), il più numeroso, è nel quadrante attività d'aula/compiti semplici. La funzione percepita è quella di supportare l'uso delle tecnologie in classe;

- il cluster 3 (*n = 11*) è nel quadrante scuola/compiti complessi. In questo caso la funzione percepita è la pianificazione e gestione dei dispositivi tecnologici nella scuola.

La diversità nel modo di pensare il proprio ruolo da coordinatori è stata ulteriormente studiata attraverso la somministrazione di interviste semi-strutturate ad alcuni coordinatori estratti da ciascun cluster con lo scopo di giungere a suggerimenti operativi per le policy e per i percorsi di formazione degli ICT coordinator.

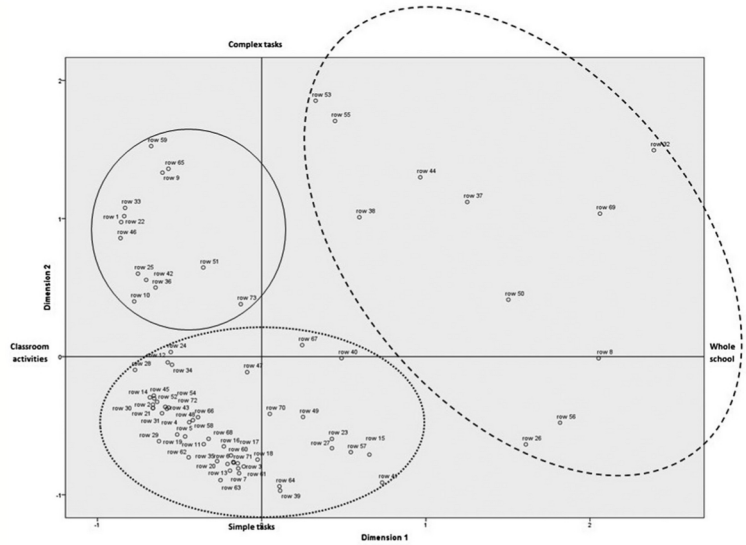


Figura 7.13 - Risultati della cluster analysis sulle posizioni degli ICT coordinator (León-Jariego et al., 2020, p. 11).

Andiamo oltre con lo studio di Ernesto López-Gómez e colleghi (2020) che mostra l'uso del MDS in uno studio psicometrico. Nel caso qui citato, gli autori intendono ragionare sul tema del tutorato in ambito universitario. L'orientamento, il placement, la pianificazione dello studio, il mentoring, il supporto tecnico sono alcuni dei servizi di tutorato più evidenti e diffusi nei contesti dell'alta formazione. Si tratta di attività dilazionate nel tempo per l'intera durata del percorso universitario e che possono essere legate a tre aree: l'area personale-sociale, l'area accademica e l'area professionale. Sulla base di queste aree e dello studio della letteratura, gli autori hanno costruito un questionario di 51 domande (17 per ciascuna area) sottoposte poi a 18 esperti per una validazio-

ne di coerenza, rilevanza, chiarezza e completezza secondo il metodo Delphi. Le domande chiedevano quanto ciascun item del modello di tutorato proposto risultasse importante (come dovrebbe essere) e applicato (come è) con valutazioni in una scala Likert a 6 livelli. Hanno risposto al questionario 569 professori e 679 studenti. I valori dell'alpha di Cronbach hanno confermato la consistenza interna dell'intero questionario e di ciascuna sezione (valori superiori a 0,90). L'analisi psicometrica del questionario è stata condotta usando in due tempi diversi la cluster analysis e il MDS. Dalla cluster analysis è stata presa in considerazione una soluzione con 6 cluster che dividono in due le tre aree con cui è stato costruito il questionario. Nel MDS PROXSCAL con stress (Normalized Raw Stress) pari a 0,06820, la vicinanza fra gli item nel piano cartesiano (Figura 7.14) ha permesso di rilevare 5 aree di cui quella con i punti in rosso potrebbe essere ulteriormente scomposta in due. La distribuzione degli item in gruppi coincide nei risultati delle due tecniche. La suddivisione in due parti degli item di ciascuna area mostra una distinzione fra item che fanno riferimento a un supporto pratico su servizi, attività, regolamenti, prospettive lavorative e item che si riferiscono alla funzione di consiglio, guida e coaching. Lo studio rileva le funzioni del tutorato all'università aspirando in seguito a costruire un modello.

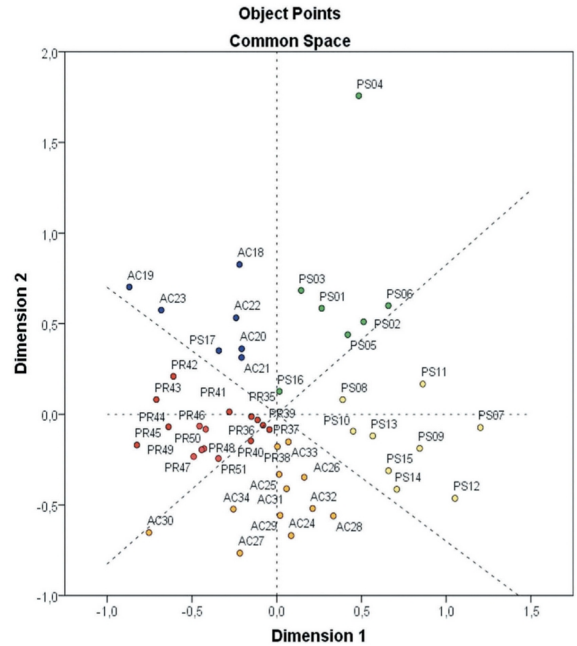


Figura 7.14 - MDS sugli item in uno studio psicometrico sui servizi di tutorato (López-Gómez et al., 2020, p. 622).

Il MDS è usato in un processo di concettualizzazione strutturata denominato *Group Concept Map* (Trochim, 1989). Si tratta di un approccio di ricerca misto, che utilizza pertanto strumenti qualitativi e quantitativi di indagine. In estrema sintesi, nel processo di Group Concept Map (GCM) le affermazioni derivate da un brainstorming fra esperti su un determinato tema vengono valutate dagli stessi, analizzate e visualizzate usando il MDS e la cluster analysis. Lo schema in Figura 7.15 riassume le sei fasi che compongono il processo: dopo aver selezionato i partecipanti, nella fase 1 si definisce il focus dell'indagine e i criteri di valutazione del tema. Gli esperti selezionati sono chiamati poi a partecipare a un brainstorming in cui ciascuno in maniera indipendente propone delle affermazioni specifiche sul tema (fase 2). La lista definitiva di tutte le affermazioni (che possono superare il centinaio) viene riordinata dagli esperti e per ciascuna affermazione a essi è chiesto di esprimere una valutazione in una scala Likert sui criteri identificati per il rating (fase 3). I risultati registrati vengono quindi analizzati (e di seguito rappresentati) attraverso un MDS non metrico e una CA gerarchica: otterremo in conclusione quindi una mappa dei punti delle affermazioni come risultato del MDS e una mappa risultante dalla cluster analysis con la suddivisione delle affermazioni in gruppi. Le analisi prodotte sui criteri di rating sono aggiunte alle precedenti e visualizzate in altri grafici (fase 4). Si procede poi all'interpretazione e all'uso dei risultati (fase 5 e 6).

Questa modalità di indagine riesce a utilizzare il linguaggio dei partecipanti nella definizione dei temi di ricerca e ha il vantaggio di restituire una visualizzazione grafica sulle questioni analizzate. Tool online (ad es. groupwisdom™) facilitano il processo di raccolta dati e analisi.

Martine Schophuizen e colleghi (2018) usano il GCM per valutare opportunità e sfide nell'attuazione di progetti di Open Online Education (OOE) nell'alta formazione nel contesto dei Paesi Bassi dove 12 programmi sono stati finanziati dal governo fra il 2015 e il 2018. Gli autori hanno scelto 59 esperti olandesi nel settore dell'OOE contattati attraverso canali come LinkedIn, Twitter, contatti personali, siti web. È stato dato loro un mese per completare questa frase: "La mia istituzione ha la seguente sfida o possibilità sull'OOE". Sono state collezionate 149 affermazioni ridotte dai ricercatori a 106. Queste sono state riordinate e valutate dagli esperti in base a due criteri: *importanza* (quanto è importante questa specifica affermazione per realizzare le OOE?) e *influenza* (quanta influenza ha l'istituzione di cui fai parte su questa specifica affermazione?). Nel tool online usato per l'analisi, un bridging value viene assegnato a ciascuna affermazione, con valori fra 0 e 1. Valori simili portano a disporre i punti (e le affermazioni) più vicini in un MDS. Lo stress che nel Group Concept Map viene

considerato adeguato con valori compresi fra 0,205 e 0,365 conferma che la mappa dei punti ottenuti dal MDS è una buona rappresentazione del riordinamento delle affermazioni.

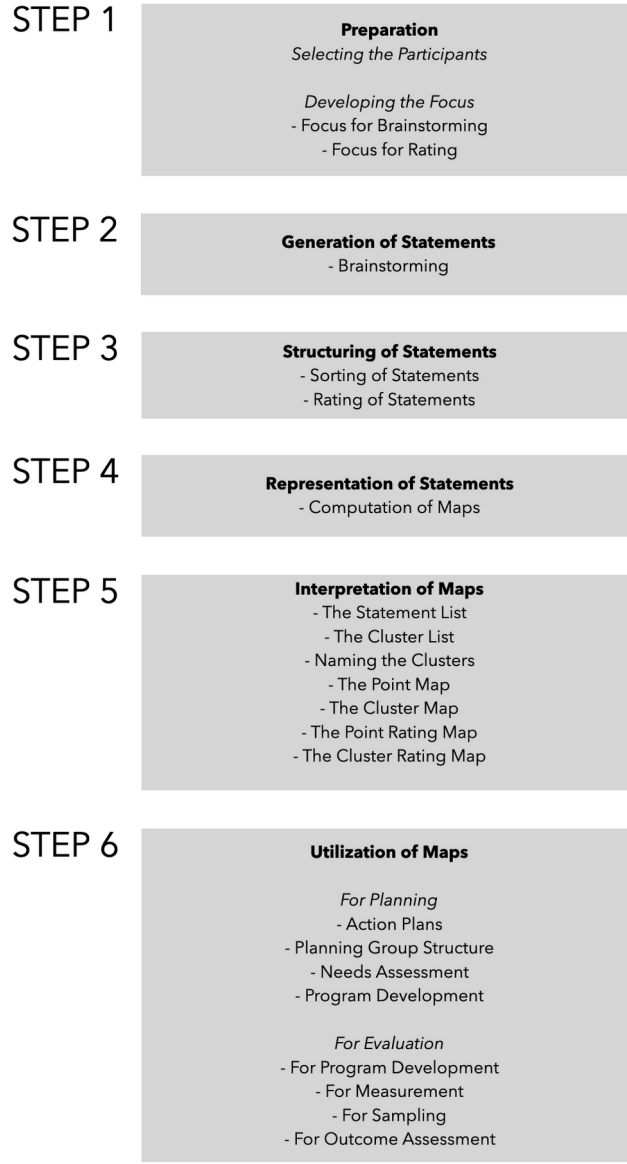


Figura 7.15 - Processo di Concept Mapping (Trochim, 1989, p. 3).

La Figura 7.16 mostra gli 8 cluster ottenuti fra le affermazioni in seguito alle procedure di riordinamento da parte degli esperti. L'area rossa contiene i cluster presentati nella maggior parte dei casi come sfide; quella verde i gruppi di item visti come opportunità. In arancione sono le affermazioni e i relativi cluster identificati come un misto di opportunità e sfida.

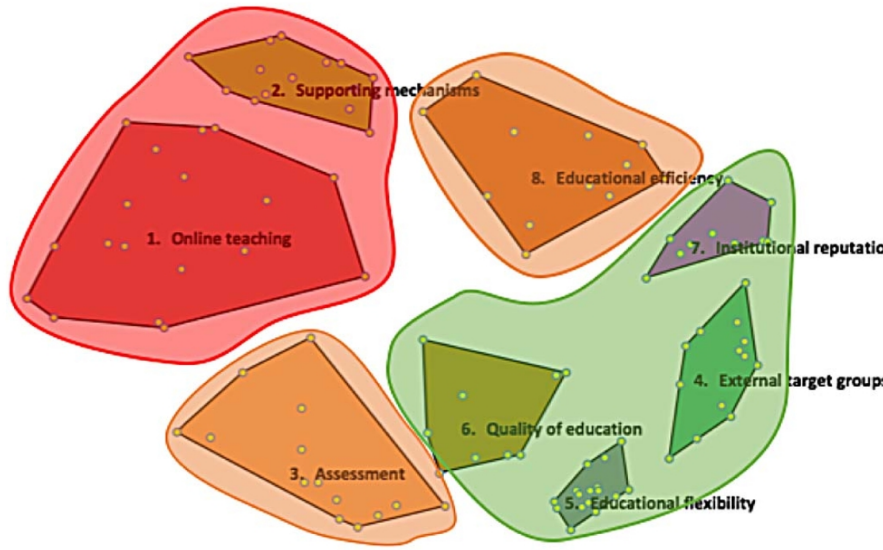


Figura 7.16 - Cluster analysis su sfide e opportunità legate all'attuazione di progetti di Open Online Education (Schophuizen et al., 2018, p. 6).

Viene restituito anche un grafico del tipo *go-zone* che mostra la distribuzione delle affermazioni in base ai criteri scelti (Figura 7.17).

Lo stesso procedimento è usato nello studio in un progetto europeo di Helen Hynes e colleghi (2015) il cui scopo è quello di definire un curriculum e i risultati di apprendimento per gli studenti di medicina sulle attività di passaggio delle consegne (handoff training). 127 esperti (accademici, medici e infermieri a vario titolo coinvolti nella formazione dei medici) sono stati invitati a partecipare per email, di questi 22 hanno concluso tutte le fasi dell'intero processo di indagine.

