Metodi e tecniche della ricerca sociale

Vol. 2
L'analisi dei dati

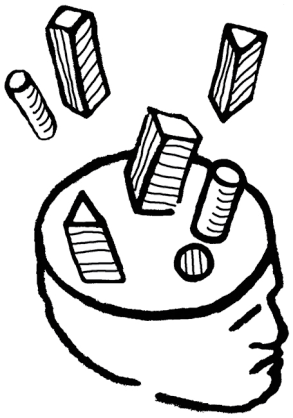
Giovanni Delli Zotti

EUT

BIBLIOTECA DELLA SOCIETÀ APERTA
DIDATTICA **2**

BIBLIOTECA DELLA SOCIETÀ APERTA

DIDATTICA **2**



BIBLIOTECA DELLA SOCIETÀ APERTA
Didattica

DIREZIONE EDITORIALE / EDITORS
Diego Abenante, Serena Baldin, Giuseppe Ieraci, Luigi Pellizzoni

COMITATO SCIENTIFICO / SCIENTIFIC BOARD
Francesco Battegazzorre (Università di Pavia), Matthijs Bogaards (Jacobs University Bremen), Bernardo Cardinale (Università di Teramo), Ian Carter (Università di Pavia), Marco Clementi (Università di Pavia), Giovanni Delli Zotti (già Università di Trieste), Paolo Feltrin (già Università di Trieste), Danica Fink-Hafner (University of Ljubljana), Damian Lajh (University of Ljubljana), Luca Lanzalaco (Università di Macerata), Liborio Mattina (già Università di Trieste), Leonardo Morlino (Luiss Guido Carli Roma), Damiano Palano (Università Cattolica Milano), Lucio Pegoraro (Università di Bologna), Franca Roncarolo (Università di Torino), Guido Samarani (già Università Ca' Foscari Venezia), Mauro Tebaldi (Università di Sassari), Michelguglielmo Torri (già Università di Torino), Luca Verzichelli (Università di Siena)

LOGO DESIGN: Pierax

Il presente volume è stato pubblicato con il contributo del Dipartimento di Scienze politiche e sociali dell'Università degli Studi di Trieste tramite il fondo di ricerca D61-RESRIC-DZOTTI intestato a Giovanni Delli Zotti.

UPI UNIVERSITY PRESS ITALIANE

Opera sottoposta a peer review secondo
il protocollo UPI – University Press Italiane

impaginazione
Elisa Widmar

© copyright Edizioni Università di Trieste, Trieste 2022.

Proprietà letteraria riservata.
I diritti di traduzione, memorizzazione elettronica, di riproduzione e di adattamento totale e parziale di questa pubblicazione, con qualsiasi mezzo (compresi i microfilm, le fotocopie e altro) sono riservati per tutti i paesi.

ISBN 978-88-5511-300-7 (print)
ISBN 978-88-5511-301-4 (online)

EUT Edizioni Università di Trieste
via Weiss 21, 34128 Trieste
http://eut.units.it
https://www.facebook.com/EUTEdizioniUniversitaTrieste

Finito di stampare nel mese di aprile 2022
da Geca Industrie Grafiche - San Giuliano Milanese (MI)

Metodi e tecniche della ricerca sociale

Vol. 2
L'analisi dei dati

Giovanni Delli Zotti

EUT EDIZIONI UNIVERSITÀ DI TRIESTE

Indice

1. Introduzione 9

2. L'analisi monovariata 17

2.1. La costruzione della matrice dei dati 17
2.2. Analisi monovariata: strumenti 20
2.3. Analisi monovariata: esempi 24

3. Analisi bivariata 35

3.1. Finalità e presupposti dell'analisi bivariata 35
3.2. Strategie dell'analisi bivariata 37
3.3. Strumenti dell'analisi bivariata 40
3.3.1. Le tabelle di contingenza 41
3.3.2. I testi di associazione tra variabili categoriali 45
3.3.3. Le percentuali e i dati mancanti nelle tabelle di contingenza 48
3.3.4. I grafici a dispersione 52
3.3.5. I coefficienti e le matrici di correlazione 56
3.3.6. La scomposizione della media 61

4. Analisi multivariata 67
4.1. Il controllo delle terze variabili 68
4.1.1. Tabelle di contingenza e grafici con variabile di controllo 69
4.1.2. Scomposizione della media con variabile di controllo 72
4.1.3. *Box plots* trivariati 73
4.1.4. Grafici a dispersione con variabile di controllo 75
4.1.5. Matrici di correlazione con variabile di controllo 77
4.2. Modelli multivariati asimmetrici 78
4.2.1. Regressione multipla e analisi dei sentieri causali 78
4.2.2. Analisi discriminante 84
4.3. Modelli simmetrici 86
4.3.1. Modelli che aggregano valori (analisi delle corrispondenze multiple) 87
4.3.2. Modelli che aggregano variabili (analisi fattoriale) 92
4.3.3. Modelli che aggregano casi (analisi dei gruppi) 95

5. Costruire e modificare variabili 101
5.1. Uso e utilità della sintassi 102
5.2. Operatori e parole riservate 105
5.3. Comandi di trasformazione dei dati 106
5.3.1. Il comando "compute" (calcola) 107
5.3.2. I comandi "if" (se) e "select if" (seleziona se) 109
5.3.3. Il comando "recode" (ricodifica) 112
5.3.4. "Recode ... into" e "compute/recode" a confronto 117
5.3.5. I comandi "count" (conta), "do if – end if" e "do repeat – end repeat" 119
5.4. Conclusioni 120

6. Categorizzare una variabile 123
6.1. Introduzione 123
6.2. Intervalli di valori di uguale ampiezza 124
6.3. Gruppi di dimensioni uguali o predeterminate 127
6.4. Discontinuità nella distribuzione 130
6.5. Valori significativi o valori soglia 132
6.6. Strategie di segmentazione a confronto 133
6.7. Come "fare a fette" (categorizzare) con Spss 135
6.8. La categorizzazione ottimale 136

7. Costruire indici e tipologie 139
7.1. Indici e/o tipologie 139
7.2. Quattro aspetti che complicano la situazione 143

7.2.1. Attribuire un diverso peso alle variabili? 143
7.2.2. Sommare o moltiplicare? 144
7.2.3. I tipi nelle tabelle di mobilità 145
7.2.4. Se manca l'informazione per una variabile 147
7.3. Quattro strategie combinatorie 148
7.3.1. Soluzione tipologica 149
7.3.2. Scelta prioritaria 154
7.3.3. Soluzione egualitaria 155
7.3.4. Soluzione elitaria 156
7.4. Conclusioni 157

8. Costruire e analizzare tabelle di contingenza multiple 161

8.1. Introduzione 161
8.2. Incrocio con domanda a risposta singola 164
8.3. Incrocio con una domanda a risposta multipla 166
8.4. Incrocio con un set di dicotomie multiple 173
8.5. Tabulazione multipla di parametri statistici 180

9. Gestire i dati mancanti 183

9.1. Introduzione 183
9.2. Variabili indipendenti vs. dipendenti 185
9.3. Pochi vs. molti dati mancanti 186
9.4. Neutralizzare i dati mancanti 188
9.5. Interpretare e legittimare il "non so" 191
9.6. Domande non pertinenti e risposte incoerenti 193
9.7. Quali risposte sono davvero mancanti? 197
9.8. I dati mancanti nell'analisi multivariata 199

10. L'esposizione dei risultati 203

11. Riferimenti bibliografici 209

1. Introduzione

"*Are you experienced?*" è il titolo, apparentemente "stravagante", di un lavoro il cui sottotitolo *Trasformazione di variabili e creazione di indici con SPSS* chiariva che in quelle pagine si parlava di un tema assai rilevante, al quale è dedicato il capitolo 5 di questo volume. Il titolo intendeva rendere omaggio a Jimi Hendrix, un chitarrista "visionario" che non si adagiava sulle sicurezze dell'esistente per battere nuove strade che a oltre cinquant'anni dalla sua scomparsa lo rendono ancora attualissimo. L'omaggio era forse pretenzioso per un *working paper* nato con finalità didattiche, ma l'idea iniziale era dare quel titolo a una pagina internet collegata alla *Memoria storica* della sezione *Metodologia* dell'*Associazione Italiana di Sociologia* (AIS). Quello spazio virtuale, purtroppo mai realizzato, ambiva a proporsi come forum di discussione e deposito di esperienze per utilizzatori di programmi di analisi dei dati che si chiedono se sono veramente *experienced*, avendo consapevolezza che c'è sempre qualcosa da imparare da intuizioni, sperimentazioni ed errori, propri e degli altri. Ci si voleva anche rivolgere a chi semplicemente utilizza i risultati delle elaborazioni statistiche effettuate da altri, perché anche per loro è utile acquisire consapevolezza dei problemi che si presentano in sede di analisi dei dati e delle soluzioni, più o meno efficaci e/o praticabili.

Quel quaderno faceva parte di una serie accomunata da un filo conduttore: illustrare e discutere i passaggi cruciali che si devono affrontare quando si devono costruire gli strumenti di rilevazione e poi analizzare i dati raccolti, due fasi del processo di ricerca strettamente connesse, perché solo se si conoscono in anticipo le opzioni e gli strumenti dell'analisi si può progettare correttamente la rilevazione. È infatti dannoso e anche imbarazzante rendersi conto in fase di analisi che, ad esempio, non si può usare la media per confrontare sinteticamente le opinioni degli intervistati, perché il "formato" delle domande ha generato variabili nominali o ordinali (che non consentono il calcolo della media). È questo il motivo che ci ha indotto a concludere il volume dedicato alla rilevazione dei dati (Delli Zotti 2021) con un paragrafo nel quale si sono illustrati i modi alternativi di porre la stessa domanda: da queste scelte dipendono infatti i tipi di variabili registrate nella matrice dei dati e, di conseguenza, gli strumenti di analisi statistica che si possono utilizzare[1]. Il paragrafo aveva finalità esemplificative di un intero capitolo dedicato alla trattazione della cruciale distinzione tra classificazioni a categorie non ordinate (variabili nominali) e ordinate (variabili ordinali) e variabili che derivano da procedure di conteggio o misurazione (variabili cardinali).

Probabilmente molti lettori ritengono sia difficile addentrarsi nella complessa articolazione dei modelli di analisi statistica di SPSS® (dalla versione 18 denominato PASW) e degli altri analoghi programmi, perché le procedure sono numerose e offrono diverse opzioni e varianti. Fortunatamente, le finestre di dialogo di questi programmi, nate proprio allo scopo di aiutare l'utente, si usano in modo intuitivo e non è perciò difficile ottenere ciò che si vuole, posto che si sappia quale modello di analisi sia più appropriato e utile per le nostre esigenze conoscitive. Inoltre, solo se si conoscono in modo sufficientemente approfondito i modelli di analisi si è in grado di interpretare correttamente i tabulati, a volte sterminati, che il software produce con una certa facilità.

In questa sede non ci limiteremo comunque a suggerire come si risolvono i problemi perché, ci si consenta il gioco di parole, intendiamo anche problematizzarli; con ciò si intende che affronteremo, oltre agli aspetti tecnici (come si fa), quelli metodologici (cos'è meglio fare, se sono disponibili soluzioni alternative).

Riguardo a "come si fa", le procedure verranno descritte in modo discorsivo, ma a volte riprodurremo le istruzioni che consentono di eseguire quelle procedure, utilizzando la sintassi del citato Spss e di Pspp (software *open access* del tutto simile e compatibile)[2]. Gli esempi potranno poi essere adattati per affrontare

---

[1] I contenuti del paragrafo derivano da un saggio (Delli Zotti 1992) il cui titolo rendeva omaggio a un altro artista, prendendo spunto dalla domanda *Il problema più importante…* usata per l'esemplificazione.

[2] Gli esempi possono essere convertiti, più o meno facilmente, nel linguaggio di altri programmi quali lo "storico" SAS e i più recenti STATA e R. Per una più completa descrizione delle

situazioni simili, cambiando i nomi delle variabili, i valori o qualche opzione; in questo modo i lettori potranno replicare quanto qui proposto e, quelli più smaliziati, sperimentare delle varianti. Si è deciso di esemplificare riportando anche le righe di programma, perché vedere le istruzioni e saper interpretare la sintassi dei comandi permette di apprezzare meglio il senso delle operazioni effettuate. Le righe di programma si possono ottenere automaticamente chiedendo che siano tradotte in comandi le scelte effettuate usando le finestre di dialogo. In tal modo rimane registrata la "storia" di trasformazioni e analisi che si possono facilmente replicare ed eventualmente modificare. Per comprendere le esemplificazioni può essere utile una conoscenza, anche solo superficiale, del software, ma riteniamo che tutti i lettori siano in grado di valutare il fondamento logico-razionale e le finalità delle procedure illustrate nel testo.

La trattazione delle tecniche di analisi dei dati inizia ovviamente dall'analisi monovariata (una variabile alla volta), che fornisce la prima risposta ai bisogni conoscitivi del ricercatore e consente il retrospettivo "controllo di qualità" sui dati raccolti. L'analisi monovariata permette inoltre di progettare in modo empiricamente fondato più approfondite analisi che si potranno realizzare utilizzando gli strumenti di analisi bivariata (associazione tra coppie di variabili) e multivariata (a partire dal ruolo delle terze variabili all'interno delle relazioni bivariate) trattati nei capitoli successivi.

In tutte e tre le sezioni le parole chiave che guidano l'esposizione sono "strategie" e "strumenti", con un necessario antefatto sintetizzabile nella parola "presupposti", perché le strategie e gli strumenti sono fortemente condizionati dalle caratteristiche dei dati da analizzare. Come si è già accennato, calcolare la media aritmetica può essere utile a sintetizzare l'età del campione di intervistati, ma la si può utilizzare solo se abbiamo chiesto l'anno di nascita o gli anni compiuti (variabili cardinali); se invece abbiamo chiesto di indicare la propria fascia d'età (variabile ordinale), la media non la potremo calcolare.

Ovviamente, la situazione si complica a livello bivariato perché, entrando in gioco due variabili, gli strumenti utilizzabili dipendono dalle caratteristiche di entrambe e tre tipi di variabili generano evidentemente nove diverse combinazioni, anche se sono possibili alcune semplificazioni. Quanto agli strumenti dell'analisi, in entrambi i casi (monovariata e bivariata), a seconda delle finalità e della natura delle variabili si utilizzano tecniche che "esibiscono" la *distribuzione* dei dati o la *relazione* tra le categorie delle variabili (tabelle di contingenza) o *sintetizzano* entrambe (parametri sintetici descrittivi della distribuzione o dell'associazione tra le variabili). In entrambi i casi è dedicato un congruo spazio anche

procedure e dei comandi si possono consultare, oltre ai manuali forniti con il software e alle spiegazioni all'interno del programma, testi introduttivi quali *L'analisi dei dati con Spss. Guida alla programmazione e alla sintassi dei comandi* (Di Franco 2010).

alla descrizione ed esemplificazione delle *Tecniche grafiche di analisi e rappresentazione dei dati*, per usare il titolo di una monografia pubblicata dall'autore di questo stesso volume (Delli Zotti 2010).

All'inizio del successivo capitolo, dedicato all'analisi multivariata, un paragrafo di "transizione" è dedicato al c.d. "controllo delle terze variabili", il primo passo nella direzione di modelli di analisi più complessi. Questi ultimi sono suddivisi in due categorie, tenendo conto del ruolo che le variabili svolgono all'interno dei modelli, oltre che delle loro caratteristiche tecniche. I modelli multivariati appartengono infatti fondamentalmente a due tipi definibili come asimmetrici o simmetrici. Nel primo tipo il focus dell'analisi è un fenomeno sociale (un comportamento, atteggiamento o un'opinione che svolge il ruolo di variabile dipendente) del quale si analizzano le cause o i fattori condizionanti con l'ausilio di altre variabili che svolgono il ruolo di variabili indipendenti. Nei modelli simmetrici si analizzano invece gruppi di variabili trattate su un piano paritario, nel senso che nessuna di esse svolge un particolare ruolo all'interno del modello. L'intendimento è infatti studiare le reciproche relazioni tra tutte le variabili, anche al fine di individuare caratteristiche sottostanti a sottogruppi di esse definibili come "fattori". Una simile operazione si può effettuare mettendo al centro dell'analisi i casi, invece delle variabili, cercando di creare gruppi di casi che presentino caratteristiche simili. Per completare il quadro dei modelli di analisi multivariata di tipo simmetrico si deve ricordare che in ogni cella di una matrice dei dati "casi per variabili" è registrato il valore rilevato per un singolo caso su una specifica variabile. Dunque, con il terzo tipo di modello di analisi simmetrico sono i valori ad essere messi in relazione reciproca, al fine di verificare quali si "attraggono" e quali invece si "respingono". Si cerca in pratica di comprendere quali tipi di intervistati, definiti sulla base delle loro caratteristiche socio-demografiche, siano attratti o respinti da specifici comportamenti, atteggiamenti o opinioni.

Completato il quadro dei modelli di analisi, caratterizzati da un crescente livello di complessità, i seguenti capitoli del volume sono dedicati ad alcuni temi specifici che, come anticipato, sono stati nel passato trattati da chi scrive in *working paper* come quello che abbiamo citato all'inizio di questa introduzione. Si tratta di approfondimenti su alcuni passaggi problematici, più o meno strettamente legati ai vari modelli di analisi. L'intendimento iniziale era quello di inframezzare, ad esempio, il passaggio dall'analisi monovariata alla bivariata con il capitolo dedicato alla trasformazione delle variabili, perché solitamente si avverte la necessità di effettuare questo tipo di operazioni dopo avere preso confidenza con le variabili tramite l'analisi monovariata e in vista dell'utilizzo di strumenti di analisi bivariata che potrebbero rendere necessarie alcune modifiche e trasformazioni. In realtà, "manipolare le variabili" è un'esigenza che si può rendere

necessaria in qualsiasi fase del processo di analisi e dunque si è preferito inserire i capitoli su questi specifici aspetti tecnici dopo avere completato la "filiera" mono-bi-multivariata.

Nel quinto capitolo dunque, all'interno dalla descrizione degli operatori logici e sintattici e dei comandi del programma Spss che, come si è accennato, hanno comunque una buona comprensibilità ed esportabilità, verranno forniti esempi di costruzione di nuove variabili e modifica di variabili già presenti in matrice che consistono, tra l'altro, nella ricodifica su base semantica di variabili categoriali e nella segmentazione (in classi) di variabili cardinali.

Il capitolo successivo è dedicato proprio a un approfondimento del tema della categorizzazione di una variabile cardinale, una delle operazioni di trasformazione delle variabili che più frequentemente è utile o necessario effettuare. Un esempio lo abbiamo appena incontrato quando si parlava di età che può essere chiesta in termini di anni compiuti oppure, preferendo essere o apparire meno "indiscreti", si può rilevare chiedendo di indicarla con riferimento ad ampie fasce. Se si chiede l'età in modo analitico può poi essere comunque utile una sua successiva categorizzazione al fine, ad esempio, del suo utilizzo nelle tabelle di contingenza. Le modalità di questa categorizzazione nel caso dell'età si appoggia a tradizioni consolidate (fasce decennali, oppure l'utilizzo di valori soglia definiti legalmente, come la maggiore età o l'età del pensionamento), ma se si devono categorizzare altre variabili cardinali (come, ad esempio, il reddito), può essere utile considerare anche criteri di segmentazione quali l'individuazione di "fratture" nella distribuzione dei redditi o la creazione di classi di reddito che contengano una uguale o simile numerosità dei casi.

Sulla stessa linea di distinzione tra tipi di variabili si muove il capitolo successivo che affronta il tema della costruzione di indici (per i quali è richiesto il livello cardinale, o almeno ordinale, delle variabili) e tipologie (con variabili categoriali). Particolarmente in quest'ultimo questo caso ci sono alcuni aspetti preliminari da considerare e poi, passando alla concreta realizzazione, almeno quattro strategie "combinatorie" tra cui scegliere, a seconda delle finalità e delle caratteristiche delle variabili utilizzate: la soluzione tipologica, la scelta prioritaria, la soluzione egualitaria e quella elitaria.

Nell'ottavo capitolo è affrontato il complesso argomento dell'analisi "esplorativa" delle tabelle di contingenza che si avvale spesso dei concetti e delle procedure di trasformazione e creazione delle variabili affrontate nei capitoli precedenti. In pratica, quando si analizzano le risposte a una singola domanda può essere comodo tenere sotto controllo in un'unica tabella di contingenza multipla diverse caratteristiche degli intervistati in modo da poter più facilmente tracciare un "identikit" o profilo socio-economico dei sostenitori dei diversi atteggiamenti o opinioni o di chi più frequentemente si comporta in un determinato modo. Le

tabelle di contingenza multipla più complesse sono però quelle dove l'insieme di variabili indipendenti (descrittive delle caratteristiche degli intervistati) vengono incrociate con domande che prevedono la possibilità di fornire più di una risposta. Questa situazione a livello di analisi dei dati presenta delle specificità tali che software come Spss hanno dedicato alla produzione di questi tipi di tabelle moduli specifici che prevedono la preventiva costruzione degli "insiemi a risposta multipla".

Particolarmente nel caso della costruzione e analisi di tabelle di contingenza multiple, ma non solo, si deve considerare con molta attenzione l'aspetto del corretto utilizzo dei dati mancanti. Il tema non è stato ovviamente ignorato descrivendo le singole tecniche di analisi dei dati, ma meritava una trattazione organica, come dimostra il fatto che, come per la tabulazione delle risposte multiple, anche alla gestione dei dati mancanti dedicano appositi moduli o funzioni i principali software di analisi dei dati.

Il capitolo finale del volume tratta un tema non sempre esplicitamente affrontato nei manuali di metodologia e tecniche della ricerca sociale: l'esposizione dei risultati o "disseminazione", se si vuole usare un termine imprescindibile da quando una consistente parte della ricerca, universitaria e non, viene finanziata dai programmi europei. La disseminazione dei risultati mediante attività di formazione o anche di semplice informazione, è infatti quasi sempre il "pacchetto di lavoro" finale di qualsiasi progetto che aspiri al finanziamento europeo. Dunque, anche se sinteticamente, si sono dedicate alcune pagine finalizzate a sollecitare l'attenzione verso questo passo conclusivo del processo di ricerca che, tra l'altro, fa parte delle attività della c.d. "terza missione" dell'università. Essa ha la finalità di diffondere le conoscenze e incidere sulla crescita economica e sul progresso sociale e viene così definita dall'Anvur (Agenzia Nazionale per la Valutazione del sistema Universitario e della Ricerca):

> Per Terza Missione si intende l'insieme delle attività con le quali le Istituzioni entrano di in interazione diretta con la società, affiancando le missioni tradizionali di insegnamento (prima missione, che si basa sulla interazione con gli studenti) e di ricerca (seconda missione, in interazione prevalentemente con le comunità scientifiche o dei pari). Con la Terza Missione le Istituzioni entrano in contatto diretto con soggetti e gruppi sociali ulteriori rispetto a quelli consolidati e si rendono quindi disponibili a modalità di interazione dal contenuto e dalla forma assai variabili e dipendenti dal contesto (2013: 559).

Anche il presente volume, come numerosi altri recenti lavori che ho realizzato o collaborato a realizzare hanno questo intendimento, essendo pubblicati in *Open Access* con la finalità di favorire il più possibile la loro diffusione. Insomma, "farlo, farlo bene e farlo sapere" è certamente una "frase fatta" ma, perlomeno nell'ambi-

to dei progetti europei, a sorvegliare che lo si faccia (abbastanza) bene ci pensano gli esigenti valutatori che poi giustamente pretendono che lo si faccia sapere[3].

Il presente lavoro completa la coppia di volumi dedicati ai metodi e tecniche della ricerca sociale. Dal momento che la ripartizione della materia è stata pensata anche con l'intendimento di rendere i due volumi relativamente autonomi, in modo che possono "viaggiare" da soli, è opportuno anche in questa sede ricordare che questo lavoro è parte di una "filiera" di attività di ricerca e didattica che parte da lontano e rende doveroso ricordare il ruolo svolto innanzitutto dagli studenti che mi hanno stimolato a cercare di applicare il più possibile criteri di chiarezza e semplicità nella trattazione. Vorrei inoltre ringraziare, altrettanto doverosamente e con molto piacere, i colleghi delle istituzioni di cui ho fatto parte nella mia ormai lunga attività di ricercatore e poi di docente: l'*Istituto di Sociologia internazionale di Gorizia* (Isig), l'*Istituto Economia, società e territorio* (Est) dell'Università di Udine e il *Dipartimento di Scienze dell'uomo* (Dsu), poi confluito nel *Dipartimento di Scienze politiche e sociali* (Dispes), dell'Università di Trieste. Ad essi si aggiunge negli ultimi anni la collaborazione didattica che, iniziata anni fa, sta proseguendo con il *Dipartimento di Studi umanistici* (Disu), dell'Università di Trieste e si è concretizzata anche nell'ambito di alcuni dei citati programmi di ricerca europei.

Molti colleghi, con i quali si sono sviluppati anche rapporti amicali, hanno contribuito in varia misura alla mia stessa formazione, mediante discussioni, incontri, convegni, oltre che con le loro pubblicazioni. Un ringraziamento particolare va ad Alberto Marradi, che mi ha trasmesso l'interesse per la metodologia della ricerca fin dal primo incontro, quando è stato docente di una nuova leva di ricercatori a un corso introduttivo di analisi dei dati nel primo anno della mia permanenza all'*Istituto universitario europeo* (Iue) di Fiesole, dove ho conseguito il dottorato di ricerca. Grazie anche a Fabio Buratto[†], Bernardo Cattarinussi, Maura Del Zotto, Luigi Pellizzoni, Ornella Urpis e altri che hanno letto con passione e attenzione versioni precedenti di questo testo e altri contributi metodologici, alcuni dei quali sono confluiti in questo volume.

[3] Alcuni dei testi alla cui redazione ha partecipato l'autore del volume sono reperibili a partire dalla pagina sul sito Academia.edu (https://trieste.academia.edu/GiovanniDelli Zotti), oltre che nei siti dei dipartimenti nei quali l'autore ha operato e degli editori, come EUT che ha pubblicato il presente volume (https://eut.units.it/It/catalogo).

2. L'analisi monovariata

2.1. La costruzione della matrice dei dati

Come anticipato nell'introduzione, in questa sede ci si occuperà dell'analisi statistica di informazioni strutturate all'interno di una "matrice dei dati" che, ci si perdoni il bisticcio di parole, è "data" (cioè, la si trova "bella e pronta") solo in alcuni casi particolari. Ci riferiamo, ad esempio, ai dati utilizzati nell'analisi secondaria o ai sondaggi online realizzati mediante software che produce la matrice dei dati contestualmente alla compilazione del questionario. Queste tecniche, trattate nel volume dedicato alla rilevazione dei dati (Delli Zotti 2021), sono sempre più frequentemente utilizzate, ma sono ancora prevalenti le tecniche di rilevazione che rendono necessari dei passi preliminari per rendere disponibili all'analisi statistica le informazioni raccolte. Solo dopo avere completato questa fase preparatoria le informazioni diventano infatti "dati" e si potranno dunque utilizzare procedure di analisi statistica selezionate in base alla natura dei dati stessi (livelli di misurazione) e agli scopi della ricerca (descrittivi o inferenziali, esplorativi o confermativi).

Per inserire le informazioni raccolte in una matrice "casi per variabili" si dovrà innanzitutto attribuire a ogni caso un codice identificativo che andrà riportato anche sulla scheda, questionario o altro supporto usato per la rilevazione, in

modo da poterlo facilmente rintracciare per individuare il motivo di eventuali anomalie o errori nei dati e, quando possibile, porvi rimedio. Se si sono utilizzati dei questionari, si dovranno anche codificare "a posteriori" le risposte a domande che non offrivano all'intervistato alternative codificate "a priori" e altre informazioni che nel questionario sono state scritte "in chiaro" per facilitare la compilazione (il comune di residenza può essere stato ad esempio scritto per esteso, ma è poi consigliabile riportarlo in matrice usando un codice numerico).

Il foglio elettronico che si utilizza normalmente per creare tabelle, calcolare somme o rapporti (percentuali) e realizzare efficaci rappresentazioni grafiche, essendo strutturato in righe e colonne è un ottimo esempio di cosa si intende per "matrice dei dati". Non è però necessariamente lo strumento più efficiente per realizzarla, anche se può essere utile, se non altro perché tutti possono disporne, anche gratuitamente. Nelle celle del foglio elettronico si possono indifferentemente scrivere testi o numeri di lunghezza variabile, ma questa "flessibilità" di utilizzo non è un vantaggio se si deve realizzare una matrice dei dati. In questo caso si deve infatti "avvertire" il programma che si è completato l'inserimento di ogni singola informazione mediante il tasto "Invio", o spostando il cursore in un'altra casella. Con programmi specifici di archiviazione dati si può invece prefissare la natura delle informazioni che si prevede di inserire nelle singole colonne (numeri, stringhe di testo, date o valori monetari) e si può anche fissare la lunghezza del campo (ad esempio, due cifre per l'età o quattro per l'anno di nascita). In questo caso il programma sposta automaticamente il cursore nel campo successivo quando si è completato l'inserimento e può così diminuire di molto il numero di battute necessarie per inserire i dati. Inoltre, una segnalazione acustica può avvertire che si è completato l'inserimento e ciò è utile se stiamo inserendo codici fiscali, di avviamento postale o qualsiasi altra informazione che abbia una lunghezza in caratteri prefissata.

I più sofisticati programmi di archiviazione possono agevolare queste operazioni mediante funzioni che consentono controlli di ammissibilità (segnalando, ad esempio, l'inserimento di lettere o altri simboli estranei in un campo definito come numerico), la congruenza dei codici immessi con dati inseriti in precedenza e anche il salto automatico di campi che non si devono compilare a causa di domande "filtro" poste in precedenza (i campi relativi alle domande non poste possono essere compilati automaticamente con il codice "non pertinente").

Anche per l'analisi dei dati è possibile utilizzare il foglio elettronico, ma professionalmente si utilizzano software specifici quali quelli citati nell'introduzione. Sono programmi complessi e completi che, fin dalla loro origine, hanno comportato una vera e propria rivoluzione nell'elaborazione statistica dei dati, in precedenza affidata ai soli specialisti che erano in grado di produrre algoritmi di calcolo usando linguaggi di programmazione che, all'inizio degli anni '70, erano assai più rudimentali di quelli ora in uso.

Per lungo tempo questi programmi potevano essere utilizzati solo da chi poteva recarsi in un centro di calcolo, dovendo perciò apprendere, oltre al linguaggio di programmazione del software per l'analisi statistica, almeno i rudimenti dei criptici linguaggi dei sistemi operativi che "facevano girare" questi programmi sui "*mainframe*" di cui i centri di calcolo erano dotati. L'avvento dei personal computer e il progressivo miglioramento delle loro prestazioni ha però ben presto reso possibile utilizzare i "pacchetti" statistici anche con questi nuovi e più accessibili strumenti; di conseguenza, si è decisamente ampliato il raggio dei possibili utenti. A quel punto si sono levate le preoccupate reazioni dei detentori del sapere statistico-informatico, allarmati dal fatto che l'utilizzo di questi strumenti da parte dei "laici" potesse portare nocumento alla qualità della ricerca: timori in buona misura ingiustificati e forse motivati anche dalla paura di alcuni di perdere parte del potere di "mandarini" dell'elaborazione statistica[1].

Un uso scorretto delle tecniche dovuto alla loro maggiore accessibilità e facilità d'uso è ovviamente possibile, ma ciò vale a tutti i livelli. Con gli strumenti informatici adatti è possibile, ad esempio, scrivere musica su uno spartito elettronico suonando un pianoforte digitale collegato al computer. Questa tecnologia avrebbe probabilmente permesso a Beethoven di scrivere diciotto sinfonie invece di nove, e gli avrebbe fatto risparmiare soldi di copisteria, ma non può aiutare chi al pianoforte non riesce ad abbozzare alcunché di decente. Allo stesso modo, un software efficiente può aiutare un sociologo preparato metodologicamente e statisticamente, ma non più di tanto un ricercatore sprovveduto.

Va infine ricordato che, dopo essere entrati stabilmente nelle memorie dei personal computer, i software per l'elaborazione statistica dei dati hanno adottato da tempo interfacce (sistemi di interazione con l'utente) "amichevoli"[2]. Inoltre, dei modelli di analisi statistica di cui veniva lamentata l'assenza nei pacchetti in voga qualche tempo fa, ora ben pochi mancano all'appello. Anche per tematiche specifiche e meno diffuse, come l'analisi dei reticoli sociali, è ormai disponibile software specifico e si sono diffusi numerosi programmi per l'analisi del contenuto di interviste o di altri documenti, alcuni dei quali sono gratuiti perché prodotti in licenza *Free/Open Source* o *Creative Commons*[3].

---

[1] Su questo tema si vedano, ad es., Ricolfi (1982) e Colombis (1989).

[2] Ad esempio, il "sistema aperto" di analisi statistica "R" utilizza principalmente un'interfaccia a righe di comando, ma sono disponibili interfacce grafiche prodotte da sviluppatori autonomi. A chi preferisce interagire con il software mediante "finestre di dialogo", si può comunque consigliare il già citato PSPP che, come già accennato, è del tutto compatibile con il più noto Spss (ora Pasw).

[3] Riguardo all'analisi dei reticoli sociali si può citare il classico lavoro di Chiesi (1999) e quello più recente di Pavan (2016), che illustra dettagliatamente uno dei software specifici finalizzati alla *network analysis*. Per le tecniche di analisi del contenuto si possono citare una serie di volumi dedicati a specifici programmi quali *Atlas-Ti* e *MAXQDA* (De Gregorio e Lattanzi 2011), *Spad* (Metastasio e Cini 2009), *T-Lab* (Lancia 2004) e *NVivo* (Pacifico e Coppola 2010; Coppola 2011).

Qualsiasi software si utilizzi, effettuare un'analisi monovariata significa esplorare il contenuto informativo di ogni singola variabile della matrice dei dati. Prima ancora di entrare nel merito dei risultati conoscitivi "sostantivi", l'analisi monovariata consente infatti di valutare caratteristiche e qualità dei dati raccolti, oltre a supportare le decisione sulle tecniche e variabili da utilizzare in eventuali successive analisi bivariate e multivariate[4].

Come sopra accennato e ampiamente illustrato nel volume precedente, dedicato alla rilevazione dei dati (Delli Zotti 2021), nella matrice "casi per variabili" ogni riga rappresenta un caso appartenente alla "popolazione" individuata dalla scelta dell'unità di analisi (individuo, famiglia, azienda, partito politico, ecc.) e di uno specifico ambito spaziale e temporale. Ogni colonna rappresenta una variabile, cioè quella particolare proprietà o attributo di cui il ricercatore ha rilevato presenza/assenza o "stato" in ognuno dei casi rilevati, e l'analisi monovariata consiste in procedure statistiche che consentono di descrivere, analiticamente o sinteticamente, l'informazione in essa contenuta.

In linea di principio è possibile effettuare una sorta di analisi monovariata anche prendendo in considerazione una alla volta le righe della matrice invece delle colonne, ma in questo caso sarebbe forse più appropriato parlare di *case study*. Esaminare gli elementi descrittivi di un caso specifico può servire a caratterizzarlo, attribuirlo a un particolare tipo o categoria e confrontarlo con altri casi presenti in matrice ed è evidente che tali operazioni sono possibili solo se il numero dei casi è limitato (come è tipico della ricerca comparata). Può essere invece difficile immaginare che si possa sintetizzare (almeno parzialmente) l'informazione presente in una riga ma, come vedremo nel capitolo 7, quando si costruisce un indice utilizzando i valori rilevati su una serie di indicatori si fa proprio questo[5].

Tornando all'analisi monovariata "tradizionale", il primo passo consiste nella realizzazione di tabelle di distribuzione di frequenza, nelle quali è riportato il risultato del conteggio del numero di casi che ricadono in ogni modalità della variabile. Come si è anticipato, ispezionare le distribuzioni di frequenza consente anzitutto di individuare eventuali anomalie: valori non previsti nel piano di codifica, presenza inaspettata o eccessivamente elevata di risposte mancanti, frequenze che si discostano molto da valori prevedibili e possono

---

[4] L'analisi monovariata è trattata con completezza in Marradi (1993). In generale, sull'analisi dei dati un ottimo testo introduttivo è costituito dal *EDS: Esplorare, descrivere e sintetizzare i dati* (Di Franco 2001).

[5] Per un approfondimento sugli aspetti metodologici e tecnici della costruzione di indicatori, anche mediante queste tecniche "combinatorie", si veda Parra Saiani (2009).

far pensare a errori di rilevazione e/o di codifica. Come pure abbiamo anticipato, spesso è possibile porre rimedio individuando il caso che ha prodotto l'anomalia e correggere il dato dopo avere controllato la relativa scheda o il questionario. Quando l'errore si è invece verificato a livello di rilevazione può essere più difficile rimediare; è perciò consigliabile, se possibile, effettuare controlli delle distribuzioni di frequenza senza attendere il completamento della rilevazione, perché più precoce è l'individuazione dell'errore, più è facile porvi rimedio.

Esaminare le distribuzioni di frequenza è importante anche con riguardo alle analisi successive. Ad esempio, ai fini della realizzazione di tabelle di contingenza (che, come vedremo, sono poco utili se di dimensioni troppo grandi perché è eccessivamente elevato il numero delle celle), si può ridurre la dispersione di una classificazione riconducendo le modalità di risposta con frequenza molto bassa alla categoria "altro", oppure aggregando categorie con contenuto semantico affine. Analogamente, si può ridurre a intervalli una variabile cardinale come l'età, costruendo classi di età quinquennali o decennali (cfr. § 6.2).

Si può anche disaggregare una categoria con un numero eccessivo di casi; un problema più complesso che si può risolvere usando una seconda variabile per discriminare all'interno della prima. Se abbiamo rilevato la professione e gli impiegati sono molto numerosi, si possono dividere tra impiegati pubblici e privati se un'altra variabile ci fornisce questa informazione.

Un importante passo successivo alla tabulazione delle frequenze assolute è il calcolo delle percentuali (frequenze relative) anche perché, per una presentazione comunque completa e senz'altro più efficace di una distribuzione, è sufficiente riportare in tabella le percentuali, corredate della *base di riferimento* per il loro calcolo (il totale in valore assoluto); in tal modo è possibile risalire al numero di casi che ricadono in ogni categoria. Tenendo presente che spesso sono presenti risposte mancanti (carenze della fonte, rifiuto, inesistenza dell'atteggiamento o comportamento, non pertinenza della domanda per alcuni rispondenti), si possono/devono poi calcolare le percentuali corrette o "aggiustate", escludendo i casi su cui non è stato rilevato il dato[6].

L'importanza delle percentuali, specialmente nelle indagini campionarie, è immediatamente percepibile: le frequenze percentuali sono infatti riportate a un'unità convenzionale e ciò permette, all'interno della stessa variabile, di relazionare ogni modalità di risposta a un totale noto e costante (100), oltre a più facili raffronti tra distribuzioni riferite a insiemi di casi di diversa numerosità.

---

[6] Si è scritto possono/devono perché l'inclusione/esclusione dei dati mancanti è questione da valutare con attenzione e, proprio per questo motivo, abbiamo dedicato al tema della gestione dei dati "mancanti" un capitolo in questo volume (cfr. capitolo 10).

Per quanto concerne i parametri statistici che forniscono informazioni sintetiche sulle caratteristiche delle variabili, quelle di tipo nominale (comune di residenza o corso di laurea frequentato) si possono descrivere sinteticamente solo utilizzando la *moda*, che indica semplicemente la categoria che raccoglie il maggior numero di casi. Con le variabili ordinali (titolo di studio o livello di gradimento per una proposta) e cardinali (età, reddito o numero di figli), si possono invece utilizzare un maggior numero di strumenti statistici che si possono suddividere in *valori caratteristici* e *misure sintetiche*.

Tra i valori caratteristici si devono citare, oltre alla moda, il *massimo*, il *minimo* della distribuzione e la *mediana* (il valore che non ha più del 50% dei casi né dall'una né dall'altra parte della distribuzione) che, per il fatto di occupare una precisa posizione nella distribuzione, sono definiti *valori posizionali*. Dividendo la distribuzione in quattro parti, si individuano altri valori posizionali: i *quartili* (il primo è il valore che ha il 25% dei casi da una parte e il 75% dall'altra, il secondo corrisponde alla mediana e il terzo ripartisce la distribuzione al contrario del primo quartile). Dividendo la distribuzione in dieci o cento parti si possono anche individuare i *decili* e i *centili*.

Gli indici di *tendenza centrale* e di *variabilità* si definiscono misure sintetiche e derivano da procedure di calcolo. Per la tendenza centrale si usa la *media*, che si ottiene dividendo la somma dei valori per il numero dei casi (ad esempio, la somma dei voti divisa per il numero di esami sostenuti). Utilizzando i valori posizionali la variabilità si misura come ampiezza del *campo di variazione* (differenza tra massimo e minimo). A partire dalla media si può invece calcolare la media degli scarti (*scostamento semplice medio*), presi in valore assoluto per impedire che la somma sia "0": il totale degli scarti positivi eguaglia infatti per definizione quello degli scarti negativi. Si può anche utilizzare la *deviazione standard* (*scarto quadratico medio* o *scarto-tipo*) che si calcola sommando gli scarti dalla media elevati al quadrato (*devianza*) che va poi divisa per il numero dei casi, ottenendo la *varianza*. L'elevazione al quadrato risolve il problema del segno (anche i valori negativi diventano positivi se elevati al quadrato), ma costringe a estrarre la radice quadra della varianza per giungere infine alla deviazione standard, espressa nella stessa unità di misura dei valori osservati.

Si può infine ricordare un'ultima interessante caratteristica delle distribuzioni: l'approssimazione alla *curva normale* (detta anche gaussiana, dal nome dello statistico Gauss). Una distribuzione è "normale" se i casi si affollano al centro e diventano meno numerosi allontanandosi verso i due estremi della distribuzione. Ad esempio, la distribuzione delle altezze in una popolazione è normale se sono numerosi gli individui di altezza vicina alla media e la numerosità scende quanto più ci si avvicina ad altezze da "nani" o "giganti". Questa caratteristica è molto importante, perché l'espressione "la distribuzione approssima la norma-

le" è un modo efficace per descriverla e perché questa forma della distribuzione è uno dei requisiti per poter applicare strumenti dell'analisi bivariata come la correlazione, su cui si si basano numerosi modelli di analisi multivariata. Purtroppo, nelle scienze sociali le proprietà che si possono rilevare tramite variabili cardinali non sono molto numerose e i fenomeni che descrivono raramente tendono ad assumere una distribuzione normale.

Le tabelle con le distribuzioni di frequenza e i parametri statistici possono spesso essere sostituiti da grafici, generalmente più "attraenti" e di facile lettura. È però alquanto riduttivo considerarli solo un modo alternativo di presentare i risultati, perché a volte costituiscono veri e propri strumenti di analisi che dovrebbero entrare a far parte del bagaglio tecnico di ogni ricercatore.

Everitt (1978) suggerisce che le tecniche grafiche si possono utilizzare principalmente per tre scopi, che sono peraltro gli obiettivi di ogni analisi dei dati:

1. rappresentazione e descrizione;
2. analisi e interpretazione;
3. sintesi ed esposizione.

In pratica, le tecniche grafiche sono utili per vari scopi che vanno:

> dalla presentazione dei dati grezzi, alla presentazione dei risultati di complesse analisi, ad *output* grafici che essi stessi costituiscono l'analisi statistica fornendo sia un sommario del contenuto informativo dei dati che un'esposizione di caratteristiche non previste, come pure possibili inadeguatezze del modello assunto. L'uso di diagrammi può spesso rivelare aspetti che potrebbero essere sfuggiti all'attenzione nel caso in cui i dati fossero tabulati unicamente in forma numerica; possono aiutare a scoprire caratteristiche dei dati completamente impreviste prima dell'analisi, o suggerire ipotesi che possano essere ulteriormente investigate (Everitt 1978: 1).

Tra le forme più utilizzate vi sono i *grafici a barre* che usano rettangoli di lunghezza proporzionale alla frequenza. Da questi bisogna distinguere gli *istogrammi*: rettangoli di altezza uguale alla frequenza e base uguale all'intervallo di classe (come già accennato e si vedrà nel capitolo 6 le classi si costruiscono riducendo a intervalli una proprietà continua come l'età). Quando le classi sono numerose gli istogrammi possono essere sostituiti da *poligoni di frequenza* congiungendo il centro degli intervalli di classe.

Se le categorie della variabile non sono ordinate è corretto utilizzare i *grafici a torta*, perché la disposizione degli spicchi non suggerisce un ordine di successione delle categorie, cosa che può accadere, anche implicitamente, a causa della disposizione delle barre da destra verso sinistra. I grafici a torta sono inoltre molto indicati per rappresentare percentuali, in quanto sono un'ottima metafora di un "intero" di cui le percentuali sono parte.

Sono infine disponibili grafici di più recente introduzione che saranno descritti in occasione del loro utilizzo in quanto, essendo meno conosciuti, è più agevole descrivere le loro caratteristiche contestualmente alla loro esemplificazione. Si tratta, ad esempio, delle *barre degli errori* che si utilizzano per rappresentare le medie delle distribuzioni associate agli intervalli di confidenza o alla deviazione standard, utilizzabili solo con variabili cardinali (con gli altri tipi di variabili non si può infatti calcolare la media) e *dei grafici a scatola* (*box plots*) che si possono utilizzare anche con le variabili ordinali.

2.3. Analisi monovariata: esempi

Passiamo ora a qualche esemplificazione utilizzando innanzitutto un campione di 2.745 maggiorenni intervistati in Italia nel 2019 nell'ambito dell'*European Social Survey* (http://www.europeansocialsurvey.org) evidenziando in questo modo anche le potenzialità dell'analisi secondaria discusse ampiamente nel volume dedicato alla rilevazione dei dati (Delli Zotti 2021)[7]. A riprova di queste potenzialità basti dire che i dati qui utilizzati sono parte di oltre 400.000 interviste realizzate negli ultimi 20 anni in tutti i paesi europei e i relativi "micro dati" (le matrici casi per variabili) sono liberamente scaricabili dal sito del progetto.

La prima distribuzione di frequenza è relativa a una domanda sul rapporto di lavoro (Tab. 2.1) e nella prima colonna sono innanzitutto distinte le risposte valide e mancanti. Le risposte valide sono quelle previste dalla definizione operativa della variabile "rapporto di lavoro" e cioè la ripartizione tra lavoratori dipendenti, autonomi e coadiuvanti familiari. L'elenco delle risposte mancanti presenti nella tabella è piuttosto esaustivo: domanda non applicabile (l'intervistato non è un lavoratore), rifiuto a rispondere, "non so" (strano in questo caso). Si potrebbe aggiungere "non risposto" per altri motivi, quali la dimenticanza di contrassegnare la risposta, come a volte accade con i questionari auto-compilati. Nella prima colonna di dati sono riportati i valori assoluti, cioè il numero di intervistati che hanno risposto (o non risposto) nei vari modi e, come si vede, la categoria modale (moda) è "dipendente" (1.593 dei 2.745 intervistati). Nella quarta colonna sono riportate le frequenze percentuali, dalle quali si evince che i 1.593 dipendenti corrispondono al 58% del totale.

Come abbiamo visto però, il 24,7% degli intervistati non ha risposto alla domanda e perciò nell'ultima colonna, nella quale le percentuali sono ricalcolate sul totale delle 2.066 risposte valide, vediamo che i lavoratori dipendenti cor-

[7] Si veda anche Bolcati-Rinaldi e Vezzoni (2012), che sul ruolo dell'analisi secondaria sono espliciti fin dal sottotitolo *Come rispondere a nuova domande con dati già raccolti*.

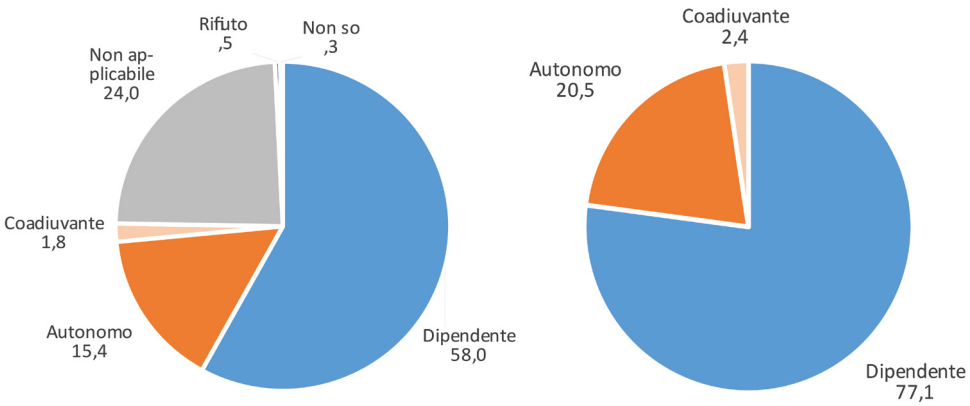
rispondono al 77,1% degli intervistati che lavorano. Tutto ciò è ben rappresentato nei due grafici a torta della Fig. 2.1 dove a sinistra è visualizzata la distribuzione percentuale dell'intero campione (compresi gli intervistati che non lavorano o non hanno indicato il rapporto di lavoro) e a destra la ripartizione dei soli occupati.

Tabella 2.1 - Distribuzione di frequenza di una variabile nominale – Rapporto di lavoro (Italia 2017)

		N	%	% valida
Validi	Dipendente	1.593	58,0	77,1
	Autonomo	424	15,4	20,5
	Coadiuvante familiare	49	1,8	2,4
	Totale	2.066	75,3	100,0
Mancanti	Non applicabile	658	24,0	
	Rifiuto	14	,5	
	Non so	7	,3	
	Totale	679	24,7	
Totale		2.745	100,0	

Fonte: European Social Survey (http://www.europeansocialsurvey.org)

Figura 2.1 - Grafico a torta della distribuzione di frequenza di una variabile nominale – Rapporto di lavoro (Italia 2017)



Fonte: European Social Survey (http://www.europeansocialsurvey.org)

Passiamo ora alla distribuzione di frequenza di una variabile ordinale: il titolo di studio degli intervistati (Tab. 2.2). L'ordinalità della variabile consente di accorpare le categorie contigue e si può perciò notare nella colonna delle percentuali cumulate che poco meno di metà del campione (44,6%) non ha studiato oltre la scuola dell'obbligo (oltre metà degli intervistati è dunque diplomato o laureato). Visto che cumulando fino al livello della scuola media non si raggiunge il 50%, la categoria mediana è "diploma", che è anche la categoria modale (meno di 10 anni fa la mediana si collocava ancora al livello della scuola media, il che significa che oltre metà degli intervistati non erano andati oltre la scuola dell'obbligo). Infine, per quanto sia poco significativa questa caratteristica per una variabile ordinale con categorie poco numerose, si può notare che la variabile approssima la distribuzione normale (sono molto più numerosi gli intervistati che hanno titoli di studio collocati al centro della distribuzione).

Tabella 2.2 - Distribuzione di frequenza di una variabile ordinale – Titolo di studio (Italia 2017)

		N	%	% valida	% cumulata
Validi	Nessun titolo	56	2,0	2,1	2,1
	Elementare	318	11,6	11,6	13,7
	Media	843	30,7	30,9	44,6
	Diploma	1.131	41,2	41,4	86,0
	Laurea	382	13,9	14,0	100,0
	Totale	2.730	99,5	100,0	
Mancanti	Rifiuto	13	,5		
	Non so	2	,1		
	Totale	15	,5		
Totale		2.745	100,0		

Fonte: European Social Survey (http://www.europeansocialsurvey.org)

Alcuni di questi aspetti sono ancora più evidenti nel caso della scala auto-ancorante rappresentata nella Tab. 2.3. La domanda era formulata in questo modo: "Complessivamente, quanto è soddisfatto della sua vita attuale? Per favore risponda usando questo cartellino, dove 0 indica totalmente insoddisfatto e 10 totalmente soddisfatto". In pratica, si chiedeva di assegnare al proprio grado di soddisfazione un punteggio da "0" a "10", invece di "ancorarlo" a termini che esprimono l'intensità di tale atteggiamento (molto, abbastanza, poco o per nulla). Questo tipo di scala si definisce auto-ancorante perché si "appoggia" semanticamente solo sulle posizioni estreme. Rinunciando a definire le categorie intermedie si può legitti-

mamente presumere che le distanze tra i punteggi siano equivalenti e, conseguentemente, trattare questo tipo di scale come cardinali. Ci fossero dubbi in merito al fatto che si tratta di un particolare tipo di misurazione, si pensi che si sarebbe potuto chiedere di esprimere un punteggio da "0" a "100" e si poteva anche tracciare nel questionario un segmento e chiedere all'intervistato di indicare con una freccia la sua posizione tra i due estremi, misurando poi la distanza.

Tabella 2.3 - Distribuzione di frequenza di una scala auto-ancorante – Quanto sei insoddisfatto/soddisfatto della tua vita attuale? (Italia 2017)

		N	%	% valida	% cumulata
Validi	Insoddisfatto	33	1,2	1,2	1,2
	1	29	1,1	1,1	2,3
	2	63	2,3	2,3	4,6
	3	78	2,8	2,9	7,5
	4	101	3,7	3,7	11,2
	5	214	7,8	7,9	19,1
	6	383	14,0	14,1	33,2
	7	616	22,4	22,7	55,9
	8	746	27,2	27,5	83,4
	9	234	8,5	8,6	92,1
	Soddisfatto	215	7,8	7,9	100,0
	Totale	2.712	98,8	100,0	
Mancanti	Rifiuto	20	,7		
	Non so	13	,5		
	Totale	33	1,2		
Totale		2.745	100,0		

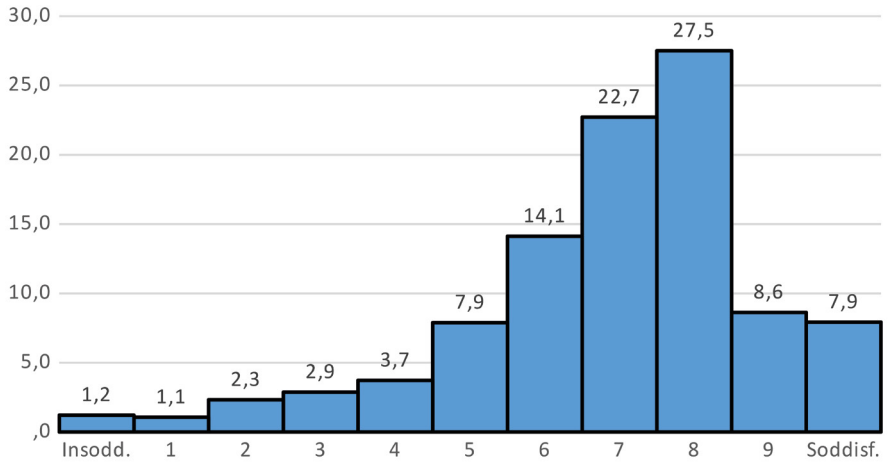
Fonte: European Social Survey (http://www.europeansocialsurvey.org)

Trattandosi dunque di una variabile che si può trattare come cardinale, è a maggior ragione legittimo cumulare le percentuali e si può vedere, ad esempio, che sommando i punteggi da "0" a "4" (che indicano diversi gradi di insoddisfazione) si arriva a un totale di poco più del 10%. Poco meno del 20% si colloca esattamente a metà strada e, per differenza, si ricava che circa il 70% si dichiara più o meno soddisfatto. Questo squilibro si può "leggere" anche osservando che la mediana è collocata sul valore "7", due punti oltre il valore centrale della scala. Trattandosi di una scala quasi-cardinale, perché la proprietà è rilevata tramite punteggi che identificano livelli di soddisfazione condivisi da numerosi intervistati, la media-

na non corrisponde al valore rilevato in un singolo caso, come si potrebbe verificare, ad esempio, rilevando con esattezza il reddito degli intervistati o il prodotto nazionale lordo di una serie di paesi. Nel nostro esempio la mediana corrisponde a una categoria che raccoglie quasi il 30% degli intervistati e, di conseguenza, a sinistra della mediana troviamo un terzo del campione e alla sua sinistra quasi il 45%.

Il riferimento a posizioni in termini di destra/sinistra allude alla rappresentazione della distribuzione di frequenza tramite il grafico a barre della Fig. 2.2. Si noti che, mentre nel caso di una variabile nominale è corretto distanziare le barre a sottolineare l'autonomia semantica delle categorie (si pensi alla regione di residenza degli intervistati), se si rappresenta una variabile quasi-cardinale è appropriato accostare le barre a sottolineare la contiguità semantica dei punteggi. La rappresentazione grafica è, come si vede, efficace nel mostrare l'asimmetria della distribuzione, e cioè l'addensamento dei casi sul versante che esprime livelli crescenti di soddisfazione per la propria vita. Ovviamente, la quasi-cardinalità della scala permette di calcolare anche la media: nel nostro caso pari a 6,89%, dunque molto vicina alla mediana, e anche questo parametro indica che la distribuzione è (fortunatamente) sbilanciata sul versante della soddisfazione.

Figura 2.2 - Grafico a barre - Quanto sei insoddisfatto/soddisfatto della tua vita attuale? (Italia 2017)



Fonte: European Social Survey (http://www.europeansocialsurvey.org)

Che i parametri sintetici siano utili per riassumere l'informazione contenuta in una distribuzione di frequenza risulta ancor più evidente quando si comparano più distribuzioni. Nel questionario dell'*European Social Survey* si chiedeva

ad esempio un giudizio sul grado di soddisfazione, oltre che per la propria vita, per alcuni aspetti della vita del paese (nelle nostre esemplificazioni ci limitiamo all'Italia). Nella Tab. 2.4 sono indicati i principali parametri statistici, con le variabili ordinate secondo valori decrescenti della media. Al primo posto si colloca la valutazione per la situazione personale con una media, come abbiamo visto sopra, di 6,89 che colloca decisamente questo aspetto sul versante della soddisfazione. Altri aspetti si collocano sul versante positivo (soddisfazione) perché hanno medie superiori a "5", anche se alquanto prossime alla "metà strada". Gli ultimi due aspetti (governo ed economia) si collocano invece sul versante dell'insoddisfazione. Non essendo possibili valori estremi, come nel caso del reddito, perché le risposte sono confinate entro un intervallo prefissato e non molto ampio, i valori della deviazione standard sono molto simili, ma c'è un po' più disparità di giudizio sui servizi sanitari e sul governo nazionale di quanta si riscontri a proposito della valutazione sulla propria vita o sul sistema educativo.

Dalle colonne di dati sulla sinistra si nota che quasi tutti gli intervistati hanno espresso un'opinione su sé stessi e sui servizi sanitari, mentre un po' più del 6% non risponde a proposito del governo. Nell'ultima riga è riportato in numero di risposte valide e mancanti "*listwise*" (per il complesso di variabili): in termini percentuali, l'88% ha risposto a tutte le domande; il restante 12% non ha risposto a tutte o anche a una sola domanda. È un aspetto di rilievo in particolare nell'ambito dell'analisi multivariata, perché in quel caso si deve decidere se si preferisce che tutte le variabili abbiano lo stesso livello di rappresentatività statistica, oppure accettare un po' di differenziazione per evitare una perdita di casi anche molto rilevante se le mancate risposte sono numerose e disposte in modo "erratico" tra le variabili, come in parte è in questo caso.

Tabella 2.4 - Parametri statistici descrittivi di una serie di scale auto-ancoranti – Grado di soddisfazione per la propria vita e per aspetti della situazione nazionale (Italia 2017)

	Validi	**Mancanti**	**Moda**	**Mediana**	**Media**	**Dev. std.**
Situazione personale	2.712	33	8	7	6,89	2,03
Servizio sanitario	2.712	33	7	6	5,77	2,29
Sistema educativo	2.648	97	6	6	5,56	2,04
Vita democratica	2.618	127	6	5	5,12	2,14
Governo nazionale	2.582	163	5	5	4,52	2,29
Stato dell'economia	2.648	97	5	4	4,08	2,09
Validi (listwise)	2.416	329				

Fonte: European Social Survey (http://www.europeansocialsurvey.org)

Anche se si è evocata l'analisi multivariata, il fatto che nella tabella siano illustrate sinteticamente una serie di variabili non deve indurre in errore riguardo al tipo di analisi effettuata: la tabella permette infatti di confrontare ad esempio le valutazioni sui servizi sanitari e sul sistema educativo, ma non rivela nulla riguardo ai rapporti tra questi due atteggiamenti. Confrontare infatti non significa analizzare congiuntamente perché solo l'analisi bivariata consentirebbe di accertare se chi è critico sul sistema sanitario lo è anche nei confronti della scuola o se, invece, le risposte sono indipendenti (parte dei critici del sistema sanitario potrebbero essere invece soddisfatti della scuola, o viceversa).

Tabella 2.5 - Anni di istruzione a tempo pieno completati (Italia 2017)

	N	%	% val.	% cum.
0	3	0,1	0,1	0,1
1	4	0,1	0,2	0,3
2	13	0,5	0,5	0,8
3	35	1,3	1,3	2,1
4	12	0,4	0,5	2,5
5	238	8,7	8,9	11,5
6	32	1,2	1,2	12,7
7	25	0,9	0,9	13,6
8	510	18,6	19,2	32,8
9	71	2,6	2,7	35,5
10	147	5,4	5,5	41
11	137	5	5,2	46,1
12	145	5,3	5,5	51,6
13	611	22,3	23	74,5
14	127	4,6	4,8	79,3
15	96	3,5	3,6	82,9
16	110	4	4,1	87,1
17	58	2,1	2,2	89,2
18	141	5,1	5,3	94,5
19	55	2	2,1	96,6
20	52	1,9	2	98,6
21	10	0,4	0,4	98,9
22	8	0,3	0,3	99,2
23	10	0,4	0,4	99,6
24	2	0,1	0,1	99,7
25	2	0,1	0,1	99,8
26	1	0	0	99,8
27	1	0	0	99,8
30	3	0,1	0,1	100
37	1	0	0	100
Tot.	2.660	96,9	100	
Rifiuto	30	1,1		
Non so	55	2		
Tot.	85	3,1		
Totale	2.745	100		

N. validi	2660	Varianza	18,76	Percentili	25	8
N. mancanti	85	Deviazione std.	4,33		50	12
Media	11,5	Intervallo	37		75	14
Mediana	12	Minimo	0			
Moda	13	Massimo	37			

Fonte: European Social Survey (http://www.europeansocialsurvey.org)

Riportiamo ora un esempio di variabile autenticamente cardinale e abbiamo scelto gli anni di istruzione, anche per mostrare come sia possibile rilevare la stessa proprietà utilizzando definizioni operative alternative (si è visto sopra che il livello di istruzione si può rilevare a livello ordinale chiedendo di indicare il titolo di studio più elevato conseguito). La Tab. 2.5 illustra con chiarezza che l'aumento del numero di valori della variabile rende sempre meno utile la rappresentazione tabellare della distribuzione di frequenza. Il caso più estremo è quello di una variabile cardinale misurata con una precisione tale che a ogni caso corrisponde un unico valore, come accade se si rileva con precisione il reddito di un campione di intervistati o il PNL di un insieme di paesi (come nella Tab. 2.6). Nel caso specifico la rappresentazione tabellare è comunque utile perché consente di notare che i valori centrali sono molto più frequenti di quelli estremi e, come era prevedibile, l'addensamento dei casi in corrispondenza del numero di anni necessari per completare ognuno dei cicli di studi (indicati dal tratteggio).

La tabella posta sotto alla distribuzione di frequenza fornisce gli elementi utili a descrivere la distribuzione utilizzando i principali parametri statistici e valori posizionali, alcuni dei quali si possono individuare facilmente anche osservando direttamente la tabella della distribuzione di frequenza. Il numero minimo di anni di istruzione è zero e il massimo 37, con un campo di variazione di 37 anni (il campo di variazione corrisponde al valore massimo solo quando, come in questo caso, il minimo è pari a zero). Osservando la colonna delle percentuali cumulate si individuano (evidenziati da un leggero retino) il 25° percentile, che corrisponde al primo quartile (8 anni di istruzione), la mediana (o secondo quartile - 12 anni) e il 75° percentile (o terzo quartile - 14 anni). In tabella si può osservare direttamente anche la moda (13 anni di studi) che, non a caso, corrisponde alla durata minima degli studi che consente di conseguire il diploma di scuola media superiore. La media deve essere invece calcolata ed è pari a 11,5 anni; devono essere calcolate, con il procedimento delineato in precedenza, anche la varianza (18,76) e la deviazione standard (4,33).

Riportiamo infine, come anticipato, un esempio di variabile cardinale misurata con precisione tale che per ogni caso si è registrato un diverso valore sulla proprietà (il prodotto interno lordo pro capite in dollari); non ci sono dunque categorie all'interno delle quali contare le frequenze. Non si può pertanto utilizzare una tabella delle distribuzioni di frequenza per riassumere i dati in matrice, a meno che non si cambi la natura della variabile da cardinale a ordinale, ricodificando i valori originari in intervalli di, ad esempio, 10.000 dollari.

Anche in questo caso la rappresentazione tabellare può essere comunque utile come strumento di analisi se non altro perché, come nell'esempio precedente, ordinando i casi a seconda del valore rilevato si possono individuare i valori posizionali. Nella Tab. 2.,5 i paesi europei sono elencati secondo il PIL pro capite e si vede anzitutto che il massimo è 115.536 dollari e il minimo 3.113; dunque il

campo di variazione (differenza tra massimo e minimo) è pari a 112.423 dollari. Si può poi individuare la mediana, cioè il PIL del caso che si colloca a metà della distribuzione; siccome i paesi sono in numero pari (46), due paesi si collocano a metà strada (Estonia e Repubblica Ceca) e la mediana corrisponde dunque alla media dei loro PIL (23.221,5 dollari). Il primo quartile corrisponde al valore registrato a San Marino (48.948) e il terzo quartile al Pil della Bulgaria (9.313).

Si può anche notare che, mentre i due paesi collocati sulla mediana occupano per definizione la posizione centrale, non è così per i loro redditi pro capite; un semplice calcolo mostra infatti che per i paesi sotto la mediana il campo di variazione è pari a quasi esattamente 20.000 dollari; i redditi sopra la mediana hanno invece un campo di variazione di 60.000 dollari, escludendo Lussemburgo e Liechtenstein che, avendo un PIL pro capite eccezionalmente alto, si possono trattare come casi "devianti" (*outliers*). Il confronto dei campi di variazione delle due metà della distribuzione mostra dunque che è asimmetrica, come indica anche la media, pari a 32.283 dollari, un valore più elevato di quello della mediana. La media infatti risente della presenza di valori molto elevati o molto bassi nella distribuzione; infatti, se si escludono dal calcolo Lussemburgo e Lichtenstein, scende a 28.650 dollari (con un calo di oltre 3.600 euro).

Tabella 2.6 – Prodotto Interno Lordo pro capite dei paesi europei (2018)

Lussemburgo	115.536	Italia	34.321	Russia	11.289
Liechtenstein	108.892	Spagna	30.733	Turchia	9.405
Svizzera	83.161	Malta	30.608	Bulgaria	9.313
Norvegia	81.550	Cipro	28.341	Montenegro	8.763
Irlanda	78.335	Slovenia	26.145	Serbia	7.223
Islanda	74.515	Portogallo	23.437	Bielorussia	6.283
Danimarca	60.897	Estonia	23.330	Macedonia Nord	6.100
Svezia	54.356	Rep. Ceca	23.113	Bosnia Erzeg.	5.755
Paesi Bassi	53.228	Grecia	20.317	Albania	5.239
Austria	51.344	Slovacchia	19.579	Kosovo	4.433
Finlandia	49.738	Lituania	18.994	Georgia	4.346
San Marino	48.948	Lettonia	18.032	Armenia	4.188
Germania	47.662	Ungheria	16.484	Moldavia	3.190
Belgio	46.696	Polonia	15.425	Ucraina	3.113
Francia	42.953	Croazia	14.870		
Regno Unito	42.580	Romania	12.270		

Fonte: Fondo Monetario Internazionale (https://www.imf.org/external/datamapper/datasets)

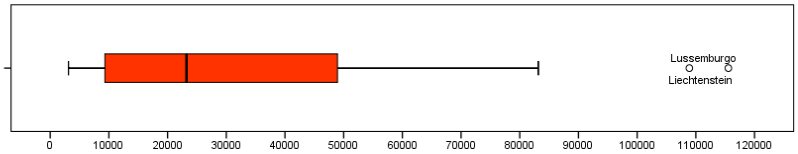
Le variabili cardinali (e le ordinali) si possono rappresentare graficamente mediante i *box plots* (grafici a scatola), sviluppati nell'ambito dell'*Exploratory Data Analysis* (Tukey 1977). A differenza di tabelle quali quella appena esaminata, i *box plots* forniscono un'efficace immagine d'assieme della distribuzione dei dati e, come si vedrà, sono particolarmente utili per effettuare analisi bivariate, specialmente se con variabili di controllo dicotomiche o con poche categorie per non rendere troppo complessa la comparazione. Inoltre, con il metodo grafico si individuano con immediatezza eventuali casi devianti (*outliers*).

> Molto spesso [...] il ricercatore non conosce in origine le caratteristiche strutturali dei suoi dati che sono più suscettibili di significative interpretazioni [...]. Perciò in molte situazioni le tecniche di cui abbiamo bisogno [...] sono esplorative e informali, piuttosto che formali e confermative come i test di specifiche ipotesi. Le tecniche grafiche sono particolarmente adatte a questo scopo in quanto tendono a essere meno formali e limitanti, e permettono al ricercatore di ottenere una conoscenza ravvicinata della struttura dei dati senza imporre legami restrittivi (Everitt 1978, 2).

Si può in effetti sostenere che le tecniche grafiche sono l'analogo di un detective, che cerca tracce o prove per scoprire i misteri dei dati. Gli strumenti numerici formali della statistica classica sono invece l'analogo del giudice, che soppesa il grado di probatorietà contenuto nelle tracce per determinare quanto ci si possa fidare delle stesse (Tukey 1977).

I *box plots* forniscono dunque un'immagine della localizzazione, dispersione, asimmetria e comportamento delle code delle distribuzioni e sono realizzati utilizzando i valori posizionali. I valori utilizzati per costruire la scatola (*box*) sono il quartile superiore e inferiore e la linea centrale rappresenta la posizione della mediana. Le linee che partono dal *box* rappresentano le code della distribuzione e si estendono fino a una distanza massima di 1,5 la differenza interquartile, mentre gli eventuali casi devianti sono marcati individualmente. L'esempio nella Fig. 2.3 conferma che Lussemburgo e Lichtenstein sono casi devianti e, nonostante siano esclusi dalla costruzione del grafico, si apprezza con grande immediatezza che la distribuzione è notevolmente asimmetrica: i redditi dei 23 paesi più ricchi si estendono infatti su un intervallo di valori molto più ampio di quello dei paesi meno ricchi.

Figura 2.3 - *Box plot* – Prodotto Interno Lordo pro capite (2018)

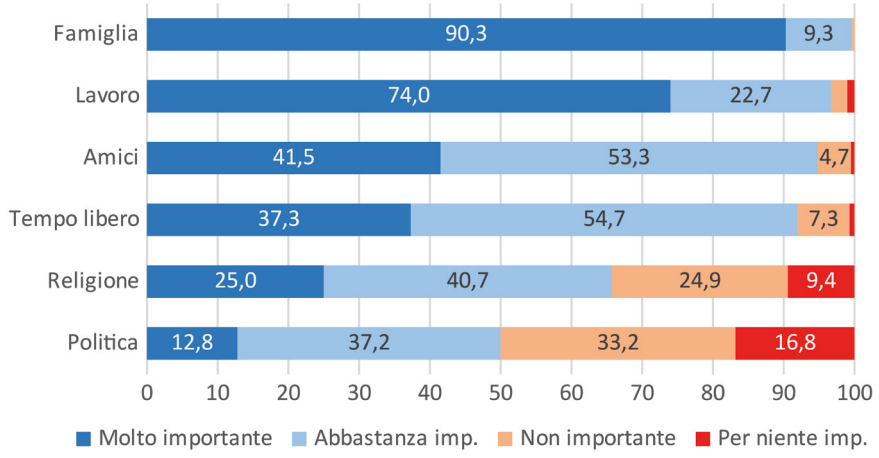


Fonte: Fondo Monetario Internazionale (https://www.imf.org/external/datamapper/datasets)

Nella Fig. 2.4 è infine rappresentata la distribuzione percentuale delle risposte a una serie di domande sull'importanza attribuita a una serie di aspetti della vita, poste a un campione di oltre 2.277 italiani intervistati tra la fine del 2018 e l'inizio del 2019 nell'ambito dell'*European Values Study* (europeanvaluesstudy.eu/)[8].

Per migliorare la leggibilità, le barre suddivise che rappresentano la distribuzione delle variabili sono state ordinate secondo il grado decrescente di importanza. È opportuno effettuare questa rappresentazione solo quando le categorie non sono molto numerose e la loro contiguità semantica permette, ad esempio, di apprezzare con immediatezza la distinzione tra chi ritiene importanti (molto o abbastanza) gli aspetti valutati e chi li ritiene non importanti o per niente importanti. La tabella non consentirebbe di percepire la situazione con altrettanta immediatezza, mentre dal grafico si vede chiaramente che la famiglia è la più importante, ma il lavoro lo è poco meno se aggiungiamo a entrambi gli aspetti anche quanti li ritengono abbastanza importanti. La religione è invece considerata molto importante da una quota minoritaria di intervistati, anche se va rilevato che si supera la maggioranza se aggiungiamo chi la considera abbastanza importante. Non è così per la politica, perché il campione è esattamente suddiviso in parti uguali tra "fan" e "oppositori" e sono comunque solo poco più del 10% gli intervistati che la considerano molto importante.

Figura 2.4 - Diagramma a barre di una serie di distribuzioni percentuali – Importanza di aspetti della vita (Italia 2019)



Fonte: European Values Study (https://europeanvaluesstudy.eu/)

[8] Come ampiamente illustrato nel volume dedicato alla rilevazione dei dati (Delli Zotti 2021), si tratta di rilevazioni facenti parte del *World Values Survey*, nato dai lavori di Inglehart (1977 e 1996) sulla mutazione dei valori nella società contemporanea.

3. Analisi bivariata

3.1. Finalità e presupposti dell'analisi bivariata

L'analisi monovariata costituisce un momento essenziale di presa di contatto con il materiale empirico nel quale il ricercatore, oltre a verificare se nei dati vi siano eventuali "anomalie", può soddisfare in parte la sua curiosità scientifica. È necessario però procedere con l'analisi bivariata di fronte a un risultato inatteso o che appare interessante. Ad esempio, le risposte a una domanda di atteggiamento possono suscitare l'interesse a verificare se sul tema divergono (e di quanto) le posizioni dei maschi e delle femmine o di persone con diverso livello di istruzione. L'interesse ad andare oltre quanto emerge dall'analisi monovariata può essere ben esemplificato anche da un esempio il quale dimostra che la necessità di approfondire l'analisi dei dati si estende oltre l'ambito della ricerca scientifica. Se i voti ottenuti da un partito variano molto da comune a comune, i responsabili politici dei partiti sono certamente interessati a individuare i comuni nei quali quel partito raggiunge i risultati migliori (o peggiori) e a comprendere le possibili cause di queste significative variazioni, specialmente se mancano dati di riferimento perché quel partito si è presentato per la prima volta alle elezioni. In questo caso l'analisi bivariata può aiutare a comprendere se si tratta di comuni

grandi o piccoli, con popolazione a elevata o bassa scolarizzazione, a economia prevalentemente agricola o industriale, ecc.; si può poi cercare di individuare i partiti che perdono più voti nei comuni in cui il nuovo partito raggiunge i risultati migliori, e così via.

L'analisi bivariata non è però solo finalizzata a rispondere a "estemporanee" curiosità scientifiche che possono nascere dall'analisi monovariata o a soddisfare i legittimi interessi conoscitivi degli stakeholder politici, economici e sociali, perché in un'indagine ben strutturata la rilevazione dei dati è principalmente finalizzata a confermare o rigettare le ipotesi formulate nella prima fase della ricerca (Delli Zotti 2021). Non si deve peraltro dimenticare che una relazione tra variabili suggerita dall'analisi bivariata potrebbe essere *spuria* (non è discernibile un nesso causale, ma le variabili co-variano perché entrambe dipendono da una terza variabile). All'opposto, l'analisi bivariata può essere inefficace, perché può non riuscire a individuare una relazione *soppressa* da una terza variabile che influenza entrambe, ma in opposta direzione (la relazione tra reddito e livello di istruzione può non apparire con evidenza per effetto dell'età, una variabile "confondente" perché influenza positivamente il reddito e negativamente l'istruzione). In questi casi è necessario ricorrere al c.d. "controllo delle terze variabili", un insieme di tecniche di analisi multivariata relativamente semplici che tratteremo all'inizio del prossimo capitolo.

Accennato ad alcune finalità, è utile ricordare un prerequisito dell'analisi bivariata che non si può dare per scontato: nella variabile su cui si concentra il nostro interesse le risposte si devono effettivamente distribuire tra due o più modalità. Se la grande maggioranza degli intervistati risponde in un unico modo, si può semplicemente constatare che ciò che reputavamo essere "variabile" è invece costante, perché sulla questione che abbiamo posto tutti (o quasi) la pensano allo stesso modo o tutti (o quasi) si comportano allo stesso modo. Il ricercatore consapevole in questi casi si pone però in "modalità falsificazione" e si chiede se il risultato non possa invece dipendere da poca accortezza nel formulare la definizione operativa della variabile. Se, ad esempio, si chiede quanto sia importante nella vita poter contare su una buona salute, è molto probabile che la stragrande maggioranza degli intervistati risponderà "molto". Se si è già conclusa la rilevazione non si può fare altro che constatare la distribuzione delle risposte, ma potremo tenere conto dell'insoddisfacente risultato e alla prossima occasione riformulare la domanda per rendere meno banali le risposte. Si potrà, ad esempio, sottoporre agli intervistati una lista di cose importanti e chiedere di indicarne solo una o due, in modo da costringerli a scegliere tra salute, denaro, soddisfazioni professionali, crescita culturale, ecc. perché sono certamente tutte importanti, ma alcune (per alcuni) possono esserlo più di altre.

Il ricercatore si trova comunque completamente "disarmato", ossia impossibilitato a condurre un'analisi bivariata, di fronte a una distribuzione che mostri un'elevata concentrazione di risposte su un'unica modalità. Se tutti o quasi tutti gli intervistai hanno le stesse opinioni o fanno le stesse cose è inutile indagare oltre, perché sono d'accordo o hanno gli stessi comportamenti maschi e femmine, giovani e anziani, istruiti e ignoranti. D'altro canto, non è ottimale nemmeno la situazione opposta, che si verifica quando in una variabile categoriale le risposte sono frazionate tra "troppe" modalità, con la conseguenza che per molte si riscontrerà una numerosità assai limitata. È infatti poco interessante cercare di comprendere a quali categorie sociali appartiene chi predilige la musica celtica e quali siano le motivazioni se questa è la scelta di solo due o tre intervistati. In questa situazione, può essere utile accorpare le modalità di risposta semanticamente contigue, riunendo ad esempio tutte le risposte che fanno riferimento alla musica folk tradizionale.

Se la variabile è di tipo cardinale è invece normale l'estremo frazionamento perché i valori, che derivano da conteggi e misurazioni, possono anche essere unici per ognuno dei casi. Come vedremo tra poco, se si utilizzano per l'analisi bivariata i coefficienti di correlazione o i grafici a dispersione non ci sono di norma problemi, anche se è opportuno verificare per lo meno che le variabili non presentino dati "estremi" che si potrebbero escludere o trattare a parte. Se si vogliono invece utilizzare le tabelle di contingenza, sarà necessario categorizzare le variabili nei modi illustrati più avanti, nel capitolo 6.

3.2. Strategie dell'analisi bivariata

L'analisi bivariata consiste dunque nell'analisi congiunta di coppie di variabili e lo scopo fondamentale è accertare se vi sia tra loro una qualche relazione, desumibile dalla co-variazione dei valori. La ricerca di tale relazione e la stima della sua forza possono essere effettuate mediante due famiglie di tecniche statistiche, analogamente a quanto si è visto per l'analisi monovariata: tecniche che *esibiscono* la relazione (tabelle di contingenza e diagrammi a dispersione) e *misure sintetiche* (coefficienti statistici).

Le variabili da analizzare sono spesso molto numerose: non è infrequente che i questionari, ad esempio, contengano decine di domande ed è probabile che alcune domande introducano una serie di "voci" (*item*) su ognuna delle quali si chiede di fornire una risposta, con conseguente generazione di altrettante variabili (si può ad esempio chiedere agli intervistati di indicare se, o quanto, considerano accettabili una lunga serie di comportamenti).

Il ricercatore che non adotta una strategia che lo guidi nella decisione su quali debbano essere le coppie di variabili di cui studiare la distribuzione congiunta

rischia di addentrarsi in una situazione alquanto difficile: incrociare le variabili in tutte le combinazioni possibili è infatti una decisione poco "saggia" perché impraticabile se le variabili sono molto numerose (con 100 variabili nella matrice dei dati dovrebbero essere esplorate 4.950 tabelle e/o coefficienti se si volesse mettere in relazione ognuna di esse con tutte le altre).

È opportuno, allora, introdurre la distinzione tra variabili che, sulla base delle ipotesi formulate, si considerano *dipendenti* (gli *effetti*) e quelle che invece nel modello di analisi svolgono il ruolo di variabili *indipendenti* (le *cause*). In alcuni casi l'attribuzione dei ruoli alle variabili poggia su un intuitivo "substrato" naturale: si può infatti presupporre che alcune caratteristiche socio-demografiche ascritte dell'individuo (sesso, età, nazionalità) possano influenzare i suoi comportamenti e atteggiamenti, mentre il contrario non appare molto verosimile. Se si tratta invece di caratteristiche acquisite, quali il livello di istruzione, la professione o l'appartenenza politica, le cose cambiano decisamente, perché il livello di istruzione, ad esempio, può essere trattato come variabile indipendente (causa), se stiamo studiando gli atteggiamenti verso l'immigrazione, ma può svolgere il ruolo di variabile dipendente se stiamo invece cercando di comprendere quanto le caratteristiche della famiglia di origine (titolo di studio dei genitori, reddito familiare, nazionalità) influiscano sul livello d'istruzione dei figli.

Oltre che per l'aspetto interpretativo dei fenomeni sociali, la distinzione tra variabili dipendenti e indipendenti è molto rilevante per alcune significative conseguenza sul piano tecnico statistico. Nelle tabelle di contingenza è ad esempio opportuno calcolare le percentuali che rendono confrontabili le categorie della variabile indipendente; esistono inoltre test statistici asimmetrici, che assumono cioè un valore diverso a seconda di quale delle due variabili sia considerata la dipendente.

La distinzione è però ancor più importante nell'impostare la strategia di analisi: una procedura consueta consiste infatti nella selezione di alcune variabili indipendenti "di base" (nei sondaggi si tratta solitamente, come sopra accennato, di alcune caratteristiche socio-anagrafiche degli intervistati) e incrociarle con le altre variabili, considerate dipendenti (atteggiamenti, opinioni, comportamenti, ecc.). Anche se agendo in questo modo (indotti dalla facilità con cui il software può produrre tabelle, grafici e matrici di coefficienti statistici) ci si può trovare ad affrontare voluminosi tabulati, non si deve necessariamente stigmatizzare questa strategia, perché il controllo "a tappeto" può permettere di trovare relazioni inattese, da esaminare accuratamente per eliminare le più banali ed evitare di scoprire l'"acqua calda". Nelle scienze sociali le relazioni troppo "strette" sono infatti sospette di tautologia, perché è probabile che le due variabili rilevino lo stesso fenomeno, pur con una diversa definizione operativa.

Che si tratti di "scansione" sistematica alla cieca o di controlli effettuati sulla base di ipotesi formulate in precedenza, la scoperta di una stretta relazione tra le variabili può essere gratificante, ma il ricercatore scrupoloso dovrà comunque «cercare di superare il pregiudizio a favore delle correlazioni (elevate) e il ripudio delle correlazioni nulle» (Galtung 1970: 396). Non bisogna infatti dimenticare che anche registrare l'assenza di una relazione tra due variabili, dove l'apparato teorico aveva predetto invece la presenza, può essere un risultato di notevole interesse scientifico.

È necessario poi progettare una serie di analisi bivariate interne al gruppo delle variabili indipendenti al fine di ottenere informazioni utili per la caratterizzazione del campione: se il genere e il tipo di istituto frequentato sono due variabili indipendenti in un'indagine sulla popolazione studentesca, è utile notare se vi è una diversa ripartizione di genere negli istituti scolastici. Avremo così una descrizione più accurata delle caratteristiche del "campione" che permetterà, ad esempio, di essere cauti nell'attribuire differenze di atteggiamento o comportamento al tipo di istituto frequentato: le differenze riscontrate potranno infatti essere attribuite, almeno in parte, alla diversa composizione per genere della popolazione studentesca nei diversi istituti. Perciò, se c'è il sospetto che le variabili indipendenti possono essere più o meno strettamente associate, per comprendere con più accuratezza il loro ruolo nella spiegazione degli atteggiamenti si dovrà studiare cosa accade se una viene messa sotto controllo con una procedura statistica che consenta di rimuovere la sua influenza: si dovrà verificare, ad esempio, se la differenze tra istituti permangono analizzando separatamente in ogni istituto gli atteggiamenti dei maschi e delle femmine (un esempio di controllo delle terze variabili su cui torneremo all'inizio del prossimo capitolo).

Può essere inoltre necessario progettare anche analisi bivariate tra le variabili dipendenti, per verificare se alcuni atteggiamenti (o comportamenti), oltre a essere "condizionati" dalle caratteristiche degli individui intervistati, sono anche più o meno strettamente associati tra di loro. L'analisi bivariata tra variabili dipendenti può anche essere il primo passo per la realizzazione di indici sintetici che "ricostruiscano" i concetti che con lo strumento di rilevazione ci eravamo proposti di misurare. Le variabili nella matrice dei dati sono infatti molto numerose anche perché, come abbiamo visto, spesso utilizziamo pluralità di indicatori (e quindi di variabili) per misurare un unico concetto (come la qualità della vita, l'autoritarismo o la fiducia nelle istituzioni).

Gli strumenti per effettuare l'analisi bivariata sono numerosi: tra i più noti le "tabelle a doppia entrata", definite anche "di contingenza" o "di incrocio" e vari tipi di grafici (a dispersione, a barre, affiancate o suddivise, ecc.). Entrambi i tipi di strumenti possono essere corredati da sintetici coefficienti di associazione tra le variabili[1]. Come nell'analisi monovariata, l'individuazione degli strumenti più utili (e corretti) dipende dalla natura delle variabili; dunque, se si devono mettere in relazione due variabili di tre tipi (cardinali, ordinali o nominali), le combinazioni possibili sono nove, stante l'esistenza di coefficienti statistici asimmetrici (Tab. 3.1).

In linea di principio, per ogni combinazione si dovrebbero individuare strumenti specifici, ma alcune non godono di sufficiente autonomia e quindi la stessa tecnica di analisi può essere applicata in più di una delle combinazioni possibili. In effetti, la distinzione veramente fondamentale è quella tra variabili categoriali (nominali o ordinali) e variabili cardinali (non importa se derivanti da conteggio o misurazione). Operata questa semplificazione, le combinazioni davvero rilevanti sono solo tre: entrambe le variabili sono categoriali, entrambe sono cardinali e, infine, una delle due è categoriale e l'altra cardinale. Nel primo caso lo strumento principale d'analisi è la tabella di contingenza, nel secondo il grafico a dispersione e nel terzo la scomposizione della media e i grafici a semidispersione. In tutte queste situazioni si possono utilizzare anche appropriati indici sintetici di associazione e, in questo caso, può essere nuovamente rilevante distinguere tra le categoriali le variabili nominali e le variabili ordinali.

Tabella 3.1 - Esempi di strumenti statistici nell'analisi bivariata

		Indipendente		
		Nominale	Ordinale	Cardinale
Dipendente	Nominale	Tabelle, grafici a barre (chi-quadro)	Tabelle, grafici a barre (chi-quadro)	Istogrammi, box-plot
	Ordinale	Tabelle, grafici a barre (chi-quadro)	Tabelle, grafici a barre (rho, tau)	Istogrammi, box-plot
	Cardinale	Istogrammi, box-plot (eta)	Istogrammi, box-plot (eta)	Grafici a dispersione (r, a+b)

[1] Per una esaustiva trattazione si veda Marradi (1997) e il più recente Marradi e Di Franco (2020), che illustra anche alcune tecniche specifiche qui non trattate.

Se entrambe le variabili sono categoriali (nominali o ordinali), si può analizzare l'eventuale loro associazione tramite le tabelle di contingenza. I valori delle due variabili intestano le righe e le colonne e pertanto queste tabelle si possono anche definire matrici di "valori per valori". Nelle celle della tabella sono riportate le *frequenze congiunte osservate* (numero di casi per ogni combinazioni di valori) e ai margini le frequenze (*marginali*) delle variabili osservate disgiuntamente (in pratica, le distribuzioni monovariate).

Oltre alle frequenze assolute all'interno delle tabelle possono essere riportate le *frequenze relative* (percentuali), calcolate *per riga* o *per colonna* a seconda che si vogliano rendere confrontabili i gruppi di casi individuati dai valori di una o dell'altra variabile. Se in un sondaggio abbiamo ad esempio intervistato un diverso numero di maschi e femmine, solo rapportando a 100 la numerosità di entrambi i gruppi, possiamo valutare se i maschi hanno atteggiamenti e/o comportamenti diversi da quelli delle femmine. Le *percentuali sul totale* sono invece meno usate e servono a indicare il peso di ogni combinazione di valori al fine, ad esempio, di costruire una tipologia (si veda la Tab. 3.4 e il capitolo 7).

La comprensione del corretto utilizzo delle percentuali richiede particolare attenzione, perciò il prossimo paragrafo sarà dedicato all'approfondimento di questo tema, ma è intanto opportuno fornire alcuni esempi. Una prima tabella (Tab. 3.2) mostra la relazione tra due variabili nominali (genere e posizione lavorativa) e le percentuali di colonna sono utilizzate al fine di confrontare la situazione lavorativa di maschi e femmine. Nella riga dei totali è indicata la numerosità di ogni categoria perché le percentuali annullano le differenze (tutte le categorie diventano grandi "100") ed è invece utile sapere se, ad esempio, le percentuali riferite a una specifica categoria sono calcolate su un insieme assai limitato di casi, il che potrebbe inficiare la rappresentatività statistica di quel segmento del campione. Se lo si ritiene utile, tramite il totale in valore assoluto si può inoltre ricostruire il numero di casi sottostante a ogni percentuale: nella Tab. 3.2 si può ad esempio calcolare facilmente che i coadiuvanti maschi sono 12 (poco meno dell'1% di 1.298), mentre serve la calcolatrice per scoprire che le femmine sono 38 (il 2,6% di 1.447).

Tabella 3.2 - Tabella di contingenza di due variabili nominali – Posizione nel lavoro per genere (Italia 2017)

	M	F	Tot.		M	F	Tot.
Dipendente	61,6	54,9	58,0	Dipendente	73,6	81,0	77,1
Autonomo	21,2	10,3	15,4	Autonomo	25,3	15,2	20,5
Coadiuvante	,9	2,6	1,8	Coadiuvante	1,1	3,8	2,4
Non applicabile	15,6	31,5	24,0	Totale	100,0	100,0	100,0
Rifiuto	,5	,6	,5		1.086	980	2.066
Non so	,3	,2	,3				
Totale	100,0	100,0	100,0				
	1.298	1.447	2.745				

Fonte: European Social Survey (http://www.europeansocialsurvey.org)

Nella parte sinistra della Tab. 3.2 sono riportate le risposte dell'intero campione e si nota una differenza notevole tra maschi e femmine, oltre a quote marginali di intervistati di entrambi i generi che non vogliono rispondere o rispondono (assai curiosamente) "non so". La domanda risulta non applicabile a quasi un terzo delle femmine (31,5%), perché non lavorano (i maschi non occupati sono solo il 15,6%). Preso atto di questa significativa differenza, si può decidere di confrontare i due gruppi (maschi e femmine) tenendo conto solo degli intervistati che sono occupati. A tal fine, nella parte destra della tabella sono riportate le percentuali ricalcolate ponendo uguale a 100 i soli occupati (1.086 maschi e 980 femmine) e si vede allora che le femmine sono lavoratrici dipendenti più frequentemente dei maschi (nella parte a sinistra della tabella la percentuale di femmine dipendenti era più bassa rispetto ai maschi, ma tutte le percentuali erano più basse prima di ricalcolare le percentuali al netto di quasi un terzo di donne e un sesto di maschi che non lavorano).

La Tab. 3.3 mostra invece la relazione tra due variabili ordinali: il titolo di studio e l'interesse per la politica. Dalla prima riga si vede che solo il 2,4% di chi non ha studiato oltre la scuola elementare si dichiara "molto" interessato alla politica, mentre tra i laureati la percentuale arriva a oltre il 13,7%. Essendo le due categorie poste agli estremi di una variabile ordinale, le percentuali consentono anche di verificare se all'innalzamento del livello di istruzione corrisponda un aumento della quota di intervistati interessati alla politica (di verificare cioè se le due variabili co-variano). In effetti, è proprio così: dal 2,4% tra i meno scolarizzati, la percentuale sale al 3,7% tra gli intervistati che hanno conseguito la licenza media, al 5,1% tra i diplomati e al 13,7% tra i laureati.

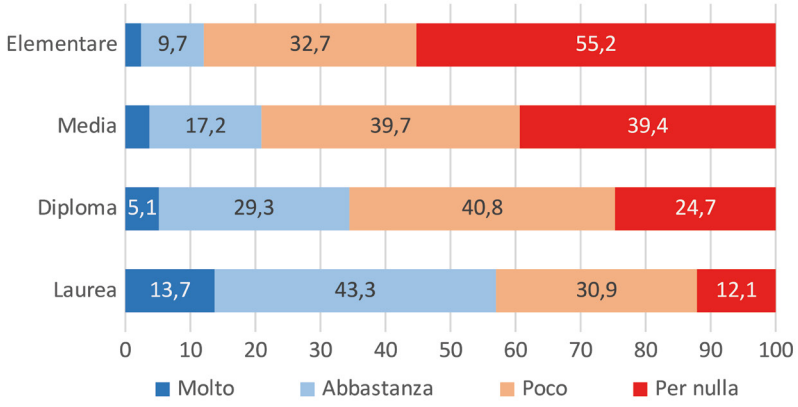
Per effetto dell'ordinalità, nella Tab. 3.3 i dati possono essere letti anche commentando le categorie che si collocano sull'altro lato della variabile, ragionando in termini di disinteresse, invece che di interesse verso la politica, se si ritiene che sia più interessante o efficace per la comunicazione dei risultati. Usando queste percentuali si osserva che non è "per nulla" interessato alla politica oltre il 50% dei "non scolarizzati", poco meno del 40% di chi ha il diploma di scuola media, un quarto circa dei diplomati, per scendere drasticamente a poco più del 10% dei laureati.

Tabella 3.3 - Tabella di contingenza di due variabili ordinali – "Quanto interessato alla politica" per titolo di studio (% di colonna) (Italia 2017)

	Elementare	Media	Diploma	Laurea	Non risp.	Totale
Molto	2,4	3,7	5,1	13,7	,0	5,4
Abbastanza	9,7	17,2	29,3	43,3	22,7	24,8
Poco	32,7	39,7	40,8	30,9	50,0	38,1
Per nulla	55,2	39,4	24,7	12,1	27,3	31,7
Totale	100,0	100,0	100,0	100,0	100,0	100,0
	373	841	1.129	372	22	2.737

Fonte: European Social Survey (http://www.europeansocialsurvey.org)

Figura 3.1 - Grafico a barre suddivise – "Consentire a immigranti di etnia diversa di venire a vivere in Italia" per titolo di studio (Italia 2017)



Fonte: European Social Survey (http://www.europeansocialsurvey.org)

L'ordinalità della variabile consente anche di accorpare le percentuali di categorie contigue (perché sono contigue anche semanticamente), dicotomizzando l'atteggiamento verso la politica (da un lato chi è molto o abbastanza interessato e dall'altra chi lo è poco o per nulla). Questo accorgimento può essere utile per una più efficace comunicazione e anche perché ci sono tecniche di analisi che richiedono questa semplificazione. La relazione con il livello di studio rimane valida e, anzi, si accentua: sommando a chi è "molto" interessato alla politica chi lo è "abbastanza" si passa infatti da poco più del 10% a oltre il 20%, fino a più di un terzo dei diplomati e poco meno del 60% dei laureati. Questa aggregazione consente dunque di constatare che quella dei laureati è l'unica categoria basata sul livello di istruzione nella quale gli intervistati complessivamente interessati alla politica sono più numerosi di quelli che non lo sono.

I dati della Tab. 3.3 sono rappresentati nella Fig. 3.1 mediante un grafico a barre suddivise; se si bada sommariamente solo alla forma il grafico è analogo a quello di Fig. 2.4 che mostrava le distribuzioni di frequenza di una serie di variabili. Le distribuzioni in quel caso erano monovariate e dunque indipendenti una dall'altra. In questo caso si tratta invece di analisi bivariata, perché ogni barra si riferisce alle categorie di un'unica variabile e il grafico mostra con grande evidenza che passando dal livello di scolarità più basso al titolo universitario si estende in modo significativo l'ampiezza delle barre che indicano "molto" e "abbastanza" interesse per la politica.

Bisogna peraltro prestare attenzione a come si commentano i dati di una tabella e di un grafico come quelli sopra presentati. Si potrebbe affermare, in modo sintetico e sbrigativo, che i più scolarizzati hanno un atteggiamento di maggiore interesse verso la politica rispetto ai meno istruiti. Esprimersi in questo modo non è però corretto, perché i dati mostrano solo che tra i laureati vi è una quota più elevata di intervistati molto o abbastanza interessati alla politica e le due affermazioni non sono equivalenti. Dalla prima si potrebbe desumere che in tutti i laureati l'atteggiamento abbia un'intensità più elevata che negli altri intervistati; con la seconda frase non si afferma invece nulla riguardo all'intensità dell'atteggiamento. Se prendiamo i laureati uno a uno, non possiamo dire che l'atteggiamento dei laureati favorevoli sia più o meno intenso di quello dei favorevoli poco scolarizzati, perché si può presumere che chi si dichiara molto interessato alla politica lo sia con la stessa intensità, a qualsiasi livello di scolarità. Cambia solo il fatto che, contando gli interessati alla politica, scopriamo che tra i laureati e i diplomati ce ne sono molti di più che tra chi ha livelli di scolarità inferiore. Vi sono dunque laureati interessati e laureati disinteressati (anche se di questi ultimi ce ne sono meno che tra i poco scolarizzati) e pertanto la spiegazione che la scolarizzazione ha effetto sull'atteggiamento di interesse verso la politica, giustificata dall'evidenza empirica, non dovrebbe far dimenticare che rimane comunque da

spiegare perché una minoranza non marginale di laureati o diplomati mostri un atteggiamento di disinteresse (totale disinteresse per alcuni).

Quella che può apparire un'oziosa puntualizzazione, senza conseguenze sul piano pratico, rende invece evidente che, in generale, è sbagliata la "gratuita" generalizzazione. Precisando che un determinato atteggiamento non è attribuibile *tout court* a un gruppo sociale, si lascia infatti spazio all'individualità. Si rifugge cioè dal determinismo riconoscendo che, pur essendo presente un qualche "tratto" in un maggior numero di appartenenti a una specifica categoria sociale, ve ne sono altri, appartenenti alla stessa categoria, che si comportano o hanno atteggiamenti diversi o anche opposti. La corretta illustrazione dei dati costringe dunque a rapportarsi agli altri in modo non pregiudiziale o stereotipato. Pur essendo empiricamente fondato che i meno istruiti comprendono un maggior numero di individui disinteressati alla politica, la persona concreta che ci si trova davanti potrebbe appartenere al gruppo, magari più ristretto, di coloro che, pur non avendo raggiunto elevati livelli di istruzione si appassiona alla politica; per converso, non ci si deve sorprendere se alcune persone formalmente colte siano invece completamente disinteressate.

3.3.2. I testi di associazione tra variabili categoriali

Le tabelle di contingenza possono essere corredate di misure sintetiche per stabilire la significatività statistica dell'eventuale relazione tra due variabili. La più conosciuta è χ^2 (chi quadrato) che utilizza le differenze tra le frequenze osservate e le frequenze "attese" (quelle ci si poteva attendere se non ci fosse alcuna associazione tra le due variabili).

Ad esempio, tenendo conto del numero complessivo di intervistati e di quanti sono i lavoratori autonomi, nella casella "maschio/lavoratore autonomo" si potevano attendere circa 222,9 casi e sono invece 275 (essere maschio dunque "attrae" la condizione di lavoratore autonomo); al contrario, nella casella "femmina/lavoratore autonomo" le intervistate dovrebbero essere 201,1 e sono invece 149 (c'è dunque "repulsione" tra l'essere femmina e il lavoro autonomo) (Tab. 3.4)[2].

Per verificare se gli scostamenti tra frequenze attese e osservate "sul campo" sono statisticamente significativi si devono sommare le differenze riscontrate in tutte le caselle e poi consultare la "tavola del chi quadro" che tiene conto delle dimensioni della tabella misurata in termini di "gradi di libertà" (in inglese, *degrees*

[2] Non è difficile calcolare le frequenze attese a mano e spiegare la procedura può servire a chiarire ulteriormente il concetto: visto che, come si legge nella Tab. 3.2, gli autonomi sono il 20,5% del campione di 2.066 casi, se le variabili fossero indipendenti i lavoratori autonomi maschi dovrebbero essere il 20,5% di 1.086 e le femmine il 20,5% di 980.

of freedom, da cui la sigla "df"), perché la somma di piccole differenze può diventare elevata per il semplice fatto che le celle sono molto numerose. Una tabella come quella dell'esempio ha due gradi di libertà perché, una volta conosciuti i valori di due delle sei celle che compongono la tabella, gli altri valori sono "obbligati", perché si calcolano per differenza dai totali. Un livello di significatività "0,00" indica che la probabilità che siano dovute al caso le differenze tra le risposte rilevate e quelle attese nell'ipotesi di assenza di associazione (e cioè di assoluta indipendenza) tra genere e posizione lavorativa è inferiore all'1%.

Tabella 3.4 - Tabella di contingenza di due variabili nominali – Posizione nel lavoro per genere (frequenze osservate e attese) (Italia 2017)

		Maschio	**Femmina**	**Totale**
Dipendente	Conteggio	799	794	1.593
	Conteggio atteso	837,4	755,6	1.593,0
Autonomo	Conteggio	275	149	424
	Conteggio atteso	222,9	201,1	424,0
Coadiuvante	Conteggio	12	37	49
	Conteggio atteso	25,8	23,2	49,0
Totale	Conteggio	1.086	980	2.066
	Conteggio atteso	1.086,0	980,0	2.066,0

Chi-quadrato di Pearson = 44,894 df = 2 Sig. asint. (2 vie) = ,000

Fonte: European Social Survey (http://www.europeansocialsurvey.org)

Per applicare correttamente la statistica χ^2 si deve anche controllare che le celle della tabella con frequenza attesa inferiore a 5 non siano più del 25% (nessuna nel nostro esempio, dove la frequenza attesa minima è 23,2). Se si supera il numero accettabile di celle vuote o quasi, è necessario accorpare le categorie di significato simile, o eliminare quelle a bassa numerosità, come le righe "rifiuto" e "non sa" nella parte sinistra della Tab. 3.2.

Se entrambe le variabili sono ordinali (Tab. 3.3), per misurare la forza della eventuale co-variazione[3] delle due variabili si possono utilizzare test (detti "non parametrici") il cui valore si può collocare tra +1 (che indica il massimo di relazione diretta) e -1 (massimo di relazione inversa tra le variabili). Si ha una relazione diret-

[3] Dal momento che nelle variabili ordinali le categorie indicano crescenti livelli o gradi della proprietà rilevata, è anche appropriato parlare di co-graduazione o contro-graduazione (se la relazione è inversa).

ta quando, al crescere dei valori di una variabile, crescono anche i valori dell'altra; la relazione è inversa quando, al contrario, al crescere dei valori di una variabile decrescono quelli dell'altra.

Tra tali test si possono ricordare *tau* di Kendall e la correlazione di Spearman che tengono conto della posizione dei casi nelle "classifiche" costituite dalle due variabili ordinali. Il primo parametro deve essere utilizzato quando un elevato numero di casi è classificato in un numero relativamente piccolo di categorie; il secondo quando è più basso il rapporto casi/categorie. I casi che occupano posizioni concordanti sulle due variabili (entrambi i valori sono bassi, intermedi o alti) contribuiscono a far crescere il valore del parametro statistico nella direzione di +1; le posizioni discordanti fanno invece crescere il valore del parametro nella direzione di -1. Nel caso in cui non si possa notare una di queste relazioni, il valore del parametro tende a "0"; ciò accade se, ad esempio, a valori bassi di una variabile corrispondono in pari misura valori bassi, intermedi o alti dell'altra e lo stesso accade a tutti gli altri livelli della prima variabile.

Nella Tab. 3.5 è possibile osservare chiaramente come si presenta una relazione diretta tra due variabili. Si può infatti notare che una parte considerevole dei casi è posizionata sulla diagonale della tabella che indica concordanza tra i due atteggiamenti (chi risponde "molti" riguardo agli immigrati della stessa etnia risponde "molti" anche per gli immigrati di etnia diversa e così via per le risposte "alcuni", "pochi" e "nessuno"). Se tutti i casi si trovassero sulla diagonale (perfetta concordanza tra i due atteggiamenti) il valore dei parametri statistici sarebbe pari a +1 (sarebbe -1 se tutti i casi si collocassero sulla diagonale opposta). Nel nostro esempio il valore dei parametri è un po' più basso dell'unità, perché c'è una forte associazione, ma non perfetta concordanza.

Tabella 3.5 - Tabella di contingenza di due variabili ordinali – Accoglienza di immigrati della stessa etnia per accoglienza di immigrati di etnia diversa (Italia 2017)

		Razza/etnia diversa				
		Molti	**Alcuni**	**Pochi**	**Nessuno**	**Totale**
Stessa razza/etnia	Molti	13,3	5,2	1,2	,2	19,8
	Alcuni	,5	29,0	9,0	1,5	40,0
	Pochi	-	1,4	23,6	4,0	29,1
	Nessuno	-	,1	,8	10,1	11,1
	Totale	13,8	35,7	34,6	15,8	100,0

Tau-b di Kendall = ,777
Correlazione di Spearman = ,820

Fonte: European Social Survey (http://www.europeansocialsurvey.org)

Il colore dello sfondo evidenzia la diagonale e un colore leggermente meno intenso le celle che completano i quadranti (in alto a sinistra e in basso a destra) che contengono risposte "coerenti" (gli intervistati rispondono alle due domande in modo simile); i quadranti in basso a sinistra e in alto a destra contengono invece risposte che vanno all'opposto. Sommando i valori delle relative celle, si vede che nel quadrante che indica apertura all'immigrazione si colloca esattamente il 48% dei casi e il 38,5% in quello che indica chiusura; rimangono dunque ben pochi casi di intervistati che sono favorevoli all'immigrazione di etnia simile e respingono i "diversi" (meno del 12%) e di "accoglienti" nei confronti di questi ultimi, ma contrari agli immigrati di etnia simile (un atteggiamento che appare contraddittorio, e riguarda infatti solo l'1,5%).

Si noti che nella tabella si sono utilizzate le percentuali sul totale allo scopo di stimare il peso di ogni combinazione di risposte che costituiscono nell'insieme una tipologia di atteggiamenti sull'immigrazione. Prima di proseguire con le altre tecniche di analisi bivariata, può essere dunque utile esplicitare maggiormente la natura e l'utilizzo delle percentuali e degli altri valori che si possono calcolare e inserire nelle tabelle di contingenza. Si potrà in questo modo tornare su aspetti, come la distinzione tra variabili dipendenti e indipendenti, che potranno così essere chiariti ulteriormente[4].

3.3.3. Le percentuali e i dati mancanti nelle tabelle di contingenza

Visto che il software consente di calcolare facilmente fino a una decina di informazioni, si può ritenere che sia utile inserire in ogni cella, oltre al conteggio dei casi, il massimo di ulteriore informazione disponibile. Ciò però contrasta con l'esigenza di produrre tabelle leggibili e facilmente interpretabili ed è dunque quanto mai opportuno "badare all'essenziale".

La prima rinuncia riguarda le frequenze attese e i relativi residui (differenza con le frequenze osservate). Nella maggior parte dei casi è infatti sufficiente il calcolo del χ^2 (chi quadro), che il programma fornisce a parte, corredato della percentuale di celle il cui valore atteso è inferiore a 5 (indispensabile per decidere se χ^2 è utilizzabile). La tabella si può dunque ridurre, senza eccessivo rammarico, in modo che contenga nelle celle le percentuali di riga, colonna e sul totale (oltre ai totali su cui le percentuali sono calcolate).

La tabella diventa così molto più leggibile, ma è possibile semplificarla ulteriormente. Le percentuali sul totale vanno tenute distinte perché, come

[4] Data l'importanza dell'argomento se ne riparlerà anche all'inizio del capitolo 8, dedicato al tema specifico delle tabelle di contingenza multiple.

abbiamo appena visto, servono a determinare la percentuale di intervistati che possiedono congiuntamente i caratteri individuati da ogni singola cella; andranno quindi richieste solo se si intende usarle a questo scopo.

Le percentuali di riga e di colonna servono invece a rendere confrontabili le categorie della variabile su cui le percentuali sono calcolate. Se si assume che il rapporto tra le variabili sia simmetrico, si possono calcolare entrambe e in fase di analisi dei dati si deciderà quale sia più conveniente utilizzare. È il caso dell'incrocio tra le variabili socio-demografiche che si potrebbero anche chiamare "strutturali" perché servono a definire le caratteristiche del campione. Analizzando la relazione tra sesso ed età, se maschi e femmine hanno diversa consistenza numerica, ponendo la loro numerosità pari a 100, i due generi diventano confrontabili e si potrà controllare se al loro interno vi sia una diversa ripartizione delle classi d'età. In alternativa, potremmo usare le percentuali per rendere confrontabili le classi d'età e verificare se e quanto varia la ripartizione tra i generi nelle diverse classi d'età.

Le cose cambiano se una delle due variabili si può considerare dipendente, perché rappresenta il fenomeno che vogliamo studiare. Ci potremmo chiedere, ad esempio, se uno specifico atteggiamento sia più presente tra i maschi o tra le femmine e la percentualizzazione serve a rendere possibile il confronto se si è intervistato un diverso numero di maschi e femmine.

La Tab. 3.6 mette in relazione le stesse variabili della Tab. 3.3, ma le percentuali sono calcolate ponendo uguale a 100 il numero di chi ha risposto "molto", "abbastanza", ecc. Nella Tab. 3.3 si vedeva facilmente l'effetto della scolarizzazione sull'interesse per la politica; se vogliamo effettuare lo stesso confronto con le percentuali calcolate nella Tab. 3.6 siamo invece costretti a percorrere una strada più lunga. Tra chi afferma di essere "molto" interessato alla politica i laureati sono il 34,2%, un po' meno dei diplomati che sono il 38,9%. Si potrebbe desumere che l'interesse per la politica si attenui un po' dopo che ci si è laureati, ma non è così, perché è laureato solo il 13,7% del campione; dunque, tra chi è molto interessato alla politica i laureati sono ben più del doppio di quanto sarebbero se la scolarità non influenzasse questo atteggiamento. I diplomati sono invece il 41,6% del campione e dunque un po' meno numerosi tra i molto interessati di quanto ci si potrebbe attendere. Al contrario, quasi un quarto (23,9%) di chi è "per nulla" interessato alla politica ha solo la licenza elementare, poco meno del doppio rispetto alla percentuale che si registra nell'intero campione (13,7%).

Tabella 3.6 - Tabella di contingenza di due variabili ordinali – "Quanto interessato alla politica" per titolo di studio (% di riga) (Italia 2017)

	Elementare	Media	Diploma	Laurea	Non risp.	Totale	
Molto	6,0	20,8	38,9	34,2	,0	100,0	149
Abbastanza	5,3	21,5	49,2	23,9	,7	100,0	673
Poco	11,8	32,4	44,7	11,1	1.1	100,0	1.032
Per nulla	23,9	38,4	32,4	5,2	,7	100,0	861
Totale	13,7	31,0	41,6	13,7	,8	100,0	2.737

Fonte: European Social Survey (http://www.europeansocialsurvey.org)

Per utilizzare correttamente queste percentuali si dovrebbe partire, ad esempio, dal fatto che i laureati sono il 13,7% dell'intero campione e poi controllare in quali categorie di risposta il loro peso è maggiore o minore di quanto loro "spetta" (la frequenza "attesa"). È meglio però evitare queste faticose "contorsioni" argomentative se, usando le percentualizza della Tab. 3.3, si può più semplicemente dire che solo il 12% dei laureati è "per nulla" interessato alla politica, percentuale che sale a oltre il 55% tra gli intervistati con il tiolo di studio elementare.

Inoltre, anche intuitivamente si comprende che è più utile e comprensibile osservare come e se si differenziano atteggiamenti o comportamenti di gruppi sociali definiti, ad esempio, dal titolo di studio, piuttosto che analizzare la distribuzione dei titoli di studio tra chi dichiara diversi atteggiamenti o comportamenti. Non si deve però essere troppo rigidi su questo aspetto, perché alcuni gruppi sociali si caratterizzano proprio per l'adesione o il rifiuto rispetto a un'affermazione, una proposta, un comportamento. Per definire i gruppi sociali da confrontare possiamo infatti utilizzare anche, ad esempio, il favore o l'opposizione verso la legislazione regolativa del divorzio o dell'aborto, perché sono problematiche largamente dibattute e chi parteggia per una o l'altra parte appartiene a veri e propri "gruppi di opinione" che, se si organizzano, possono diventare "gruppi di pressione" (ad es.: "abortisti" e "antiabortisti").

Il criterio dirimente è dunque la distinzione tra variabili dipendenti e indipendenti; una distinzione "locale", nel senso che vale per uno specifico ambito di ricerca e per una determinata ipotesi interpretativa perché, come si è già visto (§ 3.3) una variabile indipendente può assumere il ruolo di dipendente se cambia il fuoco dell'indagine. Ribadiamo però che, una volta scelto il ruolo che le variabili giocano nel modello, è utile comparare la distribuzione della variabile dipendente (*focus* dell'analisi) all'interno delle categorie della variabile indipendente e non ha molto senso fare viceversa, come bene riassume Galtung:

> [La percentualizzazione] tecnicamente può essere fatta in entrambe le direzioni, tenendo comunque conto che i confronti andranno poi effettuati nella direzione *opposta* a quella in cui è effettuata, ma in pratica una direzione di solito è più interessante dell'altra perché una variabile "produce" l'altra, e non viceversa. La regola è molto semplice: *percentualizzare sempre rispetto ai valori di quella che è vista come variabile indipendente nel modello di base sottostante* (1970: 193).

Al tema della percentualizzazione è connesso il fatto che la matrice dei dati può essere incompleta se, come si è già accennato nel capitolo sull'analisi monovariata, per alcuni casi non è possibile rilevare il valore di alcune variabili. Si può trattare di difficoltà o rifiuto a rispondere a particolari domande, di errore dell'intervistatore nel registrare la risposta o, nel caso di dati "amministrativi", di carenze della fonte che derivano dalle cause citate o da altre. La "mancata risposta" può anche dipendere dal fatto che alcune domande non sono state rivolte ad alcuni intervistati, ma il caso va trattato a parte, perché non si può parlare appropriatamente di risposte mancate. Se una "domanda filtro" individua un sotto-insieme di intervistati ai quali porre una o più domande successive, le relative tabelle dovrebbero contenere solo questi casi perché si può considerare non corretto che le percentuali siano influenzate dalla presenza di "estranei" ai quali le domande non sono stata poste[5]. Escludendo i casi "non pertinenti", la tabella può comunque contenere "veri" dati mancanti e questi ultimi dovrebbero invece essere inseriti nelle tabelle per diversi motivi.

1) Innanzitutto, se si omettono i dati "mancanti" le percentuali si calcolano utilizzando solo le risposte definite "valide". Non si vede però perché il rifiuto o la difficoltà a rispondere, specialmente se esplicitamente espressa con un "non so", non possa essere ritenuta una modalità di risposta valida al pari delle altre. Escludere questi casi significa nascondere la difficoltà insita nella domanda e attribuire a chi risponde "validamente" un'indebita rappresentanza di chi non risponde. Come è noto, nei sondaggi politici si sconta un numero elevato di rifiuti, ma ricalcolare le percentuali di gradimento per i partiti può indurre in gravi errori di stima, perché gli atteggiamenti politici di chi non si esprime possono discostarsi in misura significativa da quelli di chi si esprime liberamente.
2) Se si decide di escludere le "mancate risposte", si dovrà precisare che le percentuali si riferiscono ai soli casi "validi", senza pretendere che rappresentino anche l'opinione di chi non ha risposto. La tabella dovrà inoltre riportare i valori assoluti delle mancate risposte, in modo che si possano eventualmente ricalcolare le percentuali includendo tutte le risposte.

---

[5] Si può sostenere che sia corretto "replicare" la situazione di intervista, escludendo questi casi, ma la questione può essere "problematizzata", cosa che faremo nel capitolo 9.

3) Un altro motivo che consiglia l'inclusione dei dati mancanti nelle tabelle sta nel fatto che ciò consente la loro analisi secondo le categorie della variabile indipendente. Le mancate risposte potrebbero essere ad esempio più frequenti per i soggetti poco scolarizzati e tale "scoperta" può far sospettare che la formulazione della domanda presupponga un alto livello di scolarità per essere ben compresa.

Per calcolare un parametro statistico come la media i dati mancanti vanno invece senz'altro esclusi. Se ci si dimentica di escludere il codice "99", ipoteticamente utilizzato per registrare la mancata risposta riguardo al numero di figli, la media si innalza artificiosamente perché a chi non risponde verrebbero attribuiti 99 figli. Anche quando i dati mancanti si limitano a uno o pochi casi, ragioni pratiche ed "estetiche" possono comunque consigliare di ometterli dalle tabelle. La differenza tra le percentuali calcolate inserendo o escludendo i pochi casi per i quali l'informazione è mancante si può infatti limitare a piccole frazioni di punto, ed è un prezzo che si può pagare per ottenere tabelle di dimensioni più contenute e perciò più "leggibili".

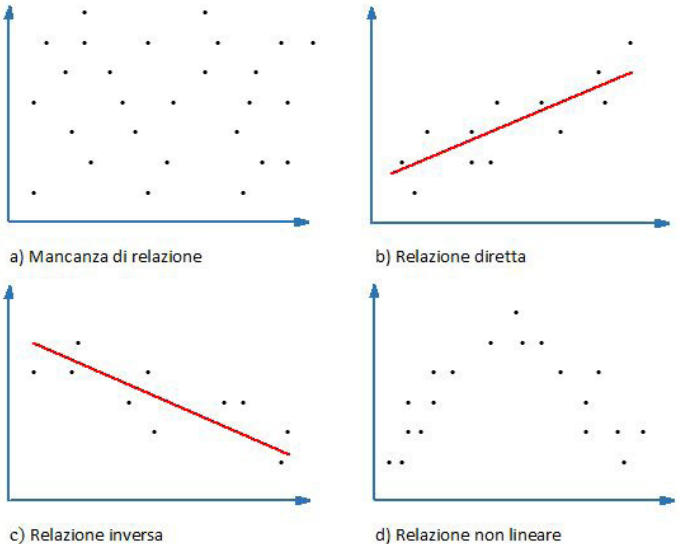
3.3.4. I grafici a dispersione

Quando entrambe le variabili sono cardinali lo strumento di analisi più efficace è il grafico (o diagramma) a dispersione. In questo tipo di grafico ogni caso è inserito in un piano cartesiano, definito dai valori delle due variabili. Non si possono infatti utilizzare le tabelle di contingenza, perché con le variabili cardinali le proprietà sono solitamente rilevate con precisione, al punto che per ogni caso si potrebbe riscontrare una specifica combinazione di valori (si pensi, ad esempio, ai voti ottenuti dai partiti in una serie di circoscrizioni elettorali). Se costruissimo una tabella di contingenza con questi valori, in ogni cella troveremmo un unico caso e la stragrande maggioranza delle celle rimarrebbe vuota. Si preferisce allora costruire un grafico a dispersione, nel quale si indicano solo delle linee di riferimento in corrispondenza di alcuni valori come, ad esempio, un incremento di dieci punti percentuali nel risultato di voto (anche un solo punto, se il campo di variazione della variabile è più piccolo)[6].

[6] Per poter usare le tabelle di contingenza si dovrebbero categorizzare le variabili cardinali, come descritto nel capitolo 6, ma si rinuncerebbe in tal modo alla precisione insita nella rilevazione delle proprietà a livello cardinale. Ovviamente, se entrambe le variabili presentassero una limitata gamma di valori (numero di figli, o esami sostenuti da uno studente in un semestre) l'utilizzo delle tabelle di contingenza sarebbe senz'altro praticabile.

L'insieme dei casi forma una "nuvola di punti" la cui ispezione visiva può suggerire l'esistenza e il tipo di relazione tra le due variabili. Gli esempi schematici della Fig. 3.2 mostrano alcune configurazioni tipiche che può presentare la nuvola di punti: a) assenza di una relazione chiaramente identificabile tra le due variabili; b) relazione diretta e c) relazione inversa, rappresentate entrambe da rette che, come si vedrà tra poco, si possono utilizzare per descrivere sinteticamente la relazione; d) relazione non lineare. Quest'ultimo può essere il caso della relazione tra reddito ed età: il reddito tendenzialmente sale fino all'età della pensione e poi decresce per effetto dell'inflazione e della non adeguata tutela del reddito da pensione rispetto alla perdita di capacità di acquisto. In questa situazione si può pensare all'utilizzo di due rette di regressione, che descrivano la relazione tra le due variabili prima e dopo la soglia dell'età della pensione.

Figura 3.2 - Esempi di grafici a dispersione con ipotetiche rette di regressione



Secondo Duverger «il problema principale dei diagrammi che adoperano le coordinate sta nello scegliere adeguatamente le unità di misura per ogni asse [...]. Unica regola generale è quella che il rapporto tra le due scale debba rendere visibile ogni variazione ritenuta significativa» (1969: 450). Può essere perciò opportuno utilizzare a volte una scala logaritmica, come nel caso della popolazione residente: nella maggior parte dei comuni non supera le 10.000 unità, in molti meno si arriva a 100.000 abitanti e sono pochi quelli che superano questa cifra. Si possono allora sostituire i valori originari con i logaritmi in base 10, per evitare un eccessivo affollamento di punti sui valori inferiori della scala che non permetterebbe di distinguere tra loro i comuni di più piccole dimensioni.

L'utilità dei grafici a dispersione è notevole perché la loro ispezione può facilmente consentire di decidere, ad esempio, se usare una retta per rappresentare la relazione tra le due variabili oppure, come nell'esempio d) della Fig. 3.2, una curva o due rette distinte che interpolino i dati al di sopra o al di sotto di una determinata soglia. L'esame del grafico può inoltre suggerire l'esclusione di eventuali casi devianti (*outliers*) e cioè casi nei quali il valore rilevato su una variabile sia molto distante da quanto si potrebbe ipotizzare sulla base del valore rilevato sull'altra variabile. Si pensi, ad esempio, a imprese con un fatturato notevolmente superiore (o inferiore) a quello che si potrebbe desumere dalla loro dimensione in termini di forza lavoro occupata. Tramite il diagramma a dispersione questi casi sono individuabili visivamente con facilità e, una volta identificati, si può decidere di escluderli e/o trattarli a parte.