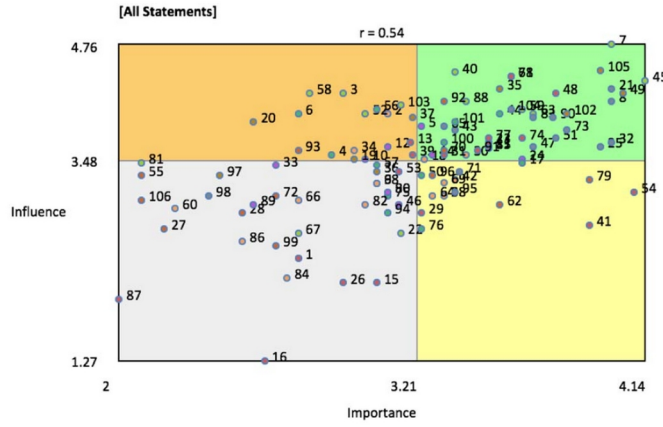


Figura 7.17 - Grafico go-zone nel quale sono rappresentate le affermazioni sui progetti di OOE sugli assi in base alle valutazioni su due indicatori: importanza e influenza (Schophuizen et al., 2018, p. 6).

La frase che nel brainstorming durato due settimane è stato chiesto di completare è: "Un risultato di apprendimento specifico del modulo handoff è". 4 esperti scelti dal campione hanno selezionato la lista finale di circa 100 affermazioni da riordinare e valutare in tre settimane sulla base di due criteri: *importanza* e *difficoltà* nel raggiungimento degli obiettivi. Anche in questo caso, lo stress ha un valore compreso nel range consentito (0,338) e per questo la soluzione è considerata una buona rappresentazione del riordinamento delle affermazioni. Fra le soluzioni con un numero di cluster compreso fra 5 e 16, le più idonee, a detta dei 4 esperti, sono risultate quelle con 9 e 10 cluster. I cluster, con i nomi ad essi assegnati rappresentano gli elementi chiave su cui costruire un efficace curriculum sull'handoff training. L'analisi è arricchita dalle valutazioni degli esperti su obiettivi più importanti e più difficili da raggiungere.

Maren Scheffel, Hendrik Drachsler e colleghi (2014) utilizzano le procedure del Group Concept Mapping per selezionare indicatori di qualità per i learning analytics tools. 74 partecipanti allo studio hanno proposto 92 idee completando l'espressione "Un indicatore di qualità specifico per valutare gli effetti dei learning analytics è". Alcuni esperti hanno rivisto l'iniziale elenco giungendo a definire 103 indicatori da sottoporre al riordinamento e alla valutazione in una scala da 1 a 7 in base all'*importanza* e alla *fattibilità*. Gli autori hanno sintetizzato gli indicatori in 8 cluster dalle procedure di riordinamento: open access, privacy, acceptance & uptake, learning outcome, teacher awareness, learning

performance, learning support, student awareness. Dall'analisi della letteratura e dei risultati in riferimento ai due criteri di rating (importanza e fattibilità) è stato infine individuato un framework composto da cinque criteri per la valutazione dei tool di LA: obiettivi, supporto a docenti e studenti nel processo di apprendimento, risultati del processo formativo ottenuti dal tool, caratteristiche dei dati, aspetti organizzativi.

CAPITOLO 8

ITEM ANALYSIS

Al termine del capitolo, il lettore sarà in grado di:

- *descrivere le procedure di item analysis;*
- *distinguere le caratteristiche degli approcci della Classical Test Theory e della Item Response Theory;*
- *elencare esempi della ricerca educativa nei quali è stata applicata l'item analysis.*

8.1 – Item Analysis

In questo capitolo ci occupiamo di tecniche di analisi statistica che hanno a che vedere con la valutazione, tema centrale nella discussione sulla didattica. Modelli di progettazione che vedono la valutazione strettamente connessa agli obiettivi di apprendimento formulati nei percorsi formativi (Biggs & Tang, 2011; Wiggins & McTighe, 2005) ne danno la giustificazione: l'efficacia di un percorso didattico può essere provata quando attraverso adeguati strumenti di valutazione siamo in grado di dire se gli studenti hanno raggiunto i comportamenti attesi espressi come obiettivi educativi del corso. Valutare ci permette di analizzare non solo gli apprendimenti degli studenti ma anche l'efficacia del percorso formativo predisposto. Gli strumenti per la valutazione andrebbero disegnati in base al tipo di apprendimenti che devono misurare: la misurazione è parte del processo di valutazione.

Come altri ambiti scientifici, per esempio la fisica e la statistica, ci hanno insegnato nell'ultimo secolo, la misurazione è affetta da un errore casuale tale che possiamo dire che ciascuna grandezza oggetto della misurazione è comunque affetta da un errore casuale, ε, che la rende una variabile casuale con tutte le conseguenze e le proprietà che ciò comporta.

Gli strumenti quantitativi di valutazione e in particolare i questionari a risposta chiusa sono quelli a cui applichiamo le tecniche descritte in questo capitolo, tecniche comprese nel termine *item analysis* le cui basi sono da rintracciare

nell'analisi psicometrica dei quesiti. Da alcuni decenni, tali tecniche sono state applicate nella produzione di scale nell'ambito della ricerca ma anche nella valutazione o meglio nei processi di misurazione degli apprendimenti nei quali usare prove quantitative è diventata consuetudine. Esse sono più veloci da somministrare, sono vantaggiose su grandi numeri di studenti, permettono di indagare un numero elevato di argomenti e sono ritenute più oggettive.

L' item analysis fornisce principalmente informazioni sulle singole domande, tuttavia alcuni degli indicatori che vedremo rilevano elementi che caratterizzano l'intero set di domande.

Perché usare queste tecniche di analisi? Attraverso di esse possiamo in primo luogo stimare le abilità degli studenti come tratti latenti di cui abbiamo già parlato nel cap. 3 e di cui non possiamo fare delle misurazioni dirette. Le risposte fornite nelle prove di valutazione dagli studenti da sole non bastano a rilevare la comprensione, l'acquisizione di un concetto. La risposta a ciascuna domanda è soltanto una manifestazione di una parte del processo di apprendimento (De Luca & Lucisano, 2011). In secondo luogo, possiamo utilizzare i risultati ottenuti in queste procedure di analisi per esaminare le prove di valutazione e verificare se le domande sono scritte correttamente, se sono troppo semplici, se aiutano a distinguere fra gli studenti preparati e quelli no, se è necessario modificare il peso che ciascuna domanda ha nel calcolo del punteggio complessivo, se tutte le domande fanno riferimento a uno stesso tratto latente, se è possibile trovare un modello che ci dica con quale probabilità gli studenti possono rispondere correttamente a una domanda in base alle loro abilità e alle caratteristiche delle domande stesse.

Su quali dati si lavora? Dopo aver prodotto un questionario, è buona regola proporlo a un campione di studenti per verificarne l'efficacia. Nella pratica capita che non sempre si riesca a realizzare una fase di tryout e quindi la prima somministrazione del questionario ne diventa la prova sperimentale. In queste fasi iniziali di utilizzazione di una scala o di un questionario può essere molto utile usare l'item analysis per i motivi che ci dicevamo. Più in generale, comunque, ogni volta che vengono collezionate risposte a una serie di item, le tecniche di cui ci apprestiamo a parlare possono essere usate per stimare l'abilità degli studenti e verificare le caratteristiche delle domande.

Gli approcci fondamentali di item analysis sono due: la *Classical Test Theory* (CTT) e la *Item Responses Theory* (IRT) proposta da Rasch negli anni Sessanta. Nel primo modello (*classical* e dunque tradizionale) si identificano alcuni indicatori per valutare gli item. L'abilità dello studente è la somma delle risposte corrette che dà al questionario. Nel secondo, sono centrali le teorie legate al

calcolo delle probabilità e alla ricerca di un modello che possa stimare l'abilità dello studente tenendo in considerazione le caratteristiche delle domande.

Entriamo nel dettaglio.

Nella Classical Test Theory (CTT) ad ogni domanda attribuiamo due indicatori, definiti *Difficulty Value = DV* e *Discriminatory Index = DI*.

L'indice di difficoltà, *DV*, in maniera controintuitiva rispetto al nome, risponde alla domanda: quanto la domanda è facile? Uno dei metodi più noti (e anche più semplice) per calcolare questo indice è quello di determinare la percentuale di studenti che ha risposto correttamente alla domanda.

(8.1)

$$Difficoltà = \frac{N.Risposte\ corrette}{N.Risposte\ totali}$$

Altra procedura, più raffinata, è quella di escludere dal campione gli studenti che hanno acquisito un punteggio attorno alla media all'intero questionario. Disponiamo quindi i punteggi in ordine decrescente e consideriamo nel calcolo del *DV* soltanto i punteggi che si trovano nelle due code. Si considera solitamente il 27% dei punteggi più alti e dei punteggi più bassi. La scelta dei due gruppi potrebbe essere adeguata ad altre percentuali o anche al primo/terzo quartile, molto dipende dai dati su cui stiamo lavorando. Il numero dei soggetti che hanno risposto correttamente in ciascuna delle due fasce viene poi suddiviso per il numero dei soggetti che fanno parte dei due gruppi. L'espressione che si utilizza è la seguente:

(8.2)

$$Difficoltà = \frac{RH + RL}{(NH + NL) - NR}$$

dove *RH* (*Right High*) è il numero di quanti rispondono correttamente alla domanda nella fascia superiore, *RL* (*Right Low*) è il numero di chi risponde correttamente nel gruppo con i punteggi più bassi, *NH* (*Number High*) è la numerosità del gruppo dei "migliori", *NL* (*Number Low*) è il numero dei "peggiori", *NR* (*Not respondent*) è il numero di quanti non hanno risposto alla domanda.

I valori dell'indice di difficoltà, come è intuibile, oscillano fra 0 e 1: 0 nel caso in cui nessuno abbia risposto correttamente alla domanda, 1 nel caso in cui tutti abbiano risposto correttamente. Quando il valore tende a 0, la domanda è difficile; se si avvicina a 1, la domanda è facile.

Solitamente domande con *DV* inferiore a 0,20 sono considerate difficili, domande con *DV* superiore a 0,80 troppo semplici. Indici di difficoltà considerati efficaci sono compresi fra 0,40 e 0,60.

Z. A. Ashraf e K. Jaseem (2020) riportano diversi criteri elaborati nel tempo per identificare i valori desiderabili del *DV*: attorno allo 0,5; fra 0,30 e 0,70; fra 0,40 e 0,70. Uno dei metodi che gli autori riportano, in base al quale definire i valori soglia per l'indice di difficoltà tiene in considerazione il numero di opzioni di risposta: al 100% dei rispondenti si somma la percentuale per la scelta di una singola domanda e in seguito si divide il risultato a metà. Quindi, ad esempio per una domanda con 4 opzioni di risposta dovremo sommare il 100% al 25% (probabilità di scegliere una sola opzione) e dividere per due: 0,625 rappresenta quindi il *DV* adeguato per i quesiti in esame.

Alle motivazioni statistiche sulla scelta dei range di accettabilità della difficoltà, vanno aggiunti anche elementi riconducibili agli obiettivi educativi e alle questioni didattiche. Spesso, ad esempio, all'inizio di un test si inseriscono volutamente domande semplici per far acquisire agli studenti confidenza e fiducia per completare la prova. O ancora, immaginiamo di voler essere sicuri che tutti gli studenti abbiano ben chiari alcuni concetti di base prima di andare avanti con le attività didattiche: in questo caso porremo domande di base con la consapevolezza che potrebbero risultare semplici. Non sarà questa la scelta in un test di ingresso per l'accesso a un corso di laurea o in una valutazione sommativa intermedia o di fine anno dove invece prepareremo domande auspicabilmente con indici di difficoltà appropriati.

Il secondo indice nella CTT, la discriminatività *DI*, risponde alla domanda: quanto questa domanda distingue gli studenti che rispondono correttamente alle altre domande del questionario da quelli che rispondono in maniera sbagliata? Guardiamo ai casi estremi: se tutti gli studenti rispondono in maniera corretta o al contrario in maniera errata a un determinato quesito, la domanda non ci aiuterà a distinguere gli studenti preparati da quelli meno preparati, non sarà cioè in grado di far trasparire il tratto latente, l'abilità di cui parlavamo.

Per calcolare l'indice *DI*, possiamo, come per la difficoltà, individuare le fasce di studenti con punteggi più alti e più bassi (27%) e calcolare la differenza fra quanti hanno risposto bene nel gruppo dei "migliori" e quanti hanno rispo-

sto correttamente nel gruppo dei "peggiori" così come nell'espressione che segue (8.3) nella quale il significato dei fattori è lo stesso dell'espressione (8.2):

(8.3)

$$Discriminatività = \frac{RH - RL}{(NH + NL) - NR}$$

Altra soluzione spesso scelta per il calcolo è quella di considerare ***DI*** come correlazione fra la risposta a un item e il punteggio totale dello studente agli altri item (correlazione punto-biseriale). Useremo allora la seguente espressione:

(8.4)

$$Discriminatività = \frac{(\bar{x}_e - \bar{x}_t)}{\sigma} \sqrt{(\frac{p}{1-p})}$$

dove $\bar{x}_e$ è la media dei risultati degli studenti che hanno risposto correttamente all'item, $\bar{x}_t$ è la media totale dei risultati, σ è la deviazione standard dei risultati al test, ***p*** la percentuale di risposte corrette (***DV***).

L'indice ***DI*** assume come la correlazione valori fra -1 e 1, vale 0 se tutti rispondono in maniera corretta o errata (siamo nel caso in cui le risposte all'intero questionario e le risposte all'item in analisi sono indipendenti). Assumerà valori positivi se tendenzialmente gli studenti preparati rispondono correttamente; valori negativi se a rispondere correttamente sono gli studenti meno preparati che non hanno punteggi molto alti nel resto del questionario.

Gli item si considerano molto buoni quando ***DI > 0,40***; ragionevolmente buoni ma da migliorare per ***DI*** compreso fra 0,30 e 0,39; necessariamente da migliorare se ***DI*** è compreso fra 0,20 e 0,29; scadenti, da eliminare o rivedere se ***DI < 0,19*** (Ebel & Frisbie, 1991, p. 232). Valori del ***DI*** superiori a 0,90 sono da verificare poiché estremamente elevati e poco verosimili.