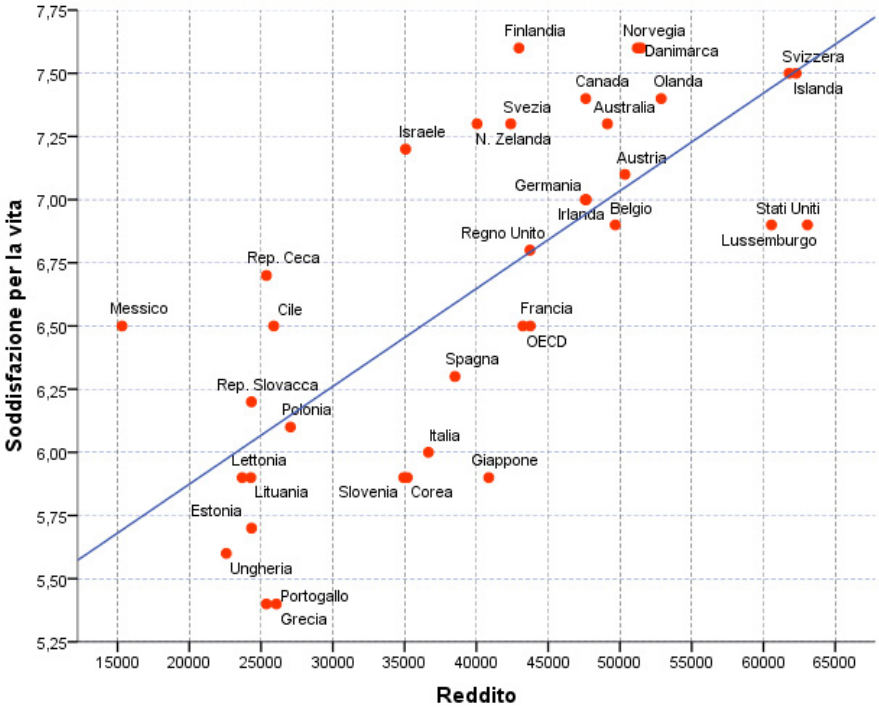
Nella Fig. 3.3 è rappresentata la relazione tra reddito e soddisfazione di vita, due indicatori utilizzati per elaborare il *Better Life Index*, lanciato nel maggio 2011 dall'OCSE (*Organizzazione per la Cooperazione e lo Sviluppo Economico*) (https://www.oecdbetterlifeindex.org). Il grafico mostra chiaramente che c'è un legame relativamente forte tra le due variabili, consentendo di apprezzare visivamente che, all'aumentare del reddito, tendenzialmente aumenta anche il punteggio medio di soddisfazione per la propria vita indicato dagli intervistati.

Con la tecnica della regressione lineare è possibile stimare i parametri "*a*" (intercetta sull'asse delle ascisse) e "*b*" (pendenza) che consentono di tracciare la retta che meglio interpola i dati, cioè quella meno distante, sull'asse delle ordinate, da tutti i punti del grafico. Senza entrare troppo nei dettagli tecnici, si costruisce (virtualmente e con il computer) un quadrato che ha come lato la distanza verticale (sull'asse delle ordinate) di ogni punto dalla retta e si sommano le aree di tutti i quadrati. La retta che meglio rappresenta la nuvola di punti è, anche intuitivamente, quella che fa registrare la minor superficie dei quadrati e per questo motivo la procedura si chiama "*metodo dei minimi quadrati*".

Nel nostro caso l'equazione della retta è: $y = 5{,}10 + 0{,}387x$. L'intercetta, cioè il valore della variabile dipendente quando quello della indipendente (reddito) è pari a zero, è 5,1 punti e ciò significa che un discreto livello di benessere soggettivo è assicurato anche quando il paese è molto povero. La pendenza della retta mostra invece che a ogni aumento di 10.000 euro di prodotto interno lordo pro capite, la soddisfazione per la propria vita tendenzialmente aumenta di un po' più di un terzo di punto (0,387). Il grafico mostra che pochi paesi si collocano esattamente sulla retta di regressione o molto vicino, il che significa che la soddisfazione per la vita in diversi paesi è sensibilmente superiore (o inferiore) rispetto al valore che ci si potrebbe attendere sulla base del reddito[7].

[7] Trattandosi di variabili cardinali, è corretto parlare al singolare di *valore* atteso, perché si fanno stime per ogni singolo caso. In precedenza si parlava invece al plurale di *frequenze* attese,

Figura 3.3 - Grafico a dispersione con retta di regressione – Soddisfazione per la propria vita per reddito pro capite (2019)



Fonte: OECD (https://stats.oecd.org/)

Mentre due dei paesi più ricchi (Svizzera e Islanda) hanno coerentemente un livello di soddisfazione per la vita di 7,5 punti (su una scala da "0" a "10"), Stati Uniti e Lussemburgo, che hanno un reddito molto simile, si collocano a oltre mezzo punto di distanza (in meno) sulla stessa scala. Su livelli di reddito molto più bassi si notano scostamenti ancora più elevati, a partire dal Messico che, con soli 15.000 dollari di reddito pro capite, mostra un livello di soddisfazione per la vita di poco inferiore a quello degli Stati Uniti e di altri paesi con un reddito di oltre 60.000 dollari. Confrontando i paesi in verticale, intorno a 25.000 dollari si nota che la soddisfazione per la vita passa da meno di 5,5 punti (Portogallo e Grecia) a ben oltre un punto in più (Cile e Repubblica Ceca). Notiamo infine che l'Italia, con esattamente 6 punti, si colloca su un livello di mezzo punto inferiore a quello (6,5) che sarebbe pronosticabile sulla base del reddito pro capite degli italiani.

In conclusione, la relazione tra ricchezza e soddisfazione per la vita è evidente, ma gli scostamenti rispetto a una relazione perfetta, che si registrerebbe se

perché si faceva invece riferimento al numero di casi all'interno di celle di tabelle di contingenza individuate da modalità di variabili categoriali.

tutti i paesi fossero collocati sulla retta di regressione, dimostrano che altre variabili svolgono un ruolo rilevante, attenuando la capacità predittiva del reddito.

3.3.5. I coefficienti e le matrici di correlazione

L'esistenza di un'eventuale relazione tra due variabili cardinali si può dunque desumere nei grafici a dispersione dalla forma della "nuvola di punti", e l'influenza della variabile indipendente (causa) sulla dipendente (effetto) può essere sinteticamente descritta dalla retta di regressione. Si può però facilmente comprendere che la retta descrive situazioni alquanto diverse a seconda di quanto i punti sono mediamente distanti da essa. In pratica, l'effetto della variabile indipendente sui valori della dipendente può essere notevole, ma poco prevedibile; è infatti sempre possibile calcolare e tracciare una retta che interpoli i dati, ma serve poco ai fini della previsione se la retta passa in mezzo a una nuvola di punti molto "dispersi". È dunque necessario determinare anche la forza della relazione, che si può desumere da quanto i punti sono disseminati nel grafico: se lo sono molto (pur tendendo a disporsi su una retta), la relazione è debole, se sono invece concentrati vicino alla retta, la relazione è forte.

Per sintetizzare la forza della relazione si usa il coefficiente di correlazione "*r* di Pearson" che, come i parametri statistici che misurano l'associazione tra altri tipi di variabili visti in precedenza, può variare da -1 a +1. Mentre il valore "0" indica assenza di relazione, "1" indicherebbe l'esistenza di una (improbabile) relazione perfetta. Il segno (+/-) del coefficiente indica invece la direzione della relazione: se è positivo, al crescere dei valori di una variabile tendenzialmente crescono anche i valori dell'altra; se è invece negativo, significa che i valori dell'altra tendono a scendere (nel caso appena esemplificato *r* è 0,727, a indicare una stretta correlazione positiva tra le due variabili).

Elevando al quadrato il coefficiente di correlazione *r* di ottiene r^2 (r-quadrato), che viene anche denominato *coefficiente di determinazione*. Questo parametro è molto significativo perché indica quanta parte della variabilità della variabile dipendente è spiegata dalla variabile indipendente; detto altrimenti, r^2 rappresenta la quantità della variazione nei valori di *y* che può essere giustificata dalla variazione di *x*. Nel nostro esempio r^2 è pari a 0,528, il che significa che poco più di metà della variabilità nel sentimento di soddisfazione per la propria vita nei paesi analizzati può essere imputabile alla ricchezza del paese, misurata utilizzando come indicatore il livello di remunerazione del lavoro dipendente.

Per esaminare contemporaneamente i rapporti all'interno di un set di variabili, i parametri si possono inserire in matrici di "variabili per variabili" definibili nel caso specifico *matrici di correlazione*[8]. Se le stesse variabili intestano le righe e anche le co-

[8] Sui diversi tipi di matrici che si utilizzano nella ricerca sociale, cfr. Delli Zotti (1985).

lonne, la matrice è quadrata, simmetrica (o speculare) e ha una diagonale principale dotata di scarsa utilità informativa se vi si registra la relazione di ogni variabile con sé stessa (la relazione è infatti necessariamente "perfetta" e il valore è dunque sempre "1"). Anche nelle matrici di *import-export* la diagonale è ipoteticamente inutile, ma la si può usare per registrare ad esempio il PNL, al fine di "rapportare" le importazioni e le esportazioni alle dimensioni economiche dei diversi paesi. Analogamente, nella diagonale di una matrice di correlazione si possono inserire informazioni relative a ogni variabile; ad esempio, le risposte mancanti, che abbassano la numerosità dei casi utilizzati per il calcolo e, di conseguenza, la significatività del coefficiente di correlazione. Se si utilizzano dati campionari è infatti opportuno associare ai coefficienti di correlazione il loro livello di significatività, cioè il margine di errore in cui si può incorrere estendendo i risultati ottenuti dal campione alla popolazione di riferimento. La significatività dipende infatti, oltre che dalla forza della relazione, dal numero di casi sui quali è calcolata e generalmente si considera accettabile un livello di significatività che indica una probabilità inferiore al 5% di incorrere in questo errore.

La Tab. 3.7 è una matrice di correlazione quadrata e simmetrica che serve anche a esemplificare la strategia delineata schematicamente all'inizio: dopo avere individuato alcune variabili indipendenti (ipotetiche cause) e una serie di variabili dipendenti (il focus dell'interesse conoscitivo del ricercatore), si inizia con un'analisi delle relazioni interne al primo di questi due gruppi di variabili.

Tabella 3.7 - Matrice di correlazione simmetrica – Serie di variabili indipendenti (Italia 2017)

	Età	**Istruzione**	**Reddito**	**Politica**	**Religiosità**
Età	1	-,41**	-,12**	-,03	,24**
Anni istruzione	-,41**	1	,39**	-,08**	-,23**
Reddito familiare	-,12**	,39**	1	-,03	-,14**
Collocazione politica (sx-dx)	-,03	-,08**	-,03	1	,11**
Grado di religiosità	,24**	-,23**	-,14**	,11**	1

** Correlazione significativa al livello 0,01.

Fonte: European Social Survey (http://www.europeansocialsurvey.org)

In questo tipo di matrici di correlazione il triangolo posto sopra la diagonale è solitamente ridondante, ma non lo è se i coefficienti sono asimmetrici. Se così fosse, in un triangolo si leggerebbe il valore che indica la forza della relazione assumendo che la variabile *x* dipenda dalla variabile *y*, nell'altro assumendo che sia invece *y* la variabile dipendente[9]. Anche quando la matrice è simmetrica può

[9] Come si vedrà nel prossimo capitolo, la specularità della matrice di correlazione simmetrica può essere "sfruttata" per inserire sopra e sotto la diagonale coefficienti di correlazione calco-

però essere utile riportare la matrice di correlazione completa, per rendere più facile la sua consultazione.

Scorrendo la prima riga (o la prima colonna) della Tab. 3.7 si vede che l'età è fortemente correlata con l'istruzione (-0,41) e il segno meno indica che con l'età tendenzialmente diminuisce il livello di scolarizzazione. È invece poco stretta, anche se non trascurabile, la relazione negativa con il reddito (-0,11), quasi nulla quella con l'appartenenza politica (-0,03) e significativa invece la correlazione positiva con la religiosità (0,24). Oltre alla stretta relazione tra istruzione e reddito familiare (0,39), si deve segnalare la relativa indipendenza della collocazione politica dalle altre variabili: è infatti assente la relazione con l'età (-0,03) e il reddito (-0,03), tenue quella con l'istruzione (-0,08) e con la religiosità (0,11).

Queste ultime correlazioni sono utili a chiarire che per ben interpretare il senso (direzione) della relazione è necessario conoscere la definizione operativa delle variabili. Nel caso di età, istruzione e reddito è ovvio che valori più elevati delle variabili significano maggiore anzianità, scolarizzazione e ricchezza e si può intuire che la scala di religiosità rileva livelli crescenti (da 0 "per niente" a 10 "molto"). Nel caso dell'auto-collocazione politica è invece indispensabile conoscere l'ancoramento semantico della scala (da 0 "sinistra" a 10 "destra") per comprendere che all'aumento degli anni di istruzione la collocazione politica si sposta moderatamente verso sinistra e verso destra se invece aumenta il livello di religiosità dichiarata.

Nella matrice di correlazione rettangolare della Tab. 3.8 le caratteristiche socio-demografiche degli intervistati sono messe in relazione con una serie di domande di atteggiamento. In presenza di valori elevati dei coefficienti di correlazione in questo caso si può parlare di un vero e proprio legane di dipendenza degli atteggiamenti dalle caratteristiche socio-demografiche perché non è plausibile che gli atteggiamenti possano influenzare le caratteristiche individuali ascritte (genere, età, estrazione sociale o appartenenza etnica). Si può invece ritenere che anche gli atteggiamenti, specie se ancorati a un solido impianto valoriale, possono svolgere un ruolo nel suggerire, motivare o in altro modo facilitare l'acquisizione di più elevati livelli di istruzione o di reddito.

Nelle matrici di correlazione rettangolari, se le variabili indipendenti intestano le righe e le dipendenti le colonne, la lettura dei coefficienti per riga è finalizzata a individuare le caratteristiche che più influenzano ogni atteggiamento; esplorando i coefficienti per colonna si può invece comprendere su quali atteggiamenti ogni caratteristica riesce a esercitare maggiormente la sua influenza. Ad esempio, dalla prima riga si rileva che la soddisfazione per la propria vita diminuisce con l'età (-0,14) e aumenta invece in relazione agli anni di istruzione (0,18) e al reddito (0,23); risultano invece ininfluenti la religiosità (0,00) e la collocazione politica (0,03). Quest'ultima variabile

lati su due segmenti contrapposti dell'intera casistica (ad esempio, maschi e femmine, giovani e anziani, paesi sviluppati e in via di sviluppo, ecc.).

è invece strettamente correlata con una domanda in merito agli effetti economici dell'immigrazione (-0,30) e, conoscendo il significato dei valori estremi della scala utilizzata per le risposte (da 0 "male" a 10 "bene"), si comprende che tendenzialmente gli intervistati di sinistra valutano positivamente l'apporto economico degli immigrati (e viceversa). Proseguendo nella lettura in colonna, notiamo poi che la collocazione politica è correlata negativamente anche con il giudizio sull'unificazione europea (in questo caso la scala procede da 0 "si è spinta troppo oltre" a 10 "dovrebbe proseguire"). Completando la lettura della colonna di coefficienti si vede infine che la collocazione politica è moderatamente, ma significativamente, correlata anche con la fiducia verso gli altri e di trovare lavoro.

Tabella 3.8 - Matrice di correlazione rettangolare – Serie di domande di atteggiamento per caratteristiche degli intervistati (Italia 2017)

	Età	**Istruzione**	**Reddito**	**Politica**	**Religiosità**
Soddisfatto della propria vita	-,14**	,18**	,23**	,03**	,00
Immigrazione male/bene per economia	-,17**	,26**	,19**	-,30**	-,06**
Mai abbastanza cauti/ci si può fidare	-,09**	,19**	,15**	-,13**	-,04*
Buone chance di trovare il lavoro che cerco	-,18**	,31**	,28**	-,06**	-,08**
Unione europea: troppo avanti/andare avanti	-,17**	,25**	,14**	-,25**	-,01

** Correlazione significativa al livello 0,01.
* Correlazione significativa al livello 0,05.

Fonte: European Social Survey (http://www.europeansocialsurvey.org)

Dal momento che, come accennato all'inizio del capitolo, è buona norma individuare anche le correlazioni deboli o assenti, specialmente se si era ipotizzata a priori l'esistenza di una relazione tra le due variabili/indicatori, è il caso di sottolineare l'assenza o quasi di correlazione della religiosità con la soddisfazione per la vita, il sostegno all'Unione Europea e la fiducia nel prossimo. Viste le prese di posizione sostanzialmente "aperturiste" della Chiesa, sembra andare in direzione opposta a quanto era forse prevedibile la relazione (peraltro debole) con la valutazione sull'immigrazione.

Anche se non inaspettata, è altrettanto utile all'avanzamento della conoscenza la conferma della capacità esplicativa del livello di istruzione, che si può ormai dare per "scontata", quasi a prescindere dal fenomeno analizzato. Scontata infatti non significa banale e, pur in presenza di correlazioni tutte più o meno elevate, è comunque utile monitorare l'eventuale minore (o maggiore) capacità esplicativa su tematiche specifiche e/o l'attenuazione (o ulteriore rafforzamento) nel tempo.

Come si è visto, accanto ai coefficienti di correlazione sono indicati i livelli di significatività, in entrambe le tabelle contrassegnati in gran parte da due asterischi i quali indicano che c'è meno di una probabilità su cento di essere in errore se si afferma che la relazione indicata dal coefficiente è riscontrabile anche nella popolazione di riferimento. La significatività statistica è però connessa, oltre che alla forza della relazione, alla numerosità del campione; pertanto, la si può ottenere anche solo perché si è intervistato un elevato numero di persone. Detto altrimenti, se gli intervistati sono molto numerosi diventa significativo, ma solo statisticamente, anche un coefficiente di correlazione molto basso. Si deve dunque distinguere tra significatività statistica e significatività "sostanziale" (o "semantica") per evitare di enfatizzare correlazioni poco significative sul piano sostanziale.

Verifichiamo ora se le variabili dipendenti, oltre a essere influenzate dalle caratteristiche socio-anagrafiche degli intervistati sono anche più o meno strettamente associate tra loro. Tutti i coefficienti della Tab. 3.9 hanno in effetti valori superiori a 0,10 che, con un campione di oltre 2.500 unità, comportano una significatività statistica molto elevata. Diventa allora interessante esaminare la forza della relazione per notare, ad esempio, che l'opinione relativa al fenomeno migratorio è fortemente correlata con quella sul processo di unificazione europea (0,61). Vi è dunque coerenza tra i due atteggiamenti: chi è restio a riconoscere gli aspetti positivi dell'immigrazione è tendenzialmente orientato negativamente anche nei confronti dell'unificazione europea (la scarsa accettazione della diversità si rivela indifferentemente nei confronti di persone e paesi "altri").

Tabella 3.9 - Matrice di correlazione simmetrica – Serie di domande di atteggiamento (Italia 2017)

	Soddisfazione	**Immigrazione**	**Fiducia**	**Lavoro**	**Europa**
Soddisfatto propria vita	1	0,14**	0,21**	0,25**	0,10**
Immigrazione bene/male per economia	0,13**	1	0,28**	0,22**	0,61**
Ci si può fidare degli altri/mai abbastanza cauti	0,21**	0,28**	1	0,20**	0,27**
Buone chance di trovare il lavoro che cerco	0,25**	0,22**	0,19**	1	0,19**
Unificazione europea: andareavanti/troppo avanti	0,10**	0,61**	0,27**	0,19**	1

** Correlazione significativa al livello 0,01.

Fonte: European Social Survey (http://www.europeansocialsurvey.org)

Anche la domanda riguardante la fiducia nel prossimo misura il grado di apertura sociale degli intervistati e il fatto che sia correlata positivamente con quelle sull'unificazione europea (0,27) e sul ruolo economico degli immigrati (0,28) suggerisce l'esistenza di una dimensione sottostante che le accumuna. Ciò suggerisce anche la possibilità di elaborare un indicatore che riassuma sinteticamente questo atteggiamento di fondo, come si vedrà nel prossimo capitolo, dedicato all'analisi multivariata, e in quello dove si tratterà dell'elaborazione di indici e tipologie.

3.3.6. La scomposizione della media

Se una variabile è categoriale (nominale o ordinale) e l'altra è cardinale si possono calcolare le misure sintetiche della variabile cardinale per categorie di intervistati individuate dai valori della categoriale. È così possibile, ad esempio, scomporre il reddito medio (e relativa deviazione standard) di un insieme di famiglie a seconda del settore in cui il capofamiglia è occupato.

Nella Tab. 3.10 è riportato il risultato della scomposizione di uno degli indicatori utilizzati nell'indagine sulla qualità della vita nelle provincie italiane realizzata annualmente dal quotidiano Sole24ore. La percentuale di giovani che non lavorano, non studiano e non sono impegnati in attività di formazione (Neet) è mediamente 21,5%, con una notevole articolazione territoriale: nel Nord-est si registra il valore più basso (13,8%) e quello più alto nelle Isole, dove la percentuale sale al 34,4% (molto più del doppio). È interessante osservare che anche la deviazione standard mostra valori assai diversificati: il minimo si registra nuovamente nel Nord-est (2,3) e valori molto più elevati nelle Isole (6,1%) e ancor più nelle province del Sud (6,5%), il che significa che la situazione all'interno di queste aree è assai più diversificata di quanto si constata nelle ripartizioni del Nord (le provincie del Centro registrano una situazione intermedia)[10].

[10] È corretto segnalare che le medie inserite nella tabella non si possono considerare stime corrette della percentuale di Neet in ogni ripartizione; per ottenerle si dovrebbe infatti attribuire a ogni dato un peso proporzionale alla popolazione della provincia. Il problema non si pone invece per i valori posizionali perché non sono frutto di un calcolo.

Tabella 3.10 - Scomposizione della media di una variabile – Percentuale di NEET per ripartizione geografica (Italia 2019)

	N	Media	Dev. std.	Mediana	Minimo	Massimo	Intervallo
Nord-ovest	25	16,4	2,9	16,1	10,6	22,5	11,9
Nord-est	22	13,8	2,3	13,4	9,7	19,5	9,8
Centro	22	18,1	4,7	18,0	11,8	29,1	17,3
Sud	24	29,3	6,5	28,1	19,0	44,9	25,9
Isole	14	34,4	6,1	36,8	21,5	42,6	21,1
Totale	107	21,5	8,8	18,8	9,7	44,9	35,2

Eta = ,854 Sig. = ,000

Fonte: Il Sole24ore (https://lab24.ilsole24ore.com/qualita-della-vita/)

Come è intuibile, oltre alla media, si possono scomporre a seconda dei valori della variabile categoriale anche gli altri parametri sintetici e valori posizionali illustrati nel capitolo dedicato all'analisi monovariata. Riguardo alla mediana non desta meraviglia che il suo valore sia simile a quello della media, particolarmente nelle ripartizioni che non registrano casi devianti i quali, come si è spiegato in precedenza, influenzano la media ma non modificano la posizione della mediana. Va anche considerato che, se si tratta di variabili che rilevano percentuali o di scale con un campo di variazione prefissato, i casi devianti non possono presentare valori estremi, come può invece accadere se il minimo e il massimo non sono fissati a priori (ad esempio, nel caso del reddito o il Pil).

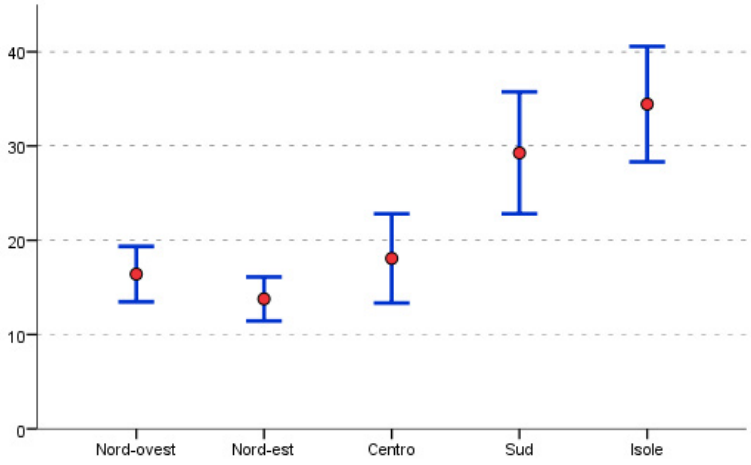
Utilizzando i valori massimi e minimi empiricamente rilevati si ottiene il campo di variazione che, analogamente alla deviazione standard ma con un'interpretazione più intuitiva, descrive sinteticamente la diversificazione tra le provincie che fanno parte di ognuna delle ripartizioni geografico/elettorali. Il campo di variazione è nel complesso di oltre 35 punti percentuali, con una situazione assai meno diversificata nel Nord-est, dove la quota di Neet passa da poco meno del 10% a meno del 20% (campo di variazione 9,8%) e molto più articolata nelle regioni del Sud dove il minimo è 19% e il massimo 44,9% con oltre 25 punti percentuali di differenza.

Per stimare la forza dell'associazione tra una variabile cardinale e una categoriale si utilizza il parametro statistico sintetico *eta* che presenta valori, solo positivi, compresi tra 0 (nessuna associazione) e prossimi a 1 (alto grado di associazione). Mancando il segno, *eta* non stima la "direzione" dell'associazione tra le due variabili, come è logico dal momento che la variabile indipendente può essere a categorie non ordinate (nominale). Se la variabile categoriale è di tipo ordinale, si può comunque individuare la direzione dell'influenza osservando se le medie salgono (o scendono) al crescere dei valori della variabile ordinale.

Il parametro *eta*, inserito in calce alla Tab. 3.9, indica che vi è una relazione assai stretta tra le due variabili (.854); di conseguenza, il livello di significatività (0,00) indica che la significatività della differenza tra le medie che si riscontra confrontando le ripartizioni rasenta la certezza.

I parametri riportati nella tabella si possono rappresentare efficacemente mediante barre degli errori costruite utilizzando in questo caso le medie e le deviazioni standard (Fig. 3.4). La denominazione di questo tipo di grafico può apparire curiosa, ma deriva dal fatto che questi parametri statistici sono nati originariamente per finalità di controllo della qualità della produzione industriale. Dei pezzi prodotti venivano infatti misurate le caratteristiche dimensionali e gli scostamenti dai valori previsti erano ovviamente considerati errori di produzione. Come nella Tab. 2.3, le variabili sono disposte secondo valori decrescenti della media e con lo strumento grafico risultano immediatamente evidenti tutte le caratteristiche delle distribuzioni della variabile per ripartizione che abbiamo descritto commentando i dati inseriti nella tabella.

Figura 3.4 - Grafico a barre di errori – Media e deviazione standard della percentuale di Neet per ripartizione (Italia 2019)



Fonte: Il Sole24ore (https://lab24.ilsole24ore.com/qualita-della-vita/)

Per visualizzare graficamente la relazione tra una variabile cardinale e una categoriale è possibile utilizzare anche i grafici a dispersione in una versione che si può definire a semi-dispersione perché i punti si disperdono solo sulla dimensione della variabile cardinale. Questo tipo di grafico è particolarmente efficace se, oltre a visualizzare la dispersione dei casi si vuole essere in grado di identificarli singolarmente; un aspetto poco rilevante se i casi sono gli anonimi intervistati di un sondaggio, ma assai rilevante se, come in questo esempio, sono ripartizioni territoriali la cui

identificazione può suscitare interessanti ipotesi che spieghino la loro specifica collocazione nel grafico. Se i casi sono poco numerosi in questo tipo di grafico è infatti possibile identificare la posizione di ognuno di essi, altrimenti si evidenzieranno solo i casi di particolare interesse e quelli "devianti". Essendo i casi devianti per definizione poco numerosi e collocati in posizioni distanti dagli altri, è infatti meno probabile che le etichette identificative si sovrappongano impedendo la loro lettura.

Nel grafico di Fig. 3.5 sono analiticamente rappresentati i dati sinteticamente descritti nella Tab. 3.9 e si vede innanzitutto che le provincie appartenenti alle diverse aree del Paese si collocano effettivamente in "fasce" diverse all'interno del complessivo campo di variazione della variabile. La possibilità di identificare i singoli casi consente inoltre di constatare che il minimo (9,7%) si è registrato nella provincia di Bolzano e il massimo (44,9%) nella provincia di Crotone. L'ispezione visiva ci consente di vedere anche che Crotone occupa una posizione alquanto isolata, come accade nella ripartizione "centro" per le provincie di Rieti, Frosinone e Latina; al contrario, nelle "Isole" sono relativamente isolate (verso il basso) le provincie di Nuoro e Cagliari. È possibile anche osservare che il valore più alto che si registra nel Nord-est a Gorizia è comunque più basso di quelli di quasi tutte le province del Sud e delle Isole (con la sola eccezione di L'Aquila).

Figura 3.5 - Grafico a semi-dispersione – Percentuale di Neet per ripartizione (Italia 2019)

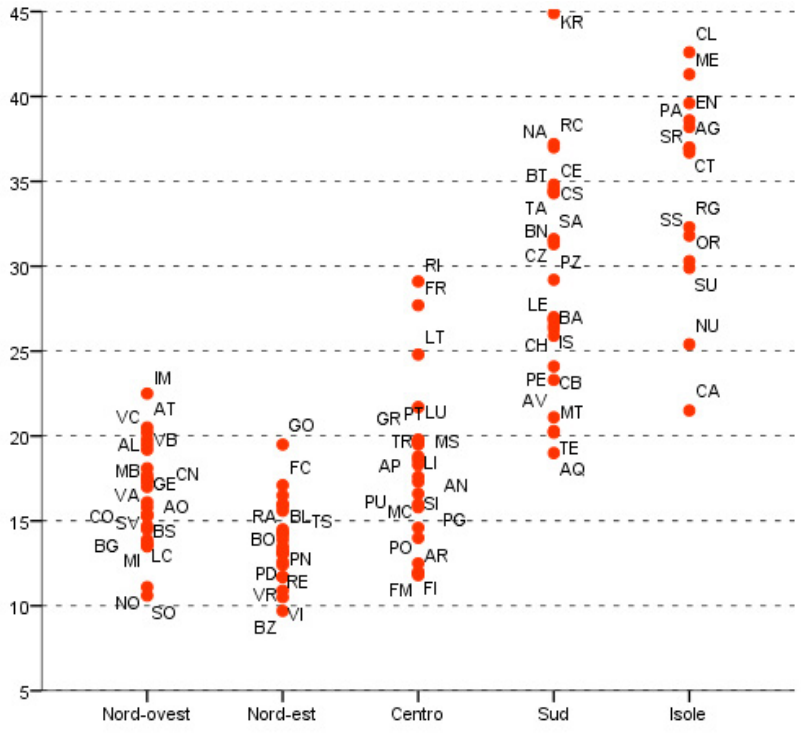
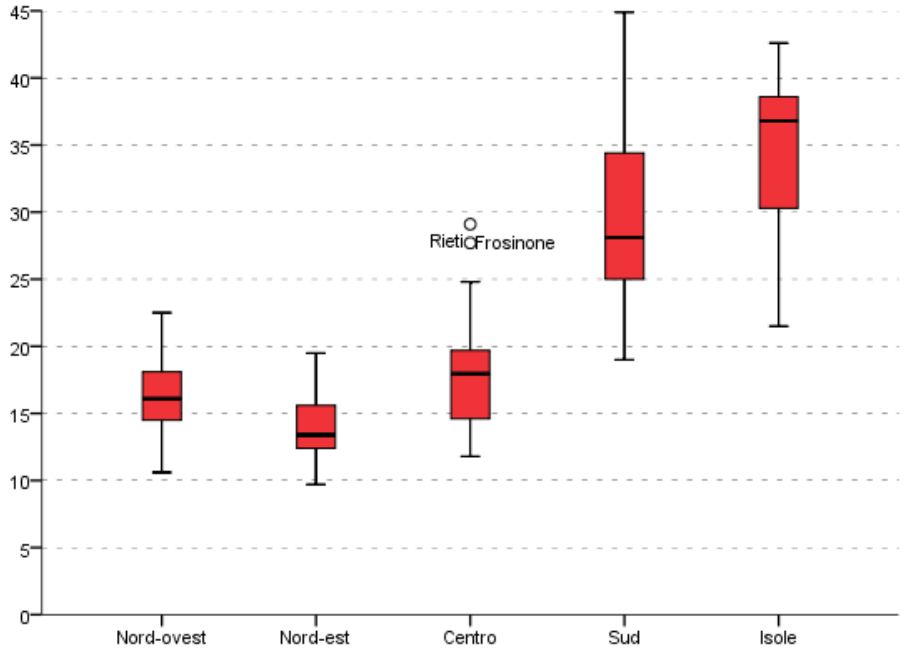


Figura 3.6 - *Box plots* di una variabile cardinale per una categoriale – Percentuale di Neet per ripartizione geografica (Italia 2019)



Fonte: Il Sole24ore (https://lab24.ilsole24ore.com/qualita-della-vita/)

Se si rinuncia alla possibilità di identificare singolarmente i casi può essere efficace usare una serie di *box plots* come nella Fig. 3.6. Anche questo grafico illustra bene l'elevata dispersione dei casi nel Sud e nelle Isole e una minore differenziazione nelle altre aree, in particolare nel Nord-est. In questa rappresentazione la grande maggioranza dei casi è contenuta all'interno del *box* e dei "baffi"; si possono dunque identificare singolarmente solo le province statisticamente "devianti" (*outlier*) perché distanti dalle altre per più di una volta e mezza la differenza interquartile (l'estensione del *box*). Nel nostro caso ciò si verifica in Centro Italia, solo per le province di Rieti e Frosinone.

Questo tipo di grafico rivela in modo efficace un ultimo aspetto che è interessante sottolineare. Utilizzando i quartili si suddividono i casi in quattro parti uguali; dunque, delle 24 province del Sud, 6 sono posizionate nello spazio tra il "baffo" di sinistra e l'inizio della scatola; altre 12 sono contenute nella scatola, metà a sinistra e metà a destra della mediana; infine, le ultime 6 sono posizionate nello spazio tra la fine della scatola e il "baffo" di sinistra. Se si osserva la distanza dalla mediana dei due "baffi" si nota un'asimmetria, che è ancora più evidente tra le province delle Isole, con una significativa differenza: nel Sud sono

più ravvicinate le posizioni delle province con una minore percentuale di Neet; al contrario, nelle Isole le provincie con una situazione occupazionale dei giovani più disagiata sono compresse in soli 6 punti percentuali, mentre quelle con una situazione occupazionale migliore si collocano in uno spazio più ampiamente dilatato (poco più di 16 punti).

4. Analisi multivariata

I modelli di analisi multivariata possono essere definiti una sorta di conseguenza (ana)logica della complessità della realtà sociale. Le tecniche di analisi bivariata si propongono di identificare le cause della variabilità riscontrata nelle distribuzioni monovariate delle variabili: da un punto di vista statistico si può dire che il ricercatore cerca una "spiegazione" della varianza delle distribuzioni. In genere però, nelle ricerche empiriche realizzate nelle scienze sociali, le correlazioni tra coppie di variabili non sono molto elevate e ciò significa che la maggior parte della varianza della variabile dipendente (effetto) che tentiamo di spiegare rimane non spiegata da un'unica variabile indipendente (causa)[1].

[1] Marradi (1997) sostiene che utilizzare il termine "spiegazione" potrebbe indurre a ritenere che non sia necessario un intervento interpretativo del ricercatore: sarebbe dunque preferibile utilizzare termini che alludano alla capacità di "predire" o "riprodurre" la varianza. In questa sede si è utilizzato però egualmente il termine "spiegazione" per la sua larga diffusione e perché negli esempi proposti si è proprio cercato, per quanto possibile, di fornire una spiegazione delle relazioni tra le variabili. Si ritiene comunque legittimo utilizzare il termine spiegazione in tutti i casi in cui si effettua una ricerca "confermativa" per sottopone a verifica empirica un modello (o delle ipotesi), Quando si trova nelle analisi un riscontro (come un'alta correlazione tra due variabili) si può a ben diritto parlare di spiegazione perché proprio da una spiegazione, per quanto ipotetica, è iniziato il lavoro.

Ciò è dovuto, oltre che alla difficoltà che il ricercatore incontra nella "misurazione" dei fatti sociali, alla loro complessa articolazione interna e alle cause altrettanto complesse, se non di più: da ciò la necessità di costruire modelli che riproducano questa complessità.

Per sottoporre a "verifica" empirica questi modelli si sono realizzate intere famiglie di tecniche multivariate di analisi dei dati. Le conoscenze tecniche necessarie per utilizzarle in maniera appropriata sono naturalmente proporzionate alla loro complessità e, dunque, non è ragionevolmente possibile affrontarle con completezza in un testo introduttivo. Cionondimeno, riteniamo sia necessario fornire almeno un'informazione generale su di esse. Sono fortunatamente disponibili diversi testi specializzati cui ci si può riferire per impadronirsi della necessaria competenza tecnico-statistica; ma, se non si conosce nemmeno l'esistenza di questi modelli e dei principali problemi di ricerca ai quali ognuno di essi cerca di offrire una risposta, le opportunità che offrono rischierebbero di rimanere inutilizzate[2]. Mentre i modelli più complessi verranno sinteticamente descritti più avanti, ci si soffermerà con maggiore dettaglio su quello che può essere definito il mattone fondamentale dell'analisi multivariata: l'utilizzo delle c.d. "terze variabili" con finalità di controllo delle relazioni bivariate[3]. Si dedicherà anche in questo caso una certa attenzione alle tecniche grafiche di analisi e rappresentazione dei dati, perché esse possono diventare ancor più utili ed efficaci proprio quando le relazioni tra le variabili sono multivariate[4].

4.1. Il controllo delle terze variabili

Benché nel campo delle scienze sociali il livello dell'analisi bivariata sia il massimo della sofisticazione statistica cui giungono la maggior parte dei lavori di ricerca, ci si dovrebbe spingere al di là di questo livello per almeno due motivi. Innanzitutto, quando trova una relazione interessante tra due variabili il ricercatore dovrebbe assicurarsi che la relazione non sia "spuria". Ognuna del-

[2] Un utile testo di riferimento per approfondire le tecniche di analisi multivariata, anche alcune non trattate in questa sede, è il volume di Di Franco (2017). Per le elaborazioni statistiche con Spss, oltre ai manuali d'uso del programma, scaricabili gratuitamente anche in versione italiana (https://www.ibm.com/support/pages/ibm-spss-statistics-26-documentation#it), è disponibile il citato testo introduttivo dello stesso Di Franco (2010).

[3] Per approfondire la conoscenza del modello più semplice di analisi multivariata, si vedano Ricolfi (2000) e Rosenberg (2003), che analiticamente illustrano ed esemplificano le funzioni che la terza variabile può esercitare nei confronti della relazione bivariata.

[4] Per una più ampia trattazione di questa tematica si veda Delli Zotti (2010).

le due variabili, infatti, può essere a sua volta fortemente correlata con una (o più) altre variabili e possono essere queste ultime le relazioni più interessanti da un punto di vista sostantivo, e non meramente statistico. Per utilizzare un noto esempio paradossale: un'alta correlazione tra numero di autopompe che accorrono sul luogo dell'incendio e gravità dei danni si può spiegare con l'esistenza della variabile "dimensione dell'incendio", correlata con le prime due, che risolve l'imbarazzo di dover concludere che l'intervento dei vigili del fuoco aggrava, anziché mitigare, le conseguenze dell'incendio. Il controllo della relazione spurie costituisce un momento essenziale e delicato dell'analisi dei dati, perché:

> Non si può sapere a priori quanto e quali siano le variabili rilevanti prima di avere raggiunto una completa conoscenza del contesto della ricerca, conoscenza che è un risultato e non una premessa della ricerca. Per questo è sempre possibile l'individuazione di relazioni spurie durante l'analisi (Gangemi 1985: 141).

Può poi accadere che la relazione tra le due variabili sia "soppressa" o "oscurata" dall'interferenza di terze variabili e solo l'introduzione di una variabile di controllo può rivelare l'esistenza della relazione o rendere palese la sua forza.

Vanno comunque considerati anche i limiti che scontano le procedure di scomposizione: aumentando il numero delle variabili poste sotto controllo, diventa sempre più difficile interpretare i risultati; inoltre, con l'aumento della frammentazione del campione alcune sotto-categorie rischiano di essere rappresentate da un numero molto limitato di casi e ciò può far diminuire di molto la significatività statistica, se non si può contare su un campione di partenza di dimensioni ragguardevoli.

4.1.1. Tabelle di contingenza e grafici con variabile di controllo

Le tecniche più semplici discendono direttamente da quelle bivariate e possono essere considerate una loro estensione finalizzata al controllo delle terze variabili. Le tabelle di contingenza possono essere, ad esempio, scomposte in sotto-tabelle che si riferiscono a sottopopolazioni individuate dai valori della terza variabile. Nella Tab. 4.1 la disponibilità ad accettare l'immigrazione in Italia di persone etnicamente diverse dagli italiani a seconda del livello di istruzione è scomposta introducendo come variabile di controllo il genere.

Tabella 4.1 - Tabella di contingenza a tripla entrata – "Permettere a persone di razza/ etnia diversa di venire a vivere in Italia" per titolo di studio e genere (Italia 2017)

		Molti	Alcuni	Pochi	Nessuno	Totale	
Obbligo	Maschio	12,4	28,8	36,9	21,8	100,0	531
	Femmina	9,9	31,2	37,1	21,8	100,0	638
	Totale	11,0	30,1	37,0	21,8	100,0	1.169
Diploma	Maschio	13,7	36,9	36,9	12,6	100,0	548
	Femmina	11,6	40,6	35,2	12,5	100,0	559
	Totale	12,6	38,8	36,0	12,6	100,0	1.107
Laurea	Maschio	22,3	43,4	28,0	6,3	100,0	175
	Femmina	28,1	46,5	20,0	5,4	100,0	185
	Totale	25,3	45,0	23,9	5,8	100,0	360
Totale	Maschio	14,4	34,4	35,6	15,6	100,0	1.254
	Femmina	13,0	37,0	34,1	15,8	100,0	1.382
	Totale	13,7	35,8	34,8	15,7	100,0	2.636

Fonte: European Social Survey (http://www.europeansocialsurvey.org)

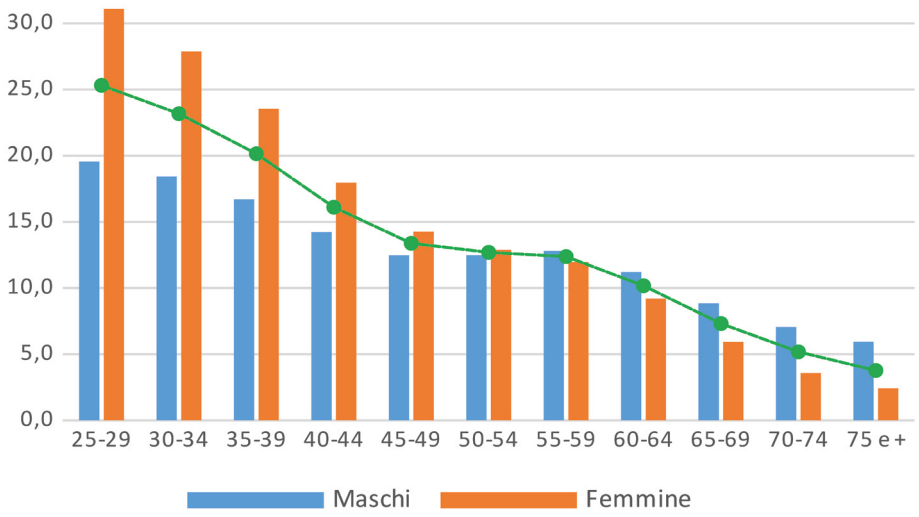
Se si guarda alla risposta "molti" per l'insieme degli intervistati in possesso dei diversi livelli di scolarità, si nota un piccolo incremento tra chi non è andato oltre la scuola dell'obbligo (11%) e chi è in possesso di un diploma di scuola media superiore (12,6%) e addirittura il raddoppio (25,3%) tra gli intervistati in possesso di un titolo di studio universitario. Suddividendo gli intervistati a seconda del genere, si vede poi nelle righe in fondo alla tabella che i maschi (14,4%) sono un po' più "accoglienti" delle femmine (13%). È interessante però vedere cosa accade se si suddividono per genere gli intervistati all'interno dei livelli di istruzione: tra chi non è andato oltre la scuola dell'obbligo si amplia leggermente la divaricazione tra maschi e femmine notata a livello generale (la differenza sale a 2,5 punti) e tra gli intervistati diplomati scende un po' (2,1 punti). Se però osserviamo i laureati, la differenza si amplia alquanto, ma è ancor più interessante notare che le percentuali si invertono; sono infatti le femmine (28,1%) a essere più accoglienti dei maschi (22,3), con una differenza che sale a quasi 6 punti.

Si può concludere che l'istruzione svolge un notevole ruolo nel plasmare gli atteggiamenti sull'immigrazione, ma particolarmente tra le femmine; tra i maschi infatti la percentuale dei favorevoli all'arrivo di "molti" immigrati quasi raddoppia se si confrontano i meno scolarizzati con i laureati, ma tra le femmine la percentuale arriva a quasi triplicare.

Rimanendo sul tema dell'istruzione, nella Fig. 4.1 è mostrata graficamente, invece che in forma tabellare, la relazione bivariata tra l'età (in classi quinquennali) e il titolo di studio, usando anche in questo caso il genere come variabile di controllo. Sono utilizzati i dati del Censimento della popolazione del 2011 perché sono più dettagliati per le classi d'età più anziane, rispetto ai dati campionari più recenti. Ciò consente di visualizzare con più completezza il cambiamento "epocale" che l'utilizzo della terza variabile (il genere) consente di rivelare.

La linea tratteggiata mostra che nel 2011 era laureato solo il 5% di chi aveva 75 anni o più e, invece, il 25% dei 25-29enni (un incremento notevole, ma inferiore a quello che si è registrato nella maggior parte degli altri paesi). Solo scorporando i dati secondo il genere si vede però che l'incremento è attribuibile principalmente alla componente femminile. Le due serie di barre mostrano infatti che tra i più anziani la percentuale di maschi laureati era più che doppia rispetto a quella delle femmine, mentre tra chi è nato circa cinquanta anni dopo sono le femmine a registrare una percentuale di gran lunga superiore (oltre il 30% a fronte di meno del 20% dei maschi). Il grafico consente anche di notare che l'inversione di tendenza ha iniziato a manifestarsi tra chi all'epoca del censimento aveva circa 50 anni e che il divario si è progressivamente ampliato. Le stime più recenti dell'Istat (2020) mostrano che il divario permane, ma si è stabilizzato: tra i 25-29enni i laureati maschi sono poco meno del 25% e le femmine poco più del 35%.

Figura 4.1 - Grafico a barre e linee – Percentuale di laureati per classe di età e genere

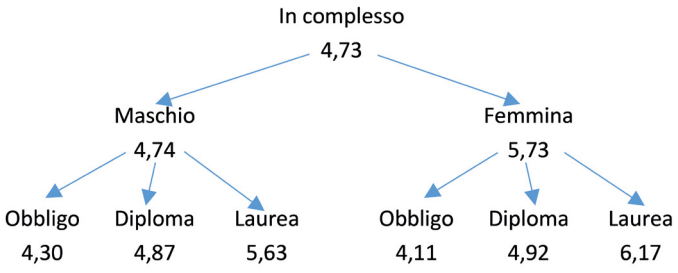


Fonte: Istat – Censimento della popolazione 2011 (http://dati.istat.it/Index.aspx)

I diagrammi ad albero sono di aspetto simile agli organigrammi o agli alberi genealogici e possono essere utilizzati per rappresentare graficamente la scomposizione dei parametri statistici, spingendo l'analisi fino a un livello di stratificazione multipla. In realtà, la metafora dell'albero è poco pertinente perché si tratta di ramificazione, ma delle radici. A parte la natura della ramificazione, è opportuno tenere presente che, dal punto di vista sostanziale, si possono realizzare modelli diversi di analisi utilizzando le stesse variabili, ma cambiando la gerarchia del loro ingresso nel modello.

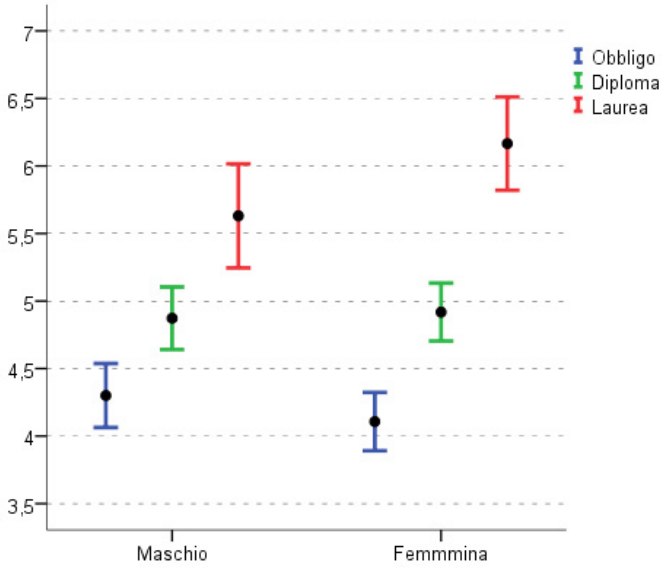
L'esempio della Fig. 4.2 mostra la scomposizione del punteggio medio di valutazione dalla presenza degli immigrati con riferimento alla vita culturale del Paese su una scala da 0 "minaccia" a 10 "arricchimento". L'analisi bivariata in questo caso mostra una differenza di circa un punto tra i maschi (4,74) e le femmine (5,73). Come si poteva prevedere, la differenza si amplia in relazione al titolo di studio, con una differenza (non mostrata nel diagramma) di poco meno di due punti tra chi non è andato oltre la scuola dell'obbligo (4,20) e i laureati (5,91). Anche in questo caso, solo scorporando le medie di maschi e femmine si scopre che il titolo di studio agisce in modo più accentuato tra queste ultime. Infatti, tra i maschi il punteggio aumenta di meno di un punto e mezzo passando da chi ha studiato poco (4,30) ai laureati (5,63), mentre tra le femmine si passa da 4,11 a 6,17, con un incremento che supera dunque i due punti percentuali. Inoltre, tra i poco scolarizzati le femmine apprezzano meno dai maschi l'apporto culturale degli immigrati, mentre tra i laureati avviene il contrario e con un distacco più ampio: in pratica, questa "compensazione" occultava una differenza di genere che l'utilizzo del livello di istruzione come variabile di controllo ha consentito di rilevare.

Figura 4.2 - Diagramma ad albero della scomposizione di una media – Vita culturale danneggiata (0) o arricchita (10) dagli immigrati per titolo di studio e genere (Italia 2017)



Fonte: European Social Survey (http://www.europeansocialsurvey.org)

Figura 4.3 - Grafico a barre degli errori – Vita culturale danneggiata (0) o arricchita (10) dagli immigrati per titolo di studio e genere (Italia 2017)



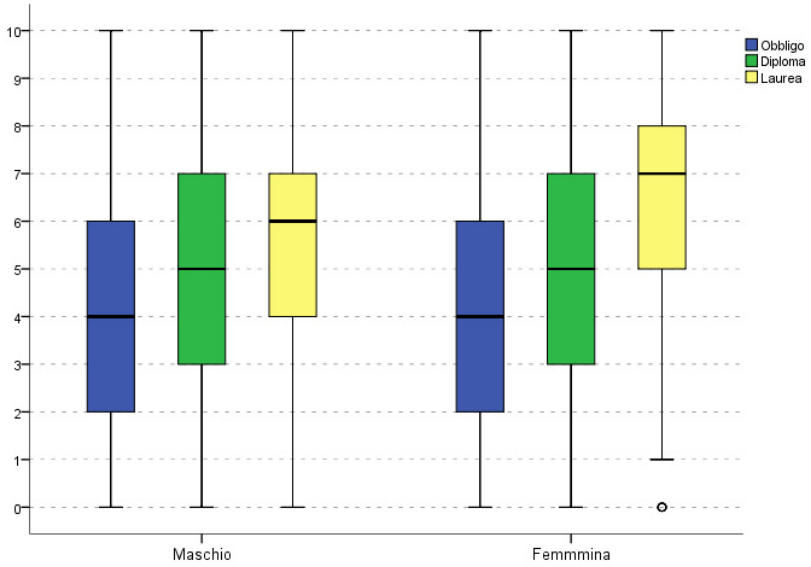
Fonte: European Social Survey (http://www.europeansocialsurvey.org)

Le medie, come si è visto nel capitolo precedente, si possono rappresentare tramite un grafico a barre degli errori e, in questo caso, dal momento che stiamo analizzando dati campionari, nella Fig. 4.3 si sono visualizzati anche gli intervalli di confidenza al 95% i quali indicano entro quali valori si stima che cada il valore vero della media relativo all'intera popolazione dalla quale il campione è stato estratto. Il grafico mostra dunque con efficacia l'impatto del livello d'istruzione che fa aumentare il grado di consenso per l'apporto culturale degli immigrati, particolarmente tra le femmine, e anche l'ampliarsi della variabilità delle opinioni tra i laureati, testimoniato dalla maggiore lunghezza delle barre.

4.1.3. Box plots trivariati

Come anticipato nel capitolo dedicato all'analisi monovariata, i *box plots* sono particolarmente utili per effettuare analisi bivariate con variabili di controllo, perché l'alternativa è esaminare tabelle di contingenza multiple assai più complesse della Tab. 4.1 esaminata all'inizio del capitolo. Ciò è particolarmente vero

Figura 4.4: *Box plots* bivariati con variabile di controllo – La vita culturale del paese è danneggiata (0) o arricchita (10) dagli immigrati per titolo di studio e genere (Italia 2017)



Fonte: European Social Survey (http://www.europeansocialsurvey.org)

quando le proprietà sono misurate con precisione e dunque, al limite, ogni valore è unico perché rilevato in un solo caso[5].

Nella Fig. 4.4 sono rappresentate mediante box-plot le risposte alla domanda sull'apporto culturale degli immigrati appena descritte sinteticamente mostrando le differenze tra le medie. Questa rappresentazione grafica è ovviamente più completa perché mostra, oltre alla tendenza centrale (utilizzando la mediana invece della media), l'intera distribuzione dei casi, e dunque anche la dispersione dei dati nello spazio semantico delimitato dalle risposte da 0 "danneggiata" a 10 "arricchita". Osservando la posizione delle mediane, al centro di ogni box, è facile rilevare l'aumento dell'apprezzamento per l'apporto culturale degli immigrati al crescere del livello di istruzione degli intervistati. Si vede poi con chiarezza che in ogni categoria di intervistati il consenso è lungi dall'essere unanime, ma si può allo stesso tempo osservare che per alcune categorie di intervistati la disparità di giudizio è più limitata, come si arguisce confrontando le lunghezze dei box. Infine, particolarmente nel

[5] Questa situazione è frequente, ma non connaturata alle variabili cardinali; chiedendo a un campione di intervistati quanti sono il loro figli o fratelli/sorelle otteniamo una variabile cardinale, ma in questo caso la gamma dei valori è limitate a dunque la variabile cardinale può essere trattata come categoriale. Ovviamente la situazione è alquanto diversa se agli intervistati chiediamo il reddito: ammesso che gli intervistati rispondano, probabilmente non troveremmo due valori uguali e dunque la tabella non sarebbe altro che una riproposizione dei redditi elencati nella matrice dei dati.

caso delle femmine laureate questo tipo di rappresentazione della distribuzione dei dati mostra con molta efficacia l'asimmetria delle opinioni, testimoniata dall'elevato valore della mediana ("7" su una scala da "0" a "10") e anche il fatto che solo per questa categoria di intervistati si registra un isolato caso che ha attribuito il valore "0" (ritenendo che gli immigrati danneggino la vita culturale del Paese).

4.1.4. Grafici a dispersione con variabile di controllo

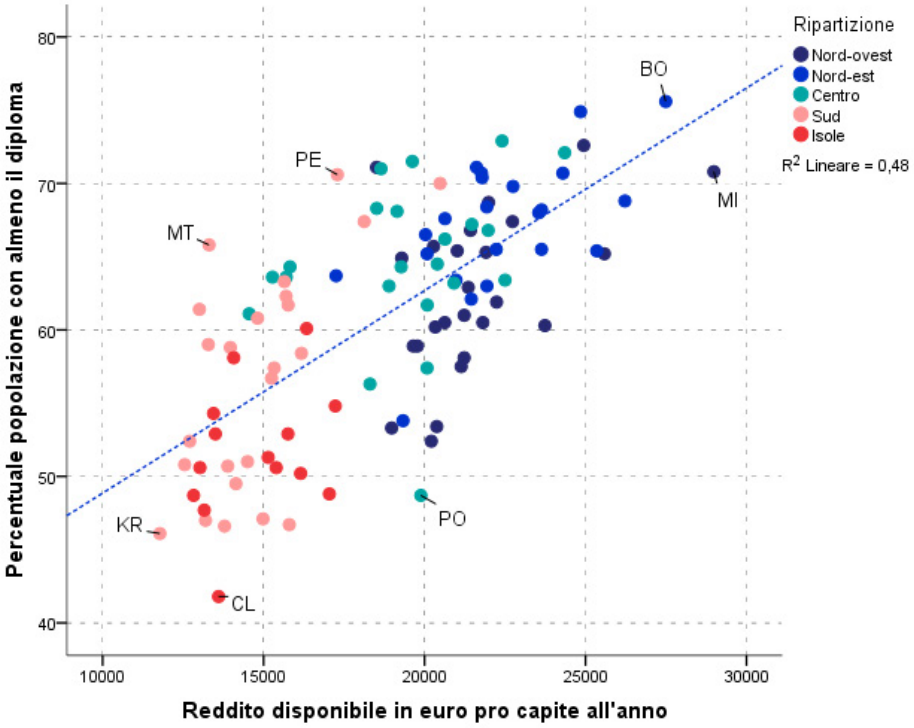
Le variabili di controllo possono essere usate anche nei diagrammi a dispersione, usando simboli specifici per contrassegnare i casi appartenenti a ognuna delle categorie. Si possono così, ad esempio, individuare eventuali effetti "compositivi" che potrebbero oscurare l'esistenza di una relazione tra le variabili (la complessiva "nuvola" può essere infatti il prodotto di sottoinsiemi di punti che, considerati separatamente, possono avere andamenti difformi da quello generale).

A questo proposito può essere interessante citare l'esempio "classico" del grafico a dispersione che mostrava una apparente relazione di segno positivo tra voti per il Movimento sociale-Destra nazionale e per l'abrogazione della legge sul divorzio nel 1974, ma nascondeva una situazione radicalmente diversa, come ben hanno illustrato Arculeo e Marradi (1985) usando come variabile di controllo la ripartizione secondo le aree di diversa tradizione politica del Paese (Nord laico, Zona bianca, Zona rossa, Centro e Sud).

Effettuata questa scomposizione, la relazione tra le due variabili risultava essere in realtà inversa: all'interno di quattro aree (con esclusione della zona rossa), dove il Msi-Dn riceveva un'alta percentuale di voti si registrava una bassa percentuale di elettori che appoggiavano la proposta di abrogazione della legislazione divorzista e, viceversa, dove aveva peso minore il suffragio per il Msi-Dn, i voti abrogazionisti erano percentualmente più presenti.

L'esempio che proponiamo (Fig. 4.5) è simile e in esso sono utilizzati i dati dell'indagine sulla qualità della vita del *Sole24ore* mettendo in relazione la percentuale di popolazione delle provincie in possesso di un titolo di studio pari ad almeno il diploma di scuola media superiore con il reddito disponibile pro capite. Come si vede chiaramente, la relazione tra le due variabili è piuttosto stretta e la pendenza della retta di regressione mostra con evidenza che all'aumento della percentuale di popolazione in possesso di un titolo di studio superiore aumenta notevolmente il reddito disponibile. La variabile di controllo utilizzata in questo caso è la ripartizione territoriale, resa visibile nel grafico attribuendo un diverso colore ai punti che contrassegnano la posizione delle provincie: due gradazioni di blu per le ripartizioni del Nord, il colore verde per le provincie del Centro e due diverse gradazioni di rosso per le provincie del Sud e delle Isole.

Figura 4.5 - Diagramma a dispersione con variabile di controllo – Percentuale di popolazione diplomata o laureata per reddito pro capite e ripartizione (Italia 2019)



Fonte: Il Sole24ore (https://lab24.ilsole24ore.com/qualita-della-vita/)

Nel grafico sono poi contrassegnate alcune delle province che occupano posizioni relativamente incongrue perché registrano una percentuale di diplomati o laureati superiore oppure inferiore a quella prevedibile sulla base della ricchezza pro capite mediamente disponibile. Ovviamente, dal momento che in questo caso la relazione di causa-effetto non è poi così evidente, si potrebbe anche ritenere che la percentuale di persone con elevato livello di istruzione è diversa da quella che farebbe supporre il reddito pro capite della provincia perché si può in effetti ritenere che l'elevato livello di scolarizzazione dalla popolazione possa avere l'effetti di una maggiore produzione di reddito; ma è anche plausibile affermare che una popolazione relativamente agiata possa permettersi di avviare agli studi superiori una più elevata quota di popolazione.

Ad ogni modo, la diversa colorazione dei punti che contrassegnano la posizione delle provincie mostra inequivocabilmente che quelle del Sud e delle

Isole occupano prevalentemente le posizioni nell'area del grafico caratterizzata da valori bassi in entrambe le variabili, mentre il contrario accade per le provincie del Centro-nord.

4.1.5. Matrici di correlazione con variabile di controllo

Anche i coefficienti di correlazione possono essere utilizzati per effettuare semplici analisi trivariate. Nella matrice della Tab. 4.2 i triangoli individuati dalla diagonale (non utilizzata) non contengono informazioni ridondanti perché il campione è dicotomicamente suddiviso per genere: sopra la diagonale sono inseriti i coefficienti di correlazione calcolati tra i maschi, sotto la diagonale quelli calcolati tra le femmine.