Un terzo indicatore, non sempre citato ma che può essere utile inserire in questa discussione, è l'efficacia dei distrattori ossia delle risposte errate (***DE***): chiunque abbia dovuto produrre dei quesiti a risposta multipla, sa che scrivere le opzioni errate è la fase più impegnativa. Motivo che spiega anche perché molto spesso si scelga di scrivere (erroneamente) domande in forma negativa (es. quale delle seguenti affermazioni è falsa? Quale dei seguenti NON è … ?).

Se un distrattore non viene scelto, potrebbe essere debole, ovvio, mal formulato. Questa formulazione inadeguata comporta delle conseguenze sui precedenti due indici: escludendo un'opzione di risposta, la domanda risulta di più semplice soluzione perché aumentano le probabilità di rispondere correttamente. I criteri standard prevedono che per considerare un item funzionale e ben scritto, almeno il 5% dei rispondenti dovrebbe selezionarlo come risposta.

Cosa facciamo una volta che abbiamo trovato degli indici non adeguati per alcuni degli item nel nostro questionario? I tre indici descritti vanno considerati contestualmente per ogni domanda dato che ciascuno di essi fornisce delle informazioni indispensabili per valutare la qualità del quesito nella sua totalità. Rileggendo la domanda insieme ad un esperto dei contenuti possiamo capire se c'è un modo per migliorarla o se la domanda va eliminata dal paniere. Frequentemente se il *DI* è basso si preferisce eliminare la domanda o sottoporla a un profondo ripensamento. Se il quesito è troppo semplice (*DV* alto), le alternative di risposta potrebbero non essere sufficientemente sfidanti oppure ovviamente errate perché magari contengono quei termini come "solo", "sempre" o opzioni come "tutte le precedenti" oppure "nessuna della precedenti" che funzionano come dei campanelli d'allarme nella lettura; in una domanda difficile (*DV* basso) potrebbe per esempio esserci un'alternativa di risposta verosimile o parzialmente vera che raccoglie un numero alto di preferenze da parte degli studenti. Domande che non discriminano (*DI* basso) potrebbero essere scritte in forma negativa. Questi sono soltanto alcuni esempi di formulazioni non corrette delle domande, si veda Domenici e colleghi (2021, pp. 148-150) per una rassegna completa.

L'indice di difficoltà e l'indice di discriminatività sono due indicatori che, come abbiamo detto, fanno riferimento a singoli item. Su un questionario intero invece, su una scala completa, possiamo calcolare l'indice α di Cronbach che abbiamo introdotto a proposito dell'analisi fattoriale. L'indice compreso fra 0 e 1 misura la coerenza interna del questionario e risponde alle domande: i quesiti misurano tutti lo stesso tratto latente? Sono coerenti fra loro? Valori di α al di sotto dello 0,60 sono solitamente considerati inadeguati. Insieme all'indice α di Chronbach si annovera l'indice Kuder-Richardson 20 (KR-20), identico al precedente ma usato per variabili dicotomiche.

Scendiamo nella pratica e vediamo come usare questi indici su dati reali con un campione composto da 865 studenti che hanno compilato un questionario fatto da 10 domande a risposta chiusa con 4 opzioni di cui una corretta.

Ciascuna delle domande ha un comportamento dicotomico: nella Tabella 8.1, come si è soliti fare, indichiamo con 1 il caso in cui lo studente abbia rispo-

sto correttamente alla domanda, con 0 il caso in cui abbia risposto in maniera errata. La somma delle righe restituisce il punteggio conseguito da ciascuno studente (compreso fra 0 e 10) e rappresentato nell'istogramma in Figura 8.1 dove notiamo uno spostamento a destra dei valori totalizzati con maggiore frequenza. La somma delle colonne fornisce il numero di risposte corrette per ciascuna domanda. Dividendo tali valori per il totale dei rispondenti, possiamo calcolare l'indice di difficoltà dei quesiti come dall'espressione (8.1).

N.	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Punti
1	0	1	1	0	0	0	1	0	1	0	4
2	1	1	1	0	0	0	1	1	1	1	7
3	1	1	0	0	1	1	1	1	0	0	6
4	1	0	1	0	0	0	1	1	1	0	5
5	0	1	0	1	0	1	1	0	1	1	6
6	1	1	1	1	0	1	1	1	1	0	8
7	0	1	1	0	0	1	1	1	1	1	7
8	1	1	0	1	1	1	1	1	1	0	8
...	...	...	...	...	...	...	...	...	...	...	...
865	1	1	1	1	1	1	1	1	1	1	10
Tot. risposte corrette	507	778	558	489	361	668	694	688	816	649	-
Difficoltà	0,59	0,90	0,65	0,57	0,42	0,77	0,80	0,80	0,94	0,75	-

Tabella 8.1 - Dataset costituito da 865 osservazioni su 10 quesiti. Nella tabella riportiamo i punteggi conseguiti da ciascuno studente (ultima colonna) e l'indice di difficoltà calcolato come rapporto fra risposte corrette e risposte totali (ultima riga).

Riportiamo nella Tabella 8.2 i valori della difficoltà appena calcolati e utilizziamo le espressioni (8.2) e (8.3) per calcolare rispettivamente la *DI* e la *DV*. Consideriamo nel gruppo dei "migliori" i punteggi superiori all'8 e nel gruppo dei "peggiori" quelli inferiori al 6. I due valori corrispondono al primo e al terzo quartile. Appartengono al primo gruppo 201 studenti (*NH = 201*), al secondo 149 (*NL = 149*). Per ciascuna domanda calcoliamo il numero di studenti che in ciascuno dei due gruppi ha risposto correttamente alla domanda stessa e applichiamo le formule.

Cominciamo con il verificare che il valore di *DV* calcolato con le due differenti procedure ci restituisce valori diversi all'incirca di 0,1. Tuttavia, le criticità

emergono in entrambi i casi. Sono ad esempio da rivedere gli item Q2 e Q9 come item troppo semplici.

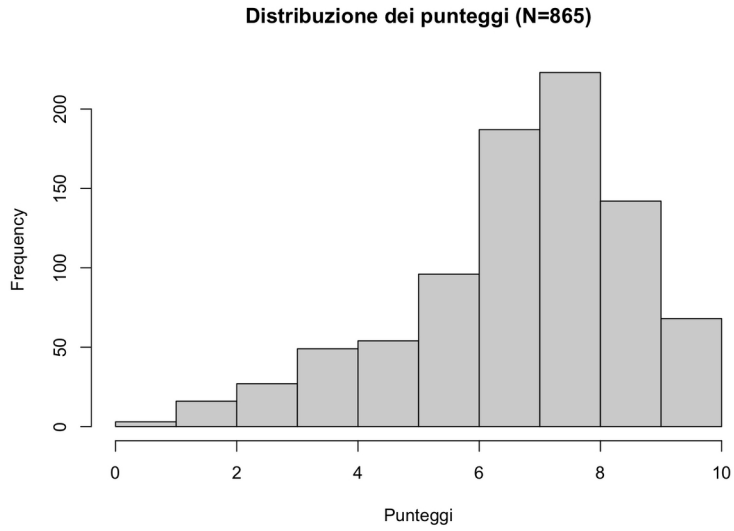


Figura 8.1 - Distribuzione dei punteggi del dataset in Tabella 8.1.

Elevati invece sono i valori della *DI* degli item Q4, Q6 e Q8 di cui per tale motivo andrebbero verificati i contenuti e i distrattori.

L'ultima riga della Tabella 8.2 contiene i valori dell'indice di discriminatività calcolato come punto biseriale. Per questo calcolo, così come per il resto del paragrafo ci serviremo di `R` e del pacchetto `ltm` che è stato realizzato espressamente per lavorare sui *latent trait model* (Rizopoulos, 2007). In questo caso abbiamo usato la funzione `biserial.cor` (avremmo tuttavia potuto usare anche la funzione di base `cor` di `R` ottenendo valori simili). I valori di *DI* delle ultime due righe sono molto diversi fra loro, siamo davanti a due procedure di calcolo differenti in cui i valori centrali assumono un ruolo importante: esclusi nella riga 3, appiattiscono i risultati nella riga 4.

Usiamo dallo stesso pacchetto `ltm` la funzione `cronbach.alpha` che come dice lo stesso nome ci restituisce il valore dell'indice α di Cronbach, per questo dataset modesto, $\alpha = 0{,}553$. Questo indica che la coerenza delle domande fra di loro non è elevata, le domande probabilmente non indagano un unico tratto latente.

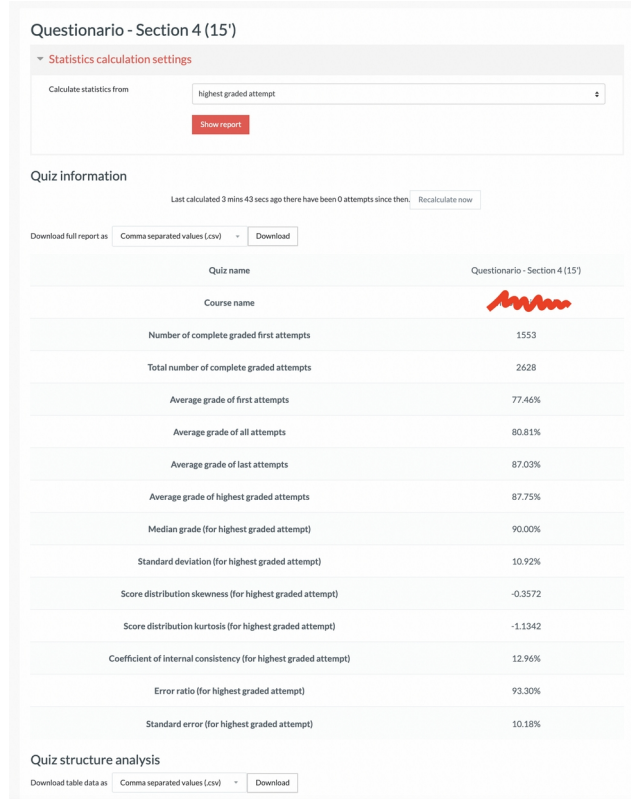
Indice	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
DV (8.1)	0,59	0,90	0,65	0,57	0,42	0,77	0,80	0,80	0,94	0,75
DV (8.2)	0,64	0,82	0,75	0,60	0,54	0,69	0,72	0,71	0,89	0,71
DI (8.3)	0,80	0,69	0,53	0,89	0,82	0,89	0,82	0,88	0,56	0,84
DI (8.4)	0,43	0,52	0,21	0,49	0,41	0,56	0,54	0,56	0,41	0,49

Tabella 8.2 - Indici di difficoltà *DV* e discriminatività *DI* per i 10 item dell'esempio calcolati utilizzando le espressioni (8.1), (8.2), (8.3), (8.4).

È bene sapere che la funzione `descript` del pacchetto `ltm` fornisce le statistiche descrittive per un dataset composto da item e l'intero set di indicatori finora descritti. Essa ci restituisce il numero delle domande e delle unità statistiche, la percentuale di studenti che hanno risposto correttamente a ciascun item, i punteggi totali, la correlazione punto biseriale, l'α di Cronbach, i *p-value* calcolati sul χ^2 delle associazioni a coppie di item: quelle problematiche ci possono fornire indicazione su item critici che hanno un alto livello di associazione fra di loro.

Anche su Moodle, uno dei più noti e utilizzati sistemi di gestione dell'apprendimento per la formazione a distanza, il sistema di *Report* per le attività *Quiz* contiene interessanti riferimenti alla Classical Test Theory e in genere all'analisi psicometrica dei questionari come si può vedere dagli screenshot nelle Figure 8.2 e 8.3. La prima sezione (Figura 8.2) fornisce una sintesi su tutte le risposte del Quiz, sul primo o ultimo tentativo, sul tentativo migliore. In particolare, otteniamo per l'intero questionario punteggi medi e mediana, deviazione standard; inoltre, i valori della kurtosis e della skewness della distribuzione dei punteggi, l'α di Cronbach, i parametri di errore.

La seconda sezione (Figura 8.3) contiene per ciascuna domanda i valori della difficoltà (qui indicata come abilità ossia percentuale di risposte corrette da parte dei rispondenti) e della discriminatività (punto biseriale), il peso previsto (in Figura, 10% poiché uguale per le 10 domande che compongono il questionario) e il peso effettivo (quello stimato come più adatto rispetto alle caratteristiche della domanda).



Questionario - Section 4 (15')

Statistics calculation settings

Calculate statistics from: highest graded attempt

Show report

Quiz information

Last calculated 3 mins 43 secs ago there have been 0 attempts since then. Recalculate now

Download full report as: Comma separated values (.csv) Download

Quiz name	Questionario - Section 4 (15')
Course name	[redacted]
Number of complete graded first attempts	1553
Total number of complete graded attempts	2628
Average grade of first attempts	77.46%
Average grade of all attempts	80.81%
Average grade of last attempts	87.03%
Average grade of highest graded attempts	87.75%
Median grade (for highest graded attempt)	90.00%
Standard deviation (for highest graded attempt)	10.92%
Score distribution skewness (for highest graded attempt)	-0.3572
Score distribution kurtosis (for highest graded attempt)	-1.1342
Coefficient of internal consistency (for highest graded attempt)	12.96%
Error ratio (for highest graded attempt)	93.30%
Standard error (for highest graded attempt)	10.18%

Quiz structure analysis

Download table data as: Comma separated values (.csv) Download

Figura 8.2 - Prima sezione di reportistica su Moodle per i questionari: Quiz > Report > Statistiche. La sezione contiene valori di sintesi sull'intera prova di valutazione.

...	...	Question name	Attempts	Facility index	Standard deviation	Random guess score	Intended weight	Effective weight	Discrimination index	Discriminative efficiency
		[redacted]	1553	88.80%	31.55%	50.00%	10.00%	9.45%	1.23%	1.91%
		[redacted]	1553	87.83%	32.70%	25.00%	10.00%	11.37%	13.09%	20.16%
		[redacted]	1553	92.21%	26.81%	25.00%	10.00%	8.49%	4.17%	8.06%
		[redacted]	1553	90.21%	29.72%	25.00%	10.00%	9.66%	6.44%	10.72%
		[redacted]	1553	91.11%	28.46%	50.00%	10.00%	8.79%	2.85%	5.08%
		[redacted]	1553	87.19%	33.44%	50.00%	10.00%	9.14%	-4.27%	-6.18%

Figura 8.3 - Seconda sezione di reportistica su Moodle per i quesiti: Quiz > Report > Statistiche. La sezione contiene nelle righe valori riferiti ai singoli item che compongono la prova di valutazione.

La Classical Test Theory vincola i risultati delle prove (i voti per intenderci) ai quesiti degli item. L'Item Response Theory, invece, introduce un approccio di natura probabilistica nell'analisi che mette in relazione l'abilità dello studente e la difficoltà degli item con la probabilità che lo studente risponda correttamente alle domande; a partire dai dati raccolti in fase di tryout o nella prima somministrazione della prova permette dunque di definire e confrontare modelli nei quali uno studente con una certa abilità ha una determinata probabilità di rispondere correttamente a un quesito con certe caratteristiche. A posteriori e in procedure che prevedono approssimazioni successive stimiamo quindi l'abilità θ di un dato studente, il tratto latente che ci poniamo l'obiettivo di misurare nei questionari di valutazione.

L'analisi fattoriale è la tecnica che ci permette di lavorare sui tratti latenti. La funzione lineare in cui abbiamo espresso la relazione fra variabili manifeste e latenti nel capitolo 3 è utile per le variabili metriche, qualcosa necessariamente cambia nel momento in cui ci troviamo a lavorare con variabili binarie come in questo caso: la relazione che lega le variabili rilevate (in questo caso le risposte alle domande) e i tratti latenti da utilizzare è invece quella logistica (Bartholomew et al., 2008) che abbiamo già incontrato parlando di regressione logistica nel capitolo 5.

Risulta quindi che

(8.5)

$$p_i = a + b_{i1}f_1 + b_{i2}f_2 + \ldots + b_{ik}f_k$$

dove p è la probabilità (con valori come è noto compresi fra 0 e 1) che uno studente risponda correttamente alla domanda i, f_k i fattori latenti.

Trasformazioni e semplificazioni conducono a definire un modello di base nella notazione:

(8.6)

$$p_i = g_i + (1-g_i)\,\frac{e^{(a_i+b_i\theta)}}{1+e^{(a_i+b_i\theta)}}$$

dove p_i è la probabilità che uno studente risponda correttamente alla domanda i, b_i è il potere discriminante dell'item i, θ l'abilità dello studente, a_i la

difficoltà dell'item i e g il parametro *guessing* ossia la probabilità che uno studente con scarsa abilità risponda correttamente all'item.

Questa formulazione si riferisce al modello più completo e più complesso a tre parametri, detto di Birnbaum.

Il modello logistico di Rasch, detto a un parametro, ipotizza che $g = 0$ e che la discriminatività assuma un valore fisso per tutte le domande. In questo caso i parametri da calcolare sono quelli legati all'abilità θ dello studente (che ricordiamo essere il tratto latente) e la difficoltà a della domanda i.

Nel modello a due parametri, solo il guessing viene escluso ($g = 0$): il secondo parametro è la discriminatività b_i.

Per applicare la IRT è necessario verificare due presupposti (Wallace et al., 2018):

1. gli item devono essere indipendenti fra loro, ossia la probabilità che lo studente risponda correttamente a un item deve essere legata solo all'abilità dello studente e non a risposte ad altri item o ad altri fattori;
2. il test deve essere unidimensionale, ossia un unico tratto latente deve spiegare completamente la performance dello studente nella prova.

Nel pacchetto `ltm` (che è solo una delle possibilità esistenti fra sofware e pacchetti di `R`) tre diverse funzioni ci permettono di calcolare i modelli a uno, due, tre parametri. Si tratta rispettivamente delle funzioni `rasch`, `ltm`, `tpm`.

Le tre funzioni restituiscono i valori dei parametri insieme a stime della bontà del modello come AIC (Akaike Information Criterion), BIC (Bayes Information Criterion) e massima verosimiglianza.

Dai valori dei parametri e attraverso un processo iterativo di ottimizzazione con le procedure di massima verosimiglianza, viene stimata l'abilità θ per ciascuno studente, elemento fondamentale come vedremo nel prossimo paragrafo quando usiamo per fini didattici la IRT. Quest'ultima, così anche come la difficoltà e la discriminatività degli item, sono espresse usando i logit, i logaritmi dell'odds ratio. In particolare, l'abilità di un dato studente risulterà espressa dalla notazione:

(8.7)

$$\theta = \log \frac{p}{1-p}$$

dove p è la percentuale di risposte corrette fornite dello studente. Nei risultati, per fare un esempio dunque, $\theta = 0$ indica che lo studente ha risposto correttamente al 50% delle domande del questionario essendo $p / (1 - p) = 1$ e $log (1) = 0$. Valori negativi di θ indicano che lo studente ha risposto a meno del 50% delle domande e viceversa per i valori positivi (si vedano le proprietà elementari della funzione logaritmo).

Graficamente la forma che la relazione individuata fra i parametri assume è quella che della curva a forma di S che abbiamo incontrato nella regressione logistica (cap. 5). Per ciascuna domanda con una data difficoltà potremo graficamente visualizzare la probabilità che uno studente con una data abilità possa rispondere correttamente. Questa curva viene definita *Item Characteristic Curve* (ICC). [Nota bene: la probabilità non raggiunge mai i valori dello 0 o dell'1, ciò significa che c'è sempre una possibilità per gli studenti con abilità molto basse di rispondere correttamente a una domanda e, viceversa, per studenti con abilità molto alta di sbagliarne una!]

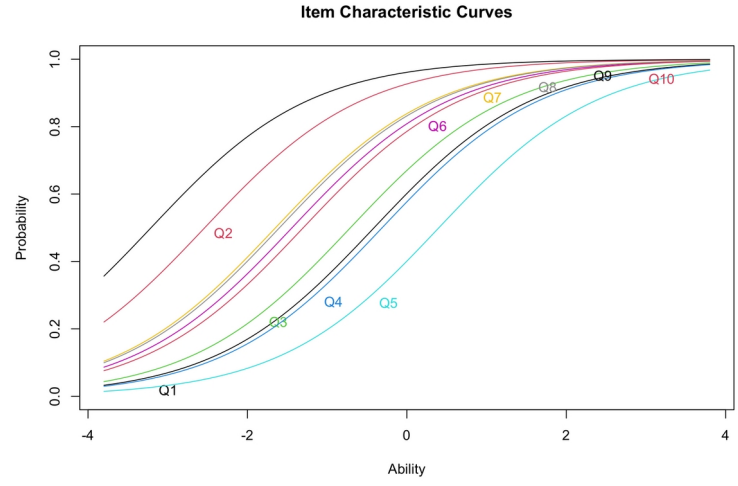


Figura 8.4 – Item Characteristic Curve dei 10 item dell'esempio in analisi.

Nella Figura 8.4 vediamo le ICC calcolate sulle 10 domande nel nostro esempio usando la funzione `plot.rasch` del pacchetto `ltm` applicata in questo caso ai risultati della funzione `rasch` dello stesso pacchetto (modello a un parametro). La probabilità che studenti con il valore di abilità più basso calcolato (-4) rispondano correttamente alla domanda Q5 tende a 0; il valore della stessa percentuale sale al 50% per la domanda Q9 che, come possiamo osservare, è più piatta. La domanda Q5 è quindi più difficile della Q9. Curve più ripide come la Q3 e la Q4 hanno un maggiore potere discriminativo perché il passaggio da livelli bassi di abilità a livelli alti è più immediato.

Riportiamo a seguire altri due grafici che si possono ottenere dalla stessa funzione: l'*Item Information Curve* (Figura 8.5) e il *Test Information Curve* (Figura 8.6). Entrambi riportano la percentuale di informazioni che i singoli quesiti nel primo caso e l'intero test nel secondo forniscono in riferimento all'abilità degli studenti. Gli item Q9 e Q10 forniscono più informazioni su abilità basse, il Q5 su abilità che tendono a valori più alti. L'intero test fornisce informazioni piuttosto su bassi livelli di abilità che su alti: il 60% delle informazioni fornite dal test è riferibile ad abilità comprese fra -4 e 0.

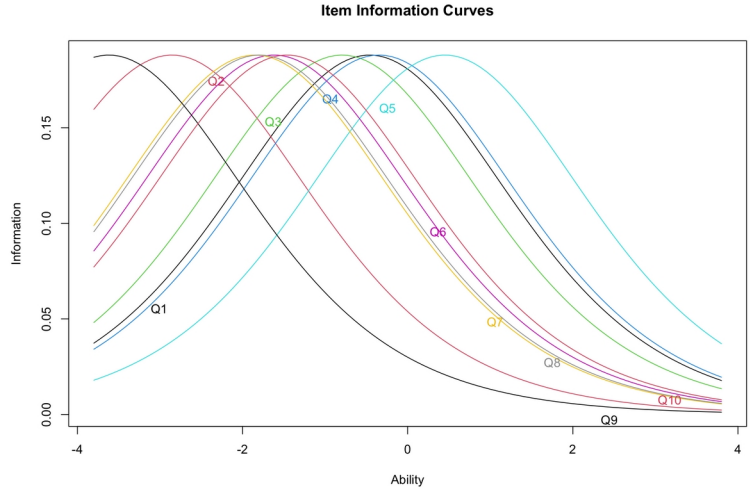


Figura 8.5 - Item information curve dei 10 item dell'esempio in analisi.

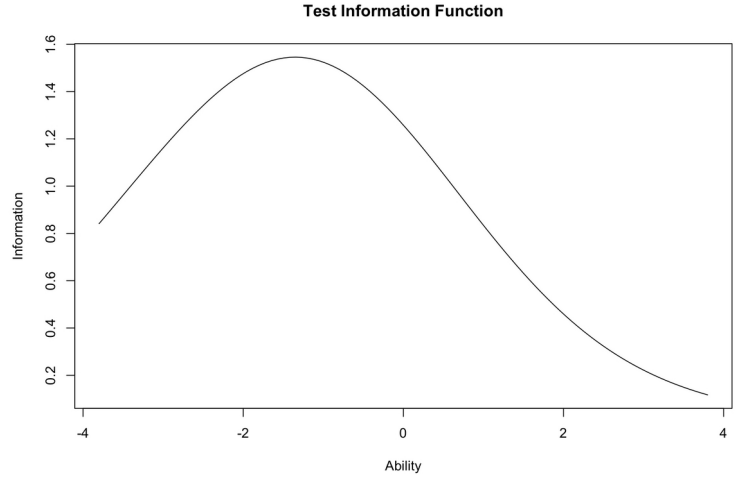


Figura 8.6 - Test information curve calcolato sull'intero questionario in analisi.

8.2 - L'uso degli approcci della Classical Test Theory e della Item Response Theory nella ricerca educativa

Kennedy Quaigrain e Ato Kwarmina Arhin (2017) fanno uso degli indici della CTT per valutare un questionario a risposta chiusa composto da 50 domande da svolgere in 65 minuti somministrato a 247 studenti in un corso post-laurea in Educazione presso il Politecnico di Cape Coast in Ghana. La prova in questione è l'esame di fine semestre proprio in un corso dedicato alla valutazione in ambito educativo (Educational Measurement course). Le domande presentavano 4 opzioni di risposta di cui soltanto una corretta; come da standard, alla domanda corretta è stato assegnato un punto e alle risposte errate 0. Pertanto i punteggi conseguibili dagli studenti variavano fra 0 e 50. Gli autori valutano il test attraverso indici di difficoltà, discriminatività, la formula Kuder-Richardson (KR-20) con un focus sulla qualità degli item e l'efficacia dei distrattori.

Calcolano la *DV* come rapporto fra numero di risposte corrette e numero di risposte date, considerando eccellenti quelle con un punteggio compreso fra 0,4 e 0,6. Per la *DI* lavorano sugli estremi al 27%, sottraggono il numero di risposte corrette dal gruppo con i punteggi più bassi dal gruppo dei "migliori",

normalizzando (dividono cioè per il numero totale dei partecipanti). Usano come valori di riferimento per la discriminatività i range riportati sopra da Ebel e Frisbie (1991).

La mediana dei punteggi dei partecipanti è pari a 30 con una differenza inter-quartile pari a 9; i valori di asimmetria e curtosi per i punteggi sono considerati accettabili; allo stesso modo è adeguato l'indice KR-20 pari a 0,77.

La maggior parte degli item ha livelli di difficoltà accettabili (compresi fra 20% e 90%). L'indice di discriminatività assume valori fra -0,22 e 0,46. Diciotto dei 50 item del questionario hanno un indice di discriminatività superiore allo 0,3 e pertanto accettabile. Al contrario 20 item, proprio in virtù del loro indice di discriminatività, necessitano di una revisione completa.

La Figura 8.7 plotta indici di difficoltà (ascisse) e di discriminatività (ordinate) degli item della prova. Vediamo graficamente che item molto difficili (*DV < 0,20*) hanno un indice di discriminatività negativo: le risposte corrette a questi item non permettono di discriminare soggetti che hanno sviluppato conoscenze sugli argomenti del corso rispetto a quanti non l'hanno fatto. Gli item con difficoltà compresa fra 0,50 e 0,70 (range più efficace) hanno i valori di discriminatività più alti. La discriminatività tende a diminuire per gli item più facili, dove la maggior parte degli studenti risponde correttamente senza, di conseguenza, garantire che l'esattezza delle risposte sia sintomo di una "buona preparazione" sui temi del corso.