Così presentate, le coppie di correlazioni parziali sono facilmente individuabili e confrontabili, ed è possibile quindi svolgere una semplice analisi trivariata. Anche un'osservazione superficiale della matrice di correlazione permette infatti di notare che i coefficienti nel gruppo delle femmine sono in quasi tutti i casi più elevati di quelli calcolati tra i maschi. In molti casi la differenza è minima, altre volte assai più elevata come nel caso del giudizio di soddisfazione per il governo nazionale e per i servizi educativi e sanitari. Ciò significa che i maschi tendono a formulare giudizi, positivi o negativi che siano, maggiormente discriminanti tra le diverse istituzioni e le femmine giudizi un po' più simili e, dunque, forse un po' più stereotipati.

Tabella 4.2 - matrice di correlazione con variabile di controllo – Soddisfazione per lo stato di una serie di istituzioni per genere (Italia 2017)

	Stato economia	**Governo nazionale**	**Vita democratica**	**Sistema educativo**	**Servizio sanitario**
Stato dell'economia	-	,47	,51	,33	,30
Governo nazionale	,51	-	,48	,19	,08
Vita democratica	,47	,58	-	,44	,34
Sistema educativo	,38	,30	,49	-	,50
Servizio sanitario	,30	,15	,39	,48	-

Sopra la diagonale = maschi

Sotto la diagonale = femmine

Fonte: European Social Survey (http://www.europeansocialsurvey.org)

4.2. Modelli multivariati asimmetrici

Oltre il livello trivariato vi sono modelli di analisi ancor più complessi, ideati per rispondere a esigenze conoscitive anche assai diverse. Questi modelli possono essere suddivisi in due ampie categorie, a seconda che introducano, o meno, un'asimmetria nei ruoli che le variabili giocano all'interno dei modelli stessi (Fabbris 1990: 22). Tra i modelli asimmetrici in questo paragrafo verranno sinteticamente descritte le tecniche di regressione multipla, con l'interessante applicazione dell'analisi dei "sentieri" causali (*path analysis*), e l'analisi discriminante.

4.2.1. Regressione multipla e analisi dei sentieri causali

Con le tecniche di regressione lineare multipla il ricercatore si propone di "spiegare" la variabilità di una variabile dipendente mediante l'impiego di più variabili indipendenti. Con tali tecniche è possibile accertare, per ogni variabile che entra nel modello, il suo specifico contributo nello spiegare (o predire) la varianza della variabile dipendente. Il modello si basa sulla constatazione che le variabili indipendenti correlate con la variabile dipendente (il fenomeno di cui cerchiamo di spiegare la variabilità), oltre a essere numerose, sono spesso anche correlate tra di loro. È allora del tutto legittimo ritenere che parte della capacità di spiegare la varianza posseduta da una variabile indipendente sia in comune con quella delle altre variabili indipendenti con le quali è essa stessa correlata.

Nell'esempio che proponiamo si potrà vedere il risultato dell'applicazione di questa tecnica ai dati dell'ultima rilevazione dell'*European Values Study* realizzata intervistando in Italia un campione di 2.277 maggiorenni. Sulla base della matrice di correlazione semplice sono state innanzitutto individuate le caratteristiche degli intervistati che era ragionevole candidare al ruolo di variabili indipendenti e cioè di predittori dell'atteggiamento nei confronti dell'aborto (da 1 "mai giustificato" a 10 "sempre giustificato"), uno dei 15 comportamenti sui quali era stato chiesto agli intervistati di esprimersi. La scelta delle variabili indipendenti era ovviamente limitata a quelle presenti nel data base, con l'ulteriore vincolo che dovevano essere tutte cardinali (come la dipendente), perché la regressione multipla si basa su correlazione bivariate che si possono calcolare solo con variabili di questo tipo. Le variabili selezionate sono l'età anagrafica, il livello di istruzione (anni di studio) e il reddito familiare. Si era ipotizzato di utilizzare anche il genere, perché una variabile dicotomica può essere trattata tecnicamente come una cardinale, ma alla fine si è deciso di toglierla dal modello di analisi perché (forse un po' sorprendentemente) è risultata assai debolmente correlata a livello bivariato con l'atteggiamento nei confronti dell'aborto.

A queste caratteristiche socio-anagrafiche si sono aggiunte come predittori altre due caratteristiche degli intervistati che si poteva supporre potessero influenzare l'atteggiamento verso l'aborto, visto che la sua introduzione in Italia è stata accompagnata da un acceso dibattito politico e prese di posizione della Chiesa che hanno poi portato a indire un referendum abrogativo. Come indicatore di atteggiamento politico abbiamo utilizzato la collocazione sul continuum da 1 "sinistra" a 10 "destra" e per l'atteggiamento religioso la risposta alla domanda "quanto è importante Dio nella tua vita"? (cui si rispondeva su una scala da 1 "per niente" a 10 "molto").

Nella Tab. 4.3 sono riportati i coefficienti di correlazione bivariata i quali mostrano innanzitutto che le cinque variabili indipendenti sono tutte significativamente correlate con l'atteggiamento nei confronti dell'aborto: come ci si poteva attendere sono positivamente correlati il livello d'istruzione e il reddito e negativamente l'età, la collocazione politica e la religiosità. Sono conferme di quanto si era ipotizzato certamente interessanti, anche se forse un po' scontate, e ci si potrebbe pertanto limitare a prenderne atto. Dal resto della matrice emergono però alcune significative correlazioni tra le stesse variabili indipendenti che, pur essendo attese, possono innescare la curiosità di andare più a fondo, per comprendere se ciò non costituisca una sovrabbondanza di spiegazione. Ad esempio, visto che l'atteggiamento nei confronti della religione, oltre a essere la variabile indipendente più correlata con il giudizio sulla liceità dell'aborto è allo stesso tempo molto correlato con l'età anagrafica, il livello di istruzione e la collocazione politica, può essere interessante cercare di capire e misurare quanto rimanga della capacità predittiva dell'atteggiamento religioso, una volta che conosciamo l'età, il livello di istruzione e la collocazione politica degli intervistati.

Tabella 4.3 - Matrice di correlazione bivariata - Caratteristiche degli intervistati e atteggiamento nei confronti dell'aborto (Italia 2019)

	Età	**Istruzione**	**Reddito**	**Politica**	**Religione**
Età completamento studi	-,37				
Reddito familiare	-,06	,36			
Collocazione politica	-,03	-,10	-,10		
Quanto importante Dio	,20	-,18	-,10	,20	
Aborto	-,20	,25	,22	-,17	-,45

Fonte: European Values Study (https://europeanvaluesstudy.eu/)

Si deve a questo punto precisare che i modelli di regressione multipla sono più di uno e tra questi quello più semplice, perché non richiede di formulare preventivamente ipotesi molto articolate, è la procedura a passi (*stepwise*). Con la tecnica

di regressione *stepwise* additiva, dopo avere calcolato i coefficienti di correlazione bivariati (Tab. 4.3) viene inserita nel modello la variabile più correlata e calcolato il coefficiente di regressione; al secondo passo viene inserita nel modello la variabile, tra quelle rimaste, che contribuisce maggiormente a spiegare la varianza residua della variabile dipendente e così via. La procedura si arresta quando non riesce a trovare un'altra variabile che faccia aumentare in maniera statisticamente significativa la capacità del modello di spiegare la varianza della variabile dipendente[6].

Vi sono altre tecniche di regressione multipla disponibili e tra queste una variante della procedura *stepwise* che prevede al primo passo l'inserimento nel modello di tutte le variabili, quindi la sottrazione di quella che meno contribuisce alla spiegazione della varianza. Il proposito è giungere alla spiegazione del massimo della varianza con un modello per quanto possibile "parsimonioso" dal punto di vista del numero delle variabili utilizzate. Il processo si arresta quando la sottrazione della variabile successiva si tradurrebbe in una significativa perdita di capacità esplicativa.

In altre situazioni il ricercatore ritiene di potere fondatamente decidere lui stesso quale sia il corretto ordine di introduzione delle variabili nel modello, sulla base di considerazioni teoriche o logiche, e perciò, abbandonando l'automatismo basato su considerazioni eminentemente statistiche potrà imporre come sequenza la gerarchia tra le variabili indipendenti, derivante dalla loro correlazione bivariata con la variabile dipendente.

La soluzione adottata per il nostro esempio è ibrida, perché si è deciso di inserire in un primo blocco le caratteristiche socio-anagrafiche (età, istruzione e reddito), lasciando all'algoritmo la decisione sulla sequenza del loro inserimento, e in due passi successivi, le altre due variabili (collocazione politica e atteggiamento religioso). Ciò è stato fatto perché, lasciando fare all'automatismo, la prima variabile che entrava nel modello era ovviamente l'atteggiamento religioso (variabile con la più elevata correlazione bivariata) che lasciava ben poco spazio predittivo a età e istruzione, se inserite nel modello successivamente. Si è preferito dunque "imporre" il preventivo inserimento nel modello delle caratteristiche socio-anagrafiche "di base" e in questo modo si è visto che la capacità predittiva dell'atteggiamento religioso rimaneva comunque assai consistente.

Nell'esempio proposto la variabile che contribuisce in maggior misura a predire l'atteggiamento verso l'aborto è il livello di istruzione, seguito dal reddito familiare e dall'età. Come si vede nella Tab. 4.4 la capacità predittiva dell'istruzione

[6] La decisione di togliere il genere dal modello di regressione multipla è stata presa anche perché, avendola inserita in alcune elaborazioni di prova, si è constatato che veniva comunque esclusa perché la sua già debole correlazione con l'atteggiamento sull'aborto si azzerava una volta che le altre variabili erano entrate nel modello.

è notevole (R= 0,25 e R^2 0,06) ma, una volta inserita questa variabile nel modello, l'apporto delle altre due variabili diventa marginale proprio perché gli anni di studio sono strettamente legati all'età anagrafica (gli anziani hanno studiato in media molto meno dei giovani) e al reddito (anch'esso strettamente legato al livello di istruzione). Ovviamente, la capacità esplicativa del livello di istruzione sarebbe diminuita se nel modello si fosse imposto di inserire per prima l'età anagrafica.

Tabella 4.4 - Regressione multipla stepwise – predittori dell'atteggiamento sull'aborto (Italia 2019)

	R	**Variazione di R**	**R^2**	**Variazione di R^2**
Anni di istruzione	,25	,25	,06	,06
Reddito familiare	,29	,03	,08	,02
Età	,31	,03	,10	,02
Collocazione politica	,35	,03	,12	,02
Quanto è importante Dio	,50	,16	,25	,13

Fonte: European Values Study (https://europeanvaluesstudy.eu/)

La tabella mostra poi che anche la collocazione politica contribuisce a incrementare la capacità predittiva, ma l'apporto decisivo giunge dall'inserimento nel modello dall'atteggiamento nei confronti della religione, che aggiunge una quota notevole di incremento a R che raggiunge un complessivo 0,50 che, elevato al quadrato, diventa 0,25. Quest'ultimo valore consente di affermare che i cinque predittori riescono a "spiegare" complessivamente poco più di un quarto della variabilità delle opinioni in merito alla giustificabilità dell'aborto, una quota che si può considerare non trascurabile, dal momento che i comportamenti sociali sono influenzati da molteplici fattori, che a volte si elidono a vicenda e sono spesso condizionati da "non razionalità" affettiva, per ricorrere a una categoria weberiana.

Un'estensione dell'utilizzo della regressione multipla si ha quando viene sottoposto ad analisi un più complesso modello che accolga esplicitamente l'idea che alcune variabili indipendenti dipendono a loro volta da altre variabili presenti nel modello. Per ognuna delle variabili dipendenti si calcola di conseguenza l'equazione di regressione e i coefficienti standardizzati sono inseriti nei c.d. diagrammi dei sentieri (*paths*), esplicitando in tal modo la complessa articolazione dei nessi causali.

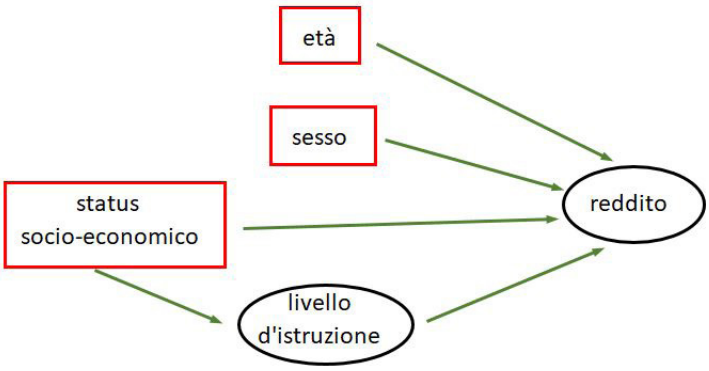
Nell'analisi dei sentieri (*path analysis*) la dimensione destra-sinistra dello spazio bidimensionale è utilizzata per rappresentare la direzione di causalità dalle

variabili antecedenti (cause) a quelle conseguenti (effetti). I singoli sentieri sono orientati secondo questa direzione generale e la varianza della variabile dipendente, spiegata da ogni variabile indipendente, viene rappresentata sotto forma di coefficiente numerico posto accanto a ognuno dei sentieri che mostrano relazioni statisticamente significative. A questa indicazione numerica più precisa si può affiancare, quale supporto visuale che aiuti ad apprezzare l'importanza relativa dei diversi sentieri, l'utilizzo di segmenti di spessore grosso modo proporzionale all'entità dei coefficienti numerici rappresentati.

Proviamo a formulare un semplice modello ipotetico che si possa esaminare con la tecnica dei sentieri causali, richiamando alcune considerazioni svolte in precedenza, quando si è accennato al tema delle associazioni tra variabili (parzialmente) spurie. Se scopriamo che il livello di istruzione è correlato con il reddito, possiamo ritenere che esso possa perciò spiegare, almeno parzialmente, i diversi livelli di reddito; detto altrimenti, se conosciamo il livello di istruzione di una persona siamo in grado di formulare una più accurata congettura riguardo al suo reddito. Possiamo però scoprire che anche altre variabili, quali il sesso, l'età o lo status socio-economico della famiglia di provenienza, possiedono una certa capacità di "predizione" del reddito. Se poi notiamo che quest'ultima variabile è correlata anche con il livello di scolarità, possiamo ritenere che vi sia una certa sovrapposizione nella capacità esplicativa delle due variabili.

Lo status socio-economico elevato della famiglia può infatti spingere i figli a ottenere alti livelli di scolarità che, in generale, sono una condizione favorevole per ottenere alti livelli di reddito. Se però un intervistato appartiene a una famiglia di status socio-economico elevato, sapere che ha un alto livello di istruzione rafforza ulteriormente la nostra convinzione che possa godere di un reddito elevato; ma non si tratta di un elemento informativo determinate, come nel caso in cui non fosse conosciuto il livello socio-economico dei genitori.

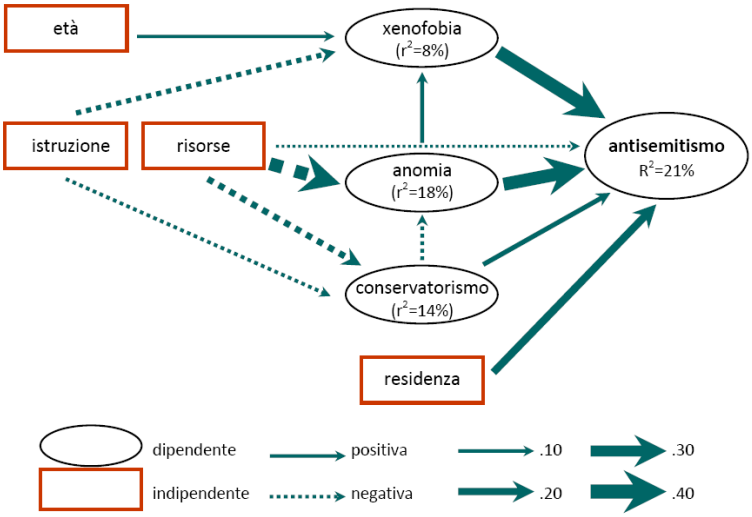
Figura 4.6 - Grafico dei sentieri causali - Modello ipotetico di spiegazione del reddito



Da quanto detto si comprende anche che la variabile "livello di istruzione" gioca un duplice ruolo: è variabile indipendente (causa) nei confronti del reddito; ma, contemporaneamente, variabile dipendente (effetto) dallo status socio-economico della famiglia. Da ciò la necessità di calcolare le diverse equazioni di regressione che ci permettono di stimare, ad esempio, l'effetto diretto dello status socio-economico della famiglia sul reddito e quello indiretto, agito attraverso il livello d'istruzione.

Nella Fig. 4.6 sono visualizzati gli ipotetici sentieri causali sopra esposti in forma verbale. La mancanza di dati adatti non permette di testare il modello in questa sede ed è solo attraverso questo test che potremmo essere in grado di appurare se le relazioni ipotizzate esistono realmente e quale sia la loro forza. Riteniamo comunque utile offrire all'attenzione del lettore un effettivo modello di *path analysis* nel quale sono esposti i percorsi causali utili a delineare una spiegazione del fenomeno dell'antisemitismo (Fig. 4.7). La ricerca di un valido esempio è stata laboriosa perché il modello deve essere parsimonioso per essere efficace: dovrebbe infatti essere in grado di riprodurre una parte significativa della varianza della variabile dipendente senza ricorrere a troppo intricati percorsi causali; d'altro canto, un modello troppo parsimonioso rischia di essere banale, perché ben difficilmente si possono descrivere i fenomeni sociali ricorrendo a un insieme di cause molto limitato.

Figura 4.7 - Grafico dei sentieri causali - Modello di spiegazione dell'antisemitismo



Fonte: Adattato da Kovacs (1999)

Il grafico è stato modificato rispetto all'originale (Kovacs 1999) per rendere il più possibile coerenti le posizioni delle variabili indipendenti con la loro distanza in

numero di passaggi necessari a raggiungere la variabile indipendente. Il grafico è stato ridisegnato anche al fine di differenziare le frecce continue (che indicano una relazione positiva) da quelle tratteggiate (relazione negativa), usando spessori delle frecce proporzionali alla forza della relazione[7]. Il ruolo delle variabili è indicato dalla loro posizione nel percorso causale, ma per renderlo più esplicito i nomi delle variabili indipendenti sono scritti all'interno di rettangoli e quelli delle variabili dipendenti all'interno di elissi. Le variabili dipendenti sono distinguibili anche perché è indicato il coefficiente di determinazione (R^2) che, come abbiamo visto, è la misura della varianza della variabile dipendente riprodotta dalle variabili indipendenti.

4.2.2. Analisi discriminante

L'analisi discriminante è una tecnica con la quale ci si prefigge di individuare un insieme, per quanto possibile limitato, di variabili indipendenti in grado di riuscire a classificare correttamente i casi come appartenenti a ognuna delle categorie di una variabile dipendente. Per certi aspetti è simile alla regressione multipla (Klecka 1976, 11), la quale ha però lo svantaggio di non poter essere utilizzata se la variabile dipendente è una classificazione a categorie non ordinate (variabile nominale). Se la classificazione ha invece le categorie ordinate si può utilizzare la regressione multipla, accettando la forzatura di considerare quasi-cardinale una variabile che in realtà è solo ordinale.

Anche nel caso dell'analisi discriminante sono disponibili diverse tecniche per l'introduzione delle variabili nel modello ed è possibile riuscire a stimare il contributo che ogni variabile indipendente riesce ad offrire alle c.d. "funzioni discriminanti", create e utilizzate per attribuire i casi alle categorie.

Un aspetto veramente interessante di questa tecnica risiede nel fatto che si possono controllare in maniera semplice e diretta le sue prestazioni e cioè la sua capacità discriminante. Dopo aver individuato le funzioni discriminanti, è possibile metterle alla prova classificando i casi sulla variabile dipendente. Siccome per la maggior parte dei casi la loro reale appartenenza alle diverse categorie è conosciuta, è possibile calcolare la percentuale di classificazioni corrette e anche, nel caso delle classificazioni erronee, se i casi non correttamente classificati si concentrino in specifiche categorie.

---

[7] Abbiamo rinunciato a riportare i valori numerici accanto ai sentieri anche perché concordiamo con Marradi che è addirittura fuorviante usare tre decimali, come nell'originale, trattandosi di dati campionari: «La tendenza a riportare percentuali e coefficienti con un numero eccessivo di decimali è endemica nell'analisi dei dati, al punto di avere un nome (*fallacy of the misplaced precision*: errore della precisione fuori posto; l'economista Morgenstern parla di *specious accuracy*: accuratezza fittizia)» (Marradi 1993: 52).

Una volta individuate le funzioni discriminanti che forniscono le migliori prestazioni, il ricercatore potrà utilizzarle per classificare casi di cui non è conosciuta la categoria di appartenenza. Il botanico potrà, ad esempio, utilizzare una serie di misurazioni di caratteristiche morfologiche o biologiche delle piante per individuare una funzione discriminante in grado di classificarle correttamente in maniera automatica. Potrà sottoporre a test queste funzioni discriminanti, verificando se siano in grado di effettuare la classificazione corretta di piante per le quali vi è un consenso degli studiosi riguardo alla loro appartenenza. In seguito, potrà utilizzare queste funzioni per tentare la classificazione di nuove specie vegetali di cui venisse in possesso.

Nel caso delle scienze sociali, un'applicazione interessante consiste nel tentativo di classificare l'appartenenza partitica di intervistati che si rifugiano spesso nel "non risposto" quando tentiamo di chiedere loro, ad esempio, per quale partito intendano votare. In questi casi non è corretto stimare l'appartenenza partitica ripartendo semplicisticamente questi individui tra i diversi partiti ricalcolando le percentuali sulla base di quanto emerge da chi ha invece risposto.

Una procedura che può dare risultati più affidabili può perciò consistere nell'individuare le variabili che sembrano essere più correlate con l'appartenenza partitica e testare la loro capacità discriminante sui casi per i quali l'appartenenza partitica è conosciuta. Trovata la soluzione più soddisfacente, si può procedere alla stima della classificazione più probabile degli intervistati che hanno preferito celare le loro preferenze politiche.

L'esempio proposto utilizza nuovamente il giudizio sull'aborto, con la scala auto-ancorante ridotta a tre categorie di risposta: "contrari", "né, né" e "favorevoli". In questo modo si potranno anche fare dei confronti tra due tecniche di analisi multivariata che hanno finalità non molto diverse: con la regressione multipla si vogliono individuare i migliori predittori della variabilità in una variabile dipendente, mentre con l'analisi discriminante gli stessi predittori sono testati per verificare se sono in grado di prevedere le risposte degli intervistati. Sempre al fine di consentire questo confronto anche le variabili utilizzate per testare la loro capacità discriminante sono le stesse viste sopra nel modello di regressione multipla. È stato anche in questo caso adottato il metodo "per passi" (*stepwise*) mediante il quale le variabili sono introdotte nel modello una alla volta, usando come criterio per l'inclusione il parametro statistico *Lambda* di Wilks, riguardo al quale accenniamo solo che è una misura del potere discriminante e cioè della capacità di separare i casi in gruppi. In pratica, a ogni passo viene introdotta nel modello la variabile che minimizza *Lambda* e il processo si arresta quando non si raggiunge il valore minimo di inserimento. Nel nostro caso (Tab. 4.5), come prevedibile, la prima variabile utilizzata è "quanto è im-

portante Dio", seguita dal reddito familiare e dalla durata in anni degli studi effettuati. Sono state invece escluse dal modello la collocazione politica e l'età anagrafica perché evidentemente non riuscivano a contribuire significativamente alla corretta classificazione dopo che nel modello erano entrate le altre tre variabili.

La percentuale generale di casi correttamente classificati supera di poco il 50%: questo è già un risultato apprezzabile, in quanto a priori, se non possedessimo alcuna altra informazione, saremmo costretti ad attribuire a caso il gruppo di appartenenza e, nell'ipotesi che avessero pari consistenza, con tre categorie la probabilità di classificare correttamente i casi sarebbe pari a solo il 33%. Interessante è poi notare che in realtà la percentuale di corrette classificazioni (evidenziate da una leggera retinatura) è più elevata nel caso delle due categorie "estreme" (53% dei favorevoli e un po' più di due terzi dei contrari). La funzione discriminante in pratica fallisce solo nel compito di classificare coloro che non si sono sbilanciati. Dal momento che le categorie della variabile sono ordinate, non sorprende però che sia più difficile classificare correttamente coloro che appartengono alla categoria intermedia, perché possiedono caratteristiche che li portano facilmente a "sconfinare" verso entrambi i due gruppi contigui.

Tabella 4.5 - Analisi discriminante - Risultati della classificazione – Atteggiamento nei confronti dell'aborto (Italia 2019)

	Gruppo di appartenenza previsto				
Gruppo originale	**Contrari**	**Nè né**	**Favorevoli**	**Totali**	
Contrari	67,7	17,8	14,5	100,0	551
Né, né	41,6	26,1	32,3	100,0	406
Favorevoli	25,6	21,3	53,1	100,0	540
Casi non raggruppati	42,1	34,2	23,7	100,0	38

Percentuale di casi "raggruppati" classificati correttamente: 51.2%

Fonte: European Values Study (https://europeanvaluesstudy.eu/)

4.3. Modelli simmetrici

I modelli simmetrici non prevedono ruoli particolari per una o più variabili perché il problema affrontato con queste tecniche di analisi è cercare tra loro delle similarità e dunque individuare gruppi di variabili il più possibile simili tra di loro

e diverse da quelle che appartengono agli altri gruppi (*analisi fattoriale*). Esistono poi modelli simmetrici di analisi multivariata che si prefiggono di effettuare un'operazione analoga, ma cercando di raggruppare casi (invece di variabili) il più possibile simili tra di loro (*analisi dei gruppi* o cluster analysis). Una tecnica di introduzione relativamente più recente è infine l'*analisi delle corrispondenze*, particolarmente utile quando la maggior parte delle variabili è categoriale ed è perciò difficile, se non impossibile, effettuare un'analisi di tipo fattoriale. Questa tecnica si propone perciò di aggregare proprio le categorie e cioè i valori delle variabili inserite nel modello. Abbiamo così tre modelli di analisi mutivariata simmetrici, utilizzabili per aggregare, sulla base di una qualche misura di vicinanza/distanza o somiglianza/dissomiglianza, ognuno degli elementi essenziali delle matrici dei dati (casi, variabili e valori).

4.3.1. Modelli che aggregano valori (analisi delle corrispondenze multiple)

L'analisi delle corrispondenze, sviluppata nell'ambito della scuola francese dell'*Analyse des données* (Benzécri *et al.* 1973), è una tecnica che è consigliabile utilizzare quando nella matrice dei dati prevalgono le variabili categoriali. Sotto questo aspetto è particolarmente preziosa perché in questa situazione il ricercatore è impossibilitato a utilizzare la maggior parte degli strumenti offerti dalla statistica multivariata. Questi, infatti, richiedono in generale un livello di misurazione almeno ordinale delle variabili (con l'eccezione, come abbiamo visto, della variabile dipendente nei modelli di analisi discriminante)[8].

Interessante è il prodotto grafico dell'analisi delle corrispondenze:

> Lo scopo principale dell'analisi delle corrispondenze è quello di rappresentare graficamente, su un numero ridotto di piani, le relazioni intercorrenti fra le modalità delle variabili analizzate: si tratta di rispondere alla necessità di una descrizione orientativa ed esplorativa dei dati di ricerca, quando ci si trovi di fronte a un grande numero di variabili qualitative e/o ad un grande numero di modalità (Mussino 1983: 69).

Per realizzare l'analisi deve essere dapprima costruita una matrice disgiuntiva completa che riporta in riga i casi e in colonna le modalità di tutte le variabili considerate. Ogni modalità corrisponde nella matrice a una nuova variabile dicotomica e in ogni cella di questa matrice a ogni caso viene attribuito un valore che è 0 oppure 1, rispettivamente per l'assenza o la presenza della modalità. La

[8] Per l'approfondimento delle tecniche dell'*Analyse des données* si vedano i lavori "storici" di Cibois (1974) e Amaturo (1988) e il più recente *Corrispondenze multiple e altre tecniche multivariate per variabili categoriali* (Di Franco (2006).

variabile "genere" produce ad esempio nella matrice disgiuntiva una variabile per la modalità "maschio" e una per la modalità "femmina". In pratica, se un intervistato è maschio, sarà registrato il valore "1" (presenza) nella prima variabile dicotomica e "0" (assenza) nella seconda.

Sulla base di questa prima matrice ne viene costruita una seconda (matrice di Burt), analizzando la quale si individuano delle dimensioni che permettono di costruire un grafico a dispersione nel quale collocare le categorie a seconda della distanza dagli assi e della loro reciproca distanza/vicinanza. «In generale, si può ritenere che più un punto è lontano dall'origine di un asse, maggiore è il contributo alla formazione dell'asse stesso, e che più elevata è la prossimità tra due variabili/modalità maggiore è l'attrazione fra loro» (Di Franco 2001: 215).

Il modello di regressione multipla, essendo basato sulle correlazioni, richiede che tutte le variabili siano cardinali e così è per l'analisi discriminante, a esclusione della variabile dipendente che deve essere categoriale; nell'analisi delle corrispondenze multiple le variabili devono essere invece tutte categoriali. Volendo confrontare tecniche di analisi che si propongono simili obiettivi conoscitivi, si potrebbero usare le stesse variabili categorizzando le cardinali, come si è fatto con la domanda sull'aborto, raggruppando le originali 10 modalità di risposta della scala auto-ancorante in tre categorie che esprimono crescente favore per l'aborto. Procedendo in questo modo, si separano però, un po' artificiosamente, posizioni che sono originariamente contigue, e dunque poco differenziate semanticamente. Si è perciò sostituita la domanda sull'importanza di Dio nella propria vita con una domanda diretta con la quale si era chiesto agli intervistati di definirsi "religiosi", "non religioso" o "atei convinti". La durata degli studi è stata sostituita dal titolo di istruzione più elevato (obbligo, diploma, titolo universitario). L'età e il reddito si potevano invece solo ridurre in classi e si è anche conservata la domanda sull'appartenenza politica (sinistra-destra), rinunciando a quella sul partito politico preferito, perché oltre il 40% degli intervistati non ha saputo indicare alcun partito o ha preferito non rispondere.

Come affermato all'inizio del paragrafo, l'analisi delle corrispondenze multiple è una tecnica esplorativa, e questa preziosa caratteristica è emersa con chiarezza lavorando all'esempio che ora esamineremo. Procedendo in modalità "*what if…*" (cosa accade se), e cioè aggiungendo o togliendo variabili per constatare come si modificava la disposizione delle categorie, è risultato infatti del tutto evidente che una tecnica basata sul calcolo delle distanze tra "oggetti" è molto sensibile all'aggiunta o sottrazione di tali oggetti. Ciò accade ovviamente quando si aggiungono o tolgono variabili, ma anche se si decide semplicemente di escludere dall'analisi una o più categorie di una

delle variabili utilizzate nel modello. Usando la variabile "condizione professionale" la configurazione del grafico può infatti cambiare, anche di molto, se si decide ad esempio di escludere dall'analisi gli inoccupati (studenti, pensionati, casalinghe, ecc.)[9].

A questo proposito, può essere utile riferire un episodio che riteniamo illuminante. Inserendo nel modello il titolo di studio e l'età (ridotta in classi) si è con sorpresa constatato che le modalità "universitario" e "16-29enne" si collocavano nel grafico a una distanza incompatibile con il fatto che i giovani sono molto più propensi degli anziani a conseguire titoli di studio elevati. Non è stato però difficile rendersi conto la maggior parte dei 16-29enni non ha l'età per potersi laureare e, di conseguenza, le due categorie non si potevano attrarre. Si è di conseguenza escluso chi aveva meno di 25 anni e, come previsto, i 25-34enni si sono immediatamente collocati accanto a "titolo universitario".

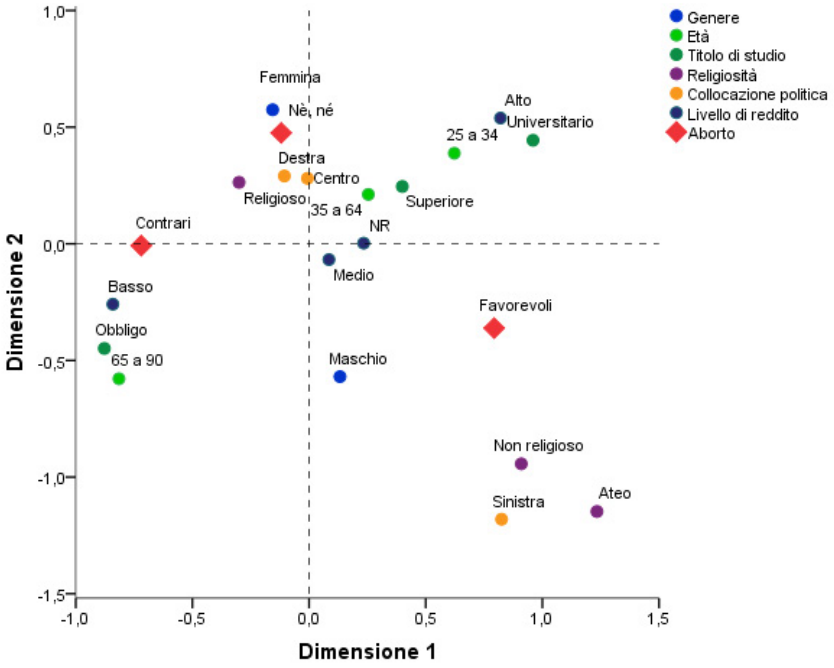
Ne è risultato il "grafico di punti delle categorie congiunte" della Fig. 4.8 del quale rinunciamo a definire semanticamente gli assi a causa dell'eterogeneità delle variabili, limitandoci a constatare che le categorie delle variabili età, reddito, titolo di studio e della domanda sulla liceità dell'aborto si dispongono sostanzialmente in orizzontale sulla Dimensione 1. La collocazione politica, il sentimento religioso e il genere appaiono invece influenzare la Dimensione 2 perché sono approssimativamente disposte in verticale.

Più interessante è constatare la contiguità spaziale di categorie che appartengono a variabili che si è constatato essere fortemente correlate come l'età, il livello d'istruzione e il reddito: i 25-34enni si collocano infatti accanto agli intervistati in possesso di un titolo universitario e con un reddito elevato. A conferma dello stretto rapporto tra le tre variabili, in basso a sinistra gli ultra-65enni sono posizionati accanto agli intervistati che non sono andati oltre l'obbligo scolastico e hanno un reddito basso. Quanto all'appartenenza di genere, mentre i maschi occupano una posizione isolata, le femmine si collocano accanto alle posizioni di centro e di destra sul piano politico e vicine a chi si dichiara "persona religiosa". I "non religiosi" e gli "atei" vanno invece a costituire un cluster a parte (in basso a destra), assieme agli intervistati che si sono definiti di sinistra.

---

[9] Anche se il riferimento è nel nostro caso estemporaneo, la tecnica di analisi *what-if* è effettivamente utilizzata per valutare in anticipo e pianificare le migliori strategie finalizzate a raggiungere determinati obiettivi. Nello specifico, è un tipo di analisi che tende a comprendere come i cambiamenti in un insieme di variabili indipendenti influiscono su un insieme di variabili dipendenti all'interno di un modello di simulazione (Rizzi 2009).

Figura 4.8 - Analisi delle corrispondenze multiple (Italia 2019)



Fonte: European Values Study (https://europeanvaluesstudy.eu/)

La collocazione degli intervistati secondo le risposta alla domanda sulla liceità dell'aborto risulta coerente con quanto è emerso dalle analisi condotte utilizzando la regressione multipla e l'analisi discriminante: la categoria "contrario" si colloca infatti sulla prima dimensione allo stesso livello dei 65-90enni e di chi ha un basso livello di istruzione e di reddito. La categoria "favorevole" è invece isolata sul piano verticale (dimensione 2), ma sul piano orizzontale (dimensione 1) si colloca alla stessa altezza di giovani, laureati, di chi ha un reddito alto, è politicamente di sinistra e non religioso o ateo. Al centro del grafico si nota infine un cluster di categorie caratterizzate da "medietà" un po' su tutti gli indicatori: sono intervistati "maturi" (35-64enni), diplomati, con un reddito medio; sono inoltre "religiosi", politicamente vicini al centro e alla destra e si collocano a metà strada riguardo all'atteggiamento sull'aborto (né decisamente contrari, né favorevoli).

Per fornire un altro esempio delle potenzialità "esplorative" di questa tecnica, si può richiamare quella che Fabbris (1990: 324) definisce la funzione di "valutazione critica dei dati": l'accostamento "spaziale" delle categorie (modalità) può infatti aiutare a individuare le risposte validamente espresse che si trovano

più vicino alle risposte "altro", "non so", "indifferente", e simili, spesso utilizzate come "scappatoia" dagli intervistati quando vengono loro poste domande imbarazzanti e non se la sentono di dichiarare che non intendono rispondere. Abbiamo provato a sperimentare questa potenzialità con la variabile "reddito" riguardo alla quale si è riscontrato, come era prevedibile, il maggior numero di mancate risposte. Come si vede dalla collocazione della categoria "NR", i "reticenti" riguardo al reddito tendenzialmente si collocano spazialmente accanto a chi è diplomato, 35-64enne, di centro-destra e ha un reddito medio (che diventa la migliore congettura sul presumibile reddito di chi il reddito non lo vuole rivelare).

Come si è sostenuto convintamente in altra sede (Delli Zotti 2010), riteniamo che questa sintetica descrizione della tecnica dell'analisi delle corrispondenze multiple illustri in modo evidente che le tecniche grafiche, oltre che strumenti adatti a "rappresentare" i dati, sono anche potenti tecniche di analisi. Pur godendo di una relativa autonomia, possono fruttuosamente rimandare ad altre tecniche statistiche, in un rapporto di complementarità, specialmente quando il loro utilizzo multivariato aveva finalità esplorative. Preso atto delle complesse interazioni tra le diverse modalità delle variabili originarie, può essere perciò utile "regredire" a tecniche più semplici, per verificare in cosa "concretamente" si traducono questi nessi in termini, ad esempio, di differenze medie o percentuali e cioè in "metriche" di più agevole comprensione e più facilmente utilizzabili per la "disseminazione" dei risultati dell'analisi.

Nella Tab. 4.6, dove è riportata la media della variabile indipendente per ogni livello che misura la "giustificabilità" dell'aborto, si vede che l'età scende di quasi 10 anni passando da chi ritiene che l'aborto non sia mai giustificato a chi ritiene lo sia sempre. Guardando agli anni di istruzione si nota una differenza di quasi tre anni e anche il reddito sale in modo pressoché monotonico, passando da chi è contrario e ha un reddito sotto alla media (tra il terzo e il quarto decile), a chi è favorevole e un reddito sopra alla media (tra il quinto e il sesto decile)[10]. Mentre la collocazione politica passa da 6,1 (moderatamente a destra) a 4,6 (moderatamente a sinistra), con una differenza dunque non molto rilevante, la risposta sull'importanza di Dio nella propria vita mostra un cambiamento molto rilevante, con una diminuzione di quasi 4 punti in una scala dove ricordiamo che 1 significa "per nulla" e 10 "molto" importante.

[10] Nel database dal quale si sono estratte le sole interviste effettuate in Italia, per consentire la comparazione dei redditi di chi risiede in paesi economicamente e monetariamente molto diversi si è usata la metrica in decili: dal primo decile, che comprende il 10% di redditi più bassi, fino al decimo, che comprende il 10% di redditi più alti.

Tabella 4.6 - Scomposizione della media – Giustificazione dell'aborto per una serie di variabili indipendenti (Italia 2019)

	Mai	2	3	4	5	6	7	8	9	Sempre
Età	58,4	55,0	55,8	55,1	55,2	53,5	52,1	51,5	53,5	49,8
Anni di istruzione	17,1	17,2	17,6	18,5	18,3	18,2	18,5	19,3	19,6	20,9
Reddito familiare	3,7	4,4	4,1	4,2	4,6	4,7	4,8	5,3	6,0	5,4
Collocazione politica	6,1	6,0	6,4	6,2	5,6	5,6	5,4	5,3	5,3	4,6
Quanto importante Dio	8,3	7,6	7,8	7,0	6,8	6,4	6,7	5,8	5,5	4,6

Fonte: European Values Study (https://europeanvaluesstudy.eu/)

4.3.2. Modelli che aggregano variabili (analisi fattoriale)

L'analisi fattoriale, una delle tecniche multivariate più conosciute, è fondamentalmente uno strumento per la riduzione della complessità della matrice dei dati. Nell'efficace descrizione dell'Enciclopedia Treccani, l'analisi fattoriale è un:

> insieme di tecniche che hanno l'obiettivo di descrivere e analizzare i legami di interdipendenza e/o dipendenza tra variabili statistiche osservate in funzione di un numero limitato di nuove variabili, più aggregate, dette fattori; tali tecniche conducono a costruire un modello che riproduca la struttura della covarianza tra le variabili considerate. Attraverso l'analisi fattoriale è possibile ridurre il numero delle variabili in esame, estraendo da un gruppo di variabili correlate fra loro un numero inferiore di fattori ortogonali che rappresentano le correlazioni esistenti tra le variabili di partenza (https://www.treccani.it/enciclopedia/ricerca/analisi-fattoriale/).

Come precedente "storico" di applicazione di questa tecnica viene ricordato il lavoro di Spearman, il quale riteneva che i risultati dei test di abilità utilizzati per il reclutamento o la selezione del personale, o a fini di orientamento scolastico, potessero misurare l'intelligenza, "identificata" con il cosiddetto "fattore g" corrispondente a una "capacità generale", correlato a specifici "fattori s" quali la memoria o altre abilità fisiche. I test risultano infatti molto correlati tra loro, ma non in maniera perfetta: dunque, accanto alla c.d. "comunalità" (varianza spiegata dall'insieme delle variabili), vi è la "specificità" di ogni variabile (o test), irriducibile a quella delle altre. Il modello di Spearman è stato comunque criticato (Gubert 1987), perché è più realistico ritenere che non basti un unico fattore (intelligenza), ma si debbano individuare più dimensioni (e cioè fattori) per rappresentare la matrice di correlazione con maggiore

aderenza alla realtà (ad esempio un fattore di abilità verbale e uno di abilità matematica).

Di questa tecnica esistono due "varianti" fondamentali, che rispondono a specifiche esigenze di ricerca. Quando viene effettuata un'*analisi delle componenti principali* (combinazioni lineari tra loro indipendenti di variabili osservate) non si pretende di "trascendere" la realtà empirica. Le componenti sono infatti delle variabili dipendenti e l'esattezza statistica di una componente consiste nel fatto che le "n" variabili empiriche riproducono la totalità della varianza della componente. Lo scopo di questo tipo di analisi è trovare un numero limitato di nuove variabili (le componenti) che siano in grado di riassumere il contenuto informativo della porzione di matrice dei dati sottoposta a questo tipo di analisi. Le componenti così individuate potranno così sostituire le variabili empiriche in successive analisi.

Con l'*analisi dei fattori* in qualche modo si trascende invece la realtà empirica, assumendo che i fattori estratti dalla procedura di analisi rappresentino delle variabili latenti (sottostanti alle variabili empiriche), che sarebbero indipendenti in quanto "causerebbero" le variabili osservate. Nel caso dei test di abilità si assume, appunto, che i risultati ottenuti dagli individui esaminati dipendano da uno, o più, fattori sottostanti che non siamo in grado di misurare direttamente[11].

L'esempio che segue è stato realizzato mediante la tecnica dell'analisi delle componenti principali e la matrice della Tab. 4.7 si riferisce alle componenti ruotate. La tecnica di analisi estrae infatti dall'insieme delle variabili componenti che sono indipendenti le une dalle altre; ma ciò si traduce nello svantaggio che quasi tutte le variabili tendono a essere correlate (si dice in questo caso "saturate") alla prima componente, che in tal modo diventa difficile da definire. Si preferisce allora rinunciare al requisito dell'indipendenza tra le componenti e si ruotano gli assi per trovare una soluzione di compromesso che riesca a distribuire le variabili in maniera più equilibrata. Con le variabili così distribuite è in genere più facile individuare ciò che accomuna le variabili presenti nelle componenti, anche se alcune variabili possono essere presenti in più di una componente. Nel definire ogni componente si terrà naturalmente conto delle variabili più saturate con essa e si trascureranno, oltre a quelle con bassa saturazione, le variabili il cui ruolo può essere considerato ambiguo perché presenti su più di una componente (anche se di norma su una hanno una saturazione più elevata).

Va anche ricordato che il ricercatore deve scegliere tra l'esclusione dei casi mancanti *listwise* (per lista) o *pairwise* (a coppie). Con la prima soluzione si utilizzano solo i casi che hanno valori validi per tutte le variabili; una soluzione che

[11] Per un confronto puntuale tra queste due tecniche e quella dello *scaling* multidimensionale (non trattata in questa sede) si veda Tommaselli (1993).

può comportare una drastica diminuzione della numerosità del campione, con conseguente significativa riduzione della sua rappresentatività statistica. Si può in alternativa preferire l'esclusione "a coppie", e cioè utilizzare di volta in volta i casi che hanno dati validi su entrambe le variabili di cui si è calcolata la correlazione. Questo metodo, "risparmioso" dei casi, comporta lo svantaggio di ottenere correlazioni basate su insiemi di dati con diversa numerosità e perciò diversa rappresentatività statistica.

L'esempio che proponiamo utilizza le risposte a una serie di 15 domande sulla liceità di comportamenti, tra le quali quella sull'aborto utilizzata fin qui per le esemplificazioni sulle tecniche di analisi multivariata. Nel nostro caso le mancate risposte alle singole domande non erano molto numerose (da una decina a meno del 5%), ma l'esclusione *listwise* avrebbe comportato la riduzione del campione di quasi il 25% perché erano disseminate "casualmente" su un numero elevato di domande/variabili. Si è perciò optato per l'esclusione *pairwise*, anche perché il numero di casi utilizzati per calcolare le correlazioni non variava molto da coppia a coppia di variabili.

Tabella 4.7 - Analisi fattoriale - Componenti principali con rotazione "varimax" delle domande sui comportamenti sociali (soluzioni a tre e a due componenti) (Italia 2019)

	Componente			**Componente**	
	1	**2**	**3**	**1**	**2**
Divorzio	,821		,212	,824	
Aborto	,760		,337	,812	
Eutanasia	,757		,268	,783	
Omosessualità	,742		,195	,737	-,131
Inseminazione o fertilizzazione artificiale	,714		,121	,671	
Evadere le tasse		,754			,703
Ottenere benefici senza diritto		,742			,708
Accettare una mazzetta		,702	,233		,744
Non pagare il biglietto		,628	,367	,121	,712
Violenza politica	-,120	,599	,298		,672
Pena di morte	,325	,565	-,208	,141	,419
Prostituzione	,184	,187	,723	,528	,375
Sesso con partner occasionali	,342		,683	,645	,252
Suicidio	,346		,609	,610	,222
Usare marijuana o hashish	,320		,599	,583	,225

Fonte: European Values Study (https://europeanvaluesstudy.eu/)

Per rendere la Tab. 4.7 più facilmente consultabile, le 15 variabili sono ordinate secondo la forza della loro saturazione con le tre componenti estratte, omettendo i valori inferiori a 0,10. Il lettore stesso può esercitare la sua fantasia sociologica per individuare che cosa accomuni le variabili presenti in ognuno dei tre fattori; ma si può comunque notare, ad esempio, che tutte le variabili saturate sulla prima componente si riferiscono a comportamenti in genere condannati dalla morale religiosa, mentre le variabili saturate sul secondo fattore si riferiscono a comportamenti attinenti principalmente alla sfera economica e politica. Sulla terza componente sono saturati comportamenti un po' eterogenei, condannati in genere dalla morale religiosa e dall'etica laica, che per la maggior parte mostrano valori di saturazione non trascurabile anche sulla prima componente. Dal momento che il software consente anche questa opzione, si è forzata la richiesta di estrarre solo due componenti e, come prevedibile, tutti i comportamenti inizialmente saturati sulla terza componente sono confluiti nella prima.

Una volta individuate le componenti, come si vedrà più analiticamente nel capitolo 7 utilizzando le variabili più significativamente correlate con ognuna di esse si potranno costruire degli indici (o macro-variabili) che in alcune analisi potranno efficacemente sostituire le variabili originarie, con un bel risparmio in termini di impiego di variabili e di contributo alla costruzione di concetti sociologici più ricchi di significato.

Come anticipato, si può realizzare anche un'analisi fattoriale confermativa e ciò avviene quando, ad esempio, vengono individuate una serie di variabili che dovrebbero, nell'intendimento del ricercatore, essere utilizzate come indicatori di un concetto più generale. L'analisi fattoriale dovrebbe in questo caso aiutare a selezionare le variabili che si potrebbero considerare indicatori validi, se risultassero accomunate in un unico fattore. Si potranno così scartare le variabili che risultano ambigue, perché saturate su più fattori o correlate con altre componenti o fattori, e perciò da considerarsi estranee rispetto al concetto che si vuole misurare, anche se utili a misurare qualche altra dimensione concettuale.

4.3.3. Modelli che aggregano casi (analisi dei gruppi)

Un'altra famiglia di tecniche multivariate ha lo scopo di creare raggruppamenti di casi che abbiano, rispetto ad alcune variabili selezionate, il massimo di somiglianza tra loro. Si tratta di tecniche di analisi dei gruppi (*cluster analysis*) che, oltre a fornire informazioni sulla bontà delle stime di somiglianza/dissomiglianza e sulla separazione dei gruppi, permettono di visualizzare le relazioni tra i casi e i gruppi che vengono formati. Anche la *cluster analysis* si avvale infatti di una rappresentazione grafica (denominata "dendrogramma", a causa della sua struttura

ad albero) per illustrare il prodotto finale dell'analisi e cioè la convergenza dei casi in aggregazioni che progressivamente includono sempre più casi e diventano di conseguenza sempre meno numerose.

Come detto, la tecnica è solitamente utilizzata per trovare aggregazioni sulla base della similarità dei casi, ma è possibile applicarla anche allo studio della similarità tra le variabili. Sarebbe possibile e interessante confrontare ad esempio questa tecnica di raggruppamento con la procedura di estrazione delle componenti principali, che in fondo ha la stessa finalità. Si giungerebbe molto probabilmente a risultati sostanzialmente simili, ma l'analisi dei gruppi applicata alle variabili aggiungerebbe la possibilità di visualizzare tutti i passaggi che hanno portato al raggruppamento, come si vedrà analizzando il dendrogramma che verrà proposto nella nostra esemplificazione.

Va notata, inoltre, la differenza tra analisi dei gruppi e analisi discriminante. L'analisi discriminante, come si è visto, richiede la preventiva conoscenza dell'appartenenza dei casi, perché è su questi gruppi (preesistenti) che viene testata la capacità predittiva delle funzioni discriminanti che successivamente si potranno utilizzare anche per classificare i casi la cui appartenenza è sconosciuta. Nella *cluster analysis* i gruppi invece non esistono a priori e non si conosce nemmeno il loro numero. Sono infatti generati dalla procedura di aggregazione che ha lo scopo di ottenere gruppi omogenei utilizzando caratteristiche dei casi (e cioè variabili) scelte dal ricercatore. Il ricercatore dovrà poi decidere anche come misurare la distanza tra i casi e anche quello da adottare per formare i gruppi, tra i diversi criteri disponibili[12]. Tutti questi "margini di manovra" concessi al ricercatore qualificano questa tecnica, che non compie alcuna assunzione "a priori", come strumento induttivo (che parte dai dati) che ha un ruolo esplorativo di ricerca di "strutture latenti" (che sono sottostanti ai dati).

Riguardo agli "oggetti" da raggruppare è fondamentale distinguere due situazioni alquanto diverse tra loro. Quando i casi possiedono un'identità chiaramente riconoscibile, è relativamente facile interpretare il significato delle aggregazioni visualizzate dal dendrogramma. Se i casi sono invece "anonimi", come nei sondaggi, si dovrà cercare di comprendere cosa "tiene assieme" chi è stato assegnato a ogni gruppo e che caratteristiche distintive possiedono i gruppi di intervistati formati dalla procedura di raggruppamento (*clustering*). A tal fine, si potranno ad esempio confrontare i gruppi sulla base dei valori medi calcolati sulle variabili utilizzate per la costruzione dei gruppi stessi. Presa confidenza con queste loro caratteristiche distintive, si potrà poi verificare se i gruppi "funzionano", se cioè il loro comportamento è tendenzialmente omo-

[12] Per una completa illustrazione delle opzioni disponibili si veda, ad es., Biorcio (1993).

geneo all'interno, e differenziato rispetto agli altri gruppi, in relazione a altri atteggiamenti o comportamenti (in pratica, si potranno utilizzare questi nuovi gruppi allo stesso modo dei gruppi definiti dal genere, dal titolo di studio, dalla classe d'età, ecc.).

La distinzione tra casi noti e casi anonimi non è poi così netta come potrebbe apparire a prima vista se si considera che intervistati anonimi possono essere chiamati a esprimersi su casi noti. Per comprendere di cosa si tratta può essere sufficiente ricordare che in un precedente volume introduttivo alla ricerca sociale (Delli Zotti 2004) proprio questo tipo di base empirica è stato usato per illustrare la *cluster analysis*. Si trattava di un'indagine sugli atteggiamenti di 237 studenti dell'ultima classe di alcuni istituti superiori del Friuli-Venezia Giulia interrogati, tra l'altro, sulla loro percezione delle caratteristiche dei paesi che facevano parte all'epoca dell'Europa comunitaria (Delli Zotti e Pellizzoni 1996). Il dendrogramma rappresentava in quel caso il risultato di un'analisi condotta su una matrice dei dati "derivata" dall'originale, nella quale i casi erano i 12 paesi valutati e le variabili i punteggi medi ottenuti da quei paesi sulle domande che riguardavano le loro caratteristiche economiche, politiche e culturali.

Questo passaggio da una matrice dei dati a base individuale (tipica dei sondaggi) a una matrice che usa "oggetti" valutati dagli intervistati è certamente un po' inconsueto, ma generalizzabile a situazioni simili: si può chiedere infatti agli intervistati di valutare istituzioni, partiti politici, oppure, in una ricerca di marketing, una serie di prodotti e finalizzare la *cluster analysis* all'individuazione di gruppi di prodotti similmente percepiti dai consumatori. Più consueto è l'utilizzo di dati statistici raccolti su base territoriale (comuni, provincie, regioni o nazioni). Sono casi riconoscibili individualmente sui quali l'aggregazione in gruppi è efficace perché innesca, al momento dell'interpretazione dei risultati dell'aggregazione, una serie di conoscenza pregresse presenti nella mente del ricercatore e poi dei fruitori del lavoro di ricerca.

Punto fondamentale è comunque la scelta delle variabili da utilizzare per la classificazione delle unità osservate; una scelta soggettiva, se rispecchia solo le convinzioni e le idee del ricercatore, che ulteriormente qualifica questa tecnica come "esplorativa". Al fine di evitare la soggettività, in questa sede si è deciso di utilizzare un database nel quale le variabili sono già state scelte con una finalità chiaramente esplicitata e finalizzata a rilevare una caratteristica ovviamente auspicabile: il grado di felicità percepita dai cittadini dei diversi paesi nei quali la rilevazione viene effettuata.

Si tratta dell'iniziativa *World Happiness Report*, nata nel 2012, una pubblicazione del *Sustainable Development Solutions Network*, sostenuta principalmente dal *Gallup World Poll* e anche dall'italiana *Ernesto Illy Foundation*. Le variabili scelte

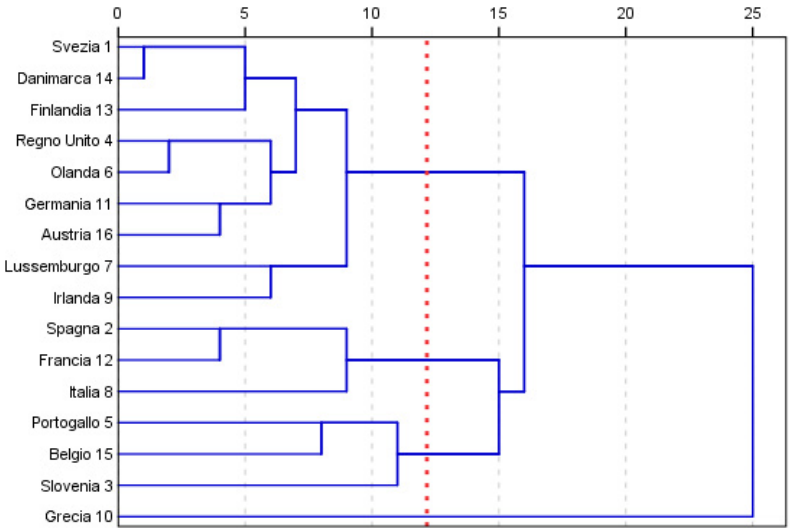
per classificare i paesi sono le seguenti: Prodotto Nazionale Lordo pro capite in termini di capacità di acquisto, aspettativa di vita in salute alla nascita, sostegno sociale (avere qualcuno su cui contare se in difficoltà), libertà nelle scelte di vita, generosità (donazioni effettuate a fini caritatevoli) e percezione della pervasività della corruzione nella vita pubblica.

Come si vede, la maggior parte delle variabili deriva dagli esiti di sondaggi effettuati sulla popolazione ed è dunque evidente che per fare questo tipo di analisi si possono usare, oltre ai consueti dati statistici reperibili sui siti dell'Istat o dell'Ocse, quelli che le principali fonti internazionali di dati demoscopici (*World Values Survey*, *European Social Survey*, *Eurobarometer*, ecc.) rendono disponibili per l'analisi secondaria (da noi ampiamente utilizzati per le esemplificazioni).

Prima di passare all'esame del dendrogramma (Fig. 4.9), è opportuno precisare che le variabili da utilizzare per la creazione dei gruppi debbono essere espresse nella stessa unità di misura. Se sono espresse in unità di misura diverse o hanno ordini di grandezza diversi sarebbe utile standardizzare le variabili, una procedura alla quale può provvedere il software, ma solo se il ricercatore si ricorda di attivare l'opzione.

Il dendrogramma mostra, da sinistra verso destra, quali gruppi si formano, partendo dai primi accoppiamenti tra i paesi più simili tra di loro, fino a formare gruppi più ampi, ma sempre più eterogenei, specialmente se la "distanza di fusione" aumenta notevolmente prima di arrivare all'aggregazione successiva. Dal grafico si nota dunque che al primo stadio si associano Svezia e Danimarca e, a breve distanza, Olanda e Regno Unito che sono dunque le coppie di paesi più simili tra loro, confrontando il profilo sui sei indicatori. Al passo successivo si crea la coppia Austria-Germania che, dopo l'aggregazione della Finlandia a Svezia e Danimarca, si uniscono alla coppia Olanda-Regno Unito. Alla stessa distanza di fusione di queste ultime aggregazioni si forma la coppia Irlanda-Lussemburgo che a pochissima distanza va alla fine a costituire il gruppo dei paesi del Centro nord europeo tra i quali manca all'appello solo il Belgio. Si infrange dunque quel Benelux che è stato protagonista dei primi passi della costruzione europea. Come si vede dalla parte inferiore del grafico, i paesi che finora non abbiamo citato si aggregano, ma a distanze di fusione un po' più elevate. Le prime a unirsi in questa area sono Francia e Spagna, alle quali si si aggiunge, un po' più avanti l'Italia. Subito sotto si forma un gruppo (Portogallo, Belgio e Slovenia) che appare alquanto eterogeneo e, infatti, l'aggregazione avviene a una considerevole distanza dall'origine. Da ultima, la Grecia costituisce un gruppo a sé, nel senso che non si unisce a nessun altro paese o gruppo di paesi fino alla fine, quando tutti i paesi, non importa quanto eterogenei, entrano a far parte di un unico gruppo.

Figura 4.9 - Analisi di gruppi di casi



Fonte: World Happiness Report 2020 (https://worldhappiness.report/ed/2020/)

Conclusivamente, come si vede nel grafico è stata tracciata una riga verticale tratteggiata la quale interseca le linee orizzontali che fanno riferimento ai gruppi che sono stati costituiti prima di arrivare a quella distanza di fusione. La decisione su dove posizionare questo discrimine è ovviamente soggettiva, ma può essere consigliata dalla considerazione che andare oltre significherebbe ridurre ulteriormente il numero di gruppi, ma al prezzo di aumentare forse un po' troppo la distanza di fusione, con la conseguenza di mettere assieme casi che rischiano di essere troppo diversi.

Il lettore attento avrà notato la presenza della Slovenia, un paese che effettivamente non faceva parte dell'Europa dei 15. Non si tratta di un errore, ma di un paese aggiunto con il proposito di mostrare le potenzialità di questa tecnica come strumento utile a effettuare delle simulazioni. In questo caso la simulazione è stata fatta retrospettivamente ma, oltre alla valutazione delle conseguenze dell'aggiunta di nuovi paesi a quelli che già fanno parte del "club europeo", si possono immaginare altre situazioni nelle quali può essere utile valutare le caratteristiche di un nuovo "oggetto" in relazione a quelle di altri dei quali sono note le differenze/somiglianze.

5. Costruire e modificare variabili

Come anticipato nell'introduzione, con il presente capitolo inizia una seconda parte del volume dedicata ad alcuni temi specifici che si è deciso di trattare a parte perché non trovano una precisa collocazione all'interno della sequenza che, a partire dall'analisi monovariata si conclude con l'illustrazione dei più complessi modelli di analisi multivariata. L'esigenza di intervenire sulle variabili con le modifiche illustrate in questo e nel capitolo seguente, dedicato alla categorizzazione delle variabili cardinali, può infatti emergere in qualsiasi fase del processo di analisi dei dati. Lo stesso si può affermare per la costruzione di indici e tipologie (trattata nel capitolo 8), che si avvale del supporto di analisi bivariate, o anche multivariate, le quali devono essere conosciute preventivamente dal lettore per poter essere trattate in modo sufficientemente dettagliato. Analogamente, il capitolo di approfondimento sul trattamento dei dati mancanti può trovare diverse declinazioni a seconda che ci si trovi in un contesto di analisi monovariata, bivariata o multivariata. Infine, il capitolo dedicato alle tabelle di contingenza multiple si poteva assegnare alla fase dell'analisi bivariata, ma la complessità del tema ha consigliato di trattarlo in un capitolo a parte.

Le tecniche di trasformazione delle variabili descritte in questo capitolo presuppongono che i dati raccolti mediante i diversi strumenti di rilevazione siano stati organizzati in una matrice "casi per variabili" (Delli Zotti 1985) che può essere realizzata utilizzando la finestra "VISUALIZZAZIONE DATI"[1] di Spss (di Pspp o di altri simili programmi). In alternativa, per "caricare" i dati si può usare un foglio elettronico (come Excel), che Spss è in grado di leggere direttamente, avendo cura di inserire nella prima riga le intestazioni delle colonne che Spss assumerà come nomi delle variabili. La matrice può anche essere creata da programmi di sondaggi online (Google Forms, Survey Monkey, ecc.) che caricano automaticamente i nuovi record in una matrice, non appena un rispondente ha completato la (auto)compilazione del questionario. Qualunque sia lo strumento utilizzato per la costruzione della matrice dei dati, per passare dai "dati grezzi" alle tabelle di contingenza o ai parametri statistici sintetici che il ricercatore dovrà esaminare per dare risposta ai propri bisogni conoscitivi si devono effettuare passi successivi utilizzando comandi e funzioni che servono a:

- *definire i dati*: aggiungere etichette ai nomi sintetici delle variabili e ai valori, indicare i codici che contrassegnano i dati mancanti, definire il tipo delle variabili (numerica o stringa) e il livello di misura (nominale, ordinale e cardinale);
- *trasformare i dati*: modificare le variabili e crearne di nuove, pesare i casi, selezionare sub-campioni, ecc.;
- *attivare procedure*: realizzare analisi statistiche o rappresentazioni grafiche, salvare i dati, esportare l'output, ecc.

Nel caso il ricercatore intenda effettuare un'analisi secondaria su basi di dati esistenti, come quelle che abbiamo ampiamente utilizzato per le esemplificazioni in questo volume, avrà a disposizione matrici dei dati già perfettamente definite. In linea di massima non dovrebbe avere bisogno del primo tipo di comandi, anche se ciò non è necessariamente vero. Potrebbe infatti avere banalmente la necessità di cambiare le etichette delle variabili e dei valori, qualora non siano scritte in italiano, o di modificare lo "status" di determinati valori (da mancante a valido), qualora decidesse di inserire in una tabella di contingenza anche questi valori. Analogo discorso si può fare per i comandi di trasformazione dei dati perché le basi di dati più strutturate offrono solitamente al lettore anche variabili generate da procedure di trasformazione dei dati originaria-

[1] I comandi, le parole chiave e i nomi delle finestre di Spss saranno scritti in maiuscoletto, al fine di renderli più facilmente identificabili.

mente rilevati (ad esempio, oltre agli anni di età dichiarati, l'età raggruppata in classi). Anche in questo caso però chi fa analisi secondaria potrebbe avere la necessità di utilizzare i comandi di trasformazione dei dati se, ad esempio, le classi d'età fossero costruite utilizzando intervalli di valori che non rispondono alle proprie esigenze.

I comandi e le funzioni sopra elencate possono essere attivati mediante il sistema a finestre, selezionando opzioni e/o scrivendo le informazioni richieste in caselle di testo aperte automaticamente dal programma. Ci dobbiamo però chiedere se sia questo il modo più efficiente di operare e chi ha esperienza nell'uso del programma sa che è assai utile, e quasi imprescindibile, conoscere la sintassi dei comandi (che, come vedremo, non significa rinunciare all'ausilio del sistema di menù, finestre di dialogo, pulsanti, caselle di selezione, ecc.). Infatti, anche se inizialmente può risultare un po' faticoso, diversi motivi consigliano la registrazione sistematica dei comandi su un foglio di sintassi.

- *Velocizzare le operazioni*: per quanto possa sembrare strano al lettore meno smaliziato, l'uso intensivo e spesso inappropriato del mouse comporta (anche) un certo degrado dell'efficienza. Ad esempio, abbandonare la tastiera mentre si scrive un testo, afferrare il mouse e selezionare la piccola icona che nella barra degli strumenti indica il corsivo, comporta maggiore perdita di tempo rispetto all'utilizzo della combinazione di tasti "CTRL + I". Allo stesso modo, è più semplice scrivere una lista di etichette di variabili o di valori in un foglio di sintassi, adattando il questionario spesso già disponibile in formato digitale, piuttosto che inserire le etichette una ad una usando menu e finestre di dialogo.
- *Documentazione interna*: le trasformazioni di variabili implicano decisioni che si possono dimenticare e non sempre sono desumibili dal lavoro finito. Se, ad esempio, non trascriviamo su un foglio di sintassi la formula usata per costruire un indice, ben presto rischiamo di non ricordarci più quali siano le variabili usate e i pesi eventualmente attribuiti. Analogamente, possiamo ridurre i valori di una variabile cardinale alle modalità "basso/medio/alto", ma in seguito potremmo non ricordare quali siano i valori usati per suddividere la distribuzione (specialmente se si tratta di punteggi di qualche astratto indice additivo).
- *Rendicontazione esterna*: documentare le convenzioni e decisioni adottate è anche un'esigenza di trasparenza e rendicontazione che può essere richiesta (come si è accennato nel capitolo precedente) per verificare e/o replicare il nostro lavoro. Si tratta ovviamente di un'esigenza imprescindibile nel caso delle basi di dati che abbiamo appena evocato che non a caso sono accompagnate da una impressionante mole di "meta-dati" e cioè di docu-

mentazione sulle caratteristiche delle variabili, dalle modalità della loro rilevazione (definizione operativa) alle modalità della loro costruzione, nel caso si tratti di nuove variabili generate da procedure di trasformazione o combinatorie dei dati originali.

- *Ripetizione delle analisi*: molto spesso il ricercatore vuole applicare le "ricette" sperimentate a nuovi file di dati o ripetere analisi e/o trasformazioni di dati su nuovi casi aggiunti in matrice. Dopo avere operato complesse selezioni di casi e/o trasformazioni dei dati, ci si potrebbe rendere conto, ad esempio, che sarebbe opportuno escludere i questionari compilati da persone troppo giovani o troppo anziane, oppure che si devono aggiungere questionari arrivati in ritardo o, infine, che è necessario ripetere le stesse analisi separatamente per gli intervistati di ognuna delle province di una regione. In questi casi, se la sequenza di procedure è registrata in un file di sintassi, replicare l'analisi su un sottoinsieme di intervistati o sulla matrice dei dati modificata diventa questione di pochi secondi.
- *Opportunità in più*: non tutte le opzioni di Spss sono disponibili mediante menu e finestre di dialogo. Ad esempio, per le correlazioni il programma presenta una sola finestra nella quale collocare le variabili e pertanto, anche volendo correlare una sola variabile dipendente con nove indipendenti, il programma produrrà una matrice di correlazione di dimensione 10x10, anche se servono solo nove coefficienti. Con la sintassi è invece possibile specificare cosa esattamente vogliamo inserendo la parola chiave WITH tra la variabile dipendente e le indipendenti (indip1 WITH dip1 TO dip9). Un altro notevole esempio è costituito dal fatto che i gruppi di variabili creati per tabulare le risposte multiple (cfr. capitolo 8), nella versione base di Spss non vengono salvati nella matrice dei dati e perciò devono essere reimpostati ogni volta che si accede al programma. Si può ovviare se si possiede il modulo opzionale TABELLE PERSONALIZZATE ma, ancor più semplicemente, salvando le istruzioni che servono a costruire le variabili multiple in un foglio di sintassi.

Come anticipato e si esemplificherà in seguito, per utilizzare i comandi non è necessario conoscere approfonditamente le regole della sintassi: si può infatti realizzare un'istruzione "tipo", usando menù e sotto-menù, e con il comando INCOLLA scriverla sul foglio di sintassi; in tal modo si ottiene un "modello-base" che si potrà in seguito adattare a nuove esigenze cambiando, ad esempio, solo i nomi delle variabili o i valori.

A proposito di sintassi e delle sue regole, si deve innanzitutto precisare che in Spss i nomi delle variabili devono iniziare con una lettera e possono essere composti da più parole, purché unite da un trattino basso; è però consigliabile usare sigle o nomi brevi e avvalersi di etichette descrittive per illustrare il contenuto. All'interno del nome si possono usare anche dei numeri, ma non si possono invece usare, oltre agli spazi, caratteri speciali (! ? ' *). Non si possono infine usare come nomi di variabili le parole, elencate nel seguito del paragrafo, che identificano operatori logici, relazionali e parole riservate a usi specifici.

Queste prescrizioni esemplificano un aspetto che va tenuto presente costantemente: per usare correttamente la sintassi si devono rispettare regole e ricordare parole chiave. Peraltro, entrambe non sono molto numerose e i nomi sono sufficientemente evocativi della funzione. Inoltre, come si è accennato e si esemplificherà in seguito, è possibile "ripassare la sintassi" costruendo il comando con finestre e pulsanti. Nonostante la facilità d'uso, è comunque opportuno illustrare sinteticamente i comandi principali, perché solo conoscendo le loro caratteristiche e funzioni si potranno usare in modo corretto e, superando la semplice imitazione, anche in modo creativo.

All'interno delle istruzioni di Spss si usano *operatori relazionali* che si possono anche combinare mediante *operatori logici*. I nomi di questi operatori possono spesso essere sostituiti da simboli (indicati nella seconda colonna della Tab. 5.1) che troviamo sulla tastiera del computer.

Tabella 5.1 - Operatori relazionali e operatori logici

Relazionali	EQ	=	Uguale a
	NE	<>	Non uguale a
	LT	<	Inferiore a
	LE	<=	Inferiore o uguale a
	GT	>	Maggiore di
	GE	>=	Maggiore o uguale a
Logici	AND	&	Entrambe le condizioni devono essere vere
	OR	|	Almeno una condizione deve essere vera
	NOT		Inverte il risultato dell'espressione

Nei comandi di definizione dei dati, di trasformazione di variabili e di attivazione di procedure si possono usare *operatori aritmetici* (+, -, /, * e ** per l'elevazione

a potenza) e altre *funzioni statistiche* o di altra natura. Si possono poi usare *parole riservate* (Tab. 5.2) di cui si scoprirà l'uso nel seguito.

Tabella 5.2 - Parole riservate

ELSE	Tutti gli altri valori
INTO	In una nuova variabile
THRU	Dal valore precedente a quello seguente
LOWEST o LO	Dal valore più basso
HIGHEST o HI	Al valore più elevato
TO	Dalla variabile precedente alla seguente
WITH	Con
BY	Per
SYSMIS	Mancante di sistema
ALL	Tutte le variabili

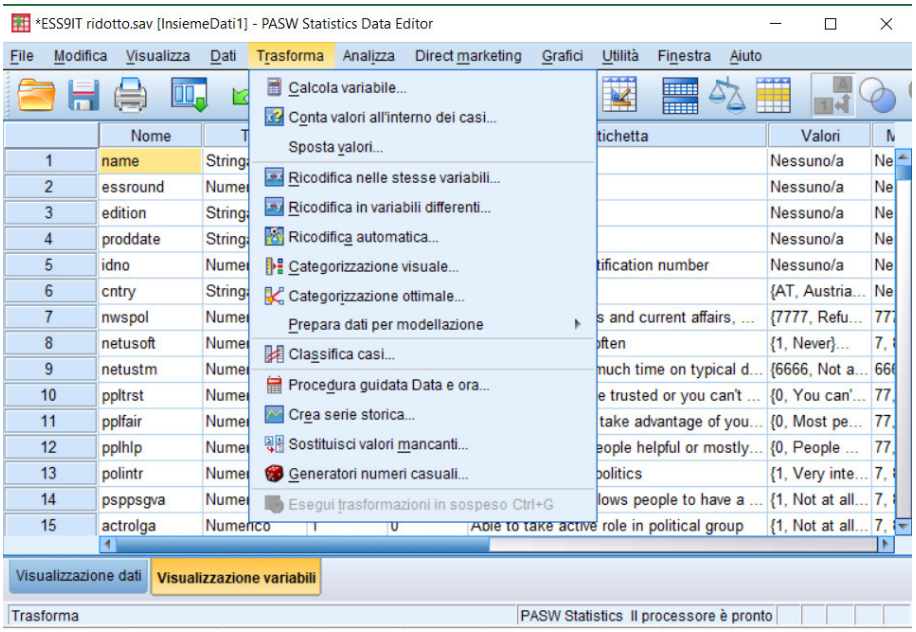
5.3. Comandi di trasformazione dei dati

Anche se i dati sono stati raccolti mediante un questionario strutturato, è spesso utile, a volte necessario, apportare modifiche o trasformazioni alle variabili. Prima di esaminare le specifiche finalità e modalità di utilizzo dei comandi di trasformazione è però opportuno richiamare nuovamente la distinzione tra tipi di variabili, perché le loro specifiche caratteristiche condizionano, ad esempio, la scelta tra l'adozione di criteri semantici o aritmetico/matematici quando si deve operare sui valori. Le variabili si distinguono dunque in:

- *nominali*: relative a proprietà i cui stati sono *discreti non ordinabili* (es. la residenza) che non consentono di utilizzare le proprietà ordinali e cardinali dei numeri;
- *ordinali*: relative a proprietà *discreti ordinabili* (es. l'istruzione) che consentono di utilizzare le proprietà ordinali, ma non quelle cardinali dei numeri;
- *cardinali*: relative a proprietà *continue misurabili* (es. l'altezza) o con *stati enumerabili* (es. il numero di figli) che consentono di utilizzare le proprietà cardinali dei numeri.

I comandi che utilizzano gli operatori e le parole riservate sopra elencate sono numerosi e rispondono a esigenze diverse che in questa sede verranno illustrate limitandoci alle funzioni principali tra quelle elencate con completezza nella finestra trasforma di Spss, riprodotta nella Fig. 5.1.

Figura 5.1 - I comandi di trasformazione in Spss



5.3.1. Il comando "compute" (calcola)

Mentre la funzione del comando COMPUTE è intuitiva, meno intuitiva è la sua sintassi: "COMPUTE variabile destinazione = espressione". Si può infatti notare l'inversione rispetto alla consueta notazione delle espressioni aritmetiche, nelle quali il risultato si scrive alla fine. Il risultato in questo caso è il valore, calcolato per ogni singolo caso, che verrà imputato nella variabile di destinazione. L'apparente incongruenza deriva dal fatto che il computer opera in modo sequenziale: solo dopo che la nuova variabile è stata creata, la si può infatti utilizzare per registrarvi gli esiti del calcolo.

Proponiamo ora qualche semplice esempio:

```
COMPUTE newvar = 0.
COMPUTE totred = redlav + redpens.
COMPUTE index = (d1 + d2 + d3 + d4) / 4.
```

Con la prima istruzione si crea la variabile "newar" e si pone il suo valore uguale a 0. L'utilità del comando non è immediatamente evidente, ma la si comprende se si considera che in questo modo alla nuova variabile viene assegnato un valore iniziale (si parla infatti di "inizializzazione della variabile") che potrà essere poi

modificato, date certe condizioni. In tal modo non è necessario esplicitare cosa fare quando queste condizioni non ricorrono, perché rimane semplicemente immutato il valore attribuito inizialmente. Con il secondo esempio si crea la variabile "redtot" sommando i redditi da lavoro (redlav) e da pensione (redpens); con il terzo si crea la variabile "index" sommando le riposte fornite alle domande da "d1" a "d4" e dividendo il risultato per 4 (in pratica, si calcola la risposta media).

Un esempio appena un po' più complesso è il calcolo dell'età (eta) a partire dall'anno di nascita (annonasc). Se si è registrato l'anno di nascita usando le sole ultime due cifre (86, invece di 1986), l'età si può calcolare facilmente sottraendo il valore della variabile "annonasc" da 120 (se si tratta di dati rilevati nel 2020).

Se si è registrato l'anno di nascita usando solo le due ultime cifre, per velocizzare la registrazione dei dati, ma poi si preferisce che l'anno di nascita sia scritto per esteso, con COMPUTE si può aggiungere 1900 al valore registrato in matrice, senza creare una nuova variabile. Per ottenere l'età, si dovrà in questo caso detrarre l'anno di nascita da 2020.

```
COMPUTE eta = 120 - annonasc.

COMPUTE annonasc = 1900 + annonasc.
COMPUTE eta = 2020 - annonasc.
```

L'ultimo esempio che proponiamo è utile a introdurre l'elemento della condizionalità nella trasformazione dei dati, non presente negli esempi precedenti. Nella Tab. 5.3 è mostrata, nella parte sinistra, la distribuzione di frequenza dei voti di diploma di un campione di studenti. Una parte dei voti è espressa in sessantesimi e si deve perciò provvedere alla loro conversione in centesimi, volendo analizzare voti espressi tutti in questa metrica. A complicare la situazione, "60" è ambiguo, perché è il voto più basso in centesimi e il più alto se in sessantesimi.

Per operare la correzione si è innanzitutto creata la nuova variabile (votocorr) e le sono stati attribuiti i valori della variabile "voto". Si è dunque inizializzata la nuova variabile attribuendo a ogni caso il valore registrato su un'altra variabile (in pratica, la si è duplicata). I valori sono poi stati divisi per 60 e moltiplicati per 100, ma solo se (IF) nella variabile "votosess" era registrato il valore 1, che indicava che quel voto era in sessantesimi. In questo modo i due 36 iniziali della Tab. 5.3 si sono trasformati in 60 che scendono a 14 nella nuova variabile perché tre dei 15 iniziali (essendo in sessantesimi) si sono a loro volta trasformati in 100 (nella nuova variabile i 100 infatti salgono a 9).

```
COMPUTE votocorr = voto.
IF (votosess = 1) votocorr = voto / 60 * 100.
```

Tabella 5.3 - Voto di diploma e voto corretto

Voto di diploma	N	%
36	2	,8
...	...	...
56	1	,4
60	15	6,3
61	6	2,5
...	...	...
99	3	1,3
100	6	2,5
Non risposto	19	7,9
Totale	239	100,0

Voto corretto	N	%
60	14	5,9
61	6	2,5
...	...	...
99	3	1,3
100	9	3,8
Non risposto	19	7,9
Totale	239	100,0

È utile infine notare che 19 studenti non hanno dichiarato il voto. Se il valore utilizzato per registrare l'assenza del dato (ad esempio "0", visto che è un voto non ammissibile in entrambe le metriche) è definito come MISSING VALUE (valore mancante), i comandi sopra descritti non avranno effetto in quei casi e il valore rimarrà dunque immutato nella nuova variabile.

5.3.2. I COMANDI "IF" (SE) E "SELECT IF" (SELEZIONA SE)

Avendo appena utilizzato il comando IF in connessione con COMPUTE con lo scopo di vincolare la sua esecuzione a una precisa condizione, è opportuno approfondire il tema, precisando innanzitutto che, come COMPUTE, anche IF può generare una nuova variabile e la sua sintassi "IF [(] espressione logica [)] variabile destinazione = espressione" ha una struttura simile: il risultato è attribuito ai casi che soddisfano l'espressione logica collocata all'inizio. Nel seguente esempio, alla variabile "status" è assegnato il valore 4 se la professione è 1.

```
IF (prof EQ 1) status = 4.
```

Dal momento che la nuova variabile non è stata preventivamente inizializzata, rimane indeterminato il livello di status per tutti gli intervistati che hanno una professione diversa da "1" finché, con istruzioni analoghe, non verranno attribuiti i livelli di status corrispondenti alle altre professioni. Se ne deduce che è sempre consigliabile inizializzare la variabile e conviene inizialmente

attribuire a tutte le professioni il livello di status condiviso dal numero più elevato di esse; in tal modo, si riduce il numero di professioni per le quali si dovrà modificare il livello di status attribuendo quello che effettivamente compete. Il numero di istruzioni da scrivere è comunque inferiore rispetto al numero di professioni alle quali si deve attribuire uno status sociale perché all'interno di una singola istruzione si possono elencare tutte le professioni che condividono lo stesso status sociale, con le modalità che illustreremo esaminando il comando SELECT IF, simile a IF per gli aspetti sintattici, ma usato per una diversa finalità. SELECT IF non serve infatti a modificare variabili, ma a selezionare casi, solitamente per effettuare specifiche analisi sui soli casi selezionati.

Per esemplificare l'utilizzo di questo comando partiamo dalla distribuzione di frequenza della Tab. 5.4 che mostra le risposte alla domanda sullo stato civile del campione di persone intervistate in Italia nell'ultima indagine dell'*European Social Survey*.

Tabella 5.4 - Stato civile (Italia 2017)

statciv Stato civile	**N**	**%**
1 Coniugato/a	1.397	50,9
2 Unito/a civilmente	29	1,1
3 Separato/a	79	2,9
4 Divorziato/a	104	3,8
5 Vedovo/a	230	8,4
6 Celibe/nubile	871	31,7
77 Rifiuto	34	1,2
88 Non so	1	,0
Totale	2.745	100,0

FONTE: EUROPEAN SOCIAL SURVEY (HTTP://WWW.EUROPEANSOCIALSURVEY.ORG)

Se vogliamo selezionare solo i celibi/nubili, oppure escluderli dalle successive analisi, possiamo usare i due seguenti comandi:

```
SELECT IF statciv EQ 6.
SELECT IF statciv NE 6.
```

I comandi non necessitano di spiegazione, se si possiede anche solo una conoscenza elementare dell'inglese e si ricorda (Tab. 5.1) che EQ significa

"uguale a" e NE "non uguale a" (o "diverso da"). È opportuno però notare che, mentre il primo comando raggiunge correttamente lo scopo di selezionare solo i celibi/nubili, con il secondo comando si selezionano tutti gli altri, ma tra questi ci sono anche intervistati che in realtà non hanno risposto alla domanda e dunque sono estranei al gruppo di intervistati che vivono o hanno vissuto un'esperienza di coppia. La fortunata circostanza che i codici che identificano la mancate risposte hanno un valore numerico superiore a quello che identifica i celibi/nubili consente di raggiungere lo scopo modificando l'istruzione nei due modi seguenti, che sono perfettamente equivalenti e si avvalgono entrambi della possibilità di selezionare valori contigui. Con la prima istruzione si chiede di selezionare tutti gli stati civili codificati con un valore inferiore a "6", con la seconda quelli codificati con un valore inferiore o uguale a 5.

```
SELECT IF statciv LT 6.
SELECT IF statciv LE 5.
```

Gli operatori relazionali si possono anche combinare, tramite l'operatore logico AND, come nell'istruzione seguente che consente di selezionare le sole femmine (codice 2 della variabile "sex") tra gli intervistati celibi/nubili:

```
SELECT IF statciv EQ 6 AND sex EQ 2.
```

Anche se l'uso degli operatori logici appare intuitivo, si possono commettere errori: volendo selezionare i 79 intervistati che all'epoca dell'intervista erano separati e i 104 divorziati, ci aspetteremmo di ottenere un insieme di 183 intervistati con il comando:

```
SELECT IF statciv EQ 2 AND statciv EQ 4.
```

Il programma emetterà invece il seguente avviso:

> Per questa procedura non sono stati specificati casi. È possibile che nel file dati di lavoro non siano presenti casi o che tutti i casi siano stati esclusi dal filtro attivo.
> Questo comando non viene eseguito.

Anche se sembra controintuitivo, il comando è corretto se si usa l'operatore logico OR invece di AND: infatti, un caso è selezionato se il risultato dell'espressione contenuta in SELECT IF è vero; dunque, se si usa AND entrambe le con-

dizioni devono essere soddisfatte e nell'esempio si pretendeva di selezionare intervistati separati e divorziati, condizione che non si può verificare perché le modalità di una classificazione sono per definizione mutualmente esclusive (non si può essere allo stesso tempo separati e divorziati). Nell'esempio precedente è invece stato possibile selezionare intervistate femmine e nubili, perché i requisiti di appartenenza all'insieme dei casi da selezionare si riferiscono a modalità di due distinte variabili (stato civile e genere).

L'errore che abbiamo esemplificato si commette se ci si dimentica che con SELECT IF si indicano le condizioni che devono essere soddisfatte e non il risultato che si vuol ottenere. Non è spontaneo usare OR perché si intende selezionare entrambe le categorie e non una *o* l'altra. Si può dunque usare distrattamente AND, dimenticando che, per selezionare entrambe le categorie, con l'istruzione si deve indicare come requisito l'essere separato *oppure* divorziato.

Per concludere, rimane solo da aggiungere che, per evitare che i casi non selezionati siano cancellati, rimanendo così disponibili per successive elaborazioni sull'intero database (o selezionare sottoinsiemi di casi sulla base di altri criteri), SELECT IF va preceduto dal comando TEMPORARY, oppure si dovrà usare il menù DATI/SELEZIONA CASI, scegliendo FILTRA CASI NON SELEZIONATI, invece di ELIMINA CASI NON SELEZIONATI.

5.3.3. Il comando "RECODE" (ricodifica)

Con il comando "RECODE lista di variabili (lista di valori = valore) ... (lista di valori = valore) [INTO lista di variabili]" si possono modificare i valori di una o più variabili. Il comando è apparentemente complesso, ma intuitivo se rileggiamo in modo discorsivo la sintassi: dopo avere indicato la variabile (o le variabili) alle quali si applicano le trasformazioni, si elenca un insieme di valori e si chiede di attribuire a tutti uno stesso valore; si prosegue poi con un'altra lista e così via, categorizzando in tal modo i valori originari.

Quella che abbiamo delineato è l'ipotesi di utilizzo del comando più complessa, ma a volte RECODE si utilizza semplicemente per accorpare modalità di risposta di significato simile. Con riferimento alla Tab. 5.4, si potrebbero ad esempio unificare le risposte "non so" e "rifiuto", dal momento che si tratta in entrambi i casi di risposte mancanti.

Con il sottocomando INTO il risultato della ricodifica dei valori viene "caricato" su una nuova variabile o una lista di nuova variabili. Le parentesi quadre segnalano che INTO è un'aggiunta opzionale; infatti, può essere superfluo creare una nuova variabile se si tratta di modifiche minime, come nel caso dell'ac-

corpamento delle risposte mancanti o se si vogliono semplicemente aggregare nella categoria residuale "altro" alcune modalità di risposta esiguamente rappresentate nel campione. Omettendo INTO si effettua invece una trasformazione permanente della variabile e non sarebbe dunque più disponibile la codifica originale. Se RECODE viene utilizzato, ad esempio, per raggruppare gli intervistati in classi l'età, con i nuovi valori (ad esempio "1" per i giovani da 18 a 35 anni) non sarà più possibile calcolare l'età media di specifiche categorie di intervistati.

Come si è accennato, è possibile usare RECODE per trasformare insiemi di variabili: con il comando che segue si trasformano ad esempio i valori da 1 a 4 (molto contrario/abbastanza contrario/abbastanza d'accordo/molto d'accordo) nei valori 0 e 1 (contrario/d'accordo), creando otto nuove variabili.

```
RECODE dom1 TO dom8 (1 2 = 0) (3 4 = 1) INTO rdom1 TO rdom8.
```

La parola riservata TO indica che si devono trasformare tutte le variabili comprese tra "dom1" e "dom8" (THRU si usa invece, come vedremo, per indicare intervalli di valori). Con TO ci si può riferire a qualsiasi intervallo di variabili consecutive, comunque denominate; se lo si usa per generare nuove variabili (come nella seconda parte del comando) i nomi devono invece avere lo stesso "prefisso" alfabetico e terminare con un suffisso numerico (nell'esempio non si poteva perciò scrivere "dom1r TO dom8r"). Ovviamente, il primo numero deve avere un valore inferiore all'ultimo.

L'utilizzo più significativo del comando RECODE è relativo all'esigenza di realizzare tabelle di contingenza compatte quando i valori da tabulare sono molto (troppo) numerosi. Una situazione frequente è l'aggregazione in classi di una variabile come l'età che, rilevata sull'intera popolazione, può comprendere un centinaio di distinti valori. Come si è visto, quando è opportuno conservare anche i valori originali della variabile, le trasformazioni dovranno essere assegnate a una nuova variabile con INTO.

Se la variabile è cardinale si procederà creando classi di valori contigui con criteri che, come si vedrà nel prossimo capitolo, sono diversi da quello della contiguità semantica, applicabile se la variabile è invece categoriale. Per ora limitiamoci a segnalare l'importanza della distinzione tra i vari tipi di variabili che abbiamo richiamato all'inizio del paragrafo, aggiungendo che nel caso delle variabili a categorie ordinate il criterio è semantico, ma il ricercatore farà meno fatica nel creare i raggruppamenti perché la contiguità semantica è già "incorporata" nell'ordinamento delle categorie, come si è appena visto a proposito della ricodifica delle variabili che rilevano il grado di contrarietà/accordo.

Per illustrare con maggiore dettaglio le caratteristiche del comando RECODE e l'utilizzo delle finestre di dialogo di Spss, proponiamo un esempio nel quale la variabile "eta" (Età) viene ridotta in classi, generando la nuova variabile "etarec" (Classe d'età). Selezionata la finestra TRASFORMA e poi RICODIFICA IN VARIABILI DIFFERENTI, la variabile da trasformare è scelta dall'elenco di sinistra e spostata nello spazio a destra (Fig. 5.2). Nella finestra VARIABILE DI OUTPUT a destra si sono poi scritti il nome e l'etichetta descrittiva della nuova variabile.

Con il bottone VALORI VECCHI E NUOVI si è aperta poi una nuova finestra (Fig. 5.3) nella quale si sono definite le modalità della nuova variabile indicando (con la parola riservata THRU) l'intervallo per la prima aggregazione (18-35 anni) cui, nella finestrella a sinistra (NUOVO VALORE), è stato attribuito il valore 1, concludendo l'operazione con AGGIUNGI. Si è poi creata la modalità "36-65" e, con il bottone INTERVALLO, VALORE AL PIÙ GRANDE, si sono attribuiti alla terza modalità i valori sopra 65 (usando la parola riservata HI).

Con INCOLLA (Fig. 5.2) è poi possibile far scrivere nel foglio di sintassi le istruzioni corrispondenti alle scelte effettuate.

```
RECODE eta (18 THRU 35 = 1) (36 THRU 65 = 2) (66 THRU HI = 3) INTO etarec.
VARIABLE LABELS etarec "Classe d'età".
```

Figura 5.2 - Ricodifica di valori in una nuova variabile (comando "recode")

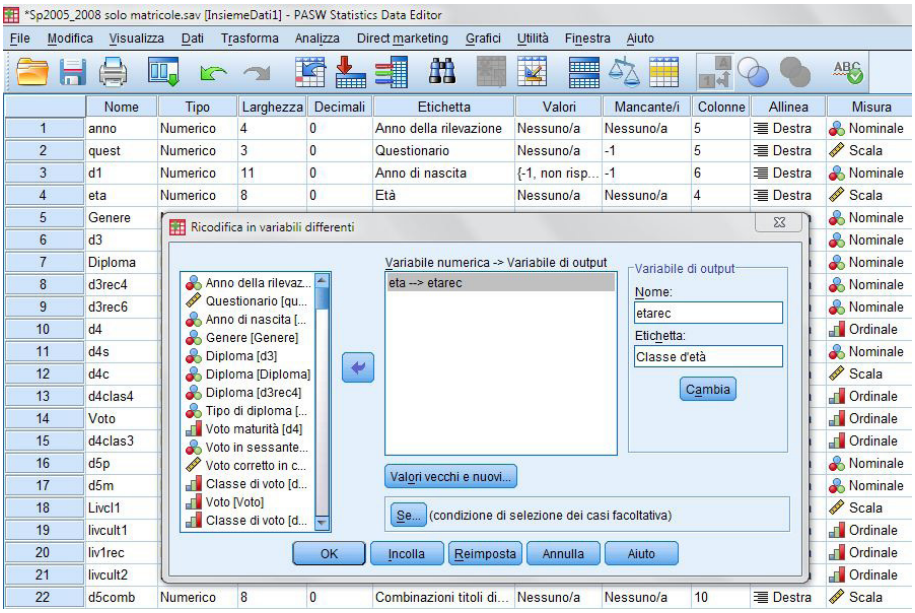
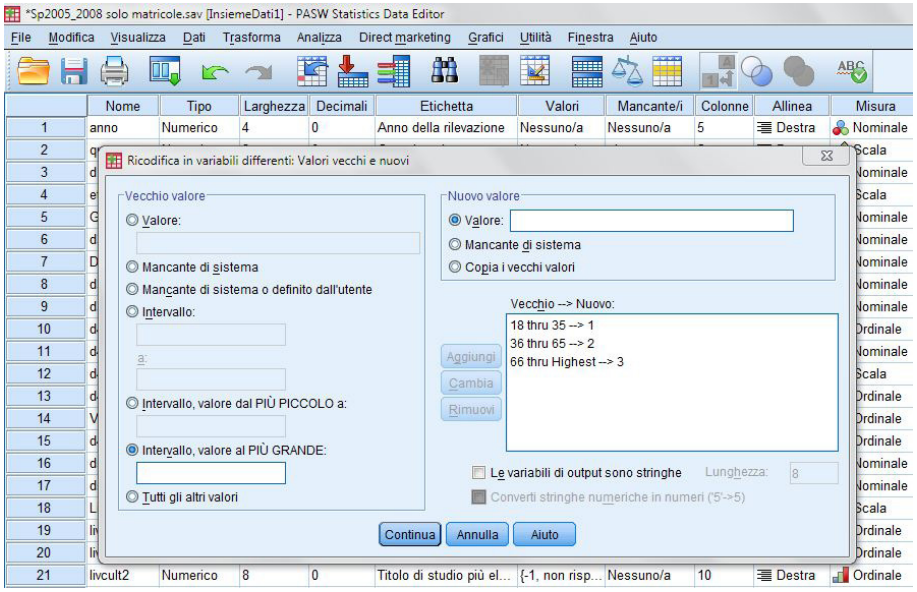


Figura 5.3 - Finestra “valori vecchi e nuovi” di “recode”



Come si può intuire, per ridurre successivamente in classi un’altra variabile basterà modificare le due istruzioni, cambiando i nomi della variabile da trasformare, della variabile di destinazione e i valori che individuano gli intervalli. Per generare un numero più elevato di classi si possono aggiungere al comando RECODE altre istruzioni del tipo “valore THRU valore = nuovo_valore”.

Nel secondo esempio che proponiamo sono stati aggregati alcuni diplomi di maturità pochi diffusi, tenendo distinto il liceo scientifico dagli altri licei, il commerciale dagli altri istituti tecnici, mantenendo separato il professionale. Un’opportuna scelta dei valori ha consentito di attribuire l’ordine voluto alle modalità della nuova variabile alle quali sono state assegnate etichette descrittive con il comando VALUE LABELS (Tab. 5.5).

RECODE dipl (1 = 1) (2 3 4 9= 2) (6 = 3) (5 7 = 4) (8=5) INTO diprec.
VALUE LABELS diprec 1 “Scientifico” 2 “Altro liceo” 3 “Commerciale” 4 “Altro tecnico” 5 “Professionale”.

Tabella 5.5 - Diploma di maturità e tipo di diploma

d3 Diploma	Freq.	Perc.
1 Scientifico	66	27,6
2 Classico	26	10,9
3 Linguistico	18	7,5
4 Psicopedagogico	21	8,8
5 Tecnico industriale	17	7,1
6 Commerciale	35	14,6
7 Altro tecnico	20	8,4
8 Professionale	15	6,3
9 Scienze sociali	21	8,8
Totale	239	100,0

d3rec Tipo di diploma	Freq.	Perc.
1 Scientifico	66	27,6
2 Altro liceo	86	36,0
3 Commerciale	35	14,6
4 Altro tecnico	37	15,5
5 Professionale	15	6,3
Totale	239	100,0

Nella ricodifica dei valori si può anche usare la parola chiave ELSE: se in un'indagine sugli iscritti all'università si vogliono distinguere gli studenti residenti in sede e fuori sede a partire da una variabile dove il comune di residenza è codificato analiticamente, basterà ricodificare in "1" il codice che identifica il comune della sede universitaria e assegnare "2" a tutti gli altri comuni di residenza indicandoli collettivamente con ELSE (che significa altrimenti, e cioè tutto il resto). L'esempio si può ovviamente generalizzare a tutte le situazioni nelle quali si intende isolare una modalità di una variabile per confrontarla con tutte le altre.

Se il comune di residenza è codificato analiticamente, la variabile può essere usata per aggregazioni alternative (se, come sopra esemplificato, si sono conservati i codici originali, attribuendo le modifiche a una nuova variabile). Se, ad esempio, si sono elencati progressivamente in ordine alfabetico provincia per provincia, è possibile usare per la ricodifica la parola chiave THRU (da ... a), come nel caso sotto esemplificato che fa riferimento l'ipotetica ricodifica del comune di residenza di studenti iscritti all'Università di Trieste.

Si deve peraltro fare attenzione, perché ELSE è una parola chiave potente, perciò potenzialmente pericolosa. A conferma della sua "pericolosità", nell'esempio che segue ELSE è stato usato scorrettamente. L'errore è stato involontario, ma si è deciso di evidenziarlo a fini didattici. Attribuiti i comuni del Friuli Venezia Giulia alle 4 Province, gli altri studenti sono stati infatti classificati come residenti altrove; l'intenzione iniziale era però di tenere separati gli stranieri e chi non ha indicato il comune di residenza. In questo caso si può però rimediare facilmente all'errore, perché l'attribuzione delle trasformazioni alla variabile "prov" ha consentito di conservare immutati i valori della variabile originale "comresid".

```
RECODE comresid (1 THRU 137 = 1) (138 THRU 188 = 2) (189 THRU 213 = 3)
    (214 THRU 219 = 4) (ELSE = 5) INTO prov.
VARIABLE LABELS prov "Provincia".
VALUE LABELS prov 1 "UD" 2 "PN" 3 "GO" 4 "TS" 5 "Altro".
```

5.3.4. "Recode ... into" e "compute/recode" a confronto

I nomi dei comandi compute e recode alludono manifestamente a funzioni diverse e vi sono tra loro differenze di sintassi, ma si sarà anche notato che entrambi possono servire a trasformare una variabile senza crearne una nuova, come nel caso dell'aggiunta con compute di 1900 all'anno di nascita registrato con le sole due ultime cifre, o costruire nuove variabili (nel caso di recode, usando into). Si può perciò ritenere che con le seguenti alternative si ottenga lo stesso risultato.

```
RECODE oldvar (1 THRU 15 = 1) (16 THRU 30 = 2) (...) INTO newvar.

COMPUTE newvar = oldvar.
RECODE newvar (1 THRU 15 = 1) (16 THRU 30 = 2) (...).
```

Con la prima soluzione si ricodificano i valori e si assegna il risultato a una nuova variabile; con la seconda si crea una nuova variabile e poi la si ricodifica. In realtà, il risultato è lo stesso solo se tutto gira per il verso giusto, ma in alcuni casi è consigliabile la seconda soluzione, anche se appare più complessa.

Ipotizziamo di aggregare i comuni di residenza tenendo conto della distanza dalla linea ferroviaria per analizzare la rilevanza di questo aspetto nella scelta della sede universitaria. La ricodifica non è semplice come per l'aggregazione in province, facilitata dal fatto che i codici sono assegnati in sequenza ai comuni di una provincia e poi a quelli della provincia successiva. Può essere perciò prudente procedere per gradi, riclassificando dapprima i comuni di più agevole collocazione e poi quelli di collocazione più incerta, tenendo anche conto della consistenza dei gruppi che si formano.

Se si usa recode con into, l'assegnazione parziale impedisce però il successivo perfezionamento della ricodifica, perché i valori non ricodificati sono tutti automaticamente assegnati alla categoria "mancante di sistema" e diventano perciò indistinguibili. Con compute si assegnano invece i valori della variabile originaria a una nuova variabile e la ricodifica parziale di quest'ultima mantiene immutati i codici dei comuni non ancora assegnati; si potrà dunque successivamente decidere se creare una terza e/o una quarta area intermedia, o ripartire il

resto dei comuni tra la zona limitrofa alla ferrovia e la categoria "altro". È comunque tranquillizzante sapere che in entrambi i casi si opera in maniera reversibile, perché anche usando RECODE ... INTO rimane immutata la variabile originale; perciò, siamo sempre in tempo a rimediare se commettiamo errori, mentre sarebbe difficile rimediare ricodificando la variabile originale senza crearne una nuova.

Concludiamo questo confronto tra i due principali comandi di trasformazione dei dati con un ultimo esempio. Se si chiede di esprimere disaccordo/accordo su una serie di affermazioni sulle conseguenze della presenza di immigrati nel paese, è opportuno alternare domande che ipotizzano effetti positivi e negativi. In questo caso, se si intende realizzare un indice additivo dell'atteggiamento nei confronti del fenomeno migratorio, si dovranno invertire i codici di risposta ad alcune di queste domande al fine di uniformare la polarità (ad esempio, dall'atteggiamento più contrario a quello più aperto nei confronti dell'immigrazione).

I codici si possono facilmente invertire con RECODE e può essere inutile creare una nuova variabile; è essenziale invece modificare le etichette dei valori per non rischiare di commentare un presunto totale disaccordo dovuto al fatto che rimarrebbe assegnata questa etichetta a una risposta che, dopo la trasformazione, significa invece "totalmente d'accordo".

```
RECODE dom1 (1 = 4) (2 = 3) (3 = 2) (4 = 1).
```

RECODE è intuitivo ed efficace ma, a conferma del fatto che si possono trovare soluzioni alternative, a volte più eleganti, lo stesso risultato si può ottenere con COMPUTE. Nell'esempio proposto, sottraendo i valori registrati in matrice da 5, si ottiene 1 per l'originario 4 (5 – 4 = 1), 2 per l'originario 3 (5 – 3 = 2) e così via.

In generale, per invertire i codici è sufficiente sottrarre i valori originali da un numero superiore di un'unità rispetto al valore più elevato. L'accorgimento funziona anche se mancano alcuni codici nella sequenza: se in una scala che usa punteggi da 1 a 4 si è attribuito il codice 9 a "non risposto", i valori della variabile originale si sottraggono da 10. I codici si invertono correttamente e la sequenza è spostata di 5 punti (il valore più basso era 1 e ora è 6, il più alto passa da 4 a 9), ma non è un problema se la variabile è ordinale; basta assegnare ai nuovi valori le etichette giuste. Se però si vuole che i punteggi ricadano nell'intervallo di valori originale, per eliminare i punti in esubero si può inserire nella formula una correzione, togliendo i 5 punti in esubero.

```
COMPUTE domnew = (10 – dom) - 5.
```

Nella Tab. 5.6, oltre ai valori originali è mostrato l'effetto dell'inversione e il risultato dopo la correzione. Il valore "-3" attribuito alle risposte mancanti non è

problematico perché, nel caso la scala avessero natura cardinale, sarebbe comunque escluso dal calcolo della media, come "9" e "1" usati nelle altre versioni della variabile.

Tabella 5.6: Inversione dei valori

Valori					**Non risp.**
Originali	1	2	3	4	9
Invertiti (10 - valore)	9	8	7	6	1
Corretti (valore - 4)	4	3	2	1	-3

In conclusione, usare COMPUTE invece di RECODE può apparire un po' complicato, ma vantaggioso se i singoli valori da invertire sono molto numerosi.

5.3.5. I COMANDI "COUNT" (CONTA), "DO IF – END IF" E "DO REPEAT – END REPEAT"

Con il comando "COUNT nome di variabile = lista di variabili (lista di valori)" si attribuisce a una nuova variabile il risultato del conteggio delle occorrenze di un valore o di una lista di valori all'interno di una lista di variabili. Con l'istruzione sotto riportata si contano ad esempio le volte che ogni intervistato ha risposto 3 o 4 (abbastanza o molta) a una serie di domande sulla fiducia verso 10 istituzioni, ottenendo così un semplice indice di fiducia generalizzata.

```
COUNT fiducia = fid_01 TO fid_10 (3 4).
```

Ricordiamo infine che COUNT e gli altri comandi fin qui visti possono essere inseriti all'interno di una struttura DO IF - END IF e in questo modo le trasformazioni si effettuano solo per i casi che soddisfano la condizione indicata in DO IF. Opzionalmente, ELSE IF indica come operare se l'espressione logica è falsa.

```
DO IF [(] espressione logica [)].
Comandi di trasformazione.
[ELSE IF [(] espressione logica [)]].
Comandi di trasformazione.
END IF.
```

Con DO REPEAT – END REPEAT si ripetono invece le stesse trasformazioni su un insieme di variabili e in questo modo si usano molti meno comandi per ottenere lo stesso risultato.

```
DO REPEAT variabile segnaposto = {lista di variabili}.
Comandi di trasformazione.
END REPEAT.
```

5.4. Conclusioni

Tra le esemplificazioni da esporre in questo capitolo si era inizialmente previsto di includerne una trasformazione di variabili un po' particolare, ottenibile con una procedura alquanto complessa. Si è alla fine deciso di escluderla, ma spiegare il motivo può offrire al lettore qualche consiglio conclusivo.

L'esempio riguardava il calcolo della distanza temporale tra due eventi. La soluzione più facile e appropriata consiste nel registrare le informazioni in una variabile "data", scegliendo in Spss tra una trentina di diversi formati. Per calcolare il tempo intercorso tra i due eventi basta in questo caso detrarre, con il comando COMPUTE, la data/ora più recente da quella del passato e il programma calcola la distanza temporale automaticamente.

Se però si sono registrati in campi separati giorno, ora e minuto, per calcolare la distanza temporale si doveva in passato fare sfoggio di abilità nell'uso dei comandi di trasformazione dei dati per "accorpare" un'informazione frazionata su più variabili. Bisognava infatti aggiungere ai minuti le ore trasformate in minuti (moltiplicandole per 60) e, nel caso di eventi accaduti in giorni diversi, bisognava rendere confrontabili i "minutaggi", aggiungendo 1.440 minuti (24 ore x 60 minuti) per ogni giorno di distanza tra le due date. A quel punto, sottraendo i minuti della fine da quelli dell'inizio dell'evento, si ottiene una durata in minuti che, volendo, si può poi trasformare nuovamente in giorni, ore e minuti.

Un problema dunque complesso, ma non ha più senso spiegare come si può risolvere, perché Spss consente ora di registrare giorno, ora e minuti in variabili separate e una procedura guidata consente poi di (ri)costruire con esattezza i momenti temporali e calcolare la distanza tra due di essi.

Il problema è dunque risolto, ma dall'esempio si possono trarre alcuni insegnamenti. Il primo è che programmi come Spss diventano sempre più completi (e complessi), perché il loro successo deriva anche dal coinvolgimento di sempre nuove frange di utenza o "nicchie di mercato" con specifici problemi conoscitivi. I problemi vengono alla luce attraverso pubblicazioni, forum di discussione o contatti diretti con il produttore che, aggiornando il programma, include nuovi moduli o opzioni che risolvono questi problemi. Quindi, chi usa il programma fa bene a controllare con sistematicità le caratteristiche delle nuove versioni, anche usando le tabelle comparative che indicano le modifiche introdotte.

Allo stesso tempo però l'esempio dimostra che, nell'attesa che il programma si adegui alle nostre esigenze, non si è costretti a rinunciare: gli strumenti di base possono spesso aiutare, magari per strade un po' tortuose e con qualche fatica, a risolvere i problemi con le variabili senza attendere che, probabilmente in modo più "elegante", l'adeguamento del software risolva il problema per noi.

Si può scoprire tra l'altro che la soluzione di Spss è spesso la stessa o è simile a quella che avevamo trovato, con la differenza che Spss, traducendola in opzione o funzione, la generalizza, mentre l'utente l'applica al caso specifico. Ad esempio, per la distanza temporale tra due eventi può essere sufficiente la precisione in minuti, ma Spss, volendo generalizzare la soluzione a tutte le eventualità, traduce gli orari nell'unità più piccola, cioè in secondi.

Si può dunque concludere, ribadendo l'utilità di apprendere le regole della sintassi: le sequenze di comandi concepite per risolvere un problema specifico possono diventare soluzioni generalizzabili, se le abbiamo memorizzate, perché spesso è sufficiente adeguare i nomi delle variabili e dei valori.

6. Categorizzare una variabile

6.1. Introduzione

Come si è visto nel capitolo 2, nell'analisi monovariata ci si trova di fronte a due situazioni alquanto diverse che dipendono dalla natura delle variabili che si devono analizzare. I tre tipi fondamentali si possono infatti dicotomizzare perché la distinzione fondamentale è tra *variabili categoriali* (a categorie *non ordinate* o *ordinate*) e *cardinali*. Nel primo caso lo strumento adatto è la distribuzione di frequenza, perché spesso, ma non sempre, le modalità sono poco numerose e pochi gli strumenti statistici utili a sintetizzare i risultati: la *moda* (modalità con la frequenza più elevata), che peraltro non "sintetizza" la distribuzione, e la *mediana*, che ripartisce in due parti uguali la distribuzione ordinata dei casi, ma non è molto utile se le modalità sono poco numerose.

Se la variabile è cardinale ci si trova spesso (ma non sempre) nella situazione opposta: la popolazione dei comuni o il fatturato delle aziende, misurati con precisione, presentano infatti quasi invariabilmente per ognuno dei casi valori diversi che la distribuzione di frequenza si limita a ordinare; un'operazione comunque utile perché permette, se non altro, di individuare agevolmente il minimo, il massimo, la mediana e di suddividere la distribuzione in quartili. La

distribuzione di una variabile cardinale si può invece agevolmente sintetizzare, anche se può essere troppo semplificante mostrare solo la *media*, pur corredata dal campo di variazione (*range*) o dallo scarto-tipo (*standard deviation*), utili a descrivere la dispersione dei dati. A volte è dunque necessario o preferibile segmentare la variabile in classi di valori perché, ad esempio, ci interessa sapere quanti intervistati ricadono nelle diverse fasce di reddito o d'età o lavorano in aziende di una determinata classe dimensionale.

Ovviamente, si può ridurre l'eccessiva dispersione dei dati anche nel caso di variabili categoriali, e in questo caso, come abbiamo visto nel capitolo precedente, si riuniscono le categorie che presentano qualche affinità semantica: una lista di istituti può essere ricondotta a "liceo/tecnico/professionale" e i comuni accorpati per contiguità geografica o appartenenza alle diverse province.

Con le variabili nominali non si tiene dunque conto del valore (codice) attribuito alle categorie, perché spesso dipende dal banale ordinamento alfabetico dei nomi. Nel caso delle variabili cardinali, si accorpano invece valori contigui ed è legittimo chiedersi quale strategia adottare (*quante* fette e *dove* tagliare?). In precedenza si è già visto qualche esempio di categorizzazione (termine più appropriato dello scherzoso "fare a fette" utilizzato nel titolo di una versione precedente del presente capitolo), ma non abbiamo approfondito e problematizzato la questione. Lo faremo in questa sede, iniziando proprio dalle strategie, che esperienza e riflessione suggeriscono essere quattro:

- intervalli di valori di *uguale ampiezza*,
- gruppi di casi di *dimensioni uguali o predeterminate*,
- fratture o *discontinuità* nella distribuzione,
- valori significativi o *valori soglia*.

La lista non pretende di essere esaustiva, ma è sufficientemente differenziata, in modo da tenere conto delle diverse esigenze che si possono presentare.

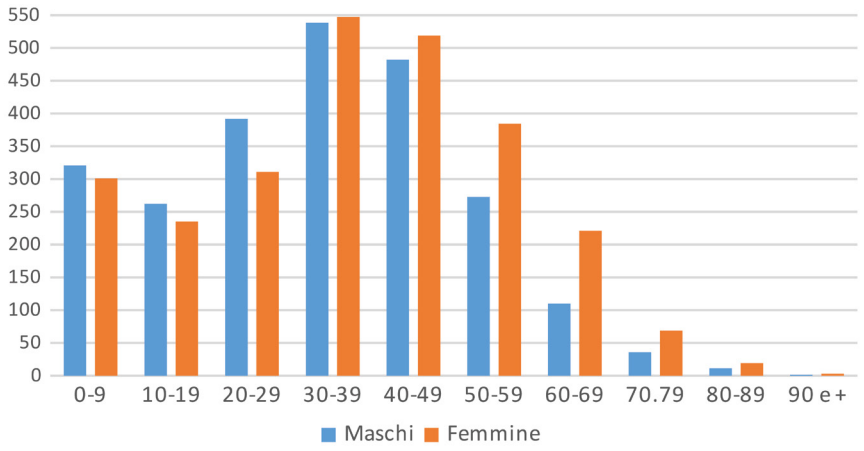
6.2. Intervalli di valori di uguale ampiezza

Suddividere il campo di variazione in intervalli di valori di uguale ampiezza è piuttosto usuale; l'esempio più noto è l'età, suddivisa in classi decennali o quinquennali al fine di costruire la c.d. *piramide d'età*. Dal momento che gli intervalli sono uguali, ciò che varia è la consistenza numerica dei casi al loro interno e infatti con essi si intende analizzare la rapidità della decrescita o il progressivo calo che ormai si registra da tempo nelle nuove generazioni. Ovviamente, queste analisi si possono effettuare anche su specifiche componenti della popolazione e in precedenza abbiamo visto un significativo esem-

pio analizzando la crescita del numero di laureati ripartendo l'età in classi quinquennali (§ 4.1.1).

Nelle piramidi d'età le barre sono contrapposte rispetto all'asse delle età ma, a nostro avviso, il grafico a barre affiancate è più efficace perché consente un più preciso confronto visivo tra le lunghezze delle coppie di barre, come si può constatare nella Fig. 6.1 che mostra la composizione per età e genere dei cittadini stranieri residenti in Italia. I dati sono un'istantanea "scattata" il 1° gennaio 2021, ma lo scorporo per classe d'età consente di leggere indirettamente la storia dell'immigrazione in Italia. In ognuna delle classi decennali 30-39 e 40-49 anni gli stranieri sono circa un milione sommando maschi e femmine (circa il 10% della popolazione italiana nelle stesse classi d'età), mentre nelle classi di età più mature la loro numerosità e l'incidenza percentuale sul totale della popolazione italiana scende drasticamente, per arrivare quasi a zero tra i più anziani. Si nota poi una maggiore consistenza tra i giovani con meno di 10 anni, rispetto agli adolescenti (fino a 19 anni), segno evidente che la presenza straniera conta quote crescenti di figli di stranieri, prevalentemente nati in Italia.

Figura 6.1 - Stranieri residenti per classi d'età e genere in migliaia (Italia 2021)



Fonte: TuttItalia (https://www.tuttitalia.it/statistiche/cittadini-stranieri-2021/)

Un altro esempio di strategia di segmentazione in classi di uguale ampiezza può essere la riduzione di una distribuzione percentuale in 10 classi, come nella Tab. 6.1 dove è analizzata la popolazione mondiale che nel 2018 viveva nelle zone urbane. Nella parte destra della tabella è indicato il numero e la percentuale di paesi che ricadono in ogni intervallo: nella quinta riga si legge, ad esempio, che in 24 paesi (10,3%) vive in città tra il 40 e il 49% della popolazione. Nella quarta

colonna sono riportate le percentuali cumulate (quota di paesi che ricadono in quell'intervallo e nei precedenti) dalle quali si rileva che nel 32,6% dei paesi la popolazione urbana non supera il 50%; ne consegue che in quasi i due terzi dei paesi prevale la popolazione urbana.

Tabella 6.1 - Percentuale di popolazione urbana in 233 paesi del mondo (2018)

	Paesi	%	% cum.	Paesi pesati	%	% cum.
0-9	3	1,3	1,3	18	,0	,0
10-19	11	4,7	6,0	137.521	1,8	1,8
20-29	17	7,3	13,3	306.080	4,0	5,8
30-39	21	9,0	22,3	2.103.339	27,6	33,4
40-49	24	10,3	32,6	476.759	6,2	39,6
50-59	33	14,2	46,8	2.158.009	28,3	67,9
60-69	28	12,0	58,8	326.219	4,3	72,2
70-79	31	13,3	72,1	706.913	9,3	81,4
80-89	33	14,2	86,3	1.167.914	15,3	96,7
90-100	32	13,7	100,0	250.047	3,3	100,0
Totale	233	100,0		7.632.819	100,0	

Fonte: World Urbanization Prospects 2018 (https://population.un.org/wup/)

Da ciò non si può però concludere che la maggior parte della popolazione del mondo vive in città: la stima non è corretta perché i paesi si devono pesare, tenendo conto della dimensione della loro popolazione. Il 34% di popolazione urbana dell'India corrisponde a oltre 460 milioni di abitanti; in Tanzania la percentuale è simile (33,8%), ma corrisponde a meno di 20 milioni di abitanti. Gli abitanti dell'Islanda che vivono in città sono molto meno numerosi (317.000), ma si tratta del 93,8% della popolazione di quel Paese.

La pesatura sulla base della popolazione porta ai risultati inseriti nella parte destra nella Tab. 6.1 dove i casi non sono più i 233 originali, ma oltre 7,6 milioni, corrispondenti a oltre 7,6 miliardi di persone (la popolazione è registrata in migliaia). Nella nuova distribuzione si vede ad esempio che i paesi con una popolazione urbana tra il 30 e il 39% sono solo il 9%, ma in essi vive più di un quarto (27,6%) della popolazione mondiale. Ancora: il 13,7% dei paesi ha una popolazione urbana che varia tra il 90 e il 100%; pesando i casi si vede però che la popolazione che vive in questi paesi è pari a solo il 3,3% del totale.

In generale, quando si categorizza una variabile si deve anche decidere se usare i minimi e massimi teorici o quelli che si sono concretamente rilevati ma,

mentre questi ultimi sono sempre presenti nella distribuzione, i limiti teorici a volte possono non essere chiaramente definiti o definibili, particolarmente se si tratta di punteggi "astratti" derivanti da qualche procedura di indicizzazione. Negli esempi proposti il campo di variazione teorico è ben definito (da "0" a "100"), ma nel caso dell'età si registra qualche gradita eccezione che può tradursi in una classe con il confine superiore "aperto" (100 anni o più). Anche per il reddito si può usare una segmentazione in intervalli costanti (ad es. 5.000 o 10.000 euro) ma, mentre l'estremo inferiore (nessun reddito) è chiaramente definito, lasciare aperto il confine superiore della classe è più artificioso perché, a differenza dell'età, non c'è quasi limite al massimo reddito percepibile.

6.3. Gruppi di dimensioni uguali o predeterminate

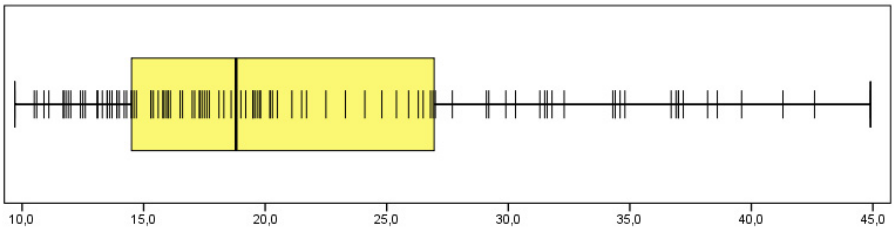
La seconda strategia di categorizzazione si basa sull'utilizzo delle percentuali cumulate, al fine di costruire classi contenenti lo stesso numero di casi. Se non ci sono ragioni che la sconsiglino, questa soluzione presenta un vantaggio non trascurabile: si è infatti certi di non ritrovarsi con una o più classi con pochi casi, come può accadere con intervalli di valori di uguale ampiezza. Come si è visto nella Tab. 6.1, la classe "0-9" contiene solo 3 paesi (1,3% del totale); la situazione in quel caso non è problematica perché si tratta di dati censuari (si prende semplicemente atto che i paesi sono solo quelli), ma la categoria sarebbe pressoché inutilizzabile se si trattasse di dati campionari, perché con soli tre casi sarebbe statisticamente poco o per nulla rappresentativa.

Per rendere più apprezzabile questa seconda tecnica di categorizzazione si può anche ricordare che alcuni indicatori sono basati proprio sulla segmentazione dei casi in gruppi di uguali dimensioni: per analizzare il reddito si possono infatti utilizzare intervalli di reddito uguali, ma più spesso si utilizza, ad esempio, il *rapporto interdecilico*, cioè il rapporto tra il reddito detenuto dal 10% più povero della popolazione e dal 10% più ricco.

La categorizzazione basata su gruppi di uguali dimensioni l'abbiamo già incontrata quando abbiamo descritto i *box plots* che, come si ricorderà, nascono proprio per rappresentare graficamente i quartili e cioè la ripartizione di una distribuzione di frequenza in quattro parti comprendenti lo stesso numero di casi. È proprio a causa di questa loro caratteristica "costitutiva" che si deve prestare particolare attenzione nell'esaminarli, per evitare equivoci interpretativi in cui può cadere il lettore disattento o non sufficientemente addestrato. Il fatto che le parti a destra e sinistra della mediana abbiano diversa estensione non significa infatti che nelle due metà del box e nei "baffi" che sono il loro prolungamento ci sia un diverso numero di casi: il numero è pari per definizione, ma è diversa la loro dispersione.

Al fine di visualizzare questo aspetto abbiamo realizzato la Fig. 6.2 nella quale al *box plot* è sovrapposto un grafico a semi-dispersione (nel quale i punti sono sostituiti da lineette per favorire la leggibilità) ed è evidente che l'estensione del box e dei "baffi" è molto diversa, ma all'interno dei loro confini è contenuto lo stesso numero di casi (una sezione ne contiene in realtà solo 26, essendo il totale di 107 province non esattamente divisibile per quattro). Nelle province dove il fenomeno è meno grave la percentuale di Neet (*Not in education, employment or training*) varia di soli 5 punti; molto più differenziata è invece la situazione all'altro estremo, dove i Neet variano da circa 27% a 45%.