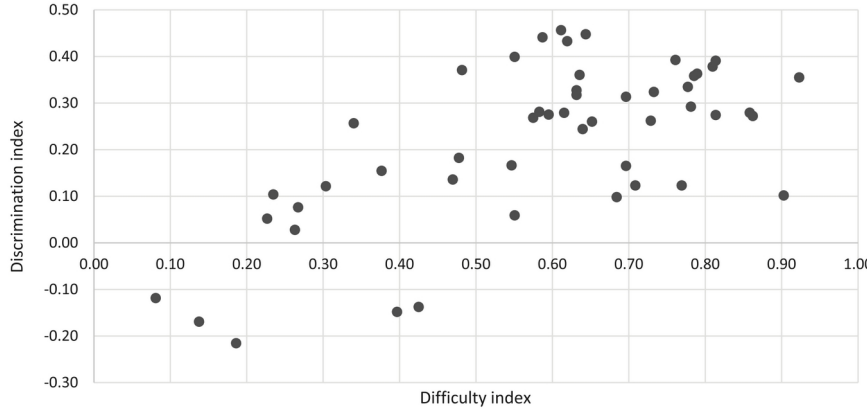


Figura 8.7 - Scatterplot fra indici di difficoltà e discriminatività dei 50 item di un questionario somministrato in un corso post-laurea in Educazione presso il Politecnico di Cape Coast in Ghana (Quaigrain & Arhin, 2017, p. 7).

Gli autori calcolano inoltre che dei 150 distrattori presenti nelle domande, ben 38 (25,3%) sono scelti da meno del 5% dei rispondenti. Tale valore può certamente aver contribuito nel computo di *DI* e *DV*. Distrattori poco plausibili, ovvi o mal scritti possono contribuire ad aumentare la facilità delle domande e compromettere la discriminatività: la scelta delle risposte non sarà dettata dalla comprensione dei concetti ma dalla cattiva formulazione degli item e dei distrattori. L'uso della CTT ha permesso di identificare item critici per i valori della *DV* e della *DI* che vanno corretti o eliminati, proprio perché potrebbero rendere la prova meno valida e anche meno appropriata a rilevare le reali abilità degli studenti. La ricerca apre la discussione sull'uso di distrattori funzionali che hanno ugualmente un ruolo nella definizione dei risultati di un questionario e pone interrogativi sulle questioni più pratiche di costruzione dei quesiti. In questo caso, ad esempio, alcuni item con un distrattore non funzionale hanno mostrato un indice di discriminatività superiore ad item con tre distrattori funzionali (rispettivamente 0,26 e 0,17).

Proseguiamo la nostra rassegna spostandoci nell'ambito dell'insegnamento dell'astronomia, dove Erin M. Bardar e colleghi (2007) hanno messo a punto la Light and Spectroscopy Concept Inventory (LSCI) progettata sin dall'inizio per produrre uno strumento di valutazione standardizzato dei temi della disciplina e per confrontare l'efficacia dei vari metodi didattici nell'insegnamento della stessa.

Gli studiosi hanno formulato il questionario a risposta multipla ritenendolo più semplice da valutare e hanno sottoposto le domande che lo compongono sia alla validazione di altri colleghi esperti della disciplina, sia a un'analisi del tryout con un gruppo di circa 50 studenti. Con i valori di *DI* e *DV* del primo tryout, sono stati già in grado di verificare 4 domande eccessivamente semplici e una risultata ambigua (Bardar et al., 2007).

In uno studio successivo (Schlingman et al., 2012) su larga scala, l'inventory è stato somministrata come pre- e post-test a circa 4000 studenti di 69 classi in 31 istituti statunitensi. Usando gli indicatori della CTT, gli autori hanno calcolato difficoltà e discriminatività nel pre- e post-test. La Tabella 8.3 mostra i valori degli indici per le prime domande. In questo caso la difficoltà è misurata come il numero di domande sbagliate sul numero di domande totali con valori accettati fra 0,2 e 0,8. Dire dunque che l'item 1 ha un livello di difficoltà pari a 0,81 nel pre-test significa dire che l'81% dei rispondenti ha scelto la risposta errata alla domanda. I valori accettati per la discriminatività sono fra 0,3 e 0,8.

LSCI item	Pre difficulty	Pre discrimination	Post difficulty	Post discrimination
1	*0,81*	0,36	0,37	0,49
2	*0,89*	*0,11*	0,76	0,53
3	*0,83*	*0,25*	0,76	*0,25*
4	*0,81*	*0,30*	0,31	0,44
5	0,64	*0,27*	0,27	0,37
6	0,77	*0,23*	0,47	0,46
7	*0,80*	*0,23*	0,55	0,46

Tabella 8.3 - Indici *DI* e *DV* per i primi item della LSCI nel pre- e post-test. In corsivo i valori esclusi nei range (Schlingman et al., 2012, p. 3/10).

Nel passaggio dal pre- al post-test, la difficoltà dell'item diminuisce come è ovvio che sia, dato che aumenta la conoscenza degli studenti dei concetti del corso. Parallelamente aumenta la discriminatività, poiché gli studenti più preparati sono in grado di rispondere correttamente alle domande e di conseguenza gli item riescono maggiormente a rilevare le differenze fra chi è preparato e chi non lo è. Alcuni item anche nel post-test hanno conservato degli indicatori non validi: ad esempio l'item 3 conserva un basso livello di discriminatività (*DI = 0,25*). Il contenuto di ciascuno di questi item viene analizzato dagli autori anche in base alla funzione che esso assume all'interno dell'inventory. Il coefficiente α di Cronbach passa da 0,37 nel pre-test a 0,78 nel post-test; gli autori giustificano questa discrepanza con il fatto che α è sensibile all'omogeneità del campione di riferimento e sicuramente nel pre-test molti studenti condividevano le stesse difficoltà e dubbi sugli argomenti.

Il grafico in Figura 8.8 confronta i punteggi del pre-test con quelli del post-test, le tonalità di blu indicano il numero di studenti che si collocano in ogni punto. La linea indica punteggi che sono rimasti identici nel pre-test e nel post-test. 82% degli studenti è al di sopra della linea. Una fetta del campione non è migliorata dopo la partecipazione ai corsi o ha conservato uno stesso punteggio iniziale. Questo gruppo di studenti è quello che da educatori e formatori deve maggiormente attrarre la nostra attenzione e richiedere un nostro intervento.

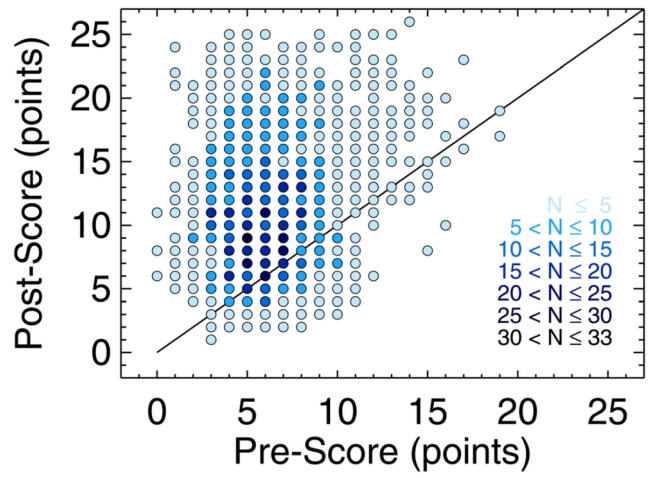


Figura 8.8 – Confronto fra punteggi del pre- e post-test per gli studenti nello studio sulla LCSI (Schlingman et al., 2012, p. 6/10).

Uno studio successivo (Wallace et al., 2018) propone un'analisi condotta con la IRT sugli item della stessa inventory LSCI nella quale sono stati analizzati i risultati di pre-test e post-test sullo stesso dataset dello studio precedente. I tentativi di adattare i dati del pre-test a modelli a due o tre parametri hanno condotto a scarsi risultati (poor fit) suggerendo che prima del percorso formativo molti studenti possedevano poco il tratto latente misurato (l'abilità). Si è scelto così di utilizzare i dati del post-test per costruire modelli a tre parametri, introdurre il parametro guessing in questo dataset contribuisce a migliorare il fit del modello ai dati. Tre item sono stati esclusi: il primo per poor fit, non si evinceva una chiara relazione fra l'abilità dello studente e la risposta corretta nell'item; il secondo perché molto difficile; il terzo perché non era indipendente dagli altri item ma sembrava essere in relazione con più di uno (nell'item, infatti, veniva chiesto di confrontare più elementi, come energia, frequenza, lunghezza dell'onda, la velocità delle onde radio e della luce). Altri 6 item che correlavano fra di loro sono stati accoppiati e trattati diversamente, proprio perché a coppie affrontavano un qualche argomento di studio. Il numero di item su cui si arriva a lavorare è in definitiva pari a 20.

L'interpretazione dei risultati sugli item conduce ad alcune riflessioni. Ad es. gli item che chiedono di applicare leggi e interpretare un grafico hanno valori di difficoltà e discriminatività più alti.

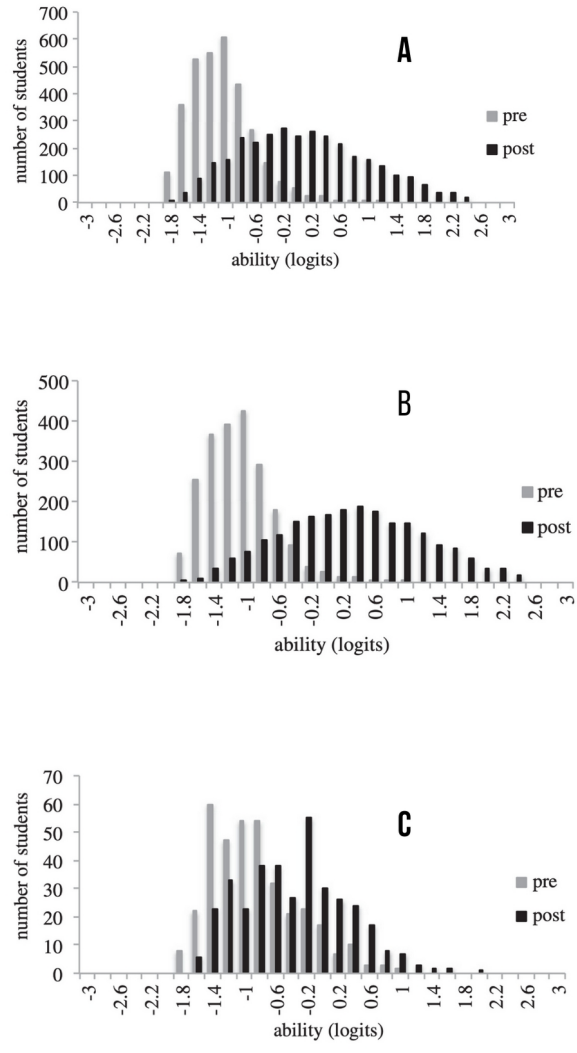


Figura 8.9 – Distribuzione della abilità nel pre- e post-test per l'intero campione dello studio (A) e i due sottocampioni con IAS superiore (B) e inferiore al 25% (C) in uno studio sulla inventory LSCI (Wallace et al., 2018).

Gli item sulle proprietà della luce sono quelli più semplici. Gli item con bassi valori di guessing devono avere dei distrattori funzionali che impediscono di rispondere correttamente scegliendo a caso la risposta corretta.

A partire dai parametri calcolati, gli autori hanno stimato l'abilità degli studenti nel pre- e post-test. Gli autori hanno inoltre utilizzato dei parametri forniti dai docenti per calcolare l'Interactivity Assessment Score (IAS) che assume valori compresi fra lo 0 e il 49% e corrisponde alla percentuale del tempo speso in classe in attività interattive dove per interattive ci si riferisce ad attività che sviluppano lo spirito critico preferibilmente lavorando con uno o più colleghi di corso (Prather et al., 2009). Le classi in cui è stato somministrato LSCI e, di conseguenza, gli studenti ad esse appartenenti sono stati suddivisi in 2 gruppi con *IAS < 25%* (*lowIAS*) e *IAS > 25%* (*highIAS*). Gli istogrammi in Figura 8.9 confrontano la distribuzione della pre- e post-abilità nel campione totale e nei due sottocampioni con IAS superiore e inferiore al 25%.

Notiamo che lo spostamento dell'abilità nei range più alti verificato per il campione totale e il gruppo di classi highIAS non è confermato per le classi lowIAS. Il t-test conferma come statisticamente significativa non solo la differenza fra i due gruppi nel post-test ma anche quella in ingresso (nel pre-test la media delle abilità è più alta per il gruppo lowIAS). Il miglioramento delle abilità degli studenti delle classi highIAS è in media superiore di quasi un intero logit rispetto alle classi lowIAS.

Pertanto, in conclusione, se un docente vuole migliorare di un logit il valore dell'abilità dei suoi studenti, deve, in base ai risultati di questo studio, dedicare almeno il 25% del tempo nelle sue lezioni ad attività interattive.

Anche lo studio di Georgianne L. Connell e colleghi (2016) mostra come si possa utilizzare la IRT per stimare e confrontare l'abilità degli studenti in pre- e post-assessment, in particolare in questo caso in un corso universitario di Biologia.

Il corso, fra quelli introduttivi con i più alti tassi di fallimento, è stato trasformato da tradizionale a student-centered nei 7 anni precedenti allo studio in modalità graduale. La struttura *Moderate*, messa a punto in un primo momento si è concentrata sulla realizzazione di lezioni interattive con attività come pause riflessive e think-pair-share.

Nella modalità successiva, definita *Extensive*, è stato introdotto un approccio flipped prendendo spunto da team based learning e collaborative learning. Sono stati creati dai docenti gruppi permanenti di lavoro composti da 4-6 studenti per completare consegne della durata di circa 30 minuti impostate se-

condo i principi costruttivisti. In questa seconda modalità di organizzazione del corso, per ciascun modulo gli studenti hanno trascorso circa 2-4 ore al di fuori dell'aula fisica, guardando videolezioni, guide, partecipando a discussioni e producendo riassunti. Il tempo d'aula è stato dedicato allo svolgimento di questionari con feedback immediato e di lezioni sugli argomenti più difficili.

Gli autori si sono chiesti dunque se vale la pena investire tempo per realizzare corsi così strutturati: è più efficace un corso organizzato con un moderato numero di attività student-centered (Moderate) o con un elevato numero di attività di questo tipo (Extensive)? In quale dei due migliora l'approccio alle scienze degli studenti?