Figura 6.2 - *Box plot* e grafico a semi-dispersione combinati – Percentuale di Neet per provincia (Italia 2019)



Fonte: Il Sole24ore (https://lab24.ilsole24ore.com/qualita-della-vita/)

Usare i valori posizionali per suddividere una distribuzione di frequenza è particolarmente utile se i valori della variabile da segmentare (o categorizzare) sono punteggi che si potrebbero definire "astratti", in quanto privi di un chiaro ancoramento semantico. È il caso dei punteggi attribuiti alle provincie sulla base di una pluralità di indicatori di qualità della vita, uno dei quali è proprio la percentuale di Neet che abbiamo appena esaminato. I singoli indicatori derivano da procedure di rilevazione empirica e si devono dunque necessariamente basare su una qualche "metrica" (unità di misura o unità di conto), mentre non è così per gli indici additivi che ne derivano perché, per poter essere calcolati, devono ricondurre i singoli indicatori a un'unitaria metrica convenzionale (ad esempio, tutte le misure o conteggi vengono convertite in punteggi da 0 a 1.000).

Conclusivamente si può aggiungere che, se i valori utilizzati per la segmentazione non fanno riferimento a una qualche metrica e/o non hanno significati condivisi o noti, dal momento che servono solo a costruire un determinato numero di fasce o classi, pare convincente denominare queste ultime semplicemente come inferiore, intermedia, superiore, oppure bassa, media, alta, come nel prospetto che segue (Tab. 6.2), dove le aggettivazioni sono adattate al loro numero.

Tabella 6.2 - Numero e denominazione delle fasce categoriali

N. categorie	Denominazione				
2	Bassa				Alta
3	Bassa		Media		Alta
4	Bassa	Medio-bassa		Medio-alta	Alta
5	Bassa	Medio-bassa	Media	Medio-alta	Alta

Per categorizzare una variabile cardinale al posto dei valori posizionali della distribuzione (i quartili e, più raramente, i decili), si può ipotizzare anche l'utilizzo di altri parametri statistici, quali la media (al posto della mediana) e la deviazione standard (al posto del primo e terzo quartile). In questo caso solo i due gruppi definiti dalla media sono tendenzialmente di dimensioni uguali, ma solo nel caso la variabile abbia una distribuzione che approssima la normale. Si può poi utilizzare lo scarto-tipo (deviazione standard) per suddividere i due lati della distribuzione e in questo caso si ottiene una ripartizione numericamente diseguale per definizione perché, in condizione di normalità, nell'intervallo definito da uno scarto-tipo in più o in meno dalla media è contenuto circa il 68% dei casi. Per questo motivo nel titolo del paragrafo si accenna a gruppi di dimensioni predeterminate, ma è più prudente parlare di dimensioni "tendenzialmente" predeterminate perché una distribuzione normale raramente si riscontra empiricamente quando si rilevano fatti sociali.

Infatti, come si vedrà più avanti nella Tab. 6.4, la distribuzione empirica della variabile "aspettativa di vita" utilizzata per mettere a confronto le strategie di segmentazione è fortemente asimmetrica e, di conseguenza, le classi create con questo metodo sono piuttosto distanti dalla ripartizione teorica: 21% e 41% dei casi nelle classi centrali a sinistra e a destra della media e 21% e 17,5% nelle due code esterne della distribuzione.

Le variabili rilevate nelle ricerche sociali sono in effetti spesso molto lontane dalla distribuzione normale e, se si preferisce una distribuzione equilibrata dei casi tra le classi, può dunque essere preferibile utilizzare per la segmentazione valori posizionali quali la mediana e i quartili. Se invece si vogliono isolare le code della distribuzione perché si ritiene interessante studiare le caratteristiche specifiche dei casi per i quali si sono rilevati valori estremi, l'utilizzo della deviazione standard è senz'altro raccomandabile e si potrebbe decidere di utilizzare per la segmentazione una distanza dalla media pari anche a due o tre deviazioni standard (o scarti-tipo). Come vedremo, se ci si vuole avvalere della possibilità di effettuare automaticamente la categorizzazione, questa è infatti una delle opzioni offerte da Spss.

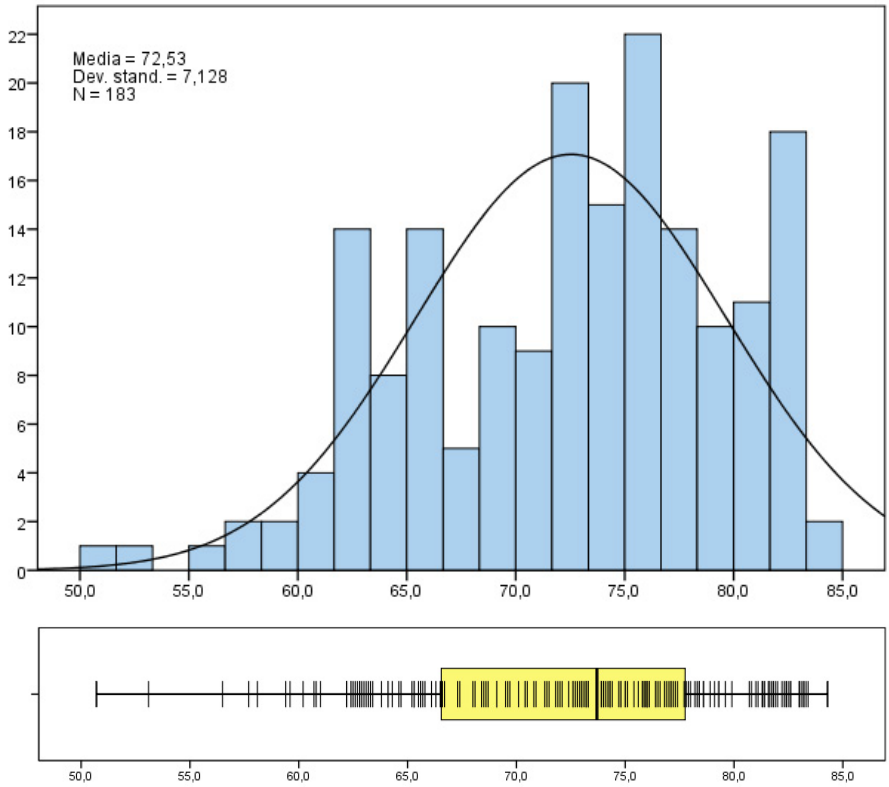
Le distribuzioni empiriche delle variabili spesso mostrano interruzioni o rarefazioni dei casi che possono costituire un interessante "pretesto" per la loro segmentazione: si taglia dunque dove si nota una rarefazione, per evitare che casi per i quali si è rilevato un valore uguale o simile siano assegnati a classi diverse. In pratica, si agisce come il tagliatore di diamanti che lavorare la pietra grezza cercando di individuare e "sfruttare" le sue linee di frattura naturali. Si può poi misurare l'efficacia dell'intervento, calcolando per ogni gruppo la media e lo scarto-tipo (deviazione standard), anche al fine di verificare, per tentativi ed errori, se una diversa segmentazione potrebbe portare a una maggiore variabilità tra le categorie (*between*), minimizzando quella all'interno di esse (*within*).

Le discontinuità o rarefazioni si possono individuare in modo relativamente facile se quella che si è rilevata è una proprietà discreta a stati enumerabili ed è relativamente piccolo il numero di potenziali valori della variabile. Esaminando la distribuzione di frequenza della variabile non è difficile in questo caso identificare valori assenti o con una frequenza relativamente bassa. Se si è invece misurata con precisione una proprietà continua, può essere persino inappropriato parlare di distribuzione di frequenza, perché per ogni valore la frequenza può essere pari a uno (come, ad esempio, nel caso del reddito di un campione di intervistati o del fatturato di un insieme di imprese). Per discontinuità o frattura in questo caso si può intendere un intervallo relativamente ampio nella successione ordinata dei valori. Si devono dunque calcolare tutti gli intervalli e si tratta poi di scorrere lunghe liste di valori per individuare gli intervalli più ampi.

Molto più semplice è in questo caso l'utilizzo degli strumenti di rappresentazione grafica dei dati che consentono invece di scorgere le discontinuità "a prima vista", come nel caso del grafico a semi-dispersione visto all'opera nel paragrafo precedente (Fig. 6.2). Quando i casi sono molto numerosi può essere però ancor più utile rappresentare graficamente i dati mediante un *istogramma*, strumento che nasce (anche) per verificare quanto la distribuzione approssimi la "curva normale" (una forma a campana con i casi concentrati al centro e una rarefazione verso i valori estremi) (vedi § 2.2). Se l'istogramma assume, ad esempio, una forma bimodale o multimodale (concentrazione dei casi intorno a due o più valori, diversi e distanti), ecco individuate le fratture.

La differenza tra i due strumenti è chiaramente individuabile nella Fig. 6.3, dove è rappresentata l'aspettativa di vita alla nascita nel 2019 (prima dello scoppio dell'epidemia) in 183 paesi di tutto il mondo (sono esclusi i mini-stati). Nel grafico a dispersione posizionato sotto all'istogramma sono rappresentati tutti i singoli casi e ognuno di essi è posizionato in corrispondenza del valore per esso rilevato sulla variabile.

Figura 6.3 - Aspettativa media di vita alla nascita in 183 paesi (2019)



Fonte: World Health Organizazion (https://apps.who.int/gho/data/view.main.SDG2016LEXREGv)

Per realizzare l'istogramma il campo di variazione dei valori viene invece suddiviso automaticamente, tenendo conto della numerosità dei casi, in un determinato numero di classi di valori di uguale ampiezza. In tal modo si realizza una distribuzione di frequenza dei casi (che originariamente non esiste) che si può poi rappresentare affiancando colonne di altezza proporzionale al numero (valore assoluto) o alla percentuale di casi che ricadono in ogni intervallo di valori. La suddivisione viene in pratica fatta "alla cieca" e, di conseguenza, alcune fratture nella distribuzione potrebbero essere occultate o create artificialmente. Come nel nostro esempio, può dunque essere utile affiancare all'istogramma un grafico a semi-dispersione che rappresenti con punti (o lineette, se ritenute più efficaci) ogni caso singolarmente, al fine di verificare se l'istogramma abbia davvero individuato le linee di frattura nella distribuzione.

Nella Fig. 6.3, osservando l'istogramma si possono individuare diverse fratture che separano gruppi di paesi con aspettative di vita diverse e la curva sovrapposta all'istogramma mostra che la distribuzione di frequenza in effetti si discosta notevolmente dall'ipotesi di normalità. A parte un paio di paesi con una speranza media di vita di poco superiore a 50 anni, si nota infatti una rarefazione dei casi intorno ai 67 anni e un'altra intorno a 80, circostanza che viene confermata dall'osservazione della distribuzione delle lineette nel grafico a semi-dispersione collocato sotto l'istogramma. Volendo ripartire la distribuzione in un numero più elevato di classi, si potrebbe valutare l'opportunità di tracciare un confine anche sulla base di una relativa discontinuità intorno ai 74 anni che si può scorgere anche nel grafico a dispersione.

6.5. Valori significativi o valori soglia

Anche la quarta strategia delineata all'inizio del capitolo parte dai valori della variabile ma, anziché costruire meccanicamente intervalli di uguale ampiezza, realizzare gruppi di casi di uguale o simile numerosità o individuare fratture nella distribuzione, si ripropone di utilizzare per la segmentazione valori dotati di un qualche rilevante significato condiviso. In genere si tratta di *valori soglia*, cioè valori che indicano una transizione, a volte fissati per legge: ad esempio, l'età lavorativa (16 anni), la maggiore età (18 anni), l'età della pensione (67 anni, o qualsiasi altra età venga stabilita in conseguenza delle discussioni in atto). Nel caso del reddito valori di questo tipo sono la c.d. "soglia di povertà", il livello di reddito che esenta dal pagamento delle tasse e le altre soglie di reddito che comportano il passaggio da un'aliquota Irpef all'altra, oppure l'esenzione o la riduzione delle tasse universitarie.

Anche nel caso di dati riferiti a unità di analisi di tipo territoriale si possono utilizzare valori soglia riferiti, ad esempio, alla numerosità della popolazione residente. Alcuni di questi valori sono molto significativi perché il loro raggiungimento può comportare rilevanti conseguenze sul piano politico. Ad esempio, se il comune supera i 15.000 abitanti il sistema elettorale per l'elezione del sindaco passa da maggioritario a turno unico a maggioritario con ballottaggio e i consigli comunali sono composti da un numero di consiglieri variabile in relazione al superamento di determinate soglie di numerosità dei residenti.

Riguardo all'ambito universitario, la misurazione più nota è ovviamente il voto che i docenti attribuiscono agli studenti che sostengono l'esame e il valore soglia più discriminante è il 18, che sancisce il superamento dell'esame. Le cose si complicano quando il voto deve "viaggiare" con lo studente che, ad esempio, passa un periodo di studi all'estero partecipando a un programma Erasmus. In questo caso è spesso necessario convertire il voto in trentesimi utilizzato nel sistema

universitario italiano al sistema ECTS (*European Credit Transfer and Accumulation System*) o ad altri sistemi nazionali, come quello vigente nelle università americane. Entrambi sono riportati nella Tab. 6.3 nella quale i voti in trentesimi vengono convertiti nelle cinque categorie da A = *Excellent* a F = *Fail*.

Tabella 6.3 - Conversione dei voti universitari al sistema ECTS e vigente in USA

Voto	**Voto ECTS**	**Definizione**	**Voto U.S.**	**Definizione.**
30 e lode 26-30	A	Excellent	30, 30 e lode: A+ 28-29: A 26-27: A-	Excellent
21-25	B	Very Good	24-25: B+ 22-23: B 21: B-	Good
19-21	C	Good	21: C+ 20: C 19: C-	Satisfactory
18	D	Satisfactory	D	Passing
14-17 0-13	Fx F	Fail	E or F	Fail

6.6. Strategie di segmentazione a confronto

Proviamo ora mettere a confronto i diversi tipi di segmentazione applicandoli a un'unica variabile in modo da chiarire ulteriormente le finalità delle diverse alternative e le loro conseguenze sulle classi di valori generate. A tal fine utilizzeremo la variabile "aspettativa di vita alla nascita" con la quale il lettore ha già preso dimestichezza quando abbiamo illustrato la strategia della ricerca delle discontinuità nella distribuzione dei valori.

Applicando la segmentazione in intervalli uguali di valori potremmo partire da 50 anni (il minimo è in realtà 50,7) e creare classi decennali. Il valore più elevato è però 84,3 anni, dunque un valore che renderebbe di fatto l'ultimo intervallo di anni molto meno ampio degli altri e ciò potrebbe consigliare di creare sette classi di ampiezza quinquennale, invece di quattro decennali. In alternativa, si può abbandonare l'idea di incardinare le classi a cifre tonde e dividere il campo di variazione in quattro sezioni di ampiezza davvero uguale. Procedendo in questo modo si ha la ripartizione riportata nella Tab. 6.4 dalla quale si vede che questa strategia comporta (nonostante il correttivo) un notevole squilibrio nella numerosità interna alle quattro classi a

svantaggio in particolare della prima. Ciò si deve alla relativa "rarefazione" nella parte inferiore della scala, come si nota chiaramente dall'istogramma e diventa ancora più evidente quando si passa alla seconda strategia. Per comprendere nella prima categoria un numero di casi pari a quello delle altre (in realtà uno in meno) bisogna infatti estendere il suo campo di variazione a più di 15 anni, mentre bastano meno di sette anni nella seconda e nell'ultima classe e solo quattro nella terza.

Tabella 6.4 - Segmentazioni alternative della variabile "aspettativa di vita alla nascita" (2019)

	Classe	Intervallo	N	%	Età media	Scarto tipo
Intervalli uguali	50,7–59,1	8,4	5	3,8	55,2	3,2
	59,2–67,5	8,4	45	30,1	63,9	2,0
	67,6–75,9	8,4	67	49,2	72,6	2,3
	76,0–84,3	8,4	66	16,9	79,7	2,6
Uguali dimensioni	50,7-66,3	15,6	45	24,6	62,6	3,3
	66,5–73,3	6,8	46	25,1	70,7	2,2
	73,7–77,7	4,0	46	25,1	76,0	1,1
	77,8–84,3	6,5	46	25,1	81,0	1,8
Media e dev. standard	50,7-65,4	14,7	38	20,8	62,0	3,3
	65,5-72,5	7,0	38	20,8	69,0	2,2
	72,6-79,7	6,2	75	41,0	75,6	2,0
	79,8-84,3	4,5	32	17,5	82,1	0,
Discontinuità	50,7–66,7	16,0	48	26,2	62,8	3,3
	67,3–73,3	6,0	43	23,5	71,0	1,9
	73,7-79,9	6,2	61	33,3	76,4	1,7
	80,7–84,3	3,6	31	16,6	82,1	1,0
Valori soglia	50,7-66,7	16,0	48	26,2	62,8	3,3
	67,3-84,3	17,0	135	73,8	76,0	4,4
	Totale	33,6	183	100,0	72,5	7,1

Fonte: World Health Organizazion (https://apps.who.int/gho/data/node.main.688?lang=en)

Se si separano le classi nei punti dove si è individuata una qualche discontinuità abbiamo la quarta segmentazione che presenta un numero di casi nelle diverse classi relativamente equilibrato e si può dunque considerare un buon compromesso. Infine, abbiamo aggiunto per completezza l'ultima ripartizione, basata sui valori soglia, e si può comunque solo notare, come semplice curiosità, che quasi i tre quarti dei paesi

considerati hanno una speranza di vita alla nascita superiore a quella del pensionamento secondo la riforma Fornero. Per differenza si ricava che in circa un quarto dei paesi del mondo si è già "mediamente" morti quando in Italia si va in pensione.

6.7. Come "fare a fette" (categorizzare) con Spss

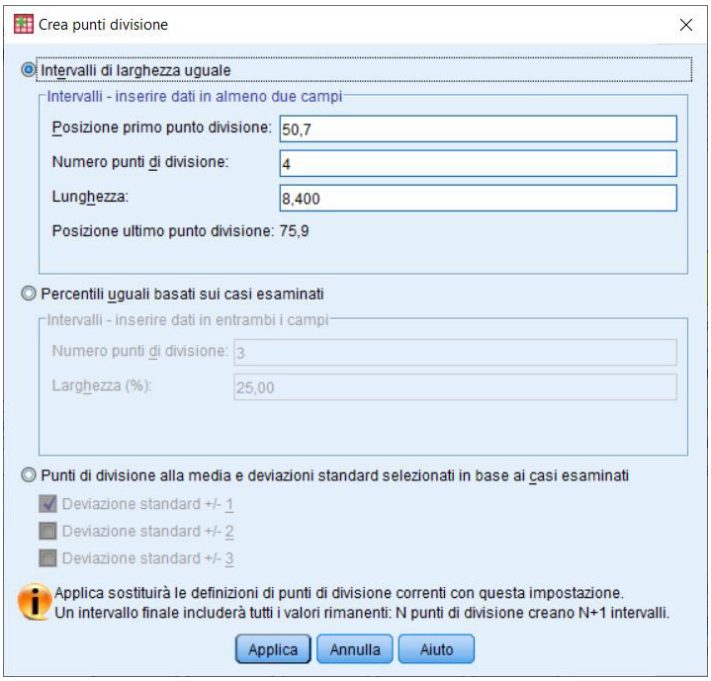
Spss ha introdotto da tempo tra le funzioni di supporto all'analisi dei dati il modulo categorizzazione visuale, nel quale si trova conferma che le opzioni fondamentali sono quelle che sono state qui esposte e che le tecniche grafiche sono un supporto molto utile per orientare la scelta. Nella videata iniziale (Fig. 6.4) è mostrato infatti l'istogramma della distribuzione della variabile selezionata, sul quale il programma traccia automaticamente delle barre verticali in corrispondenza dei valori selezionati, dopo che si è scelta la modalità di segmentazione tramite una finestra di dialogo attivata dal bottone "crea punti divisione".

Figura 6.4 - La finestra "categorizzazione visuale"

Le alternative proposte nella Fig. 6.5 sono le prime tre che abbiamo descritto in questa sede: intervalli di larghezza uguale, percentili uguali basati sui casi esaminati e punti di divisione alla media e alla deviazione standard. Inoltre, l'isto-

gramma consente di individuare le fratture o le discontinuità della distribuzione e, di conseguenza, è possibile attivare anche le altre due strategie che abbiamo descritto perché le barre si possono spostare e posizionare manualmente in corrispondenza di questi abbassamenti "locali" della frequenza o in corrispondenza di quelli che abbiamo definito "valori soglia". Presa la decisione riguardo ai valori da utilizzare per la segmentazione il programma consente, mediante il bottone CREA ETICHETTE, di contrassegnare automaticamente i valori della variabile categorizzata generata dalla procedura.

Figura 6.5 - La sotto-finestra "crea punti di divisione"



6.8. La categorizzazione ottimale

Accenniamo infine a un'ulteriore "sofisticata" strategia che trattiamo a parte perché per realizzarla entra in gioco una seconda variabile, oltre a quella da ridurre in classi. La CATEGORIZZAZIONE OTTIMALE consente infatti di determinare i gruppi scegliendo intervalli di suddivisione che massimizzano l'associazione tra la variabile raggruppata e quella usata per l'ottimizzazione. Nel caso dell'esempio che abbiamo proposto, la procedura di categorizzazione ottimale dovreb-

be garantire che qualsiasi altra ripartizione in intervalli della speranza di vita renderebbe meno forte la sua associazione con l'area di appartenenza dei paesi. Applicata la procedura, dai risultati della Tab. 6.5 si nota che quasi tutti i paesi dell'Oecd sono collocati nell'intervallo superiore (oltre 81 anni di speranza di vita media alla nascita) e quasi tutti quelli dell'Europa dell'Est nell'intervallo da circa 72 a 82 anni, una situazione simile a quella che si registra nei paesi americani (esclusi Canada e Stati Uniti che fanno parte dell'Oecd). All'opposto, quasi tutti i paesi africani si collocano nel primo intervallo, con una speranza di vita media che da poco più di 50 anni arriva fino a poco più di 67. La tabella consente anche di osservare che le situazioni più diversificate si registrano nelle due rimanenti "aree": i paesi islamici e gli "asiatici", tra i quali si sono compresi anche alcuni paesi dell'area oceanica.

Tabella 6.5 - Categorizzazione ottimale dell'aspettativa di vita per area (2019)

Intervallo	Punteggi finale		Numero di casi per Area						
	Inferiore	**Superiore**	**Oecd**	**Est Europa**	**America latina**	**Paesi islamici**	**Asia**	**Africa**	**Totale**
1	[a]	67,4	0	0	2	8	5	34	49
2	67,4	71,9	0	0	0	7	13	3	23
3	71,9	81,1	2	20	28	19	13	1	83
4	81,1	[a]	24	1	0	1	2	0	28
Totale			26	21	30	35	33	38	183

Fonte: World Health Organizazion (https://apps.who.int/gho/data/node.main.688?lang=en)

Come si è visto, nella categorizzazione ottimale entra in gioco una variabile che condiziona la ripartizione in classi della prima; siamo dunque nell'ambito dell'analisi bivariata, come si comprende anche dal fatto che i risultati della categorizzazione sono presentati tramite una tabella di contingenza nella quale è mostrata la relazione tra una variabile indipendente (l'appartenenza alle aree geografiche e socioeconomiche) e una dipendente (la speranza di vita categorizzata). In questo caso non si pone il problema della numerosità dei casi che entrano a far parte delle diverse classi nelle quali è categorizzata la speranza di vita, perché si tratta di un fatto del quale possiamo solo prendere atto.

A questo proposito, si può richiamare quanto detto a proposito della strategia che utilizza per la segmentazione la media e la deviazione standard. Si è infatti accennato al fatto che con questo tipo di categorizzazione le classi immediatamente a destra e sinistra della media tendenzialmente raccolgono un numero

maggiore di casi rispetto a quelle che si collocano all'esterno; inoltre, la simmetria teorica della distribuzione si riscontra assai raramente nei dati reali. Ciò però non deve necessariamente spingere a privilegiare la soluzione della categorizzazione basata sui quantili solo perché garantisce una simmetrica equipartizione dei casi tra le classi. Infatti, se la variabile è dipendente (oggetto di studio), la concentrazione dei casi intorno alla media (o l'estrema dispersione) e/o l'eventuale asimmetria sono semplicemente il risultato di un'analisi che può essere fatta proprio con lo scopo di verificare se, come e quanto la distribuzione dei dati si discosta dal modello teorico della curva normale.

Inoltre, in un'indagine campionaria ha poca rilevanza che le classi siano numericamente equilibrate se la categorizzazione ha lo scopo di mettere a confronto gruppi di intervistati per individuare eventuali differenze derivanti dall'appartenenza di genere, dalla scolarità o altro. In questi casi le c.d. "code della distribuzione" sono le più interessanti da analizzare ed è poco utile accorparle alle classi limitrofe solo perché la casistica è limitata, perché in tal modo non sarebbe più possibile individuare le caratteristiche distintive di questo segmento "estremo" della popolazione.

Concludiamo con un'ultima sottolineatura: come si è visto, le strategie di categorizzazione sono numerose e ognuna privilegia uno specifico criterio o metodo perché finalizzata all'ottenimento di uno specifico risultato. Ciò però non deve indurre ad applicare la strategia che si ritiene più utile in modo ottusamente acritico. Un solo esempio potrebbe essere sufficiente a spiegare cosa si intende: se si è deciso di creare gruppi di uguali dimensioni utilizzando come punti di divisione la mediana e il primo e terzo quartile, conviene osservare attentamente la distribuzione dei valori perché a volte può essere utile derogare dalla regola. Se un valore osservato che fa parte di un addensamento di casi superasse di poco il limite di classe, converrebbe comunque includerlo in quel gruppo, specialmente se ci si accorgesse che il primo caso appartenente alla classe successiva si colloca a una certa distanza. In pratica, poco importa se nella prima classe c'è un caso in più e uno in meno in quella successiva; meglio derogare dal rigoroso rispetto della regola dell'equipartizione dei casi, perché è prioritario privilegiare il criterio della similarità tra i casi appartenenti alla stessa classe, perlomeno sulla dimensione rilevata dalla variabile categorizzata.

7. Costruire indici e tipologie

7.1. Indici e/o tipologie

Come si è visto in modo articolato nel primo volume (Delli Zotti 2021), le proprietà dei casi che si intendono rilevare e poi analizzare sono sempre più o meno distanti dalla realtà sociale di cui il ricercatore e i suoi interlocutori hanno esperienza. Per definire tali proprietà si usano concetti e termini che necessitano dunque di una traduzione in indicatori, e cioè concetti più specifici e vicini alla realtà esperita che consentano la rilevazione empirica di atteggiamenti, valori, comportamenti e altre caratteristiche degli individui e degli altri tipi di unità di analisi oggetto della ricerca sociale (ripartizioni territoriali, aziende, scuole, associazioni volontarie, ecc.). Il problema della traduzione delle proprietà in indicatori, che a loro volta necessitano di una precisa definizione delle modalità della loro rilevazione empirica (definizione operativa) è generalizzabile dalla ricerca a molte altre attività che svolgiamo più o meno abitualmente nella vita quotidiana.

Ad esempio, in ambito universitario ai docenti è affidato il compito di accertare il "profitto" degli studenti, un concetto relativamente astratto che si deve però tradurre in un voto utilizzando una precisa metrica: trentesimi, nel caso di un singolo esame, o centodecimi, se si tratta della laurea. Il voto che verrà verbaliz-

zato può derivare da un insieme di valutazioni, espresse anch'esse in una qualche metrica, che si devono poi "accorpare" tramite procedure di tipo algoritmico, dal momento che gli "ingredienti" sono valori di variabili cardinali. Nel caso della laurea, al voto medio negli esami (ottenuto convertendo in base 110 il voto in trentesimi) si aggiunge un punteggio che deriva dal giudizio sulla qualità della tesi e della sua presentazione di fronte alla commissione; in alcuni regolamenti i punti assegnati dalla commissione sono integrati da bonus derivanti, ad esempio, dalla conclusione degli studi entro gli anni di corso previsti. In pratica, dopo avere effettuato delle rilevazioni su una serie di indicatori (voto agli esami, qualità della tesi e della presentazione, bonus per laurea in corso), gli esiti devono essere aggregati in un indice sintetico che fornisce una misura della proprietà che si intendeva rilevare.

Dal momento che in relazione ad alcune proprietà (genere, nazionalità, comune di residenza, professione, ecc., se si stratta di individui) possiamo solo classificare i casi all'interno di categorie, se vogliamo produrre una sintesi di questo tipo di caratteristiche dobbiamo ricorrere, invece che ad algoritmi, a una procedura di "tipizzazione", che significa comporre tali caratteristiche in un'articolata tipologia e poi assegnare ogni caso esaminato a uno dei tipi. Lo facciamo informalmente quando attribuiamo un determinato status sociale a un individuo o a una famiglia sulla base di informazioni sulle caratteristiche dell'abitazione, il titolo di studio, la professione, ecc.

Nel seguito del capitolo esporremo le soluzioni che si possono adottare se si devono combinare le informazioni contenute in due sole variabili perché può essere utile partire da una situazione semplificata, anche se molto comune, pur nella consapevolezza che la situazione si complica quando le informazioni da combinare in un indice o una tipologia sono più di due. Il compito è meno complesso se le variabili sono cardinali, perché si possono convertire i valori originali in "punteggi" da combinare in un relativamente semplice indice additivo; assai più complesso se si deve ricorrere a una procedura tipologica, perché aggiungere alle prime due variabili anche solo una variabile dicotomica porta al raddoppio delle combinazioni.

Per combinare l'informazione si possono dunque percorrere fondamentalmente due strade, a seconda delle caratteristiche delle variabili utilizzate:

- *indice*: se le variabili sono cardinali (e, con qualche limite, anche se sono ordinali) si può costruire un indice additivo, tenendo eventualmente conto del diverso "peso" che si ritiene opportuno attribuire alle variabili;
- *tipologia*: se le variabili sono categoriali, e in particolare se sono variabili nominali (classificazioni a categorie non ordinate), mediante una tabella di contingenza si esplicitano tutte le combinazioni e poi si accorpano quelle semanticamente simili, individuando in tal modo i tipi che costituiscono la tipologia.

La situazione è riassunta nella Tab. 7.1, dove sono esplicitate tutte le possibili situazioni che si verificano incrociando coppie di tre tipi di variabili (nominali, ordinali e cardinali): le combinazioni possibili sono nove, ma si possono ridurre a sei, ignorando le combinazioni speculari (*a* con *b* e *b* con *a*). Per individuare tra indice e tipologia la soluzione da adottare, si deve tenero conto che è corretto usare la tecnica appropriata alla variabile di livello inferiore; in pratica, se anche solo una delle due variabili è nominale si dovrà costruire una tipologia; se una variabile è ordinale e l'altra cardinale, si dovrà ridurre al livello ordinale la variabile cardinale e trattare i valori di entrambe come punteggi ordinali.

Per rendere concreta la spiegazione, prenderemo ad esempio l'attribuzione dello status sociale a un individuo sulla base di due sue caratteristiche rilevanti e, se si tratta di una famiglia, utilizzeremo le caratteristiche dei coniugi. La variabile nominale può essere la professione, la variabile ordinale il livello d'istruzione e quella cardinale il reddito.

Tabella 7.1 - Tecniche di combinazione di due variabili

	Nominale	**Ordinale**	**Cardinale**
Nominale	[1] indice tipologico	–	–
Ordinale	[2] indice tipologico	[4] indice additivo	–
Cardinale	[3] indice tipologico	[5] indice additivo	[6] indice additivo

Descritte sinteticamente, seguendo la numerazione indicata nella tabella, abbiamo le situazioni descritte nel seguito.

1 - Entrambe le variabili nominali

Le professioni dei coniugi andranno ricondotte a "situazioni tipo" cercando, per quanto possibile, di individuare combinazioni simili, utili a caratterizzare lo status socio-economico della famiglia.

2 - Una variabile nominale e l'altra ordinale

Per attribuire a un individuo uno status socio-economico si può usare la sua professione (variabile nominale) e il suo livello di istruzione (variabile ordinale). Sono praticabili in questo caso due soluzioni: se si ritiene che le modalità della professione non siano ordinabili, ci ritroviamo nella situazione precedente. Se è invece plausibile ordinare le categorie della variabile nominale (ad esempio, assegnando alle professioni una presuntiva capacità "bassa/media/alta" di produrre reddito), si può costruire un indice additivo.

3 - Una variabile nominale e l'altra cardinale

Se una variabile è nominale (professione) e l'altra cardinale (reddito), la situazione può essere resa simile alla precedente, perché la cardinale può essere ridotta in classi (con le tecniche viste nel capitolo precedente), perciò resa ordinale. Il risultato sarà anche in questo caso una tipologia se non si possono ordinare le professioni, un indice additivo se decidiamo di ordinarle in modo plausibile. In pratica, se la categoriale viene ordinata e la cardinale ridotta in classi, entrambe diventano ordinali e ci troviamo nella situazione che segue.

4 - Entrambe le variabili ordinali

Nel caso dei livelli di istruzione, al fine di realizzare un indice additivo si possono attribuire dei punteggi, eventualmente tenendo conto che i diversi livelli non si collocano a distanze uguali nel continuo analfabetismo/postlaurea. Si potrebbe anche tenere conto degli anni necessari per ottenere i diversi titoli di studio e di conseguenza trattare entrambe le variabili, così trasformate, come cardinali.

5 - Una variabile ordinale e l'altra cardinale

Quando una variabile è ordinale e l'altra cardinale è opportuno che le variabili siano rese "comparabili" e, visto che non si può trattare la variabile ordinale come una cardinale, la cardinale deve essere trasformata in ordinale, con un'appropriata riduzione in classi di reddito, utilizzando una delle strategie illustrate nel capitolo precedente, dedicato alla categorizzazione delle variabili. Operando in questo modo si ricade nella situazione precedente e si realizza dunque un indice additivo "sommando" i punteggi attribuiti ai livelli di reddito e ai livelli di istruzione. Dopo aver combinato reddito e istruzione di ognuno dei coniugi/genitori, si potrà procedere e costruire un indice di status familiare.

6 - Entrambe le variabili cardinali

Se entrambe le variabili sono cardinali è spesso possibile sommare i valori delle variabili (nel nostro esempio i redditi). Il reddito familiare complessivo potrà eventualmente essere ricondotto a classi di reddito, nel caso si voglia combinare questo dato con, ad esempio, il livello di scolarità della coppia di coniugi/genitori.

Si può perciò sinteticamente sostenere che, alla fine, sono sostanzialmente tre le tecniche di combinazione delle variabili:

1. se almeno una è nominale si realizza un indice tipologico, che è a sua volta una variabile categoriale;
2. se le variabili sono ordinali, o è possibile trasformare le variabili in ordinali, si può costruire un indice additivo di tipo ordinale (livelli crescenti della proprietà che si è inteso rilevare);
3. se entrambe le variabili sono cardinali è possibile costruire un indice additivo di tipo cardinale e pertanto utilizzabile per effettuare qualsiasi tipo di operazione, come il calcolo di parametri sintetici come la media.

7.2. Quattro aspetti che complicano la situazione

7.2.1. Attribuire un diverso peso alle variabili?

Nella costruzione di indici additivi la situazione è relativamente semplice se entrambe le variabili sono dello stesso tipo (ordinale o cardinale), rilevano la stessa proprietà e adottano la medesima definizione operativa; se invece le variabili rilevano proprietà diverse, vanno risolti alcuni problemi prima di poter procedere.

Se una variabile è il reddito e l'altra, utile a rilevare lo status economico, la dimensione dell'abitazione o la cilindrata dell'automobile, prima di procedere alla costruzione dell'indice si dovranno standardizzare in qualche modo le variabili. In caso contrario, il risultato finale verrà influenzato pesantemente dalle diverse unità di misura usate: il reddito mensile di un lavoratore dipendente supera normalmente il migliaio di euro, mentre la sua casa avrà probabilmente una superficie di circa un centinaio di metri quadri, quindi due unità di misura non comparabili per metrica e, in questo caso, anche con riguardo all'ordine di grandezza dei valori. Inoltre, è probabile che i redditi siano molto simili, specialmente se si conduce un'indagine ristretta a categorie con notevole omologazione salariale (ad esempio, i lavoratori dipendenti di uno specifico comparto produttivo), mentre la superficie dell'abitazione (anche se ripartita per il numero di occupanti) può essere assai più variabile, in conseguenza, ad esempio, della sua collocazione urbana o rurale.

Un modo sofisticato di operare consiste nella trasformazione dei valori originali in punteggi standard (*standard scores*), calcolando la loro distanza dalla media della distribuzione (posta pari a zero) e usando come unità di misura lo scarto-tipo (posto pari all'unità). Assai più semplicemente, si possono ridurre entrambe le variabili in classi, con le tecniche viste nel precedente capitolo. Con ciò si determina una significativa perdita di informazione di dettaglio, ma si elimina il problema del peso da attribuire alle variabili, se si ritiene che non sia molto diverso

il loro contributo nella composizione dell'indice. Si possono infatti suddividere i campi di variazione delle due variabili in tre sezioni e attribuire ad esse valori che corrispondano a livelli che convenzionalmente si possono denominare "basso", "medio" e "alto".

Le variabili si possono successivamente combinare sommando i punteggi e uno dei due punteggi potrà essere moltiplicato per un congruo coefficiente, se si ritiene che debba avere un peso diverso nella costruzione dell'indice, ritenendo sia più elevato il suo "rapporto di indicazione" con la proprietà "status socioeconomico".

7.2.2. Sommare o moltiplicare?

Ai fini della creazione dell'indice sintetico, si è finora genericamente parlato di "addizione", sottintendendo la somma di valori cardinali o di punteggi ordinali; ma i valori e i punteggi si potrebbero anche moltiplicare, e l'indice così costruito ne risentirebbe in modo significativo.

Se i valori o i punteggi vengono semplicemente sommati, si assume che le variabili debbano avere lo stesso peso nella realizzazione dell'indice, ma non solo. Infatti, non ci si pone il problema dell'eventuale incongruenza di livello: 6 punti complessivi possono infatti derivare dalla somma di 3 punti, o da una situazione assai diversa: ad esempio, 5 punti in una variabile e solo 1 punto nell'altra. Operando in modo additivo assumiamo dunque che non debba essere penalizzata l'incongruenza (nel nostro esempio incongruenza di status): ad esempio, un individuo con un alto livello di scolarità e un basso reddito (o viceversa), una situazione familiare nella quale un coniuge possiede la sola licenza elementare e l'altro è laureato, oppure uno che percepisce un reddito molto elevato e l'altro inoccupato.

Negli esempi che stiamo esaminando la somma è forse la procedura corretta, ma non sempre è così. La moltiplicazione dei punteggi non è una mera ipotesi di scuola, ma una tecnica che in alcuni casi è possibile applicare proprio allo scopo di tenere sotto controllo l'incongruenza. Qualche tempo fa i giudizi dei valutatori dei progetti di ricerca Prin (Progetti di ricerca di interesse nazionale) che entravano in competizione per il co-finanziamento del Miur (Ministero dell'università e della ricerca) venivano formulati verbalmente e poi espressi in punteggi sintetici (da 1 a 5 punti). I punti attribuiti alle due caratteristiche determinanti per la definizione della graduatoria che permetteva di accedere ai finanziamenti (originalità/contributo al progresso delle conoscenze scientifiche e chiarezza/verificabilità degli obiettivi), venivano poi moltiplicati (e non sommati).

Tabella 7.2 - Somma e moltiplicazione di punteggi

A	B	A+B	A*B
5	5	10	25
5	4	9	20
5	3	8	16
4	4	8	15
5	2	7	12
4	3	7	10
5	1	6	9
4	2	6	8
3	3	6	6
4	1	5	5
3	2	5	4
3	1	4	4
2	2	4	3
2	1	3	2
1	1	2	1

Guardando la Tab. 7.2 (nella quale non sono visualizzate le combinazioni di punteggi equivalenti: solo 1 e 2 e non anche 2 e 1), è possibile verificare l'effetto della moltiplicazione. A parità di somma, se le due valutazioni differiscono di uno o più punti, la moltiplicazione fa perdere posizioni in classifica. L'esempio più evidente è dato dal punteggio complessivo di 6: se si ottiene con 5 punti su un indicatore e 1 solo sull'altro ci si colloca piuttosto in basso nella classifica dei punteggi moltiplicati: con 4 e 2 si guadagnano 2 posizioni e con 3 e 3 un'altra.

In pratica, moltiplicando i punteggi si vuole scoraggiare l'incongruenza: nel caso del finanziamento della ricerca si vuole evitare di premiare progetti estremamente deboli su un indicatore, anche se buoni o eccellenti per un altro aspetto. Si potrebbe applicare questa tecnica anche alle valutazioni del profitto, se si vuole evitare che lo studente ignori (o quasi) parti del programma d'esame, contando di raggiungere comunque (mediamente) la sufficienza.

7.2.3. I tipi nelle tabelle di mobilità

L'incrocio di due variabili ordinali che rilevano la stessa proprietà genera una tabella con una diagonale che individua le celle con lo stesso valore su entrambe le

variabili. Nel caso le variabili siano cardinali, per metterle in relazione si potrà utilizzare un diagramma a dispersione e i casi per i quali si è rilevato lo stesso valore in entrambe le variabili si collocheranno sulla bisettrice.

Usando le variabili fin qui usate per l'esemplificazione, sulla diagonale si collocano famiglie nelle quali i coniugi hanno stessa professione, livello di istruzione o reddito. Nella Tab. 7.3 il triangolo sotto la diagonale consente di individuare le famiglie nelle quali il padre possiede un livello di scolarità più elevato; sopra la diagonale sono collocate invece le famiglie nelle quelli è la madre ad aver raggiunto un più elevato livello di istruzione[1]. Le fasce di celle parallele alla diagonale indicano un progressivo aumento di disparità nell'istruzione e le due che si trovano alla distanza massima dalla diagonale (in basso a sinistra e in alto a destra) indicano la differenza più elevata (uno dei due genitori è laureato e l'altro possiede solo la licenza elementare o non ha alcun titolo di studio).

Tabella 7.3 - Titolo di studio del padre e della madre

	Madre					
Padre	**Elementare**	**Media inf.**	**Diploma prof.**	**Media sup.**	**Laurea**	**Totale**
Elementare	2,2	2,2	0,4	-	-	4,8
Media inferiore	,9	10,0	4,8	7,0	1,3	23,9
Diploma profess.	,9	5,2	5,2	7,4	1,3	20,0
Media superiore	-	7,0	2,6	17,8	7,8	35,2
Laurea	-	-	,9	5,2	10,0	16,1
Totale	3,9	24,3	13,9	37,4	20,4	100,0

L'esempio si può generalizzare ad altre tabelle nelle quali l'identificazione della diagonale e dei due opposti triangoli può essere anche più significativa dal punto di vista sostantivo. Il caso più noto è quello delle tabelle di mobilità, che permettono di identificare le situazioni di immobilità (sulla diagonale si collocano i figli che appartengono alla stessa classe sociale dei padri o gli individui che non hanno mutato collocazione sociale nel tempo) e di mobilità verticale ascendente o discendente (intergenerazionale se cambia lo status tra genitore e figlio; intragenerazionale se cambia nel tempo lo status di uno stesso individuo).

Quello qui evidenziato non è perciò un problema da affrontare, ma un'opportunità in più offerta dalla particolare natura delle tabelle di contingenza quadrate e simmetriche che suggeriscono di costruire una semplice tipologia accorpan-

[1] I dati esposti nella tabella hanno solo valore esemplificativo e si riferiscono a un'indagine, che non è essenziale identificare, svolta da chi scrive diverso tempo fa.

do le celle in tre gruppi: quelle che compongono la diagonale e i due opposti triangoli. Nel tabelle di mobilità i tre gruppi individuano immobilità, mobilità ascendente e discendente; nel caso dello status familiare congruenza di status e situazioni di maschio o di femmina dominante (dal punto di vista professionale, culturale o economico); se si sono rilevati i redditi in due momenti diversi, le tre posizioni corrispondono al mantenimento dello stesso livello, all'aumento o al calo; e così via, tenendo conto che si può parlare di mobilità anche quando si tratta di comportamento di voto, di atteggiamenti o di opinioni.

7.2.4. Se manca l'informazione per una variabile

Un problema non banale riguarda la soluzione da adottare se manca, con riferimento al nostro esempio, una delle persone coinvolte (famiglie di singoli o stima dello status di una famiglia monoparentale). Si tratta in questi casi di decidere se rispettare il fatto che vi è un solo membro (o un solo genitore), attribuendo alla famiglia il punteggio di status di quell'unico componente, o ritenere che valga il doppio lo status dell'unica persona esistente o della quale possediamo le informazioni.

Il problema dunque si può risolvere tramite due opposti approcci:

1. *deprivazione*: le "risorse" mancanti non sono surrogabili;
2. *surrogazione*: le "risorse" del singolo possono supplire quelle del membro mancante.

Quale scelta sia la migliore dipende sostanzialmente dalla natura dell'indicatore che vogliamo creare; in pratica, da ciò che riteniamo che l'indicatore debba prioritariamente indicare. Se vi è un solo genitore (o una persona vive da sola) e si tratta di un laureato, appare corretto classificare la famiglia nella fascia dell'alta istruzione e perciò si può raddoppiare il punteggio per non penalizzarla rispetto alle famiglie nelle quali si possono sommare le lauree di entrambi i coniugi/genitori. Infatti, probabilmente non conta molto quanti siano i laureati in casa per attribuire a una famiglia un livello culturale di tipo universitario. Se ad esempio ai livelli di scolarità fondamentali nel sistema italiano (elementare, media, superiore, laurea) si attribuiscono i punteggi da 1 a 4, una famiglia nella quale è presente un solo genitore laureato riceverebbe complessivamente 4 punti, pari a quelli di una famiglia dove i coniugi possiedono solo un titolo di scuola media inferiore, oppure uno è diplomato e l'altro ha solo la licenza elementare. Per riportare i punteggi familiari all'interno del campo di variazione delle variabili originali, il totale va diviso per due, ottenendo il livello di scolarità medio, e la compensazione per le famiglie monoparentali si otter-

rebbe dividendo il punteggio totale per uno (cioè non dividendo affatto). Di fatto, si divide per il numero di addendi (coniugi/genitori) e tutto si sistema automaticamente.

Non è convincente operare allo stesso modo nel caso del reddito perché, moltiplicando per due il reddito dell'unico percettore, verrebbero attribuite artificiosamente alla famiglia monoparentale (o al singolo) risorse economiche in realtà non disponibili. Nelle famiglie (come nelle aziende) sono però realizzabili economie di scala: la valutazione dello status può perciò essere fatta utilizzando ad esempio i coefficienti suggeriti dall'Istat per individuare le soglie di povertà a seconda della numerosità del nucleo familiare. Analogamente, tornando all'istruzione, se non è corretto penalizzare l'unico laureato, è forse eccessivo moltiplicare per due il suo punteggio; perciò può essere plausibile semplicemente aggiungere un "bonus" (comunque determinato) al punteggio dell'unico componente. In effetti, è indubitabile che la socializzazione ai valori dell'eccellenza culturale può essere più efficace se i laureati in famiglia sono due; se questo è ciò che l'indicatore si prefigge di rilevare, non si può non tenerne conto.

7.3. Quattro strategie combinatorie

Il tema principale che intendiamo affrontare è quello dell'appropriato trattamento della simmetria/asimmetria tra gli indicatori che nel nostro esempio sono le caratteristiche dei coniugi/genitori, utili ad attribuire un livello di status socioeconomico o culturale alla famiglia. Si può infatti preferire che essi siano trattati come dei pari ma, oltre alla soluzione egualitaria, se ne possono ipotizzare altre, parimente legittime, riassumibili nel seguente elenco, che segue l'ordine di trattazione:

- soluzione tipologica,
- scelta prioritaria,
- soluzione egualitaria,
- soluzione elitaria.

Alcune soluzioni implicano problemi tecnici di realizzazione non sempre facili da superare, perciò i comandi di Spss che servono per ottenere il risultato deciso a livello teorico saranno illustrati anche al fine di mostrare come tradurre le nostre scelte in istruzioni tecnicamente corrette e, allo stesso tempo, quali sono le implicazioni di scelte tecniche prese a volte con poca consapevolezza.

7.3.1. Soluzione tipologica

Per illustrare la soluzione tipologica è utile riportare un articolato esempio tratto da AlmaLaurea, il consorzio interuniversitario che cura le indagini sui laureati per la maggior parte delle università italiane. AlmaLaurea elabora un indicatore di efficacia della laurea che sintetizza due importanti aspetti desunti dalle risposte fornite alle domande sul reale utilizzo delle competenze acquisite e sulla necessità (formale e sostanziale) del titolo universitario per svolgere l'attività lavorativa.

> Secondo la chiave interpretativa proposta nello schema sotto riportato (Tab. 7.4), si possono distinguere cinque livelli di efficacia:
> - "molto efficace", per gli occupati la cui laurea è richiesta per legge o di fatto necessaria, e che utilizzano le competenze universitarie in misura elevata;
> - "efficace", per gli occupati la cui laurea non è richiesta per legge ma è comunque utile e che utilizzano le competenze acquisite in misura elevata, o il cui titolo è richiesto per legge e che utilizzano le competenze in misura ridotta;
> - "abbastanza efficace", per gli occupati la cui laurea non è richiesta per legge, ma di fatto è necessaria oppure utile, e che utilizzano le competenze acquisite in misura ridotta;
> - "poco efficace", per gli occupati la cui laurea non è richiesta per legge né utile in alcun senso e che utilizzano in misura ridotta le competenze acquisite, oppure il cui titolo non è richiesto ma utile e che non utilizzano assolutamente le competenze acquisite;
> - "per nulla efficace", per gli occupati la cui laurea non è richiesta per legge né utile in alcun senso, e che non utilizzano assolutamente le competenze acquisite (AlmaLaurea 2003: 19-20).

Tabella 7.4 - Definizione dell'efficacia della laurea

Utilizzo competenze	Richiesta per legge	Necessaria	Utile	Non richiesta né utile	Non Risponde
Elevato	ME	ME	E	NC	NC
Ridotto	E	AE	AE	PE	NC
Per niente	NC	NC	PE	NE	NC
Non risponde	NC	NC	NC	NE	NC

Legenda: ME *Molto efficace* E *Efficace* AE *Abbastanza efficace*
PE *Poco efficace* NE *Per nulla efficace* NC *Non classificabile*

Fonte: Almalaurea (Condizione occupazionale dei laureati. Indagine 2020. Note metodologiche)

Riteniamo che l'esempio sia sufficientemente chiaro: le diverse combinazioni sono ricondotte ai tipi ragionando con buon senso sulla somiglianza delle situazioni (contiguità semantica). Inoltre, esporre analiticamente le scelte effettuate, come nella tabella, rende trasparente la procedura e consente ad altri di criticarle e di decidere eventualmente di operare in modo diverso. Infine, l'esempio mostra l'utilità di costruire una tipologia, che consente di individuare tutte le combinazioni possibili, salvo considerarne alcune residuali e/o improbabili.

La tipologia proposta da AlmaLaurea configura una situazione in effetti abbastanza agevole: le variabili sono ordinali e si possono dunque tendenzialmente assegnare allo stesso tipo celle contigue o celle che cumulano un analogo punteggio combinando le due informazioni. La situazione diventa invece ben più complessa se si deve creare una variabile che, combinando le professioni, indichi lo status socioeconomico di una famiglia, perché in questo caso le celle della tabella sono solo accidentalmente contigue (le modalità si potevano infatti disporre in un ordine diverso).

Tabella 7.5 - Professione del padre e della madre

	Padre		**Madre**	
	N	**%**	**N**	**%**
Imprenditore, libero prof.	58	24,3	20	8,4
Autonomo	26	10,9	15	6,3
Impiegato	63	26,4	71	29,7
Insegnante	11	4,6	31	13,0
Operaio	35	14,6	18	7,5
Casalinga	-	-	60	25,1
Altro	39	16,3	17	7,1
Totale	232	100,0	232	100,0

Tabella 7.6 - Combinazioni delle professioni del padre e della madre

				Madre				
Padre	**Imprend.**	**Auto-nomo**	**Impieg.**	**Insegn.**	**Ope-raio**	**Casa-linga**	**Altro**	**Tot.**
Impr., l. pro.	4,4	1,8	4,8	4,4	,4	7,9	1,8	25,6
Autonomo	,9	2,2	2,6	1,3	1,3	3,1	-	11,5
Impiegato	1,8	,4	15,0	2,6	1,3	5,3	,4	26,9
Insegnante	,4	-	,4	2,2	-	-	,9	4,0
Operaio	,4	1,3	4,0	1,8	4,0	3,1	,4	15,0
Altro	,9	,9	3,5	1,3	,9	5,7	4,0	17,2
Totale	8,8	6,6	30,4	13,7	7,9	25,1	7,5	100,0

Nelle variabili originarie le professioni sono classificate in modo piuttosto analitico (7 modalità per le femmine e 6 per i maschi, mancando per loro la categoria "casalinga"), come si vede dalle due distribuzioni di frequenza (Tab. 7.5). L'incrocio delle due informazioni in una tabella di contingenza crea perciò una tipologia iniziale articolata in 42 celle, alcune delle quali sono vuote (Tab. 7.6). Ciò permette di identificare con una certa precisione situazioni analoghe ed eventualmente, con l'ausilio di ulteriori informazioni (ad esempio il titolo di studio), di definire ulteriormente lo status sociale delle famiglie degli studenti.

Ad esempio, si potrebbero distinguere gli impiegati laureati e attribuire loro uno status più elevato rispetto agli impiegati che hanno solo il titolo di scuola media inferiore o superiore. Ciò esula largamente dai propositi di questo scritto; perciò accorperemo le modalità delle variabili in modo da creare una tipologia semplificata. Con RECODE i valori di "profp" e "profm" sono riassunti in 4 modalità e le trasformazioni attribuite (INTO) alle nuove variabili "profprec" e "profmrec".

```
RECODE profp profm (1 2 = 1) (3 4 = 2) (5 = 3) (6 7 = 4) INTO profprec profmrec.
```

Tabella 7.7 - Professione del padre e della madre

	Madre				
Padre	**1 Autonoma**	**2 Impiegata**	**3 Operaia**	**4 Inoccupata**	**Totale**
1 Autonomo	9,3	13,2	1,8	12,8	37,0
2 Impiegato	2,6	20,3	1,3	6,6	30,8
3 Operaio	1,8	5,7	4,0	3,5	15,0
4 Inoccupato	1,8	4,8	,9	9,7	17,2
Totale	15,4	44,1	7,9	32,6	100,0

Nella Tab. 7.7 si possono individuare tutte le combinazioni di situazioni professionali e la nuova variabile può essere costruita mediante una serie di comandi IF. Sette combinazioni comprendono almeno un lavoratore autonomo ed è sufficiente un comando per attribuire a tutte il codice 2; si può infatti porre semplicemente la condizione che sia autonomo il padre oppure (OR) la madre.

```
IF (profprec = 1 OR profmrec = 1) profgen = 2.
```

Tra queste famiglie ci sono anche quelle dove *entrambi* i genitori sono autonomi; ad esse si può voler attribuire il codice 1 per indicare che si collocano al primo posto nell'indicatore di status.

```
IF (profprec = 1 AND profmrec = 1) profgen = 1.
```

Può sembrare contro-intuitivo che si sia proceduto "a ritroso", ma invertendo i comandi non si ottiene il risultato auspicato. Se si definiscono prima le famiglie dove entrambi i coniugi sono lavoratori autonomi e poi quelle dove lo è almeno uno, si vanifica il primo comando perché alla categoria "almeno un autonomo" appartengono anche quelle dove lo sono entrambi; se invece si definiscono prima tutte le famiglie con almeno un autonomo, con il successivo e più restrittivo comando si cambia lo status solo a quelle in cui lo sono entrambi.

Seguono cinque tipi di famiglie nelle quali c'è almeno un impiegato e nessun lavoratore autonomo; si tratta di situazioni un po' più complesse da definire e, anche per evitare di fare errori, può essere più semplice scrivere cinque comandi, accomunati dal codice "2" per la professione di almeno uno dei coniugi/genitori.

```
IF (profprec = 2 AND profmrec = 2) profgen = 3.
IF (profprec = 2 AND profmrec = 3) profgen = 3.
IF (profprec = 2 AND profmrec = 4) profgen = 3.
IF (profprec = 3 AND profmrec = 2) profgen = 3.
IF (profprec = 4 AND profmrec = 2) profgen = 3.
```

Spesso i comandi si possono semplificare facendo riferimento a insiemi di celle e/o escludendone altre con un uso accorto degli operatori AND e OR. I cinque comandi potrebbero infatti essere sostituiti dai due seguenti: con il primo si assegna a "profgen" il valore 3 se la professione del padre è 2 (impiegato) e quella della madre non uguale a 1 (non è un'autonoma); con la seconda istruzione si assegna il valore 3 se invece il padre non è un autonomo e la madre è impiegata.

```
IF (profprec = 2 AND profmrec NE 1) profgen=3.
IF (profprec NE 1 AND profmrec = 2) profgen=3.
```

Con ulteriori comandi si attribuisce infine il codice "4" alle famiglie nelle quali almeno un componente è operaio e nessuno è impiegato o autonomo e il codice "5" alle famiglie nelle quali entrambi i componenti sono inoccupati, ottenendo alla fine la variabile la cui distribuzione di frequenza è mostrata nella Tab. 7.8.

Tabella 7.8 - Status professionale dei coniugi/genitori

	N.	%	% cumulate
Entrambi autonomi	21	9,3	13,8
Almeno un autonomo	77	33,9	43,2
Almeno un impiegato	88	38,8	81,9
Almeno un operaio	19	8,4	90,3
Entrambi inoccupati	22	9,7	100,0
Totale	227	100,0	

Come si è visto, si ottiene il risultato voluto con una serie piuttosto lunga di comandi che devono essere scritti prestando particolare attenzione. È possibile però ottenere lo stesso risultato ricorrendo a una tecnica proposta in altra sede (Delli Zotti, 1999), che risulta ancor più utile quando si devono individuare combinazioni di tre o più variabili. Si tratta in pratica di realizzare una lista di tutte le combinazioni riscontrate nella matrice dei dati che sono più facilmente aggregabili se sono elencate in un'unica variabile. Invece di partire da una tabella a doppia entrata, si costruisce una nuova variabile "accostando" la professione del padre e della madre; in pratica, con COMPUTE si crea la nuova variabile "combgen" sommandolo il valore attribuito alla professione della madre a quello del padre, moltiplicato per 10 al fine di spostarlo sulla colonna delle decine.

```
COMPUTE combgen = profprec * 10 + profmrec.
```

Procedendo in questo modo si ottengono nella variabile "combgen" valori quali, ad esempio, "11", che corrisponde alle 21 famiglie in cui entrambi i genitori sono lavoratori autonomi; "12" e "21" che corrispondono alle 30 famiglie (20+5) nelle quali un genitore è lavoratore autonomo e l'altro impiegato; e così via fino a "44", che indica famiglie nelle quali entrambi i coniugi/genitori sono inoccupati (22 casi). Questi valori si possono successivamente accorpare con un'istruzione RECODE per ottenere alla fine le stesse situazioni familiari della Tab. 7.8), ma in modo più semplice e intuitivo che con la lunga serie di IF sopra descritta.

```
RECODE combgen (11 = 1) (12 13 14 21 31 41 = 2) (22 23 24 32 42 = 3) (33 34 43
    = 4) (44 = 5) INTO profgen.
```

7.3.2. Scelta prioritaria

Se si ritiene che la professione del padre sia prioritaria nel definire lo status della famiglia, si può tenere conto solo di questa e utilizzare la professione della madre solo in caso di mancanza del padre (famiglia monoparentale). La soluzione si può definire "maschilista" (o del maschio dominante) e può apparire inaccettabile ma, per specifiche finalità e in alcune situazioni, può essere semplicemente realistica. È appena il caso di ricordare che solo pochi decenni fa era questa la situazione vigente in Italia, anche a causa di un minore accesso al lavoro da parte delle donne. Anche oggi, del resto, basterebbe spostarsi in qualche altro paese del mondo o presso altre culture per trovare assetti sociali nei quali il maschio è "*dominus*" nel definire lo status sociale della famiglia e sarebbe dunque appropriato tenere conto prioritariamente della sua professione.

Che la soluzione della "scelta prioritaria" possa essere neutrale dal punto di vista valoriale lo testimonia il fatto che può rispondere alle finalità di ricerca una soluzione "femminista" se, ad esempio, si intende creare un indice che tenga conto del livello di istruzione cui è stato esposto un bambino, visto che il rapporto con la madre può essere considerato più influente sulla socializzazione dei figli, in particolare nei primi anni di vita. Anche in questo caso, in mancanza della madre, si può usare il livello d'istruzione del padre (o magari della nonna o del nonno).

Al fine di mantenere l'unitarietà dell'esemplificazione, che consentirà in conclusione di confrontare i risultati ottenuti con le varie soluzioni, opereremo anche in questo caso utilizzando le professioni. Con il comando COMPUTE si pone innanzitutto la nuova variabile "profgen2" pari alla professione del padre (profp). Con il comando IF si adotta poi il correttivo: se la professione del padre è pari al valore "4" (inoccupato o risposta mancante), la professione di riferimento per definire lo status (profgen2) è quella della madre.

```
COMPUTE profgen2 = profp.
IF (profprec eq 4) profgen2 = profmrec.
```

Ovviamente, nel caso si decida che deve avere priorità la professione della madre, i comandi andranno modificati di conseguenza. Notiamo infine che la soluzione "scelta prioritaria con sostituzione" presenta un non trascurabile vantaggio: si riescono a classificare anche le famiglie per le quali manca l'informazione per uno dei due genitori.

7.3.3. Soluzione egualitaria

Anche se può essere a volte accettabile tenere conto prioritariamente del profilo socio-economico di uno dei due coniugi/genitori, in generale è più convincente trovare una soluzione, in linea con la sensibilità sociale e il vigente diritto di famiglia, che metta la coppia su un piano di uguaglianza. Per costruire un indice di status cui ognuno dei coniugi/genitori contribuisce con le sue "risorse" è necessario però che le variabili siano almeno ordinali per poter legittimamente sommare (ed eventualmente dividere per due) i valori o i punteggi attribuiti alle modalità ordinate delle variabili, altrimenti la soluzione egualitaria rischia di essere tecnicamente impraticabile e poco convincente dal punto di vista sostantivo.