Usando un disegno quasi-sperimentale, hanno proposto a due classi di Biologia, ciascuna composta all'incirca da 180 studenti, un corso basato sugli stessi contenuti, tenuto nello stesso semestre e condotto nello stesso giorno della settimana da uno stesso docente con i due approcci, Moderate ed Extensive.

Oltre ai punteggi degli esami e all'uso di una scala per valutare le attitudini all'apprendimento della scienza degli studenti, sono stati somministrati un pre- e un post-test composto da 40 domande di cui 28 riprese da altre inventory e 12 prodotte in maniera originale dagli autori. Dei 40 item, 5 non hanno superato la verifica dei prerequisiti (criterio dell'indipendenza degli item) e sono stati eliminati; altri 6 sono stati fusi in coppie perché correlati fra loro.

Attraverso un metodo a posteriori, gli indicatori calcolati nel post-assessment sono stati usati per calcolare l'abilità degli studenti nel pre- e post-test. I modelli di IRT a due e tre parametri sono stati usati per aumentare la comprensione della performance degli studenti. La Figura 8.10 riporta le percentuali delle abilità degli studenti nel pre- e post-test nei due gruppi, possiamo osservare (anche se con effetti meno marcati del precedente studio) che nella sezione Extensive le percentuali di abilità sono aumentate di più rispetto alla sezione Moderate. Gli autori spiegano tali risultati con l'uso di metodologie didattiche attive, attività in gruppi cooperativi, uso della valutazione formativa e di processi metacognitivi.

Per verificare l'esistenza di differenze demografiche fra i due gruppi, gli autori utilizzano una regressione lineare dove il punteggio del post-test è la variabile dipendente e le variabili indipendenti sono il punteggio del pre-test, l'appartenenza a una delle due sezioni (Moderate/Extensive), il numero di corsi di biologia frequentati, all'high school o all'università e l'anno di corso. Propongono anche una regressione lineare per vedere se le abilità stimate per il pre- e post-test sono collegate ai fattori demografici e in particolar modo alla sezio-

ne di appartenenza. Nella prima regressione lineare, il coefficiente della variabile sezione (Moderate/Extensive) è di 2,5: tenendo fisse tutte le variabili, lo score è quindi più alto di 2,5 punti per gli studenti della sezione Extensive. Nella regressione sull'abilità il valore del coefficiente della variabile sezione è di 0,18 e non è statisticamente significativo. Questo potrebbe essere attribuito al fatto che la IRT introduce un grado di incertezza nelle stime che non è comune nella teoria classica della CTT. La combinazione dell'analisi dei punteggi e delle abilità (con riferimento ai modelli tradizionali e probabilistico) sembrano evidenziare che nella sezione Extensive il miglioramento degli studenti è stato superiore ma gli effetti non sono stati così dirompenti come la prima regressione lineare sugli score ha lasciato ipotizzare.

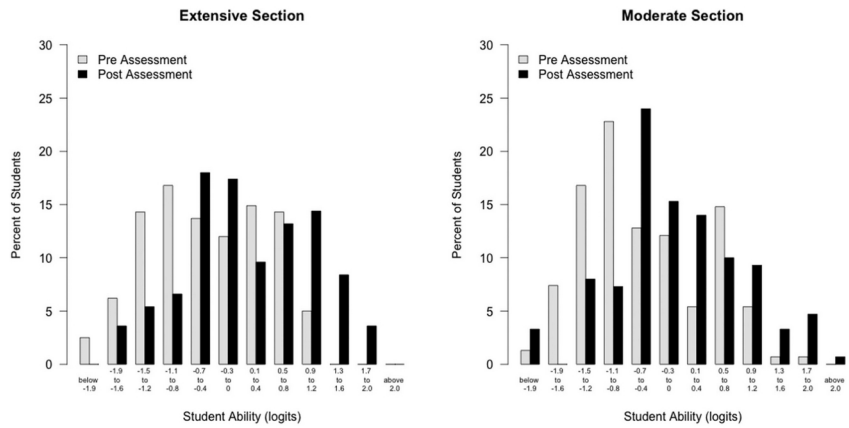


Figura 8.10 – Confronto fra i risultati del pre- e post- test nella sezione Extensive e Moderate in un corso di Biologia (Connell et al., 2016, p. 9).

Il prossimo e ultimo lavoro che presentiamo, per osservare come CTT e IRT siano utilizzati nella ricerca educativa, riguarda uno studio condotto in Nuova Zelanda in un insegnamento universitario di Psicologia educativa erogato nel primo anno di corso (Brown & Abdulnabi, 2017). Il contributo analizza due questionari composti ciascuno da 50 domande con 4 opzioni di risposta, di cui uno intermedio e uno finale che valgono il 50% del punteggio finale d'esame. Il dataset consta delle risposte fornite ai questionari da circa 380 studenti. Scopo degli autori è determinare la qualità dei 100 item e verificare possibili impli-

cazioni nel superamento degli esami in base al punteggio totalizzato nelle prove di valutazione. Gli autori sottolineano infatti come nonostante esistano delle linee guida per la scrittura degli item, talvolta queste non sono seguite e item malscritti conducono a valutazioni che non sempre corrispondono alle reali conoscenze acquisite dagli studenti, anzi forniscono feedback non del tutto attendibili agli studenti sul loro grado di preparazione e agli stessi docenti sulla preparazione degli studenti e sull'andamento generale del corso.

Gli autori rilevano e confrontano le caratteristiche psicometriche degli item usando la CTT e i modelli della IRT a 1, 2, 3 parametri. L'applicazione di ciascuna delle soluzioni conduce all'eliminazione di un determinato numero di item. Il modello che presenta il miglior fit con i dati è quello a due parametri. Il valore dell'AIC risulta essere quello inferiore se confrontato con i modelli a uno e tre parametri sia nella prova intermedia che in quella finale. Un numero maggiore di item nella prova intermedia (come si può vedere nella Tabella 8.4) non è considerato adeguato, la qualità degli item in questa prova appare quindi meno elevata. Le Item Characteristic Curve riportate nella Figura 8.11 mostrano per alcuni item un andamento inverso o traiettorie piatte, entrambi segnali di valori poco efficaci della discriminatività e della difficoltà.

	Midterm test					**Final exam**				
	Raw	**CTT**	**Rasch**	**2PL**	**3PL**	**Raw**	**CTT**	**Rasch**	**2PL**	**3PL**
k	50	22	24	26	30	50	34	31	47	32
M	23,78	10,53	11,36	13,02	15,69	32,10	21,76	19,95	31,66	20,54
SD	5,96	7,40	3,10	5,07	4,94	6,64	5,99	4,10	6,64	4,50
SEM	3,21	1,81	2,21	2,09	2,52	2,82	2,40	2,28	2,74	1,86
α	0,71	0,94	0,49	0,83	0,74	0,82	0,84	0,69	0,83	0,83

Tabella 8.4 – Statistiche della prova intermedia e dell'esame finale realizzati nell'insegnamento di Psicologia educativa distinti per modello statistico (k = numero di item, M = media, SD= deviazione standard, SEM = errore medio, α = alpha di Cronbach, Raw = dati grezzi, CTT = Classical Test Theory, Rasch = Modello di Rasch a un parametro, 2PL = Modello a due parametri, 3PL = modello a tre parametri) (Brown & Abdulnabi, 2017, p. 8).

Riducendo il numero degli item secondo quanto emerso dal modello IRT a 2 parametri, ricalcolando i voti conseguiti dagli studenti nelle due prove, gli autori hanno verificato che il voto di ben due terzi degli studenti coinvolti nell'indagine è cambiato, per 123 studenti sarebbe incrementato e per la stessa porzione di studenti diminuito.

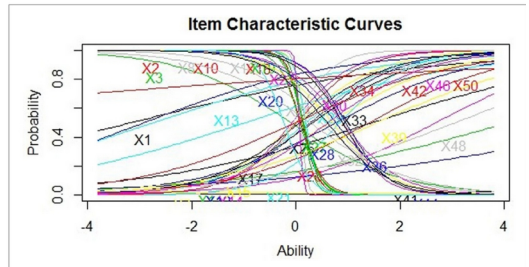


FIGURE 1 | Mid-Term Test IRT 2PL Item Characteristic Curves.

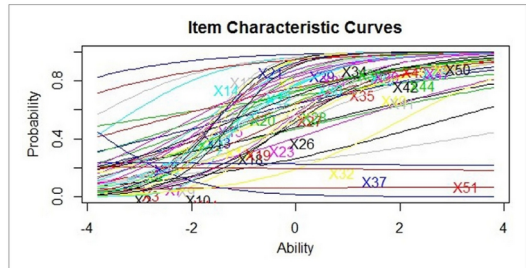


FIGURE 2 | Final Exam IRT 2PL Item Characteristic Curves.

Figura 8.11 – Item Characteristic Curve per la prova intermedia e finale (Brown & Abdulnabi, 2017, p. 8).

Fra i limiti dello studio c'è sicuramente l'assenza dell'intervento di esperti della disciplina nella revisione dei quesiti, con il loro contributo sarebbero probabilmente stati identificati item indispensabili per verificare la copertura della totalità degli argomenti della disciplina o possibili equivoci. Tuttavia l'articolo pone l'attenzione su quanto prove mal strutturate possano incidere sulle decisioni sul superamento di un esame: corsi, prove e metodi di analisi delle stesse meritano un rinnovamento.

POSTFAZIONE

Pietro Lucisano

Il metodo e le tecniche di cui disponiamo ci mettono di fronte a grandi responsabilità, per questo come postfazione a un testo che affronta con serietà e rigore metodologie di analisi dei dati e prospettive di ricerca, da vecchio artigiano della ricerca quantitativa, vorrei proporre ai lettori alcune riflessioni che invitano alla prudenza.

I dati che rileviamo sono sempre elementi di informazione tratti da un contesto, non a caso Dewey preferisce definirli "fatti del caso" per sottolineare il loro legame con il contesto e la loro natura legata a un "qui ed ora" che al momento dell'analisi è quasi sempre irrimediabilmente passato e perduto. Il ricercatore nell'avvicinarsi ai dati dovrebbe costantemente cercare di richiamare alla sua memoria il contesto, le interazioni che ne hanno determinato la rilevazione e le scelte che ne hanno definito le misure. Il metodo, come diceva Cartesio, è quella strada vecchia che consente di non commettere errori grossolani e che ci aiuta a riconsiderare, sintetizzare, controllare gli elementi di cui disponiamo per la ricerca. E il metodo, come bene illustrato in questo manuale da Annamaria De Santis, va seguito con estremo rigore, senza concedersi scorciatoie. Le scelte in ogni situazione competono al ricercatore che, tuttavia, deve tenere presente che non esistono dati in sé al di fuori della loro relazione con noi, essi sono in relazione a che cosa pensava chi li ha rilevati, a che cosa vuole vedere chi li osserva e a che cosa si aspetta di poter affermare o fare sulla base della loro analisi. Le stesse proprietà dei dati esistono solo in una dimensione relazionale. Queste considerazioni ci rimandano alle riflessioni critiche che emergono dalla fisica quantistica che Rovelli descrive bene nel suo breve saggio *Helgoland*: *"Le variabili fisiche non descrivono le cose: descrivono*

il modo in cui le cose si manifestano le une alle altre. Non ha senso attribuire loro un valore, se non nel corso di una interazione." Concetti che dovrebbero essere familiari a chi ha esperienza di relazioni educative. La natura di un ragazzo, le sue qualità non sono "dati" che una volta presi possono essere assunti come conoscenza certa. Un ragazzo è in un modo con un insegnante, in un modo diverso con un altro, risponde in un modo a un certo test e in un altro modo a un test differente. A noi sembra di guardare ogni giorno lo stesso ragazzo, ma ogni giorno è diverso: non ci si bagna mai due volte nello stesso fiume. Il passato non determina il presente e il futuro. Le tendenze mutano rapidamente come le intenzioni di voto nei sondaggi.

La ricerca quantitativa va difesa perché ci aiuta nella comprensione della realtà, anche se forse è eccessivo pretendere di trarne evidenze normative data l'ampia varietà dei fenomeni che studiamo e la grande complessità del rapporto tra evidenze e decisioni nella vicenda umana.

I dati, è necessario raccoglierne molti, raccoglierli bene e poi interagire con i loro portati consapevoli delle nostre ipotesi e della nostra equazione personale, bisogna guardarli e riguardarli, contestualizzarli, ruminarli come San Girolamo diceva della parola di Dio, senza fretta di trarre conclusioni. Infine, è bene tenere conto del fatto che metodo e atteggiamento scientifico sono in un rapporto dialettico. Poiché se il metodo è la strada vecchia, sicura, con ridotte probabilità di errore è anche vero che la conoscenza si è costruita attraverso trasgressioni del metodo che di volta in volta hanno portato a nuovi paradigmi. Bisogna dunque padroneggiare le tecniche di analisi che vengono illustrate in questo lavoro, essere in grado di applicarle con rigore, di leggerne gli esiti e poi cominciare a riflettere.

Riflettere adottando quell'atteggiamento scientifico che Dewey definisce come "il desiderio di ricercare, esaminare, discriminare, tracciare conclusioni solo sulla base dell'evidenza, dopo essersi presi la pena di raccogliere tutti i dati possibili. È l'intenzione di raggiungere credenze, e di provare quelle che risultano accettabili, sulla base dei fatti osservati, riconoscendo al tempo stesso che i fatti sono privi di senso a meno che non indichino idee. È, d'altra parte, l'atteggiamento sperimentale che riconosce come, mentre le idee sono necessarie per l'organizzazione dei fatti, esse sono al tempo stesso ipotesi di lavoro da verificare sulla base delle conseguenze che producono". Le conseguenze

come le intenzioni diventano allora il banco di prova della validità che non appartiene a uno strumento, ma alle interazioni che questo realizza nei diversi contesti.

Così vorrei concludere questa breve postfazione che segue la lettura e incoraggia alla rilettura del testo di Annamaria De Santis con alcune parole di Meirieu in cui mi riconosco: *"Essere un esperto invita alla modestia: si tocca con mano la complessità delle cose e si sperimenta tutti i giorni l'abisso che corre tra la generosità delle nostre intenzioni e la mediocrità delle nostre pratiche. Nessuno meglio del ricercatore conosce la distanza incolmabile tra il dire e il fare. Nessuno conosce meglio di lui l'infinita presunzione delle menti forti che credono sia sufficiente sapere cosa bisogna fare per farlo e che basti tentare per riuscire "*.