L'operazione è senz'altro difendibile se si intende utilizzare l'istruzione, perché è possibile attribuire un punteggio crescente ai diversi titoli di studio, tenendo ad esempio conto del numero di anni necessari per il loro conseguimento (cinque anni/punti per la licenza elementare, otto per la licenza media e così via). In questo modo l'indice di status familiare si potrebbe ottenere facilmente, semplicemente sommando i due punteggi.

Nel caso delle professioni l'operazione sarebbe meno convincente perché, aggregandole in poche generiche categorie, otterremmo eterogenee aggregazioni viziate da un'eccessiva variabilità interna. Se comunque accettiamo la relativa forzatura, e riteniamo plausibile che i lavoratori autonomi (imprenditori, liberi professionisti e artigiani) siano in grado di acquisire maggior reddito (perciò più elevato status sociale) dei lavoratori dipendenti (operai e impiegati) e degli inoccupati, possiamo, ad esempio, attribuire 2 punti agli autonomi, 1 punto ai dipendenti (senza distinzione tra impiegati e operai) e 0 agli inoccupati. Sommando i due punteggi, si otterrà poi la distribuzione di frequenza della variabile nella parte destra della Tab. 7.9, con valori da "0" (entrambi i coniugi/genitori sono inoccupati) a "4" (entrambi lavoratori autonomi), che si potrà poi ricodificare in 3 livelli di status ("basso", "medio" e "alto") nella variabile ordinale "statusgen".

Tabella 7.9 - Punteggi di status dei genitori prima e dopo la ricodifica

	N	%	% cum.
0	22	9,7	9,7
1	36	15,9	25,6
2	104	45,8	71,4
3	44	19,4	90,7
4	21	9,3	100,0
Totale	227	100,0	

	N	%	% cum.
Basso	58	25,6	25,6
Medio	104	45,8	71,4
Alto	65	28,6	100,0
Totale	227	100,0	

```
RECODE profprec profmrec (1 = 2) (2 3 = 1) (4 = 0) INTO puntiprofp puntiprofm.
COMPUTE profgen3 = puntprofp + puntprofm.
RECODE profgen3 (0 1 =1) (2=2) (3 4=3) INTO statusgen.
```

7.3.4. Soluzione elitaria

A parte le semplificazioni nell'attribuzione dei punteggi, la soluzione egualitaria potrebbe essere ritenuta pienamente soddisfacente, anche perché equa, ma l'equità non dovrebbe essere considerata un criterio di scelta tra le diverse strategie perché dovrebbe essere molto più importante chiedersi cosa l'indicatore dovrebbe, per l'appunto, indicare, e trovare una soluzione valida che dia un risultato attendibile.

Che si tratti di professione o di scolarità, riteniamo più convincente un'ultima soluzione, che si potrebbe denominare "elitaria", rischiando di urtare qualche sensibilità: si tiene conto della professione di status più elevato ed è indifferente che sia quella del padre o della madre. Questa soluzione è la più difendibile perché, in caso di asimmetria, si può ritenere che la posizione sociale possa essere maggiormente influenzata dal coniuge socialmente dominante (da cui la possibilità di definirla "soluzione della dominanza")[2].

Se un imprenditore ha sposato una casalinga, è poco convincente diminuire lo status della famiglia calcolando la media tra il punteggio dei lavoratori autonomi e degli inoccupati. Si pensi anche al livello di istruzione: se una laureata dovesse sposare un uomo con un basso livello di istruzione, è ragionevole che la famiglia possa essere collocata al top dello status socio-culturale, specialmente se, ad esempio, si intende utilizzare la variabile come predittore delle scelte scolastiche dei figli. È infatti probabile che la madre laureata preveda un futuro da studente universitario per i figli e che il padre con più basso livello di istruzione non si opponga alla scelta, piuttosto che una mediazione che abbia come esito una poco aurea *mediocritas* nelle scelte scolastiche.

Si tratta del resto di alternative presenti anche in altri ambiti: in alcuni sport si calcola la media delle prestazioni, come il punteggio totale ottenuto in una serie di tuffi dal trampolino; in altri si cerca l'eccellenza, e non importa la lunghezza degli altri salti se, come nel salto in lungo, si viene classificati sulla base di quello più lungo che si è riusciti ad effettuare.

La soluzione della dominanza è abbastanza facile da realizzare tecnicamente, ma i due comandi IF vanno considerati con attenzione. Per ottenere l'effetto vo-

[2] Il principio della dominanza, proposto da Erikson (1984) e utilizzato da Cobalti e Schizzerotto nelle loro indagini sulla mobilità (1994) è stato poi adottato anche dall'Istat (ad es., nelle *Indagini Multiscopo sulle famiglie italiane*) e dalla stessa AlmaLaurea (2015).

luto, abbiamo infatti dovuto modificare le istruzioni rispetto a quanto è intuitivo perché la codifica da autonomo (1) a inoccupato (4) attribuisce un valore numerico più basso allo status più elevato. Perciò, con le istruzioni si deve scegliere il genitore con la professione cui è stato attribuito il valore numerico inferiore (cui però corrisponde uno status più elevato): con il primo IF si attribuisce a "profgen4" la professione del padre, se ha un valore inferiore o uguale a quello della madre; con il secondo si attribuisce la professione della madre, se quella del padre ha un valore superiore (ma di status inferiore).

```
IF (profprec LE profmrec) profgen4 = profprec.
IF (profprec GT profmrec) profgen4 = profmrec.
```

Tabella 7.10 - Professione più elevata dei genitori

	N	%	% cumulate
Autonomo/a	98	43,2	43,2
Impiegato/a	88	38,8	81,9
Operaio/a	19	8,4	90,3
Inoccupato/a	22	9,7	100,0
Totale	227	100,0	

Di ciò si era tenuto conto anche nella soluzione precedente: alla professione "1" sono stati attribuiti due punti di status, mentre agli inoccupati (codice "4") sono stati attribuiti 0 punti. Ciò è accaduto perché nelle variabili nominali i valori possono essere attribuiti liberamente alle categorie; con le variabili ordinali i valori devono invece essere congruenti con la dimensione semantica sottostante all'ordinamento.

7.4. Conclusioni

Per controllare che le trasformazioni siano state effettuate correttamente è sempre opportuno realizzare un riepilogo dei casi che visualizzi le due variabili di partenza e la tipologia e/o gli indici di status risultanti dall'esecuzione delle istruzioni. Nel nostro caso il quadro riassuntivo servirà anche a confrontare le differenze negli esiti finali derivanti dalle diverse soluzioni.

Osservando la prima riga della Tab. 7.11 si vede una famiglia in cui entrambi i coniugi/genitori sono lavoratori autonomi che, ovviamente, qualunque genitore si consideri, viene classificata tra le famiglie di lavoratori autonomi e i 4 punti

di status le attribuiscono la categoria di status "alto". Analogamente la famiglia costituita da due inoccupati, che riceve un punteggio di status pari a 0 (definito "basso"), viene classificata tra le famiglie di inoccupati, qualunque dei due genitori venga considerato.

Nella penultima riga troviamo invece una famiglia nella quale il padre è operaio e la madre impiegata: la famiglia è assegnata al tipo "(almeno) un impiegato" con la prima soluzione; classificata come "operaia" con la soluzione del maschio dominante; riceve un punteggio di status pari a 2, poi riclassificato in scala ordinale come "medio", con la soluzione egualitaria; infine, è collocata nella classe impiegatizia, se si tiene conto della professione più elevata in termini di status.

Tabella 7.11 – Prospetto della situazione familiare risultante dalle varie opzioni

Padre	Madre	Tipologia status prof.	Combi-naz. prof.	Prof. padre/ madre	Punti	Status	Più Elevata
Auton.	Auton.	2 auton.	11	Auton.	4	Alto	Auton.
Auton.	Impieg.	1 auton.	12	Auton.	3	Alto	Auton.
Auton.	Impieg.	1 auton.	12	Auton.	3	Alto	Auton.
Auton.	Inocc.	1 auton.	14	Auton.	2	Medio	Auton.
Auton.	Inocc.	1 auton.	14	Auton.	2	Medio	Auton.
Auton.	Operaia	1 auton.	13	Auton.	3	Alto	Auton.
Impieg.	Auton.	1 auton.	21	Impieg.	3	Alto	Auton.
Impieg.	Impieg.	2 impieg.	22	Impieg.	2	Medio	Impieg.
Impieg.	Inocc.	1 impieg.	24	Impieg.	1	Medio	Impieg.
Impieg.	Inocc.	1 impieg.	24	Impieg.	1	Medio	Impieg.
Inocc.	Auton.	1 auton.	41	Auton.	2	Medio	Auton.
Inocc.	Inocc.	2 inocc.	44	Inocc.	0	Basso	Inocc.
Operaio	Auton.	1 auton.	31	Operaio.	3	Alto	Auton.
Operaio	Impieg.	1 impieg..	32	Operaio	2	Medio	Impieg.
Operaio	Inocc.	1 operaio	34	Operaio	1	Medio	Operaio

Sebbene quest'ultima tabella sia stata inserita principalmente per mostrare al lettore le conseguenze delle diverse strategie classificatorie, riteniamo sia utile anche al fine di suggerire una soluzione pratica ai lettori che hanno poca dimestichezza con l'utilizzo delle istruzioni di Spss che, come abbiamo visto, a volte sono perfino controintuitive.

Nelle prime due colonne della Tab. 7.11 sono riportati i valori originali delle variabili utilizzate per costruire la tipologia o l'indice di status familiare e la so-

luzione pratica potrebbe proprio consistere nell'ispezione visiva delle coppie di professioni e nell'attribuzione manuale di un codice o un valore a ognuna delle combinazioni che può essere inserito manualmente in una colonna opportunamente inserita a destra delle prime due. In pratica, dopo avere proceduto a un ordinamento della professione del padre, e poi della madre (come abbiamo fatto nel nostro esempio, limitato ai primi 15 casi della matrice), e dopo avere deciso quale strategia adottare, nelle prime righe ci si troverà di fronte a tutte le famiglie dove, ad esempio, entrambi i genitori sono lavoratori autonomi. Deciso il codice da attribuire, verrà digitato nella matrice dei dati e, con un semplice "copia e incolla", assegnato a tutte le altre famiglie che si trovano nella stessa situazione, per poi passare alle famiglie dove il padre è autonomo e la madre impiegata, e così via. In tal modo, con pochi passaggi verranno classificate manualmente e/o sarà attribuito un punteggio di status a tutte le famiglie.

8. Costruire e analizzare tabelle di contingenza multiple

8.1. Introduzione

Nella ricerca empirica può essere utile esplorare il contenuto informativo di una matrice dei dati anche senza una teoria che guidi nel controllo di relazioni tra le variabili che il ricercatore abbia ipotizzato a priori. Ad esempio, può essere opportuno incrociare sistematicamente atteggiamenti, opinioni o comportamenti con informazioni socio-demografiche quali genere, età, livello di scolarità, ecc., procedendo in pratica a una sorta di "*fishing expedition*". Questo modo di operare non serve solo a rimediare a eventuali carenze di impostazione teorica, perché talvolta la ricerca può semplicemente ambire a descrivere un fenomeno mal conosciuto sul quale si raccolgono il maggior numero di informazioni possibili. Successive ricerche potranno poi essere maggiormente focalizzate, proprio perché guidate da ipotesi formulate sulla base di quella preliminare analisi esplorativa.

Del resto, controllare e comparare un numero elevato di tabelle di contingenza, di medie o di coefficienti di correlazione può avere un notevole valore euristico anche quando le ipotesi di partenza dell'indagine sono ben specificate. Come insegnano Merton e Barber (2002), la *serendipity* (scoperta casuale)

ha spesso svolto un ruolo importante nello sviluppo e progresso delle scienze (sociali e non). Il ricercatore parte con l'intendimento di cercare qualcosa di specifico, ma nel corso della ricerca possono emergere, in maniera del tutto inaspettata e a volte provvidenziale, nessi fra variabili che il mero controllo di relazioni ipotizzate a priori non avrebbe forse consentito di rilevare. Del resto, proprio il riconoscimento di queste esigenze e opportunità ha fornito la spinta allo sviluppo di tecniche di analisi esplorativa ben più sofisticate, come l'analisi delle corrispondenze multiple (si veda, ad esempio, Di Franco 2006) e l'analisi esplorativa di Tukey (1977), che ha tra l'altro proposto i grafici box-plot che abbiamo descritto e utilizzato nei capitoli precedenti.

In effetti, quando si fa ricerca si possono commettere errori del I e del II tipo (Schwartz e Jacobs 1987: 345): la pratica più consueta prevede che, dopo aver individuato una possibile relazione statistica tra variabili, si effettuino controlli per evitare di enunciare una relazione tra due variabili (fenomeni) che non esiste (errore del I tipo). Cercando di minimizzare la possibilità di fare errori del I tipo si rischia però di farne del II tipo (non scoprire una relazione che invece esiste). Da ciò discende dunque la possibilità di distinguere tra:

> … ricerche *descrittive* e ricerche *esplicative*, riconducibili l'una al "contesto della scoperta", l'altra al "contesto della giustificazione": nel primo il ricercatore, osservando la realtà sociale, induce delle ipotesi (usualmente di tipo causale) che richiedono poi di essere validate attraverso ricerche esplicative che rientrano nel secondo contesto, quello della giustificazione (Palumbo 2012: 105).

In realtà, i due tipi di ricerca possono essere compresenti e spesso «la ricerca descrittiva fornisce la base per la formulazione di ulteriori ipotesi di natura interpretativa o esplicativa» (Agnoli 2003: 87). Effettuare analisi descrittive/esplorative può perciò essere utile, anche se non legittima la negligenza perché, se possibile, è sempre opportuno formulare un ipotetico modello esplicativo, utile se non altro a orientare la costruzione di un appropriato strumento di rilevazione dei dati (questionario o altro).

Se il livello di misurazione delle variabili lo permette, l'esplorazione può essere condotta, oltre che con tecniche più sofisticate, quali la citata analisi delle corrispondenze multiple, mediante compatte matrici di coefficienti di correlazione. Se le variabili sono cardinali l'analisi esplorativa si può effettuare anche scomponendo le medie secondo i valori di alcune variabili categoriali, perché si possono produrre tabelle altrettanto "compatte", come si è visto, ad esempio, con la Tab. 4.6 nel capitolo dedicato all'analisi multivariata.

Se però le variabili da mettere in relazione sono categoriali, si è costretti a condurre l'analisi esplorativa mediante tabelle di contingenza e ciò implica la

produzione di tabulati "sterminati", perché le variabili si possono incrociare solo a coppie[1]. La procedura RISPOSTE MULTIPLE di Spss, e il modulo opzionale TABELLE PERSONALIZZATE, permettono però di riassumere in un'unica tabella le informazioni contenute in numerose tabelle singole; un miglioramento, apparentemente solo quantitativo, che si può tradurre in un vero e proprio salto di qualità nell'analisi. Incrociare in un'unica tabella le risposte a una batteria di domande a scelta dicotomica con una serie di variabili indipendenti può infatti consentire confronti tra i dati realizzabili con maggiore difficoltà se si dovessero analizzare pagine e pagine di singole tabelle a doppia entrata.

Come si può intuire, le tabelle multiple possono essere molto complesse ed è dunque necessario selezionare accuratamente le informazioni da tabulare per non correre il rischio di renderle inadatte all'analisi esplorativa. Come si è già visto (§ 3.3.3), se si utilizzano le percentuali come strumento di analisi si possono tralasciare le differenze (residui) tra frequenze assolute attese e rilevate. Questa è la scelta più consueta, ma può essere altrettanto "illuminante" analizzare i residui che, nel caso di un sondaggio, aiutano a individuare caratteristiche socioeconomiche degli intervistati che attraggono determinati loro comportamenti o atteggiamenti (se i residui sono ridotti al minimo) o li respingono (residui elevati). Si deve comunque scegliere e, volendo, si possono anche provare entrambe le strategie (usare le percentuali oppure i residui), ma separatamente.

A seguire, verranno illustrati i quattro tipi fondamentali di tabelle di contingenza multiple e le modalità per la loro corretta realizzazione. L'insieme delle variabili indipendenti (ipotetiche cause dei fenomeni analizzati) possono essere infatti incrociate con:

1. una domanda a risposta singola;
2. una domanda a risposta multipla;
3. un set di domande a risposta dicotomica;
4. un insieme di parametri statistici.

[1] Nel paragrafo dedicato al controllo delle terze variabili (§ 4.1.1) abbiamo in effetti descritto tabelle "a tripla entrata", ma la terza variabile era in quel caso "annidata" e non "affiancata", come nelle tabelle qui illustrate che rispondono a una diversa finalità di analisi. Si trattava in quel caso di modelli semplificati di analisi multivariata, mentre qui ci si propone di comparare gli esiti di analisi che, prese singolarmente, hanno natura bivariata.

Il primo tipo di tabella multipla è un po' diverso dagli altri che esamineremo nel seguito perché la "molteplicità" si presenta in questo caso solo sul versante delle variabili indipendenti. Questo tipo di tabella non presenta alcuna complessità concettuale, perché si realizza collocando semplicemente una sopra l'altra una serie di singole tabelle di contingenza a doppia entrata. Nell'esempio che proponiamo (Tab. 8.1) nelle colonne sono collocate le risposte alla variabile dipendente (l'interesse per la politica) e nelle righe i gruppi sociali individuati da cinque variabili indipendenti: genere, età, livello di scolarità, posizione lavorativa e appartenenza politica.

Questo tipo di tabelle si possono realizzare con relativa facilità usando un software specifico per l'analisi statistica come Spss. Si può però arrivare allo stesso risultato anche producendo separatamente le singole tabelle a doppia entrata con, ad esempio, il modulo analisi dati di Excel[2]. Si userà poi lo stesso foglio elettronico per "accostare" le tabelle, dopo avere eliminato la riga dei totali (lasciando solo l'ultima, perché i totali non variano da una tabella all'altra).

Dal semplice esempio che proponiamo (Tab. 8.1) è facile notare il vantaggio di questo tipo di tabulazione compatta: se ci si concentra, ad esempio, sulla ristretta minoranza di intervistati che si sono dichiarati molto interessati alla politica, scorrendo dall'alto in basso la relativa colonna della tabella si constata facilmente che sono relativamente più numerosi tra i maschi, i laureati e i lavoratori autonomi. Per quanto concerne l'età, si nota un maggiore interesse tra i più anziani, ma la differenza non è molto ampia e si attenua ulteriormente se si sommano le risposte "molto" e "abbastanza". Nel caso della collocazione politica sulla dimensione sinistra-destra, l'interesse si attenua tra chi si colloca al centro ed è più accentuato, confrontando i due poli, tra chi si colloca a sinistra. La tabulazione compatta consente anche di notare che la caratteristica più discriminante è il livello di scolarità, come si può constatare anche osservando l'estrema divaricazione delle percentuali che si registra sul versante opposto (risposte "poco" e "per nulla").

[2] Excel contiene in effetti un modulo di analisi statistica che consente di produrre distribuzioni di frequenza, tabelle di contingenza e calcolare statistiche descrittive mono e bivariate. Il modulo non viene installato automaticamente, ma è presente e si può facilmente "attivare" facendo clic sulla scheda File, quindi su Opzioni/Componenti aggiuntivi.

Tabella 8.1 – Tabulazione di variabili indipendenti con una dipendente (Italia 2017)

		Quanto interessato alla politica						
		Molto	Abbast.	Poco	Nulla	NR	Totale	
		%	%	%	%	%	%	N
Genere	Maschio	7,4	29,3	38,1	25,0	,2	100,0	1.298
	Femmina	3,7	20,6	37,9	37,5	,3	100,0	1.447
Età	16-34	3,5	21,3	39,8	35,2	,2	100,0	633
	35-64	5,6	28,3	39,2	26,5	,4	100,0	1.305
	65-90	6,9	21,5	34,7	36,9	,0	100,0	786
	NR	,0	23,8	28,6	38,1	9,5	100,0	21
Scolarità	Obbligo	3,3	14,9	37,5	44,1	,2	100,0	1.217
	Diploma	5,1	29,3	40,8	24,7	,2	100,0	1.131
	Laurea	13,6	43,0	30,7	12,0	,5	100,0	374
	NR	,0	21,7	47,8	26,1	4,3	100,0	23
Occupazione	Autonomo	8,2	29,2	39,1	23,5	,0	100,0	473
	Dipendente	5,8	27,6	38,8	27,5	,4	100,0	1.593
	Non occupato	2,7	14,7	35,0	47,3	,3	100,0	658
	NR	,0	19,0	47,6	33,3	,0	100,0	21
Collocazione politica	Sinistra	11,4	35,7	33,4	19,3	,2	100,0	446
	Centro	4,9	24,7	41,6	28,6	,2	100,0	935
	Destra	7,1	30,6	42,4	19,7	,2	100,0	620
	NR	1,1	13,2	32,5	52,7	,5	100,0	744
Totale		5,4	24,7	38,0	31,6	,3	100,0	2.745

Fonte: European Social Survey (http://www.europeansocialsurvey.org)

Mantenere nella tabella la categoria "non risposto" (NR), presente in tutte le variabili indipendenti a esclusione del genere, può essere utile perché di questi intervistati "reticenti" riusciamo in questo modo a sapere almeno quanto sono interessati alla politica (in larga maggioranza "poco" o "nulla"). Se però si decidesse di togliere le mancate risposte quando i casi sono pochi, non sarebbe necessario ricalcolare le percentuali perché ogni categoria socio-anagrafica è indipendente dalle altre. Eliminando la mancate risposte dalla variabile dipendente, le percentuali si dovrebbero invece ricalcolare, perché si ridurrebbe la loro base di calcolo. Nel nostro esempio si tratta però di solo otto casi su 2.745 (0,3%) e pertanto il ricalcolo comporterebbe un cambiamento delle percentuali del tutto trascurabile. Come già anticipato nel § 3.3.3, e si argomenterà più analiticamente nel prossimo capitolo, eliminando questi pochi dati mancanti si ottiene il vantaggio

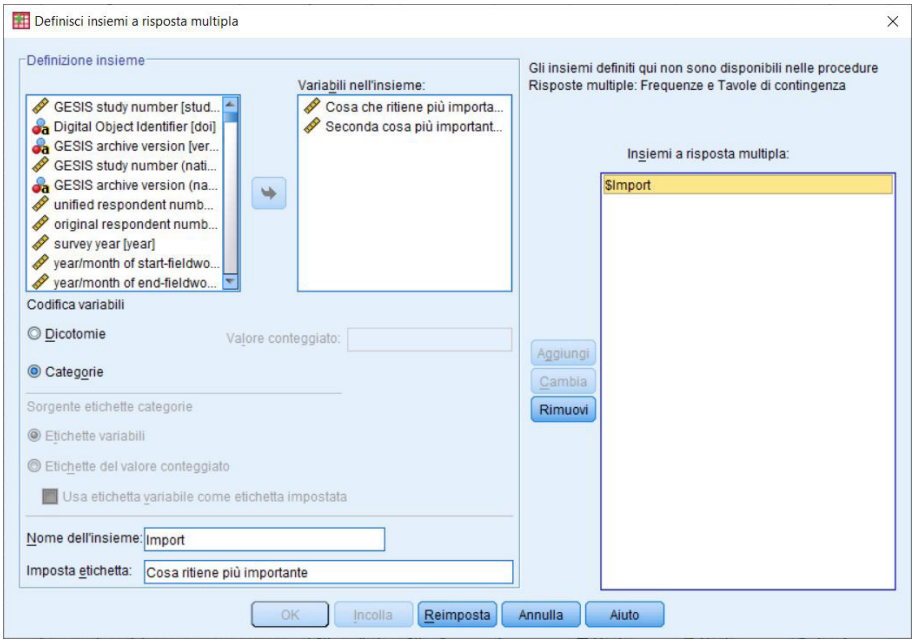
della semplificazione perdendo molto poco in termini di precisione. Il ragionamento è valido a maggior ragione nel caso di dati derivano da un'indagine che, essendo campionaria, comporta non trascurabili margini di incertezza statistica anche in campioni di cospicue dimensioni, quali quelli utilizzati nelle nostre esemplificazioni.

8.3. Incrocio con una domanda a risposta multipla

In questo paragrafo esamineremo una tabella nella quale la molteplicità si manifesta anche sul versante della variabile indipendente, un "insieme a risposta multipla" che così si definisce perché deriva da una domanda che consente di fornire più di una risposta tra quelle elencate in una lista. Anche se talvolta l'intervistato è lasciato libero di indicare anche tutte le risposte, solitamente viene fissato un limite: ad esempio, al massimo tre risposte che poi si registrano in tre colonne della matrice dei dati. Come è intuibile l'analisi diventa più complessa, anche se si vuole semplicemente sapere quanti intervistati hanno fornito ognuna delle risposte (analisi monovariata). Dal momento che ognuna delle risposte può essere registrata nella prima, seconda o terza colonna, a seconda che sia stata indicata per prima, seconda o terza, per conoscere la sua frequenza si dovrà contare quante volte è stata registrata in ognuna delle colonne e poi sommare. Si dovranno dunque elaborare le distribuzioni di frequenza delle tre variabili e poi sommare le occorrenze di ogni risposta con una calcolatrice o un foglio elettronico. Ovviamente, le elaborazioni diventano più complesse se si analizzano le risposte secondo le categorie di una variabile indipendente (analisi bivariata), perché si devono sommare i contenuti di tutte le celle di due, tre (o più) tabelle.

Con software di analisi statistica come Spss si possono invece produrre direttamente le distribuzioni di frequenza e le tabelle dopo avere costruito un insieme a risposta multipla tramite il modulo risposte multiple, contenuto all'interno del programma base, oppure con il modulo opzionale tabelle personalizzate. Tramite la finestra definisci insiemi a risposta multipla (Fig. 8.1) si scelgono le variabili da tabulare congiuntamente che poi si spostano nella colonna a destra. Bisogna poi selezionare il bottone categorie, dare un nome e un'etichetta all'insieme e "cliccare" aggiungi. Si possono poi costruire altri insiemi (anche del tipo "dicotomie" che esamineremo nel prossimo paragrafo) da utilizzare al posto della domanda a "risposta singola" in tabelle simili a quella dell'esempio precedente.

Figura 8.1 – Tavole di contingenza e opzioni per gli insiemi a risposte multiple



Il modulo TABELLE PERSONALIZZATE (Fig. 8.2) è molto versatile e consente un controllo totale su aspetti quali la scelta e la disposizione delle variabili, l'ordinamento delle categorie, il tipo e il formato delle statistiche riassuntive, il trattamento dei dati mancanti, ecc. Tramite il bottone CATEGORIE E TOTALI si accede a una sotto-finestra (Fig. 8.3) attraverso la quale si può decidere se aggiungere dei totali e sub totali (di riga e/o di colonna) e dove posizionarli. Si può anche decidere per ogni variabile se tabulare tutti i valori, escluderne alcuni o, al contrario, tabulare anche i VALORI MANCANTI per errore, rifiuto a rispondere o risposta "non so". Si può anche scegliere di rappresentare nella tabella eventuali CATEGORIE VUOTE, e cioè valori che sono stati etichettati perché erano previsti come risposte ammissibili, ma sono risultati privi di referenti empirici. Può infatti essere interessante constatare che nessun intervistato ha indicato una risposta che pure era ritenuta plausibile, essendo stata proposta nel questionario. In una fase preliminare di "ripulitura dei dati", può anche essere opportuno scegliere che vengano mostrati ALTRI VALORI RILEVATI DURANTE L'ESAME DEI DATI perché ciò consente di individuare eventuali errori di registrazione che dovranno essere coretti.

Figura 8.2 – Tabelle personalizzate

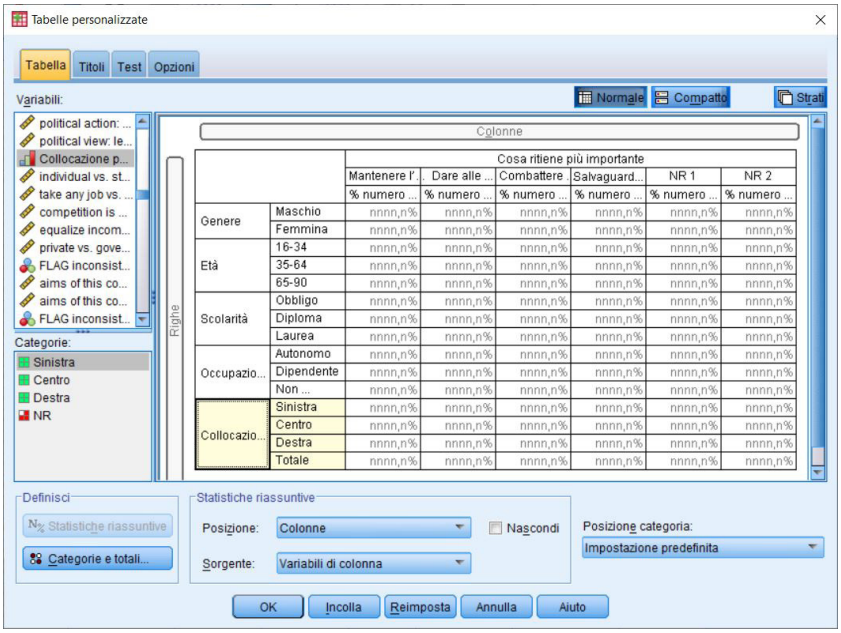
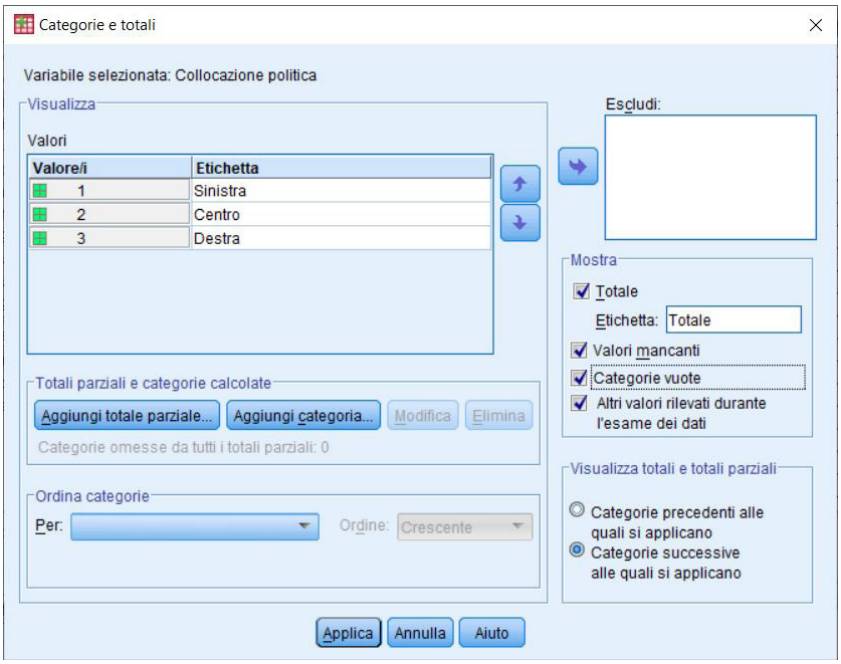


Figura 8.3 – Categorie e totali



Per esemplificare questo tipo di tabella abbiamo deciso di utilizzare una domanda da anni riproposta nelle rilevazioni del *World Values Study* al fine di classificare gli intervistati come portatori di valori moderni o post-moderni (Inglehart 1977 e 1996). In questo caso le alternative di risposta erano solo quattro e si chiedeva agli intervistati di indicare quale obiettivo ritenessero prioritario per l'Italia e quale il secondo per importanza. Nella teorizzazione di Inglehart scegliere "mantenere l'ordine nella nazione" e "combattere l'aumento dei prezzi" indicherebbe un orientamento di valore definibile come moderno e scegliere invece "dare alle persone più voce in capitolo nelle decisioni governative importanti" e "salvaguardare la libertà di parola" un atteggiamento post-moderno.

Le risposte rilevate nella più recente tornata del sondaggio mostrano innanzitutto che nel complesso degli intervistati le percentuali non variano molto da un obiettivo all'altro (Tab. 8.2). È peraltro curioso che la lotta all'aumento dei prezzi raccolga i maggiori consensi (55,4%), perché nel periodo della rilevazione (da settembre 2018 a gennaio 2019) non si registravano significative avvisaglie di problemi inflattivi; evidentemente questo aspetto è considerato prioritario anche quando l'aumento dei prezzi è solo un pericolo potenziale. Poco sotto al 50% si collocano i due obiettivi rivelatori di un atteggiamento post-moderno e, infine, poco più del 40% degli intervistati considera prioritario (al primo o al secondo posto) il mantenimento dell'ordine nella nazione.

Tabella 8.2 – Tabulazione di variabili indipendenti con una domanda a risposta multipla (Italia 2019)

		Ordine	**Voce in capitolo**	**Aumento prezzi**	**Libertà di parola**	**NR 1**	**NR 2**	**Totale**
		%	**%**	**%**	**%**	**%**	**%**	**N**
Genere	Maschio	42,8	49,0	52,2	48,6	2,1	5,2	1.143
	Femmina	42,8	46,3	58,6	46,4	1,4	4,5	1.134
Età	16-34	36,2	54,5	46,9	52,8	1,9	7,6	475
	35-64	42,1	46,9	56,9	49,2	1,4	3,4	1.191
	65-90	49,1	43,9	59,1	40,1	2,3	5,6	611
Scolarità	Obbligo	45,3	46,2	63,9	37,7	1,7	5,3	875
	Diploma	40,6	48,9	52,9	52,5	1,4	3,8	1.008
	Laurea	43,0	48,3	42,5	57,5	2,4	6,3	381
Occupazione	Autonomo	41,1	47,9	48,2	55,7	1,4	5,7	280
	Dipendente	39,2	49,0	56,4	50,1	1,2	4,1	775
	Non occupato	45,4	46,9	56,4	44,1	2,2	5,1	1.205
Collocazione Politica	Sinistra	27,3	58,3	40,4	68,3	,9	4,7	319
	Centro	40,7	49,8	53,9	48,8	1,7	5,1	980
	Destra	60,5	41,5	56,2	38,4	,3	3,1	354
	NR	43,9	42,3	65,1	40,1	3,0	5,6	624
Totale		42,8	47,7	55,4	47,5	1,8	4,9	2.277

Fonte: European Values Study (https://europeanvaluesstudy.eu/)

Scorporando questi valori a seconda dei gruppi di intervistati individuati dalle variabili socio-economiche si possono effettuare interessanti analisi confrontando le percentuali in orizzontale, oppure in verticale. Con ciò si intende che ci possiamo concentrare su una variabile indipendente e individuare l'obiettivo o gli obiettivi sui quali quella variabile mostra di avere più potere discriminante; possiamo però anche confrontare i diversi gruppi sociali per individuare quelli che hanno opinioni maggiormente differenziate su ognuno degli obiettivi da perseguire.

Si può dunque osservare, ad esempio, che il genere mostra un limitato potere discriminante, con una differenza tra maschi e femmine che supera di poco sei punti percentuali solo riguardo al controllo dell'inflazione. Più interessante è il ruolo attribuibile all'età, con differenze poco oltre il 10% per tutti gli obiettivi: all'aumento dell'età corrisponde un crescente "gradimento" per gli obiettivi materialisti (ordine e prezzi) e un interesse relativamente minore per i post-materialisti (partecipazione e libertà di parola). Il ruolo svolto dal livello di istruzione dipende molto dall'obiettivo considerato: è assai limitato per due degli obiettivi (ordine e voce in capitolo), e assai rilevante (oltre 20 punti di differenza) se si tratta di controllo dell'aumento dei prezzi (il maggior favore si riscontra tra i meno scolarizzati) e di libertà di parola (obiettivo maggiormente condiviso dai più scolarizzati). Anche il tipo di occupazione mostra una debole capacità di differenziare riguardo ai primi due obiettivi; riguardo agli altri due, un minore interesse per l'aumento dei prezzi da parte dei lavoratori autonomi e per la libertà di parola da parte dei non occupati.

La maggiore differenziazione delle opinioni si riscontra però suddividendo gli intervistati a seconda del loro orientamento politico e gli atteggiamenti sono conformi a quanto si poteva presumere: non sorprende infatti che il mantenimento dell'ordine sia considerato un obiettivo prioritario da oltre il 60% degli intervistati che si collocano a destra e da solo il 27% di quelli che si definiscono di sinistra. Una differenza simile, ma a parti invertite, si riscontra riguardo alla salvaguardia della libertà di parola. Per gli altri due obiettivi si riscontrano differenze un po' più limitate (intorno ai 15 punti), ma comunque coerenti (sono più materialisti gli elettori di destra e più post-materialisti quelli di sinistra).

In verticale si può effettuare una lettura dei dati complementare: concentrando l'attenzione sulla salvaguardia della libertà di parola si vede ad esempio che questo obiettivo è maggiormente supportato da intervistati ad elevata scolarizzazione, giovani, lavoratori autonomi o simpatizzanti per la sinistra. La lotta all'inflazione è invece una priorità per un maggior numero di intervistati anziani, poco scolarizzati e simpatizzanti per la destra o per il centro.

A differenza delle risposte elencate nel questionario, che si possono indicare una sola volta (o nessuna), le mancate risposte possono essere più di una

(nel nostro caso due). Adottando un codice diverso per le due situazioni è però possibile distinguere le mancate risposte alla prima e alla seconda domanda. La tabella mostra che l'1,8% degli intervistati non ha indicato il primo obiettivo e il 4,9% il secondo (per differenza, il 3,1% ha indicato il primo, ma non il secondo).

Passando al versante delle variabili indipendenti, mentre per semplificare la tabella le poche mancate risposte riguardo al livello di istruzione e all'occupazione le abbiamo omesse, abbiamo conservato la riga che contiene il rilevante numero di intervistati che non hanno indicato la loro "simpatia" politica (27,4%). La maggior parte di essi ha infatti risposto alla domanda sugli obiettivi prioritari del paese e ciò consente di notare che il loro profilo di risposta è molto simile a quello degli intervistati che si sono dichiarati di destra. Questa constatazione potrebbe suggerire di approfondire la questione mediante tecniche di analisi più sofisticate che, come abbiamo visto quando abbiamo descritto l'analisi discriminante (§ 4.4.2), consentono di formulare congetture, fondate su base probabilistica, sulla loro ipotetica classificazione.

Riunendo le due risposte in un insieme a risposta multipla non si è tenuto conto che nella prima variabile è registrato l'obiettivo ritenuto più importante e nell'altra la "seconda scelta". In pratica, si sono trattate le due variabili come se si fosse semplicemente chiesto agli intervistati di scegliere due dei quattro obiettivi elencati nel questionario. La semplificazione era in parte giustificata dal fatto che la successiva analisi era finalizzata a individuare le differenze di atteggiamento valoriale tra i diversi gruppi sociali, più che a stabilire un ordine gerarchico tra gli obiettivi. Ci si potrebbe però chiedere come si potrebbe operare per tenere conto di questo aspetto.

Un primo esempio può prendere spunto dal fatto che nell'ambito dell'*European Values Study* la combinazione delle due risposte viene utilizzata per elaborare un indice di materialismo/post-materialismo senza tenere conto dell'ordine tra le due risposte. È infatti classificato come materialista (o postmaterialista) l'intervistato che indica due obiettivi dello stesso tipo e "misto" chi sceglie due obiettivi di tipo diverso. Si potrebbe invece tenere conto dell'ordine di citazione dei due obiettivi, anche al fine di "scorporare" la categoria intermedia che raggruppa oltre il 50% degli intervistati. Si potrebbe perciò definire semi-materialista chi indica per primo l'obiettivo materialista (e per secondo quello postmaterialista) e semi-postmaterialista chi indica per primo un obiettivo postmaterialista. Seguendo questa logica, la combinazione "aumento dei prezzi e libertà di parola" verrebbe classificata "semi-materialista" e come "semi-postmaterialista" la combinazione "libertà di parola e aumento dei prezzi", che è simmetrica, ma non uguale (Tab. 8.3).

Anche al fine di evidenziare il fatto che il ricercatore, particolarmente quando si tratta di definire operativamente una variabile, ha un margine di discrezionalità che va rivendicato, ma allo stesso tempo esplicitato e "giustificato", abbiamo apportato all'indice un'altra modifica che tende a massimizzare l'utilizzo dell'informazione raccolta. Se l'intervistato ha indicato il primo obiettivo da perseguire e non il secondo, si può presumere che ritenga quell'obiettivo assolutamente prioritario e può essere dunque ragionevole classificare quell'intervistato come materialista o post-materialista a seconda del tipo di obiettivo che ha indicato. Agendo in tale modo si recupera la maggior parte di quelli che hanno fornito una sola risposta e, di conseguenza, le mancate risposte scendono da quasi il 5% a meno del 2%.

Tabella 8.3 – Indice di materialismo/post-materialismo (Italia 2019)

		N	%		N	%
Validi	Materialista	461	20,2	Materialista	497	21,8
				Semi-materialista	757	33,2
	Misto	1278	56,1			
				Semi-postmaterialista	521	22,9
	Postmaterialista	427	18,8	Postmaterialista	462	20,3
	Totale	2.166	100,0	Totale	2.237	100,0
Mancanti	NR	111	4,9	NR	40	1,8
Totale		2.277	100,0	Totale	2.277	100,0

Fonte: European Values Study (https://europeanvaluesstudy.eu/)

Un modo alternativo per valorizzare la distinzione tra prima e seconda scelta, consiste nel registrare le risposte dedicando a questo scopo quattro colonne della matrice dei dati, invece di due. Ogni colonna viene intestata a uno dei quattro obiettivi e la risposta si codifica con "2" se quell'obiettivo è indicato per primo, "1" se indicato per secondo e "0" se non è stato scelto. In pratica, si attribuiscono dei punteggi che si possono trattare come variabili cardinali, il che consente di utilizzare il punteggio medio come parametro statistico riassuntivo al fine di ordinare gli obiettivi secondo l'importanza loro attribuita complessivamente, come visualizzato nella Tab. 8.4.

Tabella 8.4 – Indice di materialismo/post-materialismo (Italia 2019)

	I scelta	II scelta	Non selez.	Totale	
	%	%	%	%	Media
Combattere l'aumento dei prezzi	30,4	25,0	44,6	100,0	0,86
Dare alle persone più voce in capitolo	22,5	25,1	52,3	100,0	0,70
Salvaguardare la libertà di parola	20,6	26,9	52,5	100,0	0,68
Mantenere l'ordine nella nazione	24,7	18,1	57,2	100,0	0,67

Fonte: European Values Study (https://europeanvaluesstudy.eu/)

8.4. Incrocio con un set di dicotomie multiple

Se a una domanda si può rispondere selezionando da una lista un limitato numero di risposte è spontaneo utilizzare per la registrazione nella matrice dei dati un numero di colonne pari al massimo di risposte consentite. In alternativa, come si è similmente fatto qui sopra, si può usare per ogni voce della lista una colonna e registrare le risposte in forma dicotomica ("sì/no", intendendo "menzionato/non menzionato"). Spesso è proprio quest'ultima la soluzione utilizzata, come è poi la scelta obbligata se non ci sono limiti al numero di risposte affermative: se si chiede, ad esempio, quali prodotti alimentari vengano acquistati abitualmente, è possibile che qualche intervistato acquisti anche tutti i prodotti elencati nel questionario. Il ricercatore si trova in questo caso ad analizzare un numero elevato di dicotomie e cioè di variabili di natura un po' particolare: se sappiamo quanti intervistati hanno risposto "sì", sappiamo anche quanti hanno risposto "no", perché lo possiamo ricavare per differenza dal totale (in valore assoluto o percentuale). Di conseguenza, la tabulazione di una sola delle risposte a una domanda posta in forma dicotomica ("sì/no", "d'accordo/contrario", "italiano/straniero", "celibe/nubile", ecc.) non comporta alcuna perdita di informazione (se risponde "sì" il 47%, i "no" sono per certo il 53%).

Si tratta di una caratteristica preziosa, tanto più se si considera che il campo di applicazione si può ampliare di molto tabulando, ad esempio, una sola delle risposte a una scala di atteggiamento, che solitamente ne comprende quattro o cinque. In questo modo si dicotomizza "artificialmente" una variabile categoriale; operazione che comunque comporta una rilevante perdita di informazione. A differenza di quanto accade nel caso di una dicotomia "na-

turale", il complemento a 100 della percentuale di risposta che viene analizzata (ad esempio, "molto") si distribuisce infatti in quantità non note tra le modalità di risposta escluse dall'analisi ("abbastanza", "poco" e "per nulla"). Inoltre, se si accorpano le risposte "molto" e "abbastanza" (nel caso le risposte "molto" siano poco numerose), non è più possibile distinguere il contributo di ognuna di queste due diverse risposte. Ciononostante, ai fini di un'analisi *esplorativa*, la semplificazione può essere pragmaticamente accettabile anche perché, per verificare la tenuta di quanto emerge dalla prima esplorazione, si potrebbe ripetere l'analisi adottando un diverso "punto di vista": la tabulazione della risposta "per nulla", collocata all'altro estremo della scala.

Nel primo esempio utilizziamo la batteria di domande che abbiamo già incontrato alla fine del capitolo sull'analisi monovariata, con le quali si è chiesto agli intervistati di indicare quanto sia importante nella loro vita ciascuno di una serie di ambiti. I valori analizzati sono quelli relativi alla sola risposta "molto" che per alcuni degli ambiti raccoglie la stragrande maggioranza dei consensi.

Il programma di analisi costruisce la tabella in modo incrementale, aggiungendo un caso ogni volta che trova la risposta che si è deciso di tabulare; di conseguenza, se un intervistato non fornisce mai quella risposta rimane escluso e le percentuali sono calcolate sul totale di intervistati che hanno risposto in quel modo almeno una volta. Per far sì che il totale rimanga uguale al numero di interviste effettuate è dunque necessario aggiungere chi ha risposto "nessuno" alla domanda virtuale: "quanti ambiti di vita sono per te molto importanti?".

La nuova variabile si può costruire con il comando COUNT, illustrato nel capitolo 5 (§ 5.3.5), e avrà un *range* di valori da "0" (nessun ambito molto importante) a "6" (tutti gli ambiti molto importanti). Siccome il valore utilizzato per costruire l'insieme di dicotomie è "1", con RECODE si dovrà operare un'ulteriore trasformazione ponendo "0" uguale a "1" (con il significato di "sì, nessuno ambito è molto importante") e gli altri valori uguali a "0" (almeno uno è molto importante). Come si vedrà in modo ancor più convincente con il prossimo esempio, la variabile che contiene il conteggio analitico delle risposte "molto" può essere essa stessa interessante da analizzare ed è perciò opportuno attribuire alla sua versione dicotomica un nome diverso, in modo da poter disporre di entrambe le versioni.

Come nel caso delle domande a risposta multipla trattate nel precedente paragrafo, anche con le variabili dicotomiche (o dicotomizzate) si inizia utilizzando la finestra DEFINISCI INSIEMI A RISPOSTA MULTIPLA (Fig. 8.1) per scegliere le variabili da raggruppare, compresa quella che registra i casi che non hanno mai indicato la risposta che si intende tabulare. A differenza dell'e-

sempio precedente, si deve in questo caso selezionare DICOTOMIE, indicare il valore da conteggiare e, per completare, attribuire un nome all'insieme, etichettarlo e "cliccare" AGGIUNGI.

Visto che il programma consente di indicare un unico valore da conteggiare, se si intende analizzare la percentuale di intervistati che hanno risposto "molto" oppure "abbastanza", le variabili si dovranno preventivamente dicotomizzare, attribuendo il codice "1" a questi due valori e "0" a tutti gli altri in una nuova serie di variabili (per conservare anche quelle con i valori originali).

In questo tipo di tabella ogni percentuale è indipendente da tutte le altre perché il complemento a 100 di ognuna va inteso come la percentuale di intervistati di quella categoria socio-demografica che non considera "molto" importante quell'ambito di vita. Trattandosi di dicotomie "artificiali", è bene comunque ribadire che, se il 12,6% degli intervistati considera la politica "molto" importante, agli altri non sono completamente disinteressati, perché alcuni hanno risposto "abbastanza", altri "poco" e solo una parte "per nulla".

Prendendo spunto da qualche altro dato interessante che emerge dall'esame della Tab. 8.5, notiamo che anche la religione riscuote un debole interesse (24,6%). Invece, come era prevedibile, oltre alla famiglia (90,3%), anche il lavoro riscuote un notevole interesse, con quasi tre quarti di intervistati (73,7%) che lo considera un ambito di vita molto importante; gli amici (41,5%) e il tempo libero (37,2%) si collocano invece a metà strada.

Confrontando le categorie socio-demografiche, si constata che gli intervistati che considerano molto importante la religione sono circa il 15% tra i giovani fino a 34 anni e ben più di un terzo (36,8%) tra chi ha 65 anni o più. Analizzando le percentuali per riga, si può notare che tra i giovani quote più consistenti della media considerano molto importante l'ambito amicale e la disponibilità di tempo libero. Quanto alle differenze di genere, le femmine apprezzano più dei maschi la famiglia e la religione, mentre tra i maschi c'è maggior interesse per il lavoro, il tempo libero e la politica, con differenze peraltro non molto elevate.

Solo il 3,1% degli intervistati considera non "molto" importanti tutti gli ambiti di vita elencati o non ha risposto a queste domande. Nella colonna dedicata a questa variabile "residuale" (etichettata sinteticamente "NR") si può notare qualche differenza tra tipi di intervistati, ma la loro bassa numerosità rende poco interessante una più approfondita analisi.

Tabella 8.5 - Tabulazione di variabili indipendenti con una serie di dicotomie (Italia 2019)

		Famiglia	Lavoro	Amici	Tempo lib.	Religione	Politica	NR
		%	%	%	%	%	%	%
Genere	Maschio	88,4	77,4	41,7	39,5	18,4	14,4	3,6
	Femmina	92,3	70,0	41,3	35,0	31,4	10,8	2,6
Età	16-34	86,1	74,1	55,4	46,7	15,2	14,5	4,4
	35-64	91,9	75,6	40,1	38,0	22,6	11,3	2,4
	65-90	90,7	69,9	33,4	28,3	36,8	13,7	3,3
Scolarità	Obbligo	90,3	72,5	34,4	30,5	30,9	9,5	3,4
	Diploma	90,7	75,4	43,8	41,8	21,3	14,9	2,8
	Laurea	90,6	73,5	51,7	40,7	21,0	13,9	2,4
Occupazione	Autonomo	89,6	79,6	47,5	43,9	18,6	15,7	1,8
	Dipendente	91,5	79,1	45,5	42,8	17,3	10,6	2,5
	Non occup.	89,8	69,1	37,7	32,0	31,4	13,2	3,7
Collocazione politica	Sinistra	91,5	77,7	51,1	45,8	19,1	22,9	1,3
	Centro	89,5	72,9	41,4	34,5	24,6	12,0	3,5
	Destra	89,5	76,6	37,9	38,1	25,4	15,5	3,1
	NR	91,5	71,5	38,8	36,7	27,9	6,7	3,4
Totale		90,3	73,7	41,5	37,2	24,9	12,6	3,1

Fonte: European Values Study (https://europeanvaluesstudy.eu/)

La situazione cambia se la categoria, oltre a essere più corposa, la si può definire concettualmente come è il caso, passando al versante delle variabili indipendenti, degli intervistati che non hanno risposto alla domanda sulla collocazione politica. La numerosità è infatti elevata (oltre il 25%) ed è anche definibile come "area del non voto", molto appetibile per i partiti che ambiscono ad allargare il loro consenso. Lasciando in tabella questo gruppo di intervistati "reticenti" è possibile confrontare i loro atteggiamenti con quelli degli altri gruppi e notare inoltre che tra di loro la quota che ritiene "molto" importante la politica è molto più bassa della media, cosa che indica, se non altro, coerenza di atteggiamento.

Quando si rilevano opinioni o atteggiamenti solitamente le risposte si distribuiscono su tutto l'arco delle possibilità offerte dal questionario anche perché, se una domanda non riesce a discriminare a sufficienza tra le varie posizioni, significa che è banale o mal formulata. Ovviamente, se si chiede quanto sia importante la famiglia, si può prevedere che la risposta "molto" riceva un largo consenso (90,3%) e quasi l'unanimità sommando la risposta "abbastanza" (99,6%). Era però ugualmente interessante porre la domanda per avere un termine di confronto con gli altri ambiti di vita e individuare qualche deviazione, seppur piccola, dal generalizzato atteggiamento di favore. Inoltre, nell'insieme delle domande erano

comprese alcune, come quelle sulla politica e la religione, per le quali una certa quota di unanimismo era prevedibile che si sarebbe concentrato sul lato opposto della scala, e dunque il complesso delle domande copriva in modo sufficientemente completo le varie posizioni valoriali.

Si deve anche considerare che se si rilevano atteggiamenti o opinioni è plausibile presumere che tutti (o quasi) gli intervistati abbiano perlomeno un orientamento di massima su qualsiasi questione. Di conseguenza, nella batteria di domande appena esaminata si sono registrate nel complesso molto poche mancate risposte. Nel caso si analizzino comportamenti, è invece possibile che numerosi intervistati rispondano negativamente a tutte le domande e vengano di conseguenza esclusi dall'analisi. Se si chiede che tipo di musica piace o quali giornali si acquistano o leggono online, in un campione di popolazione si troveranno probabilmente numerosi intervistati che non ascoltano musica o non leggono giornali. In queste situazioni la variabile "residuale" conterrà una casistica numericamente rilevante e diventa perciò interessante aggiungere questa variabile nell'insieme a risposta multipla perché in tal modo si potranno analizzare, oltre alle caratteristiche degli amanti dei diversi generi musicali e dei lettori dei vari giornali, anche le caratteristiche di chi non ascolta musica e non legge i giornali.

Per illustrare quanto possa essere cruciale in queste situazioni la decisione sull'inclusione/esclusione, utilizzeremo una batteria di domande sull'attività associativa; tema che abbiamo scelto proprio perché è lecito attendersi che molti intervistati non abbiano alcun tipo di vita associativa. I dati sono rappresentati nella Tab. 8.6 dove nella prima colonna si legge per ogni tipo di associazione quanti intervistati hanno affermato di farne parte. Nel complesso si sono registrate 871 risposte affermative da parte di 638 individui che costituiscono il gruppo di intervistati "associati". Sulla base di questa componente minoritaria dell'intero campione (28%) sono calcolate le percentuali riportate nella seconda colonna di dati il cui totale assomma a 136,5% perché una parte degli intervistati ha dichiarato di far parte di più di una associazione (in media 1,37). Come si vede, le associazioni sportive o ricreative registrano una percentuale del 27,4% e si collocano intorno al 20% le benefiche/umanitarie, le religiose e le culturali. Tali percentuali appaiono elevate, ma solo se non si tiene conto che sono riferite ai soli intervistati (un po' più di un quarto) che hanno dichiarato di far parte di almeno un'associazione. Per valutare correttamente la pervasività della vita associativa non si deve però dimenticare che il 72% degli intervistati non fa parte di alcuna associazione.

Di ciò si tiene conto nelle percentuali dell'ultima colonna, dove le risposte sono rapportate alla dimensione dell'intero campione (2.277 intervistati). Di conseguenza, i 175 intervistati che hanno dichiarato di far parte di società sportive o ricreative, che corrispondono al 27,4% di quelli che fanno parte di almeno un'associazione, scendono al 7,7% se rapportati all'intero campione. A

nostro avviso, in una indagine campionaria rappresentativa della popolazione italiana è questo il dato da considerare se si vuole apprezzare correttamente l'impatto del fenomeno associativo sulla società. Evidentemente, se avessimo realizzato un'indagine sull'associazionismo, potrebbe essere corretto ragionare all'interno di questo fenomeno e appropriato utilizzare la prima serie di percentuali.

Tabella 8.6 - Partecipazione ad associazioni (Italia 2019)

	N. risposte	% associati	% campione
Sport/ricreative	175	27,4	7,7
Benefiche/umanitarie	137	21,5	6,0
Religiose	125	19,6	5,5
Culturali	122	19,1	5,4
Professionali	70	11,0	3,1
Sindacali	61	9,6	2,7
Ambientali	59	9,2	2,6
Politiche	35	5,5	1,5
Auto/mutuo aiuto	18	2,8	,8
Consumeristiche	9	1,4	,4
Altre associazioni	60	9,4	2,6
Totale	871	136,5	38,3
Nessuna	1.639	-	72,0
Totale casi		638	2.277

Fonte: European Values Study (https://europeanvaluesstudy.eu/)

Passando all'analisi bivariata, vista la loro ridotta numerosità, nella tabella di contingenza multipla si sono ulteriormente accorpate le associazioni, facendo confluire nella categoria "altra associazione" anche quelle su cui era stata posta una domanda esplicita, ma riscontrata una consistenza inferiore al 3% (Tab. 8.7). La tabella mostra con tutta evidenza che, a differenza dell'analisi sull'importanza degli ambiti di vita, nel caso del comportamento associativo la numerosità della casistica consiglia di iniziare l'analisi proprio dalla categoria che non si può considerare "residuale" per la sua consistenza numerica e per il fatto che si tratta di una componente della popolazione dotata di una chiara identità. Chi non fa parte di alcuna associazione costituisce infatti un gruppo sociale, ancorché informale, come chi non va al cinema, non legge giornali o libri o non va mai in vacanza.

Come si vede dalla relativa colonna di percentuali, il gruppo di intervistati (72%) che non fanno parte di alcuna associazione ha dimensioni relativamente

minori in particolare tra le persone più scolarizzate (55,9) e chi ha dichiarato di collocarsi a sinistra in ambito politico (55,5%). Le differenze più significative si riscontrano però quando si osservano le percentuali di partecipazione ai diversi tipi di associazioni, a partire da quelle sportive (7,7%) dove la percentuale tra i maschi (10,1%) è più che doppia rispetto a quella che si riscontra tra le femmine (4,9%) ed è quasi tripla se confrontiamo chi non è andato oltre la scuola dell'obbligo (4,3%) con i laureati (12,1%).

Tabella 8.7 - Partecipazione ad associazioni per caratteristiche socio-demografiche (Italia 2019)

		Sport/ ricreat. %	Volon- tariato %	Reli giosa %	Cultu- rale %	Profes- sionale %	Altra assoc.	Nessuna %	Tot. N
Genere	Maschio	10,4	6,4	3,9	4,6	3,7	11,3	70,5	1.143
	Femmina	4,9	5,6	7,1	6,1	2,5	10,0	73,5	1.134
Età	16-34	10,1	5,3	5,1	6,5	1,9	8,4	71,4	475
	35-64	7,1	6,2	5,3	4,8	3,9	12,5	71,2	1.191
	65-90	6,9	6,2	6,2	5,6	2,3	8,7	74,0	611
Scolarità	Obbligo	4,3	3,3	4,1	2,2	0,5	6,3	82,5	875
	Diploma	8,8	7,2	6,1	6,3	2,3	11,0	69,1	1.008
	Laurea	12,1	9,2	7,3	10,2	11,3	18,6	55,9	381
Occupa- zione	Autonomo	8,9	7,9	3,9	3,9	12,9	14,3	66,1	280
	Dipendente	8,9	7,0	5,2	6,7	3,1	13,5	68,8	775
	Non occupato	6,5	5,1	6,1	4,7	0,8	7,7	75,7	1.205
Collocaz. Politica	Sinistra	14,4	11,0	5,6	10,7	5,0	20,1	58,0	319
	Centro	8,2	5,3	6,2	5,6	3,1	11,6	69,8	980
	Destra	5,9	7,1	4,8	3,4	3,4	8,8	74,9	354
	NR	4,5	4,0	4,6	3,4	1,9	5,3	80,9	624
Totale		7,7	6,0	5,5	5,4	3,1	10,6	72,0	2.277

Fonte: European Values Study (https://europeanvaluesstudy.eu/)

Procedendo anche in questa tabella con una lettura dei dati in orizzontale, possiamo osservare come si esplicita la divaricazione dei comportamenti a seconda del livello di istruzione che abbiamo individuato osservando la percentuale di intervistati non associati. Mettendo a confronto le opposte categorie dei laureati e chi si è fermato alla scuola dell'obbligo, si notano differenze notevoli e coerenti per tutti i tipi di associazioni (solo un po' attenuate per le associazioni di tipo religioso o ecclesiale). La situazione più estrema si registra a proposito delle associazioni culturali, delle quali fanno parte solo il 2,2% dei meno scola-

rizzati e invece il 10,2% dei laureati. Ciò era prevedibile anche perché la laurea è a volte un requisito indispensabile, come nel caso delle associazioni professionali (0,5% di chi ha al massimo il diploma di scuola media superiore e 11,3% dei laureati).

Infine, con questa tabella abbiamo una ulteriore prova dell'utilità di analizzare le mancate risposte quando non sono banalmente residuali, come nel caso di chi non ha inteso rivelare il proprio orientamento politico (forse anche perché non ne c'è l'ha o è relativamente "fluido"). Si tratta di oltre un quarto degli intervistati (27,2%) e analizzarli consente di delinearne ulteriormente le loro caratteristiche. Tra essi la quota di "non associati" è simile a quella che si riscontra tra i meno scolarizzati, appartenenti alle categorie più deboli socialmente (anziani, inoccupati) e si dichiara di destra (e ciò è di aiuto se si vuole stimare la collocazione politica di chi preferisce non rivelarla).

8.5. Tabulazione multipla di parametri statistici

Se si ritiene che la tabulazione di un unico valore tratto da una scala si traduca in un'inammissibile perdita di informazione, è possibile utilizzare, in alternativa alle percentuali, parametri statistici sintetici come la media che è calcolata utilizzando tutti i valori della distribuzione (escluse le mancate risposte). Si deve però avere consapevolezza che, pur utilizzando tutti i valori della distribuzione, riassumerla in un unico dato comporta comunque una notevole perdita d'informazione. Nell'esempio qui proposto è usata solo la media, ma è possibile aggiungere la deviazione standard, in modo da poter valutare anche l'aspetto della dispersione dei dati, come si è visto ad esempio nella Tab. 2.3. Si possono anche utilizzare, al posto della media, altri parametri statistici come la mediana, più corretta per variabili ordinali, ma poco utile se i valori della scala sono poco numerosi.