E tuttavia la nostra ricerca vale la pena.

Roma, agosto 2022

Pietro Lucisano
Sapienza Università di Roma
Presidente della SIRD
Società Italiana di Ricerca Didattica

RINGRAZIAMENTI

Ringrazio di cuore il prof. Tommaso Minerva per tutto quello che mi sta insegnando e per aver riacceso in me la "passione per i numeri". Lo ringrazio perché mi ha sostenuto in tutte le fasi di stesura di questo volume e continua a supportarmi (e sopportarmi) con preziosi suggerimenti, incoraggiamenti, ironia e risate. Senza la sua guida paziente e la sua costante presenza questo progetto non avrebbe trovato inizio né conclusione.

Ringrazio le mie amiche-colleghe Claudia, Katia e Luisa (in rigoroso ordine alfabetico!) per ogni cosa che condividiamo quotidianamente: pezzi di vita, attività lavorative, progetti, idee, dubbi, scritture …e poi pranzi, merende, aperitivi e cene al Republic.

Ringrazio Cinzia e i miei colleghi passati e presenti del Centro Edunova che fra sorrisi, vassoi di frutta e non solo hanno saputo e sanno insegnarmi con grande competenza ogni giorno qualcosa di nuovo.

Ringrazio e penso agli studenti del corso di laurea in Digital Education, i primi con cui ho potuto condividere idee e tecniche contenute in questo volume, scritto anche per loro.

BIBLIOGRAFIA

Abate, M. (2000). *Algebra Lineare*. Milano: McGraw-Hill Education.

Agresti, A. (2013). *Categorical Data Analysis* (3rd Ed.). John Wiley & Sons: Hoboken (NJ).

Akaike A. (1969). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, *22*, 203-217.

Akaike A. (1978). Bayesian analysis of the minimum AIC procedure. *Annals of the institute of Statistical Mathematics*, *30*, part A, 9-14.

Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye tracking research on multimedia learning. *Computers & Education*, *125*, 413-428.

Ary, D., Jacobs, L.C., Irvine, C.K.S., & Walker, D. (2010). *Introduction to Research in Education* (8th ed.). Belmont (CA): Wadsworth Cengage Learning.

Ashraf, Z. A., & Jaseem, K. (2020). Classical and Modern Methods in Item Analysis of Test Tools. *International Journal of Research and Review*, *7*(5), 397-403.

Bardar, E. M., Prather, E. E., Brecher, K., & Slater, T. F. (2007). Development and validation of the light and spectroscopy concept inventory. *Astronomy Education Review*, *5*(2), 103-113.

Bahri, S., Zoghlami, N., Abed, M., & Tavares, J. M. R. (2018). Big data for healthcare: A survey. *IEEE access*, *7*, 7397-7408.

Barnard, L., Lan, W. Y., To, Y. M., Paton, V. O., & Lai, S.-L. (2009). Measuring self-regulation in online and blended learning environments. *Internet and higher education*, *12*(1), 1-6.

Bartholomew, D.J., Steele, F., Moustaki, I., & Galbraith, J.I. (2008). *Analysis of multivariate social science data* (2nd ed.). Boca Raton (FL): CRC press, Taylor & Francis Group.

Bellini, C., De Santis, A., Sannicandro, K., & Minerva, T. (2019). Data Management in Learning Analytics: Terms and Perspectives. *Je-LKS, Journal of e-Learning and Knowledge Society*, *15*(3), 133-144.

Berthoz, A. (2011). *La semplessità*. (trad. it a cura di F. Niola, 2019). Torino: Codice Edizioni.

Biggs, J., & Tang, C. (2011). *Teaching for Quality Learning at University. What the Student Does* (4th ed.). McGraw-hill education (UK).

Brown, G. T., & Abdulnabi, H.H. (2017). Evaluating the quality of higher education instructor-constructed multiple-choice tests: Impact on student grades. Frontiers in Education, 2, 24.

Buscema, M., & Pieri, G. (2004). *Ricerca scientifica e innovazione. Le parole chiave*. Catanzaro: Rubettino editore.

Cafarelli, B., & Crocetta, C. (2016). An Evaluation of the Student Satisfaction Based on CUB Models. In G. Alleva & A. Giommi (Eds.), *Topics in Theoretical and Applied Statistics*. Springer, Cham.

Centoni, M., & Maruotti, A. (2021). Students' evaluation of academic courses: An exploratory analysis to an Italian case study. *Studies in Educational Evaluation*, *70*, 101054.

Chao, P.-Y. (2016). Exploring students' computational practice, design and performance of problem-solving through a visual programming environment. *Computers & Education*, *95*, 202-215.

Chatti, M.A., Dyckhoff, A.L., Schroeder, U., & Thüs, H. (2012). A Reference Model for Learning Analytics. *International Journal of Technology Enhanced Learning (IJTEL)* – Special Issue on "State-of-the-Art in TEL".

Choi, M., Cristol, D., Gimbert, B. (2018). Teachers as digital citizens: The influence of individual backgrounds, internet use and psychological characteristics on teachers' levels of digital citizenship. *Computers & Education*, *121*, 143-161.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). London and New York: Routledge.

Connell, G.L., Donovan, D.A., & Chambers, T.G. (2016). Increasing the use of student-centered pedagogies from moderate to high improves student learning and attitudes about biology. *CBE—Life Sciences Education*, *15*(1), ar3.

Crocetta, C., D'Ovidio, F., Mancarella, R., Mariella, L., & Toma, E. (2016). Relationships between user characteristics and customer satisfaction about university websites. *Interdisciplinary Journal of Research and Development*, 16, 55-60.

Csikszentmihalyi, M. (2014). The Experience Sampling Method. In M. Csikszentmihalyi, *Flow and the Foundations of Positive Psychology. The Collected Works of Mihaly Csikszentmihalyi* (pp. 21-34). Dordrecht: Springer Science+Business Media.

de Lillo, A., Argentin, G., Lucchini, M., Sarti, S., & Terraneo, M. (2007). *Analisi multivariata per le scienze sociali*. Milano: Pearson Education.

de Lima, C. R. M., Soares, T. C., de Lima, M. A., Veras, M. O., de Andrade, J. B. S. O., & Guerra, A. (2020). Sustainability funding in higher education: a literature-based review. *International Journal of Sustainability in Higher Education*, *21*(3), 441-464.

De Luca, A. M., & Lucisano, P. (2011). Item analisi tra modello e realtà. *Italian Journal of Educational Research*, (7), 85-96.

De Santis, A., Bellini, C., Sannicandro, K., & Minerva, T. (2020). Students' perception on e-proctoring system for online assessment (pp. 161-168). *Enhancing the Human Experience of Learning with Technology: New challenges for research into digital, open, distance & networked educationEuropean Distance and E-Learning Network (EDEN) Proceedings 2020, Research Workshop | Lisbon, 21-23 October, 2020*. Retrieved from: https://www.eden-online.org/proc-2485/index.php/PROC/article/view/1770

De Santis, A., Sannicandro, K., Bellini, C., & Minerva, T. (2019). Predictive Model Selection for Completion Rate in Massive Open Online Courses. *Je-LKS, Journal of e-Learning and Knowledge Society*, *15*(3), 145-159.

De Santis, A., Sannicandro, K., Bellini, C., & Minerva, T. (2021a). Cluster analysis for tailored tutoring system. *Q-TIMES WEBMAGAZINE*, *XIII*(3), 265-277.

De Santis, A., Sannicandro, K., Bellini, C., Cadamuro, A., & Minerva, T. (2021b). Students' academic achievements: clusters based on metacognition, literacy and numeracy skills. *Giornale italiano di educazione alla salute, sport e didattica inclusiva*, *5*(2), 89-101.

Delgado-Algarra, E.J., Román Sánchez, I.M., Ordóñez Olmedo, E., & Lorca-Marín, A.A. (2019). International MOOC trends in citizenship, participation and sustainability: Analysis of technical, didactic and content dimensions. *Sustainability*, *11*(20), 5860.

Demchenko, Y., Membrey, P., Grosso, P., & de Laat, C. (2013a). *Addressing Big Data Issues in Scientific Data Infrastructure*. First International Symposium on Big Data and Data Analytics in Collaboration (BDDAC 2013). Part of The 2013 Int. Conf. on Collaboration Technologies and Systems (CTS 2013), May 20-24, 2013, San Diego, California, USA.

Demchenko, Y., Ngo, C., & Membrey, P. (2013b). *Architecture Framework and Components for the Big Data Ecosystem* [Draft Version 0.2]. Retrieved from http://www.uazone.org/demch/worksinprogress/sne-2013-02-techreport-bdaf-draft02.pdf

Domenici, G., Lucisano, P., & Biasi, V. (2021). *Ricerca sperimentale e processi valutativi in educazione*. Milano: Mc Graw Hill.

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285-296.

Doyle, K., Sainsbury, K., Cleary, S., Parkinson, L., Vindigni, D., McGrath, I., & Cruickshank, M. (2017). Happy to help/happy to be here: Identifying components of successful clinical placements for undergraduate nursing students. *Nurse education today*, *49*, 27-32.

Drabowicz, T. (2017). Social theory of internet use: Corroboration or rejection among the digital natives? Correspondence analysis of adolescents in two societies. *Computers & Education*, *105*, 57-67.

Drachsler, H., & Greller, W. (2016). Privacy and analytics: it's a DELICATE issue a checklist for trusted learning analytics. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 89-98).

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). New Delhi: Prentice Hall of India.

Fabbri, L. (2012). Ricerca didattica e contesti di apprendimento. Nuovi costrutti epistemologici. In P.C. Rivoltella e P.G. Rossi (Eds.). *L'agire didattico. Manuale per l'insegnante*. Brescia: Editrice La Scuola.

Felini, D., & Zobbi, E. (2022). University climate in distance education contexts: developing an assessment instrument. *Journal of e-Learning and Knowldge Society*, 18(1), 75-86.

Ferguson, R. (2014). Learning analytics: fattori trainanti, sviluppi e sfide. *TD Tecnologie Didattiche*, *22*(3), 138-147 (Trad. It. a cura di Davide Taibi e Giovanni Fulantelli).

Ferrari, S. (2013). Berthoz e il paradigma della semplessità. In P.C. Rivoltella (ed.), *Fare didattica con gli EAS* (pp. 63-68). Brescia: Editrice La Scuola.

Fisher, H.E., Boone, W., & Neumann, K. (2014). Quantitative Research Designs and Approaches. In N.G. Lederman & S.K. Abell (Eds.), *Handbook of Research on Science Education*, Vol. 2 (pp. 32-51). New York and London: Routledge.

Forrest, J., Lean, G., & Dunn, K. (2015). Challenging racism through schools: teacher attitudes to cultural diversity and multicultural education in Sydney, Australia. *Race Ethnicity and Education*, *19*(3), 618-638.

Galli, M., & Minerva, T. (1999). Algoritmi genetici per l'evoluzione di modelli lineari. Metodologia ed Applicazioni [Working paper]. Retrieved from: http://morgana.unimore.it/materiali_discussione/0284.pdf

González-González, C. S., Infante-Moro, A., & Infante-Moro, J. C. (2020). Implementation of E-proctoring in Online Teaching: A Study about Motivational Factors. *Sustainability*, *12*(8).

Greller, W., & Drachsler, H. (2012). Translating Learning into Numbers: A Generic Framework for Learning Analytics. *Educational Technology & Society*, *15*(3), 42–57.

Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., & Tatham, R.L. (2014). *Multivariate data analysis* (7th ed.). Edinburgh Gate, Harlow, Essex (GB): Pearson.

Han, I., & Shin, W.S. (2016). The use of a mobile learning management system and academic achievement of online students. *Computers & Education*, *102*, 79-89.

Hynes, H., Stoyanov, S., Drachsler, H., Maher, B., Orrego, C., Stieger, L., Druener, S., Sopka, S., Schröder, H., & Henn, P. (2015). Designing learning outcomes for handoff teaching of medical students using group concept mapping: findings from a multicountry European study. *Academic Medicine*, *90*(7), 988-994.

Hussein, A. A. (2020). Fifty-Six Big Data V's Characteristics and Proposed Strategies to Overcome Security and Privacy Challenges (BD2). *Journal of Information Security*, *11*(4), 304-328.

Istat (2016). *Aspetti della vita quotidiana 2016. Aspetti metodologici della ricerca* [Aspects of daily life 2016. Survey Methodology]. Retrieved from: http://www.istat.it/it/archivio/129956

Iten, N., & Petko, D. (2016). Learning with serious games: Is fun playing the game a predictor of learning success?. *British Journal of Educational Technology*, *47*(1), 151-163.

Kassambara, A. (2017). *Practical Guide to Principal Component Methods in R*. STHDA, Statistical tools for high-throughput data analysis.

Kay, R. H., & Loverock, S. (2008). Assessing emotions related to learning new software: The computer emotion scale. *Computers in Human Behavior*, *24*(4), 1605-1623.

Khalil, M. & Ebner, M. (2015). Learning Analytics: Principles and Constraints. In P*roceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2015 (*pp. 1326-1336). Chesapeake, VA: AACE.

Khan, N., Alsaqer, M., Shah, H., Badsha, G., Abbasi, A. A., & Salehian, S. (2018). The 10 Vs, issues and challenges of big data. In *Proceedings of the 2018 international conference on big data and education* (pp. 52-56).

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1), 1-10.

Kmiotek-Meier, E. A., Skrobanek, J., Nienaber, B., Vysotskaya, V., Samuk, S., Ardic, T., Pavlova, I., Dabasi-Halázs, Z., Diaz, C., Bissinger, J., Schlimbach, T., & Horvath, K. (2019). Why is it so hard? And for whom? Obstacles to intra-European mobility. *Migration Letters*, *16*(1), 31-44.

Kolski, T., & Weible, J. (2018). Examining the relationship between student test anxiety and webcam based exam proctoring. *Online Journal of Distance Learning Administration*, *21*(3).

Kovanović, V., Joksimović, S., Poquet, O., Hennis, T., de Vries, P., Hatala, M., Dawson, S., Siemens, G., & Gašević, D (2019). Examining communities of inquiry in Massive Open Online Courses: The role of study strategies. *The Internet and Higher Education*, *40*, pp. 20-43.

Lang, C., Wise, A.F., Merceron, A., Gašević, D., & Siemens, G. (2022). What is Learning Analytics? In C. Lang, G. Siemens, A. Friend Wise, D. Gašević & A. Merceron (Eds.). *Handbook of Learning Analytics* (2nd. ed.). Vancouver: SoLAR.

Laurillard, D. (2012). *Teaching as a Design Science: Building Pedagogical Patterns for Learning and Technology*. New York and London: Routlegde (Trad. it. di P. Rossi, 2014, FrancoAngeli, Milano).

León-Jariego, J. C., Rodríguez-Miranda, F. P., & Pozuelos-Estrada, F. J. (2020). Building the role of ICT coordinators in primary schools: A typology based on task prioritisation. *British Journal of Educational Technology*, *51*(3), 835-852.

López-Gómez, E., Leví-Orta, G., Medina Rivilla, A., & Ramos-Méndez, E. (2020). Dimensions of university tutoring: a psychometric study. *Journal of Further and Higher Education*, *44*(5), 609-627.

Lucisano, P., & Salerni, A. (2002). *Metodologia della ricerca in educazione e formazione*. Roma: Carocci.

Lynch & J.L., von Hippel. P.T. (2016). An education gradient in health, a health gradient in education, or a confounded gradient in both? *Social Science & Medicine*, *154*, 18-27.

McCandless, D. (2010). *The beauty of data visualization*, TED talk. In https://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization

McCune, B., & Grace, J.B. (2002). Data transformation. In B. McCune & J.B. Grace, *Analysis of Ecological Communities* (pp.67-79). Gleneden Beach, Oregon: MjM Software Design.

Mecatti, F. (2015). *Statistica di base. Come, quando e perché*. Milano: McGraw-Hill Education.

Ministero dell'Istruzione, dell'Università e della Ricerca (2015). *Piano Nazionale Scuola Digitale*. Retrieved from: https://www.istruzione.it/scuola_digitale/allegati/Materiali/pnsd-layout-30.10-WEB.pdf

Morgan, P.L., Farkas, G., Hillemeier, M. & Maczuga, S. (2017). Replicated Evidence of Racial and Ethnic Disparities in Disability Identification in U.S. Schools. *Educational Researchers*, *46*(6), 305-22.

Narushima, M., Liu & J., Diestelkamp, N. (2016). Lifelong learning in active ageing discourse: its conserving effect on wellbeing, health and vulnerability. *Ageing & Society*, *38*, 615-675.

Olshannikova, E., Ometov, A., Koucheryavy, Y., & Olsson, T. (2016). Visualizing Big Data. In *Big Data Technologies and Applications* (pp. 101-131). Springer International Publishing.

Nascimbeni, F. (2020). *Open Education. OER, MOOC e pratiche didattiche aperte verso l'inclusione digitale educativa*. Milano: FrancoAngeli.

Ortogus, J.C. (2016). From the periphery to prominence: An examination of the changing profile of online students in American higher education. *Internet and Higher Education*, *32*, 47-57.

Papadakis, S., Vaiopoulou, J., Kalogiannakis, M., & Stamovlasis, D. (2020). Developing and Exploring an Evaluation Tool for Educational Apps (E.T.E.A.) Targeting Kindergarten Children. *Sustainability*, *12*(10).

Parola, A., & Donsì, L. (2018). Sospesi nel tempo: inattività e malessere percepito in giovani adulti NEET. *Psicologia della salute*, *3*, 44-73.

Paterlini S., & Minerva, T. (2001). Evolutionary Clustering Analysis. *Soft Computing*, *8*,165-176.

Paterlini S., & Minerva, T. (2010). Regression Model Selection using Genetic Algorithms. In V., Munteanu, R., Raducanu, G., Dutica, A., Croitoru, V.E., Balas, & A. Gravilut (Eds.), *Recent Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing* (pp. 19-28). USA: WSEAS Press.

Pattarin, F., Paterlini, S., & Minerva, T. (2004). Clustering fnancial time series: an application to mutual funds style analysis. *Computational Statistics & Data Analysis*, *47*, 353-372.

Perna, L.W., & Leigh, E.W. (2017). Understanding the Promise: A Typology of State and Local College Promise Programs. *Educational Researcher*, *XX*(X), 1–26.

Piccolo, D. (2000). *Statistica*. Bologna: Il Mulino.

Piras, V., Reyes, M.C., & Trentin, G. (2020). *Come disegnare un corso online. Criteri di progettazione didattica e della comunicazione*. Milano: FrancoAngeli.

Plonsky, L., & Ghanbar, H. (2018). Multiple Regression in L2 Research: A Methodological Synthesis and Guide to Interpreting R2 Values. *The Modern Language Journal*, *102*(4), 713-731.

Porta, M., & Rastelli, S. (2013). Lo studio dei tracciati oculari (eye-tracking) nella ricerca sul linguaggio. In S. Rastelli (Ed.), *La ricerca sperimentale sul linguaggio: acquisizione, uso, perdita*. Pavia: Pavia University Press.

Prather, E. E., Rudolph, A. L., Brissenden, G., & Schlingman, W. M. (2009). A national study assessing the teaching and learning of introductory astronomy. Part I. The effect of interactive instruction. *American Journal of Physics*, *77*(4), 320-330.

Puertas, R., & Marti, L. (2019). Sustainability in Universities: DEA-GreenMetric. *Sustainability*, *11*.

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1), 1301013.

Reisbig, A.M., Danielson, J.A., Wu, T.F., Hafen Jr, M., Krienert, A., Girard, D., & Garlock, J. (2012). A study of depression and anxiety, general health, and academic performance in three cohorts of veterinary medical students across the first three semesters of veterinary school. *Journal of veterinary medical education*, *39*(4), 341-358.

Reisenwitz, T. H. (2020). Examining the Necessity of Proctoring Online Exams. *Journal of Higher Education Theory and Practice*, *20*(1).

Rivoltella, P.C. (2018). Pedagogia e razionalità scientifiche. *Scholé: rivista di educazione e studi culturali*, *LVI*, 1/2, 45-63.

Rivoltella, P.C. (Eds.) (2022). *Apprendere a distanza. Teorie e metodi*. Milano: Raffaello Cortina Editore.

Rivoltella, P.C., & Rossi, P.G. (Eds.) (2012). *L'agire didattico. Manuale per gli insegnanti*. Brescia: Editrice La Scuola.

Rizopoulos, D. (2007). ltm: An R package for latent variable modeling and item response analysis. *Journal of statistical software*, *17*, 1-25.

Saarikoski, M., Isoaho, H., Warne, T., & Leino-Kilpi, H. (2008). The nurse teacher in clinical practice: developing the new sub-dimension to the clinical learning environment and supervision (CLES) scale. *International journal of nursing studies, 45* (8), 1233–1237.

Saarikoski, M., & Leino-Kilpi, H. (2002). The clinical learning environment and supervision by staff nurses: developing the instrument. *International journal of nursing studies*. *39*, 259–267.

Sancassani, S., Brambilla, F., Casiraghi, D. & Marenghi, P. (2019). *Progettare l'innovazione didattica*. Milano: Pearson.

Scheffel, M., Drachsler, H., Stoyanov, S., & Specht, M. (2014). Quality indicators for learning analytics. *Journal of Educational Technology & Society*, *17*(4), 117-132.

Schlingman, W.M., Prather, E.E., Wallace, C.S., Rudolph, A.L., & Brissenden, G. (2012). A classical test theory analysis of the Light and Spectroscopy Concept Inventory national study data set. *Astronomy Education Review*, *11*(1).

Schmidt, J.A., Rosenberg, J.M., & Beymer, P.N. (2018). A Person-in-Context Approach to Student Engagement in Science: Examining Learning Activities and Choice. *Journal of Research in Science Teaching*, *55*(1), 19-43.

Schophuizen, M., Kreijns, K., Stoyanov, S., & Kalz, M. (2018). Eliciting the challenges and opportunities organizations face when delivering open online education: A group-concept mapping study. *The Internet and Higher Education*, *36*, 1-12.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary educational psychology*, *19*(4), 460-475.

Schwendimann, B.A., et al. (2017). Perceiving Learning at a Glance: A Systematic Literature Review of Learning Dashboard Research. *IEEE Transactions on Learning Technologies*, *10*(1), 30-41.

Simpson, A. (2017). The surprising persistence of Biglan's classification scheme. *Studies in Higher Education*, *42*(8), 1520-1531.

Song, M. K., Lin, F. C., Ward, S. E., & Fine, J. P. (2013). Composite variables: when and how. *Nursing research*, *62*(1), 45. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5459482/

Srinivasa K.G., & Muralidhar Kurni (2021). *A Beginner's Guide to Learning Analytics*. Cham (Svizzera): Springer.

Stites, A.A., Berger, E., Deboer, J., & Rhoads, J.F. (2019). A Cluster-Based Approach to Understanding Students' Resource-Usage Patterns in an Active, Blended, and Collaborative Learning Environment. *International Journal of Engineering Education*, 35(6/A), 1738–1757.

Stokes, D.E. (1997). *Pasteur's Quadrant. Basic Science and Technological Innovation*. Washington, D.C.: Brooking Institution Press.

Tabachnick, B.G., & Fidell, L.S. (2013). Cleaning Up Your Act: Screening Data. In B.G. Tabachnick & L.S. Fidell, *Using Multivariate Statistics* (6th ed., pp.60-114). Boston: Pearson.

Telea, A. C. (2014). Introduction. In A.C. Telea (2014), *Data visualization: principles and practice*. Boca Raton, FL: CRC Press.

Tomietto, M., Saiani, L., Palese, A., Cunico, L., Cicolini, G., Watson, P., & Saarikoski, M. (2012). Clinical learning environment and supervision plus nurse teacher (CLES+ T) scale: testing the psychometric characteristics of the Italian version. *Giornale italiano di medicina del lavoro ed ergonomia*, *34*(2 Suppl B), B72-80.

Trinchero, R. (2002). *Manuale di ricerca educativa*. Milano: FrancoAngeli.

Trinchero, R., & Robasto, D. (2019). *I mixed methods nella ricerca educativa*. Milano: Mondadori Education.

Trochim, W.M.K. (1989). An introduction to concept mapping for planning and evaluation. *Evaluation and Program Planning*, 12, 1-16.

Uddin, M. F., & Gupta, N. (2014). Seven V's of Big Data understanding Big Data to extract value. In *Proceedings of the 2014 zone 1 conference of the American Society for Engineering Education* (pp. 1-5). IEEE.

Unwin, A., Chen, C., & Härdle, W.K. (2008). Introduction. In C. Chen, W.K. Härdle, & Unwin, A., *Handbook of Data Visualization*. Heidelberg, Germania: Springer.

Vannini, I. (2009). Ricerca empirico-sperimentale in Pedagogia. Alcuni appunti su riflessione teorica e sistematicità metodologica. *Ricerche di Pedagogia e Didattica*, *4*(1). Retrieved from: https://rpd.unibo.it/article/view/1549/922

Vizcaya-Moreno, M.F., Pérez-Cañaveras, R. M., De Juan, J., & Saarikoski, M. (2015). Development and psychometric testing of the clinical learning environment, supervision and nurse teacher evaluation scale (CLES+ T): The Spanish version. *International journal of nursing studies*, *52*(1), 361-367.

Wallace, C. S., Chambers, T. G., & Prather, E. E. (2018). Item response theory evaluation of the Light and Spectroscopy Concept Inventory national data set. *Physical Review Physics Education Research*, *14*(1), 010149.

Wiggins, G., & McTighe, J. (2005). *Understanding by design*. Alexandria, VA: Association for Supervision and Curriculum Development.

Zacharis, N.Z. (2015). A multivariate approach to predicting student outcomes in web-enabled blended learning courses. *Internet and Higher Education*, *27*, 44–53.

Zupic, I., & Čater, T. (2015). Bibliometric methods in management and organization. *Organizational research methods*, *18*(3), 429-472.

Pearson

ANALISI MULTIVARIATA E LEARNING ANALYTICS: METODI E APPLICAZIONI

Negli ultimi decenni il crescente uso di ambienti online per la formazione ha reso disponibili grandi quantità di dati relativi ai processi educativi che, analizzati in rigorosi processi di indagine attraverso tecniche di analisi statistica, possono contribuire a descrivere, conoscere e, in definitiva, migliorare l'apprendimento, l'insegnamento, la pianificazione e l'organizzazione della didattica.

Il volume, delineando il legame esistente fra ricerca educativa, analisi statistica e tecnologie digitali, presenta otto tecniche di analisi multivariata seguite da applicazioni di tali metodi nell'ambito educativo nel contesto internazionale.

Esso intende fornire un contributo per sollecitare l'uso di approcci quantitativi alla ricerca educativa e l'avvicinamento di due comunità, quella degli statistici e quella dei pedagogisti, due "culture" scientifiche e metodologiche la cui collaborazione, in un arricchimento reciproco, è fondamentale per non perdere l'opportunità di usare i tanti dati prodotti oggi per ampliare la conoscenza sul modo in cui apprendiamo e migliorare i nostri sistemi formativi.

***Annamaria De Santis**. PhD in Pedagogia e Scienze dell'Educazione, è docente di "Tecniche per l'analisi dei dati in ambito educativo" e "Instructional design nei contesti digitali" nel corso di laurea in Digital Education dell'Università di Modena e Reggio Emilia.*

Dal 2015 è Instructional Designer presso il Centro Interateneo Edunova della medesima università.

I suoi interessi di ricerca si focalizzano, in particolare, sulle applicazioni di approcci quantitativi alla ricerca educativa e sulla progettazione e valutazione di percorsi formativi mediante l'utilizzo di ambienti e soluzioni digitali.

9788891932419

OPEN ACCESS