Facendo riferimento a Spss, la tabulazione compatta dei parametri statistici è possibile solo con tabelle personalizzate, perché la procedura risposte multiple permette di inserire nelle celle della tabella solo i valori assoluti e le percentuali. Ovviamente, chi non possedesse tale modulo, e nemmeno il programma Spss (il suo analogo Pspp o programmi simili), può calcolare le medie utilizzando altri strumenti e poi disporre i dati nel modo che esemplifichiamo nella Tab. 8.8.

Tabella 8.8 - Tabulazione di un set di domande di atteggiamento con un set di variabili indipendenti. Valori medi su scala da 1 = mai giustificato a 10 = sempre giustificato (Italia 2019)

		Divorzio	**Omosex**	**Inseminaz.**	**Eutanasia**	**Aborto**	**Sesso**	**Suicidio**
Genere	Maschio	6,7	6,1	6,1	6,0	5,3	4,3	3,2
	Femmina	6,6	6,5	6,3	5,7	5,2	3,4	3,0
Età	16-34	7,0	7,2	6,6	6,3	5,8	4,8	3,5
	35-64	7,0	6,6	6,5	6,1	5,5	4,0	3,2
	65-90	5,7	5,0	5,3	5,1	4,4	2,9	2,5
Scolarità	Obbligo	5,9	5,1	5,7	5,2	4,4	3,2	2,6
	Diploma	7,0	6,7	6,4	6,2	5,7	4,1	3,2
	Laurea	7,5	7,7	6,7	6,6	6,2	4,6	3,9
Occupazione	Autonomo	7,5	7,0	6,7	6,7	6,0	4,4	3,5
	Dipendente	7,2	6,8	6,8	6,4	5,8	4,4	3,2
	Non occup.	6,1	5,7	5,7	5,4	4,8	3,4	2,8
Collocaz. politica	Sinistra	7,7	7,4	6,9	6,9	6,5	4,7	4,1
	Centro	6,7	6,5	6,4	6,0	5,4	3,9	3,1
	Destra	6,5	5,6	5,9	5,7	4,8	4,1	2,7
	NR	6,1	5,6	5,7	5,3	4,6	3,2	2,6
Totale		6,6	6,3	6,2	5,9	5,3	3,9	3,1

Fonte: European Values Study (https://europeanvaluesstudy.eu/)

Le variabili analizzate tramite i valori medi sono una parte della batteria di giudizi sulla liceità dei comportamenti che abbiamo analizzato tramite la tecnica dell'analisi fattoriale nel paragrafo ad esse dedicato all'interno del capitolo sull'analisi multivariata (§ 4.3.2). Per ragioni di spazio si sono inserite in tabella solo le sette variabili (su nove) maggiormente saturate sul primo nella soluzione a due fattori illustrata nella parte destra della Tab. 4.7 dove le variabili sono comunque ordinate secondo il valore della media (a sinistra i comportamenti più giustificati).

Anche con questa tabella si può procedere a un'analisi per riga o per colonna. Come esempio di lettura dei dati per riga possiamo notare che per tutti i comportamenti i giovani sono più permissivi degli anziani, con una differenza che si dilata a due punti riguardo alla liceità dell'omosessualità e del sesso occasionale. Simile è il ruolo dell'istruzione (i laureati sono più permissivi degli altri) e, per quanto riguarda l'atteggiamento politico, riguardo all'aborto e al divorzio ritro-

viamo le stesse posizioni valoriali che hanno ispirato il voto in occasione dei due relativi referendum.

Procedendo in verticale, si può prendere in considerazione il tema dell'eutanasia, di cui si discute animatamente e si sta cercando di legiferare in Italia al momento in cui vengono scritte queste pagine, e scorrere la tabella per colonna, al fine di individuare le categorie sociali maggiormente inclini a giustificarla di fronte a una malattia inguaribile. Le differenze in verità non sono molto elevate e si aggirano su poco più di un punto, confrontando le categorie opposte; sono comunque un po' più spostati sul versante della giustificazione i maschi, i più giovani, i laureati, i lavoratori (rispetto ai non occupati) e gli intervistati che si collocano a sinistra (in questo caso la differenza è poco sotto i due punti). Facendo nuovamente dei confronti in orizzontale, concludiamo osservando che l'eutanasia (5,9) riceve in generale un giudizio di liceità molto più elevato rispetto al suicidio (3,1).

9. Gestire i dati mancanti

9.1. Introduzione

La matrice dei dati è spesso incompleta perché alcune informazioni non sono state rilevate a causa di difficoltà o rifiuto a rispondere, errori nella registrazione o, nel caso di dati "amministrativi", carenze della fonte che possono derivare dalle cause citate o da altre.

La matrice dei dati può essere incompleta anche perché alcune domande non sono pertinenti per specifici soggetti. Il problema delle modalità di utilizzo di questo ultimo tipo di dati mancanti apparentemente non si pone: se una *domanda filtro* individua i soggetti ai quali deve essere posta la domanda successiva, le analisi su tale domanda dovrebbero riferirsi solo a loro. Un esempio del precedente capitolo ci ha messo in contatto con una situazione simile, perché le percentuali calcolate sulla base dei soli intervistati che hanno dichiarato di fare parte di almeno un'associazione riproducevano implicitamente la situazione che si sarebbe verificata se si fosse posta la domanda "fa parte di qualche associazione?", chiedendo poi di indicare il tipo di associazione solo a chi aveva risposto affermativamente. Agendo in tal modo sarebbe stato spontaneo escludere dall'analisi chi aveva risposto negativamente alla

prima domanda, tenendo comunque conto, come si vedrà anche nel seguito, delle conseguenze dell'esclusione di intere categorie di intervistati sulla percezione della pervasività degli atteggiamenti e comportamenti all'interno dell'intera società.

I casi "sopravvissuti" possono comunque contenere "veri" dati mancanti che si devono escludere dal calcolo dei parametri statistici. Il loro inserimento nelle tabelle di contingenza può essere invece consigliabile per diversi motivi che abbiamo anticipato nel § 3.3.3 e nel resto del capitolo tratteremo con esaustività.

1. Se una tabella contiene dati completi si possono successivamente escludere alcuni valori e ricalcolare le percentuali, ma non è vero il contrario.
2. Se l'intervistato ha difficoltà a comprendere la domanda o gli manca l'informazione, la risposta "non so" si può/deve considerare valida.
3. Inserire nelle tabelle di contingenza le mancate risposte consente di analizzare le caratteristiche degli intervistati che non hanno risposto.
4. Escludendo chi non risponde, qualsiasi sia il motivo, si attribuisce a chi risponde un'indebita rappresentanza di chi non lo fa.
5. Le comparazioni sono più omogenee e agevoli se ci si riferisce costantemente allo stesso campione di N intervistati, invece che a chi di volta in volta è stato in grado o ha deciso di rispondere a singole domande.

Con riferimento all'ultimo punto, si deve comunque ricordare che in ogni rilevazione campionaria si producono stime affette da un errore statistico la cui entità dipende dal rapporto tra le dimensioni del campione e della popolazione; perciò, escludere (pochi) dati mancanti provoca un minimo aumento dell'errore che si può considerare ininfluente se i dati mancanti sono distribuiti in modo "accidentale". Se però alcuni intervistati non hanno risposto per qualche motivo specifico (come la ritrosia a rivelare la propria opinione), l'esclusione delle risposte mancanti può generare una qualche distorsione nelle stime. Ovviamente, la distorsione sarà modesta se i dati mancanti sono pochi e si può dunque ribadire che la numerosità è un criterio utilizzabile per decidere come trattarli.

In questo capitolo si è deciso di utilizzare i dati di un'indagine che l'autore ha realizzato personalmente per fornire un'esemplificazione in "presa diretta" dei nodi problematici che si devono concretamente affrontare. L'indagine, svolta nell'ambito di un progetto europeo sul tema del bullismo in ambiente scolastico, consentirà anche di evidenziare quanto possa impattare sulla comunicazione dei risultati il corretto trattamento dei dati mancanti. Tale esigenza deve essere tenuta presente in ogni tipo di ricerca sociale, ma è ancor più cruciale nel caso di una problematica particolarmente delicata e se si

devono comunicare i risultati, oltre che a un pubblico di specialisti, ai giovani studenti e alle loro famiglie. La rilevazione dei dati è stata effettuata in istituti scolastici del Nord-est e il campione era composto da alunni di quinta elementare e studenti di quarta di istituti medi superiori che hanno compilato il questionario in classe con la supervisione di un membro del gruppo di ricerca[1].

9.2. Variabili indipendenti vs. dipendenti

Nella Tab. 10.1 è messo in relazione il genere degli studenti intervistati con la domanda sul titolo di studio del padre e le percentuali sono calcolate per colonna, trattando in tal modo il genere come variabile indipendente e il titolo di studio come variabile dipendente[2]. Se i dati mancanti sono poco numerosi si possono escludere anche perché, nel caso specifico, non ha grande utilità conoscere il titolo di studio del padre dei soli cinque studenti che non hanno indicato il loro genere. Inoltre, l'esclusione è innocua, perché l'eliminazione della colonna "non risposto" non comporta il cambiamento delle percentuali relative ai maschi e alle femmine, perché rimarrebbe immutata la loro base di calcolo (il totale dei maschi e delle femmine).

Al contrario, se si togliessero le mancate risposte dalla variabile "titolo di studio", tutte le percentuali cambierebbero radicalmente perché si ridurrebbe la loro base di calcolo di oltre 150 casi (21,7% del totale). Eliminare le mancate risposte a questa variabile impedirebbe inoltre di osservare che sono più frequentemente i maschi (24,2%), rispetto alle femmine (19%), a non saper rispondere riguardo al titolo di studio del loro padre. D'altro canto, il ricalcolo farebbe sparire l'anomalia di percentuali che tra i maschi sono inferiori, rispetto a quelle delle femmine, per ogni livello di scolarità. Non essendo ragionevole che i titoli di studio dei genitori possano essere in qualche relazione con il genere dei figli, l'apparente anomalia si spiega infatti con il più elevato numero di mancate risposte registrato tra i maschi.

---

[1] I risultati dell'indagine svolta in Italia sono pubblicati nel volume *Children's Voices. Etnicità e bullismo nella scuola* (Delli Zotti 2014) e quelli dell'intero progetto europeo, che ha coinvolto gruppi di ricerca sloveni, austriaci, ciprioti e inglesi, in *Children's Voices: Studies of interethnic conflict and violence in European schools* (Sedmak *et al.* 2014).

[2] Non essendo ragionevole assumere una dipendenza "causa-effetto", è corretto parlare di variabili *trattate* rispettivamente come indipendente e dipendente. Come si è visto nel capitolo dedicato all'analisi bivariata può essere utile, come in questo caso, mettere in relazione due variabili per individuare possibili "anomalie" nella composizione del campione.

Tabella 10.1 – Titolo di studio del padre per genere (variabile indipendente)

	Maschio	Femmina	Non risposto	Totale
Primaria	18,8	20,1	20,0	19,5
Secondaria	36,2	39,4	20,0	37,7
Università	19,1	21,0	-	19,9
Altro	1,7	,6	20,0	1,3
Non so/NR	24,2	19,0	40,0	21,7
Totale	100,0	100,0	100,0	100,0
	356	353	5	714

In conclusione, le mancate risposte alla variabile indipendente si possono escludere senza problemi, ma spesso non è utile farlo; se invece si tolgono alla variabile dipendente, dal momento che cambiano le percentuali ci si deve chiedere se sia corretto e/o opportuno farlo. Se stiamo analizzando, ad esempio, i dati di un pre-test finalizzato alla messa a punto del questionario, o si sta operando nella prima fase di analisi, utile anche a individuare incongruenze o anomalie nei dati, è meglio essere conservativi e tabulare dunque dati completi.

9.3. Pochi vs. molti dati mancanti

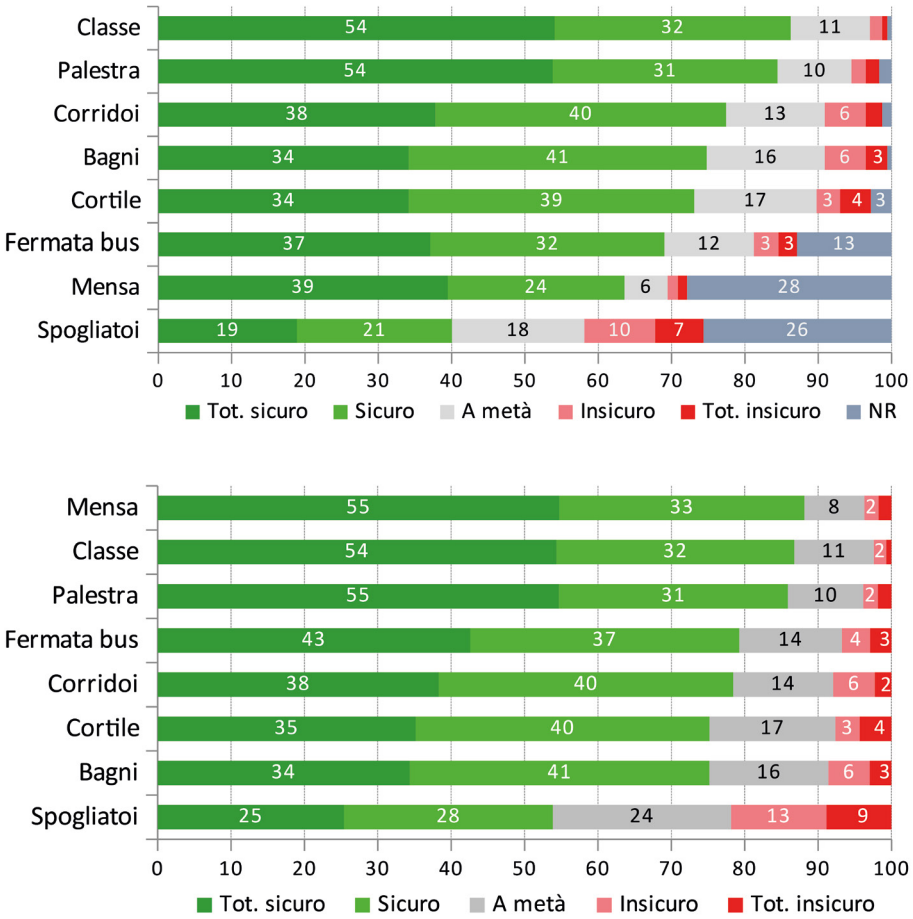
In una serie di domande sul bullismo nella propria scuola i dati mancanti sono risultati di modesta entità (tra 1 e 2%); pertanto, eliminandoli le percentuali ricalcolate si modificano di poco. La situazione cambia un po' per gli insulti via SMS o Facebook: le mancate risposte superano in questo caso il 6% e il ricalcolo provoca dunque modifiche un po' più consistenti (gli studenti che denunciano episodi di "bullismo informatico" passano da 21 al 22,4%). La numerosità un po' più elevata fa anche sospettare che le mancate risposte dipendano, almeno in parte, da cause "strutturali": per rispondere con cognizione di causa è infatti necessario conoscere strumenti di comunicazione poco usati dai bambini di quinta elementare (in questo gruppo le mancate risposte raggiungono infatti quasi il 10%, mentre tra gli studenti delle superiori sono meno del 3%).

Riguardo alla domanda su "altri" (non specificati) episodi di bullismo, il 90% degli intervistati non risponde, ma è inappropriato parlare di "mancate risposte" perché più che di una vera e propria domanda si tratta di un'opportunità di risposta; perciò, se "altri" episodi non accadono, alcuni rispondono "mai", ma la maggior parte lascia la risposta in bianco. Se si escludono le risposte mancanti senza tenerne conto, ricalcolando le percentuali emergerebbe una situazione paradossale: gli

episodi di bullismo più frequenti sarebbero proprio quelli che non sono esemplificati nel questionario. Perciò, sarebbe semmai corretto riclassificare le "false" mancate risposte come un implicito "mai".

Nel prossimo esempio esamineremo una situazione nella quale parlare di cause "strutturali" è particolarmente pertinente, perché riguarda gli ambienti nei quali si sono eventualmente verificati gli episodi di bullismo. Nella prima versione del grafico di Fig. 10.1 si vede che gli spogliatoi sono "sicuri" o "totalmente sicuri" per solo il 40% degli studenti. Più di un quarto degli intervistati però non risponde, perché gli spogliatoi non sono presenti in tutte le scuole. Ovviamente, escludendo le risposte mancanti, ci sentiamo più sollevati (gli studenti sicuri sono ora più di metà), ma il sollievo è illusorio, perché allo stesso tempo aumentano anche i più o meno insicuri (che complessivamente passano da 17 a 22%).

Figura 10.1 - Quanto ti senti sicuro: domande con e senza le mancate risposte



Comparando i due grafici si osserva che il ricalcolo delle percentuali cambia la gerarchia tra gli ambienti scolastici: se si tolgono le scuole dove non è presente, la mensa, che nel primo grafico si colloca al penultimo posto, diventa il luogo più sicuro. In realtà, è più corretto dire che la mensa "appare" il posto più sicuro perché, come abbiamo appena sottolineato, ci si dimentica che i dati dovrebbero essere considerati complessivamente (tenendo conto anche delle risposte negative). Infatti, ordinare le barre secondo la somma delle risposte di chi si sente sicuro è solo una delle opzioni; se, ad esempio, si fossero ordinate tenendo conto di quanti si sentono insicuri (usando la somma delle risposte collocate all'estremo opposto della scala), si sarebbe ottenuto un risultato in parte diverso.

9.4. Neutralizzare i dati mancanti

Per evitare di realizzare un ordinamento basato su criteri soggettivi, e dunque opinabili, si possono usare strumenti statistici che consentono di "neutralizzare" le mancate risposte, attenuando il rischio di distorsione. Si può ad esempio cercare di rispondere alla domanda virtuale: "quanti studenti si sentono insicuri per ogni studente che si dichiara sicuro?". Per rispondere si calcola il rapporto tra la somma delle risposte "totalmente sicuro" e "sicuro" e delle risposte "insicuro" e "totalmente insicuro": in questo modo si neutralizzano (cioè, non se ne tiene conto) le mancate risposte e anche la risposta "a metà strada". Il calcolo di questo rapporto non consente di stimare quanti studenti si sentono sicuri (o insicuri); si può però ad esempio affermare che, per ogni studente che si dichiara insicuro, 36 si sentono sicuri in classe. Per quanto concerne gli spogliatoi il rapporto è molto più basso: per ogni studente insicuro, solo 2,5 si sentono sicuri (Fig. 10.2).

Il calcolo del rapporto tra sicuri e insicuri è stato possibile nel nostro esempio perché la scala ordinale utilizzava aggettivi in opposizione, ma si sarebbe potuto operare similmente anche se si fosse utilizzata una scala Likert, calcolando il rapporto tra le somme di "molto/abbastanza" e di "poco/per nulla" sicuro. Analogamente, si sarebbe potuto calcolare il rapporto tra la somma delle risposte da "0" a "3" e delle risposte da "7" a "10" in una scala dove si poteva indicare quanto ci si sente insicuri da "0" (totalmente sicuro) a "10" (totalmente insicuro), ignorando le risposte intermedie (da "4" a "6"), oltre a eventuali mancate risposte.

Anche se non è appropriato statisticamente nel caso di una scala ordinale, si può anche calcolare una sorta di "punteggio medio di sicurezza", operazione che sarebbe stata senz'altro corretta se si fosse utilizzata la scala auto-ancorante con punteggi da "0" a "10" (vedi § 2.3). Per il calcolo della media si devono ovviamente ignorare le mancate risposte e si utilizza invece, a differenza del calcolo del rapporto tra risposte in opposizione, anche la risposta intermedia.

Figura 10.2 – Rapporto tra studenti che si dichiarano insicuri e sicuri

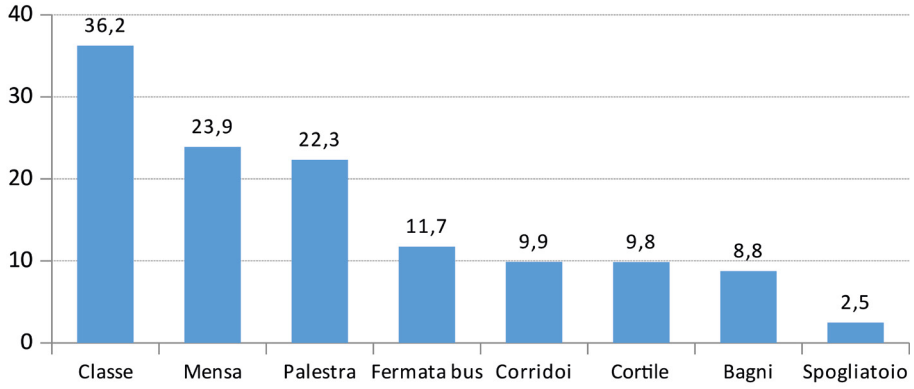
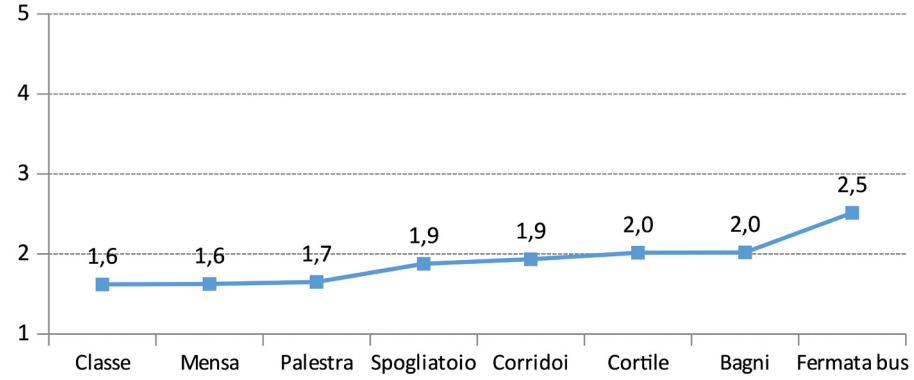


Figura 10.3 – Valori medi su scala da 1 "totalmente sicuro" a 5 "totalmente insicuro"

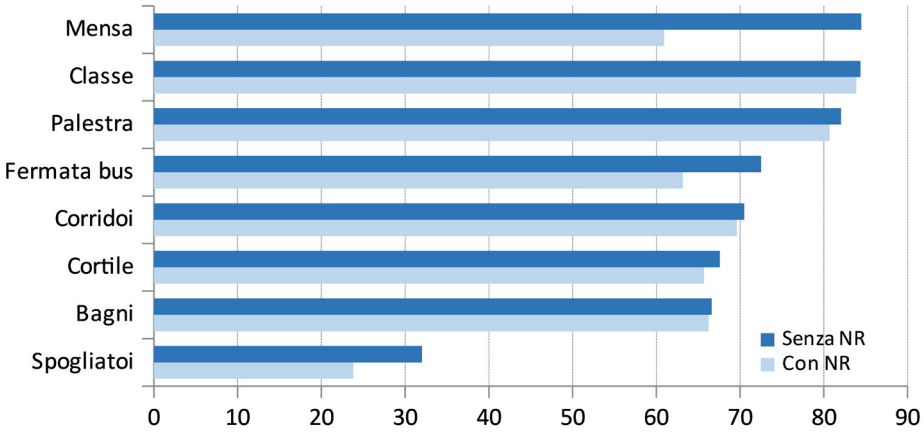


Ordinando gli ambienti scolastici secondo il punteggio medio di "insicurezza", il risultato cambia ancora (Fig. 10.3): per le prime posizioni si mantiene la stessa graduatoria (classe, mensa e palestra), ma migliora la posizione dello spogliatoio, perché la media tiene conto anche dei numerosi studenti che si colloca a metà strada tra sicuro e insicuro. Scende invece all'ultimo posto la fermata dell'autobus che, con un punteggio medio pari a 2,5, si colloca poco sotto la metà della scala i cui valori vanno da "1" (totalmente sicuro) a "5" (totalmente insicuro).

Per confrontare gli ambienti scolastici si può anche calcolare, invece del rapporto, la differenza tra le risposte che indicano sicurezza ("totalmente sicuro" e "sicuro") e insicurezza ("insicuro" e "totalmente insicuro"). Il calcolo è semplice e il risultato intuitivo: la differenza quasi in tutti i casi supera (anche di molto) i 60 punti, con l'eccezione dello spogliatoio, per il quale si riduce a poco

più di 20 o di 30 punti percentuali, a seconda che il calcolo sia fatto sulle percentuali originali o ricalcolate escludendo le mancate risposte (Fig. 10.4).

Figura 10.4 – Differenza tra "totalmente sicuro + sicuro" e "Insicuro + totalmente insicuro" con e senza NR



In generale, usare la differenza tra risposte opposte non è peraltro consigliabile, perché, paradossalmente, pur in presenza di una consistenza delle risposte escluse dal calcolo (mancate risposte e risposte intermedie) radicalmente diversa, si può arrivare a un identico risultato: con 50% di studenti sicuri, 10% "a metà" e 40% insicuri si ottiene la stessa differenza di 10 punti, che si registra con distribuzioni assai diverse (ad es., 20% sicuri, 70% intermedi e 10% insicuri).

Riassumendo, i risultati di una rilevazione si possono divulgare usando due strumenti che differiscono riguardo alle finalità e al modo di trattare i dati mancanti: percentuali e parametri statistici. Le percentuali sono ancorate alla popolazione di riferimento dell'indagine e, siccome spesso siamo interessati a conoscere la diffusione di un comportamento, di un atteggiamento, ecc., ci consentono di operare delle stime campionarie. Se questa è la finalità principale, la modalità di trattamento dei dati mancanti è cruciale, perché includerli o escluderli modifica la stima, anche in modo consistente se sono numerosi.

I parametri statistici sintetici sono invece per loro natura "astratti", in quanto ancorati alla "misura" della proprietà e non alla popolazione di riferimento come le percentuali; possono dunque essere privilegiati se si vogliono stimare "andamenti" (la diffusione o la gravità di un determinato fenomeno è aumentata nel tempo?) o fare comparazioni (in quale settore o gruppo sociale il fenomeno è più presente?).

Come abbiamo visto, è anche importante stabilire se i dati mancanti hanno natura "accidentale": in questo caso escluderli, e ripartire la loro quota tra le risposte valide, può essere relativamente innocuo. Se la causa è invece "strutturale", l'esclusione deve essere perlomeno problematizzata. Se gli intervistati che "sopravvivono" hanno un'identità sociale "debole" perché la si può descrivere solo con locuzioni generiche quali "chi ha risposto affermativamente alla domanda precedente", meglio includere nell'analisi l'intero campione. L'esclusione è più giustificabile se invece chi ha risposto ha un'identità chiaramente identificabile, cui fare riferimento per chiarire che le stime non si riferiscono all'intera popolazione (chi va al cinema, le donne, i maggiorenni o chi è iscritto a qualche associazione, come nell'esempio alla fine del precedente capitolo).

9.5. Interpretare e legittimare il "non so"

Esaminiamo ora un'altra domanda per la quale l'elevato numero di mancate risposte può far pensare a una causa "strutturale". In generale, tutti possono rispondere a domande di atteggiamento e, infatti, nell'indagine che stiamo utilizzando per le esemplificazioni si erano registrate poche mancate risposte per quasi tutte le domande di questo tipo. Nell'esempio che proponiamo era stata però esplicitamente proposta la risposta "non so", una sorta di legittimazione della mancata risposta che può essere il motivo "strutturale". Infatti, di fronte alla possibilità di utilizzare una "scappatoia", le mancate risposte arrivano a poco meno del 30% del campione.

Meno di due terzi dei ragazzi (62,3%) accetterebbe come compagno di banco un ragazzo di religione diversa dalla sua; però, togliendo "non so" e le poche mancate risposte, e ricalcolando le percentuali, i favorevoli salgono a quasi il 90% (Tab. 10.2). Il risultato è confortante per chi crede nell'interculturalità, ma lo si raggiunge mediante un notevole azzardo, come quando in un sondaggio elettorale si redistribuiscono le mancate risposte di chi non si è espresso. È dunque prudente fermarsi ai risultati della prima versione della tabella, perché si può presumere che, per apparire "*politically correct*", molti studenti non propensi ad accettare la vicinanza di studenti di altra religione abbiano preferito rifugiarsi nel "non so", che non offende nessuno, piuttosto che dichiarare esplicitamente la loro contrarietà.

Tabella 10.2 – "Ti piacerebbe sedere accanto a un ragazzo di diversa religione" per tipo di scuola, genere e nazionalità con e senza "non so" e "non risposto"

	Tipo di scuola		Nazionalità			
	Primaria	**Secondaria**	**Italiana**	**Europea**	**Non Euro**	**Totale**
Sì	73,5	62,3	57,8	78,8	80,6	62,3
No	6,6	7,0	7,8	2,5	6,0	7,0
Non so	16,6	28,7	32,6	16,3	10,4	28,7
Non risposto	3,3	2,0	1,8	2,5	3,0	2,0
Totale	100,0	100,0	100,0	100,0	100,0	100,0
	362	352	567	80	67	714

	Tipo di scuola		Nazionalità			
	Primaria	**Secondaria**	**Italiana**	**Europea**	**Non Euro**	**Totale**
Sì	91,7	87,3	88,2	96,9	93,1	89,9
No	8,3	12,7	11,8	3,1	6,9	10,1
Totale	100,0	100,0	100,0	100,0	100,0	100,0
	290	205	372	65	58	495

Le tabelle mostrano un altro rischio che si corre escludendo le risposte di questo tipo: nella seconda tabella le risposte positive si aggirano sullo stesso livello (intorno al 90%) per tutti i tipi di studenti. Una tale uniformità di atteggiamento non si riscontra però nella prima tabella, nella quale si legge che rispondono "sì" meno del 60% degli italiani e sono invece molto vicini all'80% gli studenti stranieri (europei e non europei); inoltre, sono per il "sì" solo metà degli studenti delle superiori e quasi tre quarti di quelli della primaria. Eliminando le risposte "non so", spariscono quasi completamente le differenze e viene anche occultato un fatto importante: gli italiani che non sanno/vogliono rispondere sono il doppio di quelli di altre nazionalità europee e il triplo degli extra-europei.

Passiamo ora a una domanda che fornisce una sorta di controprova sul ruolo del "non so". Si chiedeva se qualche alunno fosse stato preso in giro perché non parlava o parlava male l'italiano e solo pochi studenti (0,7%) non hanno risposto. È poi presente "non so" con una frequenza (15,1%) che anche in questo caso sarebbe stata presumibilmente più bassa se la risposta non fosse stata stimolata. Definendo anche questa risposta "mancante", possiamo verificare l'effetto della sua esclusione nell'ultima colonna della Tab. 10.3. In questo caso l'esclusione di "non so" dal calcolo delle percentuali, oltre che arbitraria (la risposta è presente nel questionario), è fuorviante perché aumenta "artificialmente" la percentuale di studenti i quali affermano che nella loro scuola si sono verificati episodi di bul-

lismo. È invece probabile che tali episodi non siano accaduti (dunque, le risposte si potrebbero/dovrebbero sommare a "no, mai") perché in successive domande su altre forme di bullismo, nelle quali "non so" non era suggerito, la percentuale di "mai" è stata sempre pari o superiore alla somma dei "mai" e "non so" conteggiati in questa domanda.

Tabella 10.3 – "Qualche alunno è stato preso in giro durante lo scorso anno scolastico perché non parlava bene l'italiano?"

		N	%	% valide
Valide	Sì, spesso	92	12,9	15,3
	Sì, qualche volta	233	32,6	38,8
	Sì, una volta	75	10,5	12,5
	No, mai	201	28,2	33,4
	Totale	601	84,2	100,0
Mancanti	Non so	108	15,1	
	Non risposto	5	,7	
	Totale	113	15,8	
Totale		714	100,0	

Come si vede, i risultati possono in parte cambiare a seconda del modo di porre la domanda e di trattare statisticamente le risposte mancanti (o ritenute tali) e non solo, come dovrebbe essere, in ragione delle opinioni o comportamenti degli intervistati. Anche in questo caso la "distorsione percettiva" è un po' attenuata perché, escludendo le risposte "non so", aumenta chi complessivamente afferma che gli episodi di bullismo sono accaduti (da 56 a 66,7%), ma aumenta allo stesso tempo chi lo nega (da 28,2 a 33,3%). Rimane dunque immutato il rapporto tra chi afferma e chi nega, ma non è giusto che rimanga immutato se è ragionevole legittimare il "non so" e sommarlo a "mai".

9.6. Domande non pertinenti e risposte incoerenti

Nel questionario alcune domande erano tra loro collegate e ciò ci consente di affrontare il tema della coerenza nelle risposte che non si può dare per scontata, particolarmente in un'indagine, come quella che stiamo usando per le esemplificazioni, nella quale il questionario era auto-compilato.

Come anticipato, una serie di domande del questionario riguardavano gli episodi di bullismo accaduti nella scuola e l'analisi delle risposte ha consentito di

suddividere gli intervistati tra chi aveva affermato che tali episodi erano accaduti e gli altri. Si era poi chiesto agli studenti di indicare la propria reazione a questi fatti e, con qualche sorpresa, si è visto che numerosi studenti, nonostante avessero risposto sempre "mai", hanno ugualmente risposto alla domanda sulle loro reazioni a tali situazioni. Si può ovviamente presumere che questi studenti abbiano risposto ugualmente a tali domande perché si sono immedesimati e hanno immaginato la propria reazione nel caso avessero assistito a episodi come quelli che erano stati prospettati nelle domande precedenti.

Comunque, se può essere labile il nesso tra episodi dei quali spesso non si ha una conoscenza diretta e ipotetiche reazioni, ci si può attendere maggiore coerenza da risposte a domande che riguardano episodi di bullismo dei quali gli intervistati sono stati (eventuali) vittime. Anche in questo caso si chiedeva allo studente se nel corso dell'anno gli fosse successo di essere preso in giro (insultato sui social, ignorato o evitato, colpito o maltrattato, ecc.) e gli studenti che hanno sempre risposto "mai" sono stati distinti dagli altri e i due gruppi ulteriormente suddivisi tra italiani e stranieri, perché le domande facevano esplicito riferimento all'origine etno-nazionale (ad es.: "mi hanno preso in giro o insultato a causa della mia diversa lingua, religione o cultura"). Come era forse prevedibile, ha dichiarato di essere stato vittima almeno una volta il 14,5% degli studenti italiani e il 50,3% degli stranieri. Incrociando i dati con la domanda che chiedeva di indicare il genere degli aggressori, si osserva con sorpresa che una quota non trascurabile degli intervistati (oltre il 10%) ha risposto in modo contraddittorio, indicando il genere degli aggressori pur avendo affermato di non avere mai subito episodi di bullismo.

Tabella 10.4 – "Ti sei trovato a subire episodi di bullismo?" per "Chi di solito ti fa queste cose?"

		Chi di solito ti fa queste cose						
		Maschi	Femmine	Entrambi	NR/NP	Totale		
Subìto bullismo		%	%	%	%	%	N	%
Italiano	No	5,6	,8	6,4	87,2	100,0	485	85,5
	Sì	30,5	11,0	31,7	26,8	100,0	82	14,5
	Totale	9,2	2,3	10,1	78,5	100,0	567	100,0
Straniero	No	5,5	1,4	6,8	86,3	100,0	73	49,7
	Sì	24,3	10,8	36,5	28,4	100,0	74	50,3
	Totale	15,0	6,1	21,8	57,1	100,0	147	100,0

Analogamente, incrociando la domanda sugli episodi di bullismo subiti con quella sulle proprie reazioni emergono numerose risposte incongrue di studenti

che non hanno compreso che si chiedeva di riferire la reazione di fronte a fatti effettivamente accaduti (Tab. 10.5). I dati sono interessanti perché mostrano, ad esempio, maggiore "reattività" da parte dagli studenti stranieri. In questa sede ci interessa però sottolineare il fatto che, forse perché ci si riferisce a reazioni personali e non ad aspetti fattuali (il genere dei persecutori), aumentano notevolmente i casi di studenti che rispondono ugualmente alla domanda anche se non li riguardava: paradossalmente ha risposto oltre la metà (due terzi tra gli stranieri) di chi non ha subito alcun episodio di bullismo. Si sono probabilmente "immedesimati" e hanno risposto alla domanda virtuale: "se ti fosse accaduto, come avresti reagito?".

Tabella 10.5 – "Ti sei trovato a subire episodi di bullismo?" per "Come reagisci?"

		Come reagisci							
		Reagisco	Sopporto	Chiedo aiuto	Altro	NR/NP	Totale		
Subìto bullismo		%	%	%	%	%	%	N	%
Italiano	No	25,2	16,1	4,3	24,7	42,7	100,0	485	85,5
	Sì	39,0	19,5	7,3	46,3	9,8	100,0	82	14,5
	Totale	27,2	16,6	4,8	27,9	37,9	100,0	567	100,0
Straniero	No	37,0	16,4	2,7	17,8	34,2	100,0	73	49,7
	Sì	59,5	6,8	5,4	35,1	9,5	100,0	74	50,3
	Totale	48,3	11,6	4,1	26,5	21,8	100,0	147	100,0

Nella colonna contrassegnata "NR/NP" le mancate risposte (studenti vittime di bullismo che non hanno risposto come reagiscono) sono riportate assieme alle risposte "non pertinenti" (studenti che correttamente non hanno indicato le loro reazioni dal momento che non sono stati vittime di queste situazioni).

In questi casi il ricercatore potrebbe decidere di annullare le risposte incoerenti, introducendo un nuovo tipo di dati mancanti: le risposte annullate. Con la loro eliminazione cambia però ancora la base di calcolo delle percentuali e, di conseguenza, il risultato della rilevazione; diventa perciò sempre più difficile stimare correttamente quanti studenti agiscono in un certo modo, quanti hanno una determinata opinione, ecc.

Nei questionari a volte si pone una domanda "filtro" (ad es. "vai al cinema?"), seguita da una o più domande di approfondimento (con che frequenza, in che giorni, con chi, quali tipi di film ti piacciono) e ciò è stato fatto, non a caso, affrontando il tema "critico" del comportamento degli insegnanti. Alla domanda "gli insegnanti intervengono quando un alunno è trattato male" si è registrato il 15% di mancate risposte e il 6% ha risposto "mai", in linea con il fatto che poco più del

20% afferma che non si sono verificati episodi di bullismo nella loro scuola (per il 40% gli insegnanti intervengono "qualche volta" e per il 38,4% "sempre").

Alla successiva domanda su "perché gli insegnanti non intervengono" avrebbe dovuto rispondere solo chi ha risposto che intervengono 'qualche volta' o 'mai' (nel complesso 46,6 %), ma le risposte "valide" sono state meno del 25%. Vi è dunque un'eccedenza di mancate risposte perché evidentemente non ha risposto anche chi non vuole/non sa esprimere un giudizio o ritiene sufficiente che gli insegnanti intervengano "qualche volta", e cioè nei casi più gravi.

Nella Tab. 10.6 sono visualizzate solo le risposte affermative (si poteva indicare più di un motivo): nella prima colonna le percentuali sono calcolate sull'intero campione, nella seconda escludendo le risposte "mancate" (se vogliamo chiamarle così). Confrontando i dati, la valutazione del fenomeno cambia radicalmente, perché i valori nella seconda colonna si sono più che quadruplicati.

Tabella 10.6 – Perché gli insegnanti non intervengono (risposte affermative per tutti gli studenti e con esclusione delle mancate risposte)

	Tutti gli studenti	**Senza mancate risposte**
Non sanno cosa sta accadendo	12,6	55,9
Non vogliono intervenire	2,8	12,5
Non è loro compito intervenire	1,5	6,9
Gli studenti non li ascoltano	2,7	11,9
Non gliene importa	5,3	23,8
Hanno paura	4,5	20,0
Altro	2,1	9,7

Di fronte a differenze di questa entità è legittimo chiedersi quali valori rappresentino più "fedelmente" la situazione: gli insegnanti ai quali "non importa" sono il 5,3 o il 23,8%? Entrambe le serie di percentuali sono tecnicamente corrette, ma usando la seconda si deve prestare molta attenzione alla comunicazione dei risultati. Innanzitutto, anche se non si può negare che per un quarto degli studenti gli insegnanti si disinteressano di ciò che accade loro, oltre tre quarti di loro non è di questo avviso. Inoltre, e ancor più importante, alla domanda ha risposto meno di un quarto degli studenti; dunque, anche se sembra uno scioglilingua, meno di un quarto di meno di un quarto degli studenti pensa che agli insegnanti non importa cosa accade. Ci si può però chiedere se sono davvero necessarie tutte queste precisazioni, visto che la prima colonna di percentuali della Tab. 10.7 descrive con chiarezza la situazione: oltre tre quarti degli intervistati (77,6%) non risponde, probabilmente perché non si sono verificati episodi di

bullismo o gli insegnanti sono intervenuti e un altro 17,1% pensa che il mancato intervento sia dovuto ad altre cause (come si è visto nella Tab. 10.6, per la maggior parte degli intervistati gli insegnanti non sono al corrente). Dunque, solo per il 5,3% degli studenti gli insegnanti non intervengono perché non gliene importa.

Tabella 10.7 – Gli insegnanti non intervengono perché "non gliene importa"

		N	%	% valide
Valide	Sì	38	5,3	23,8
	No	122	17,1	76,3
	Totale	160	22,4	100,0
Mancanti	Non risposto	554	77,6	
Totale		714	100,0	

9.7. Quali risposte sono davvero mancanti?

Facciamo ora riferimento a una questione sintetizzabile nella domanda: la risposta è davvero mancante? Con ciò si intende dire che a volte è necessario che il ricercatore operi una sorta di esegesi del testo prima di decidere che l'intervistato davvero non ha risposto o che la sua risposta è incoerente. La necessità di interpretare i segni che l'intervistato ha lasciato (o omesso) sul questionario ovviamente si pone prevalentemente nel caso di questionari auto compilati, perché intervistatori addestrati e diligenti dovrebbe garantire una compilazione corretta, che non richieda esercizi interpretativi.

L'esperienza insegna però che anche gli intervistatori esperti (anzi, particolarmente loro) per velocizzare l'esecuzione si limitano a segnare le caselle "Sì" per indicare ad es. le associazioni di cui gli intervistati fanno parte, sembrando superfluo contrassegnare le altre con '"No". Ovviamente, questa interpretazione è credibile di fronte ad almeno un "Sì" o una risposta valida; se invece non ci fosse alcun segno di risposta per tutte le domande, può essere invece realistico ritenere che l'intervistato abbia deciso di non rispondere.

Questa ricostruzione regge però a patto che non ci sia il ragionevole dubbio che, pur avendo risposto ad alcune domande, l'intervistato non abbia voluto o potuto esprimersi sulle altre perché, ad esempio, presuppongono conoscenze che l'intervistato non possiede, si riferiscono a comportamenti da lui non praticati o a questioni sulle quali l'intervistato preferisce mantenere il riserbo.

Tutto ciò rivela che nelle batterie di domande ci sono almeno due tipi di risposte mancanti: la mancata risposta "globale" (rifiuto o difficoltà di rispondere all'insieme delle domande) e la mancata risposta "locale" (rifiuto o difficoltà di rispondere

su una singola questione perché non conosciuta, non praticata, considerata fatto privato, ecc.). C'è poi la successione "domanda filtro – domanda condizionata" rispetto alla quale si possono verificare quattro situazioni (Tab. 10.8). Se un intervistato che ha risposto affermativamente alla domanda filtro risponde anche alla domanda seguente, la compilazione è corretta; se non risponde, si registra invece una mancata risposta alla seconda domanda. Se l'intervistato risponde negativamente alla domanda filtro, è corretto che non risponda alla domanda seguente perché per lui non è pertinente; se risponde, si pone un problema di incoerenza che può essere risolto annullando la seconda risposta, o correggendo la prima.

Tabella 10.8 – Combinazioni sequenza domanda filtro – domanda condizionata

		Domanda condizionata (quali specifici episodi sono accaduti)	
		Risponde	Non risponde
Domanda filtro (c'è bullismo a scuola)	Si	Corretto	Mancata risposta
	No	Incoerente	Corretto (non pertinente)

"Istintivamente" si tende ad adottare la prima soluzione, ma non è detto che sia una saggia decisione, perché non c'è un solido motivo per considerare corretta la prima risposta, se non il fatto che viene per prima. Generalmente la prima domanda è molto sintetica, perché il compito di precisare maggiormente la questione è demandato alle domande successive; dunque, può capitare che l'intervistato risponda "no" alla prima domanda in buona fede, solo perché non ha ben compreso ciò che gli si chiede. Se le domande successive fanno venire in mente qualcosa che non aveva considerato, o comprende meglio cosa il ricercatore intendesse, può accadere che risponda alla domanda successiva, senza ricordarsi di correggere quella precedente; il ricercatore però, se ritiene valida questa ricostruzione, può legittimamente farlo. Ciò è tanto più sensato quanto più dettagliate sono le risposte fornite alle domande successive alla prima e, a maggior ragione, se si tratta di domande aperte che richiedono una più o meno estesa verbalizzazione della risposta.
Un secondo esempio di situazione nella quale può essere consigliabile "correggere" le risposte è mostrato nella Tab. 10.9 dove sono illustrati due modi simili di porre la stessa domanda. Con il primo modo (a sinistra) si propone una lista e l'intervistato è invitato a contrassegnare le associazioni cui è iscritto (anche più di una) con una spunta sulle caselle; nel secondo modo (a destra) la lista è proposta nella forma di una serie di domande dicotomiche.

Tabella 10.9 – Domande a risposta multipla

A quali associazioni sei iscritto: (possibili più risposte)			No	Sì
- ambientalistiche	☑	- ambientalistiche	0	~~1~~
- assistenziali	☐	- assistenziali	0	1
- culturali	☐	- culturali	0	1
- musicali	☐	- musicali	0	1
- religiose	☑	- religiose	0	~~1~~
- sportive	☑	- sportive	0	~~1~~
- altro (_ _ _ _ _ _)	☐	- altro (_ _ _ _ _ _)	0	1

Siccome anche alla domanda posta nel primo modo si possono dare più risposte, è opportuno anche in questo caso dedicare una variabile dicotomica a ogni associazione usando i codici "0" o "1"; strutturata in questo modo la procedura di registrazione non si viene sfiorati dal dubbio che le caselle senza la spunta possano essere considerate mancate risposte.

Nel secondo esempio la forma della domanda è esplicitamente dicotomica, al fine di suggerire con ancor più evidenza che si può rispondere affermativamente anche per tutte le associazioni; però, come si vede dall'esempio e suggerisce la pratica di ricerca, accade che gli intervistati, anche solo per velocizzare la compilazione, si limitino a indicare le associazioni a cui sono iscritti e trascurino di segnare per le altre la risposta "no". È comunque evidente che non si tratta di risposte mancate, perché non è ragionevole che non si voglia rispondere riguardo alle associazioni musicali, culturali o assistenziali, dopo avere dichiarato l'appartenenza ad associazioni religiose, sportive e ambientalistiche (ed è altrettanto improbabile che non si sappia se si è iscritti o meno). Se, attenendoci scrupolosamente a quanto si vede nel questionario compilato, decidessimo di codificare le associazioni non contrassegnate come mancate risposte, ossequieremmo la forma, ma tradiremmo la sostanza.

9.8. I dati mancanti nell'analisi multivariata

Come si è visto, nell'analisi monovariata la "gestione" dei dati mancanti è relativamente semplice: si deve decidere se includerli o no nel calcolo delle percentuali e li si deve invece escludere per calcolare la media o individuare la mediana, anche se a volte si possono sostituire con stime. Se, ad esempio, vogliamo sapere con buona approssimazione quanti saranno i nuovi immatricolati in Italia, e tardano ad arrivare i dati da uno o due atenei, possiamo operare una stima usando i dati dell'anno precedente, eventualmente corretti applicando il tasso di variazione che si constata nelle università di cui possediamo i dati.

Nell'analisi bivariata, i dati mancanti possono essere presenti in una sola o in entrambe le variabili e, come si è visto, nelle tabelle di contingenza si può decidere se includerli o escluderli, tenendo conto della loro numerosità e del ruolo delle due variabili (dipendente o indipendente). Vanno poi esclusi dal calcolo dei parametri statistici, ma si può calcolare il reddito medio (variabile cardinale) anche di chi non ha indicato il partito per cui vota (variabile categoriale), perché "non risponde" è una categoria. Ovviamente, per fare il calcolo si possono usare solo i redditi dichiarati, sostituendo eventualmente i dati mancanti con una stima. Infine, analizzando l'associazione tra variabili, vanno esclusi i casi che presentano dati mancanti anche su una sola variabile, ma fa eccezione il chi quadrato, perché si utilizza con variabili categoriali e anche "non so" o "non risponde" sono categorie (di risposta).

Nell'analisi multivariata il problema si complica, perché possono mancare dati in una, più di una o tutte le variabili presenti nel modello. In questi casi si offrono al ricercatore due alternative, la cancellazione per lista (*listwise*) o a coppie (*pairwise*). Come si è già visto nel § 4.3.2, nell'analisi fattoriale, tecnica multivariata che si basa sul calcolo dei coefficienti di correlazione bivariata tra tutte le variabili presenti nel modello, adottando la prima alternativa si elimina un caso se manca l'informazione anche per una sola variabile. Nell'eliminazione a coppie, se manca l'informazione anche per una sola delle due variabili il caso viene eliminato dal calcolo di quel coefficiente, ma se entrambe le informazioni sono disponibili rientra in gioco nel calcolo di altri coefficienti,.

Per decidere come sia meglio operare si deve ovviamente tenere conto di come i dati mancanti sono distribuiti: se lo sono in modo casuale, la cancellazione *listwise* può comportare la perdita di un numero elevato di casi; infatti, al crescere del numero di variabili, aumenta la probabilità che manchi l'informazione anche su una sola variabile e si rischia perciò che alla fine l'analisi si basi su un numero molto limitato di casi. Nell'esempio fittizio della Tab. 10.10, i cinque dati mancanti (codice "9") su un totale di 50, comportano l'eliminazione di cinque casi e dunque il dimezzamento del "campione" (con la perdita di 20 informazioni rilevate correttamente). L'eliminazione a coppie è più conservativa, perché con gli stessi dati i coefficienti di correlazione bivariata si possono calcolare usando da un minimo di 6 casi (V1 con V4) a tutti (correlazione tra V2 e V3). Si paga perciò un prezzo (i coefficienti, calcolati su un numero di casi variabile, hanno diversa significatività statistica), ma non è molto elevato se i dati mancanti sono poco numerosi in rapporto al totale. Se però si concentrano su una o poche variabili, può essere saggio eliminare quelle variabili, anche perché le numerose mancate risposte potrebbero indicare specificità che non consentono di considerarle omogenee alle altre (sono ad es. pertinenti solo per parte degli intervistati). Nell'esempio di Tab. 10.10, si potrebbe eliminare la variabile V4, responsabile della maggior parte dei dati mancanti.

Tabella 10.10 – Cancellazione listwise e effetto della cancellazione pairwise sulla matrice di correlazione

Caso	**V1**	**V2**	**V3**	**V4**	**V5**
C1	1	2	1	**9**	1
~~C2~~	~~**9**~~	~~1~~	~~1~~	~~2~~	~~1~~
~~C3~~	~~1~~	~~2~~	~~1~~	~~1~~	~~2~~
C4	1	1	2	1	**9**
~~C5~~	~~2~~	~~1~~	~~1~~	~~**9**~~	~~1~~
~~C6~~	~~1~~	~~2~~	~~1~~	~~1~~	~~1~~
C7	1	2	1	2	2
C8	2	2	2	**9**	1
~~C9~~	~~2~~	~~1~~	~~1~~	~~1~~	~~2~~
C10	1	2	2	2	2

Casi usati per le correlazioni

	V1	V2	V3	V4	V5
V1	-				
V2	9	-			
V3	9	10	-		
V4	6	7	7	-	
V5	8	9	9	6	-

Anche in questi modelli più complessi si può considerare la possibilità di sostituire il dato mancante con una stima. Siccome le variabili che si utilizzano in questi modelli sono cardinali, il dato mancante si può sostituire con la media calcolata sui casi per i quali il dato è disponibile o, più accuratamente, con la media calcolata all'interno della categoria sociale di appartenenza di chi non ha risposto.

10. L'esposizione dei risultati

Molti testi di metodologia della ricerca sociale, anche se non tutti, terminano senza ritenere che possa essere utile dare qualche consiglio metodologico e tecnico finalizzato a fare in modo che il processo di ricerca si concluda con un "lieto fine". È un vero peccato, perché il lavoro del ricercatore non termina certamente con l'analisi dei dati, che è un passaggio essenziale, ma pur sempre strumentale alla verifica delle ipotesi formulate all'inizio del processo di ricerca oppure, nel caso di ricerche con finalità prevalentemente applicative, a fare il punto riguardo al raggiungimento o meno degli scopi conoscitivi che ci si era prefissi o richiesti dal committente della ricerca. Anche se il ricercatore non si prefigge la verifica di ipotesi chiaramente esplicitate, dovrà comunque giungere a una qualche conclusione e alla sintesi dei principali risultati conseguiti mediante il lavoro d'indagine. Di fatto, come sostiene Ricolfi: «l'*esposizione dei risultati* costituisce l'ultimo livello e, di norma, anche l'ultima fase di una ricerca empirica. In genere ha almeno tre compiti: a) rendere il più trasparente possibile l'itinerario di ricerca; b) comunicare i risultati più importanti ottenuti in sede di analisi; c) stabilire un raccordo con la lettura precedente e, eventualmente, suggerire nuove linee di ricerca» (1995: 396-397).

Che sia utile portare a termine al meglio questi compiti è fuori di dubbio, ma non si tratta solo di utilità, perlomeno da quando larga parte della ricerca sociale

(e non) è finanziata da progetti, nazionali e specialmente europei, che enfatizzano molto quella che con un termine ormai entrato stabilmente a far parte del lessico del ricercatore, è definita "disseminazione". Qualsiasi sia la natura del progetto finanziato dai programmi europei non c'è da dubitare che uno dei WP (*work packages* o pacchetti di lavoro) sarà dedicato alla comunicazione, principalmente finalizzata alla disseminazione dei risultati del lavoro di ricerca. Inoltre, a proposito di termini entrati ormai da qualche tempo a far parte del lessico universitario, emerge con sempre maggiore forza la necessità di fare fronte alla c.d. "terza missione" che si aggiunge alla scontata missione didattica e all'attività di ricerca che è, in linea di massima, il tratto che distingue l'università dagli altri livelli del sistema educativo. Come ci ricorda l'Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca (ANVUR), richiamando la VQR e cioè il sistema di Valutazione della Qualità della Ricerca:

> Nella VQR 2004-2010, ANVUR ha introdotto il concetto di Terza Missione come "apertura verso il contesto socio-economico mediante la valorizzazione e il trasferimento delle conoscenze", in una concezione più attuale che include oltre alle attività di valorizzazione economica della ricerca, anche iniziative dal valore socio-culturale ed educativo (https://www.anvur.it/attivita/temi/).

Chiarito che è ormai cruciale, ma in fondo lo è sempre stato, portare a una degna conclusione il processo di ricerca, aggiungiamo che non si è parlato a caso di "lieto fine", perché una ricerca condotta con correttezza e rigore si conclude comunque "positivamente" anche nel caso in cui le ipotesi di partenza non risultino suffragate dalle risultanze empiriche. Infatti, ai fini della cumulabilità della conoscenza scientifica, arrivare alla conclusione che una determinata ipotesi di ricerca non è stata confermata è comunque utile, perché la scienza si avvantaggia anche di evidenze empiriche - sia pure poco gratificanti per il singolo ricercatore - che dimostrino che in futuro si dovranno percorrere altre strade per raggiungere risultati migliori. Bisognerebbe, come accennato in precedenza, tentare di contrastare la tendenza a «... privilegiare le conferme anziché le smentite alle proprie aspettative (secondo il principio baconiano per cui "errore peculiare e perpetuo dell'intelletto umano è quello di essere mosso e stimolato più dalle affermazioni che dalle negazioni")» (Stanzione 1981: 56).

Su questo punto mi si consenta la citazione di un autore che ha poco a che fare con la metodologia della ricerca scientifica e, ciò nonostante, pone il problema in maniera arguta e intelligente e spiega l'atteggiamento che deve avere il ricercatore consapevole del suo ruolo:

> Uno scienziato da baraccone che, con una attrezzatura da Frankestein, faccia sensazionali "esperimenti", sa in anticipo il risultato dei suoi tramestii e quindi non fa affatto

> un lavoro scientifico. In compenso, un meccanico che suona il clacson della moto per vedere se la batteria è carica fa, in modo informale, un vero e proprio esperimento scientifico, poiché verifica un'ipotesi facendo la domanda direttamente alla natura. Lo scienziato televisivo che borbotta tristemente "L'esperimento è un fiasco; non siamo riusciti ad ottenere quello che speravamo" è vittima di un copione scadente. Un esperimento che non ottenga i risultati previsti non è un fiasco. Lo è solo quando non fornisce alcuna conclusione valida, in un senso o nell'altro, rispetto alle ipotesi di partenza (Pirsig 1981: 360).

È dunque una buona consuetudine esplicitare chiaramente nelle conclusioni di un lavoro di ricerca empirica la "sintesi delle risultanze" e cioè un riassunto, elaborato per precisi punti, in cui vengono evidenziati i più rilevanti risultati conseguiti. Si tratta di quelle che nell'ambito del processo di ricerca si definiscono "generalizzazioni empiriche", che hanno natura induttiva perché, come chiarisce il termine "empiriche", sono generalizzazioni che derivano dalla conoscenza acquisite "sul campo". Queste generalizzazioni vanno formulate in ogni caso perché, oltre a dare delle risposte al committente della ricerca (che può essere il ricercatore stesso), consentono ad altri ricercatori di valutare a prima vista la rilevanza della ricerca per il loro lavoro e di decidere se valga la pena ripercorrere la strada che ha portato a quei risultati per, ad esempio, metterne a prova la congruità rispetto agli strumenti utilizzati e alle informazioni raccolte o per formulare diverse ipotesi interpretative.

In alcuni casi eventuali nuove ipotesi possono essere testate dallo stesso ricercatore che ha condotto la ricerca, perché le matrici dei dati solitamente sono piuttosto ricche di informazioni. Quando progetta il questionario, accanto alle domande che generano variabili dipendenti e indipendenti da utilizzarsi nell'ipotetico modello interpretativo, il ricercatore inserisce molto spesso altre domande, perché ritiene che queste informazioni, pur marginali rispetto al modello stesso, possano rivestire un qualche interesse di per sé. Inoltre, vengono spesso poste domande relative a caratteristiche socio-demografiche e socio-economiche degli intervistati anche se nelle ipotesi di partenza non è previsto che tutte abbiano un ruolo nei modelli di spiegazione.

Questa ricchezza di informazioni può in alcuni casi dimostrarsi preziosa per apportare le necessarie correzioni di rotta ed è opportuno, e istruttivo per il lettore attento, che il ricercatore esponga con sufficiente dettaglio il percorso che lo ha portato al modello finale: non diminuisce per nulla il suo valore svelare che il percorso è stato accidentato.

Se si parla di ricerche condotte con correttezza e rigore non è necessario dissimulare, riformulando ad arte le ipotesi di partenza in modo da far sembrare che fin dall'inizio la strada imboccata fosse quella giusta. Non vi è nulla di particolarmente sconveniente nell'ammettere che ipotesi plausibili "sulla carta" non si

sono dimostrate corrette. Molto spesso il ricercatore formula infatti le ipotesi di partenza sulla base di teorie accreditate dalla letteratura scientifica e, dunque, se le ipotesi sono state dedotte correttamente dalla teoria, può essere riconosciuto come un suo merito l'essere riuscito a falsificarle.

In altri casi la matrice dei dati non offre informazioni in grado di aiutare il ricercatore a testare ipotesi alternative a quelle formulate in partenza, ma è possibile effettuare un supplemento di raccolta di informazioni che permettano di integrare quelle già presenti nella matrice dei dati, in particolare se si tratta di dati reperibili da fonti statistiche. Una integrazione delle informazioni può essere invece difficilmente realizzabile nel caso di un'indagine condotta mediante questionario, specialmente se anonimo. A volte però è possibile ricontattare, magari telefonicamente, alcuni intervistati, oppure si possono riascoltare brani di interviste registrate per rilevare aspetti che si erano trascurati. Se si tratta invece di un'analisi basata sull'uso di documenti (testi, immagini, video, ecc.) si possono riesaminare per effettuare una più puntuale analisi del contenuto.

Nel caso in cui l'indagine sia stata svolta attraverso un questionario, un accorgimento utile a migliorare la qualità della comunicazione dei risultati è riprodurre in appendice una copia del questionario stesso. Nel corso della trattazione non è infatti sempre possibile, o opportuno, trascrivere con completezza il testo delle domande formulate e cioè esplicitare puntualmente la definizione operativa adottata. Riportare il testo integrale del questionario permette invece questo controllo sulla correttezza della formulazione delle domande. Inoltre, conoscere la collocazione di ogni singola domanda nel contesto del questionario permette di valutare, ad esempio, la possibilità che la risposta fornita a una specifica domanda possa essere stata influenzata da quanto chiesto in precedenza. Se invece di riprodurre il questionario "anastaticamente" in allegato, viene proposta una versione nella quale si inseriscono i valori percentuali rilevati (le distribuzioni di frequenza), scrivendoli al posto dei codici o delle caselle utilizzate per contrassegnare le risposte, si ottiene un vantaggio aggiuntivo, perché in tal modo è anche possibile apprezzare e confrontare con un'occhiata i principali risultati.

Infine, se il ricercatore ha utilizzato complessi modelli di analisi multivariata, è raccomandabile, per aumentare la controllabilità delle operazioni effettuate e l'utilità del proprio lavoro per altri ricercatori, pubblicare la completa matrice di correlazione bivariata (ovviamente, solo per le coppie di variabili il cui livello di misurazione ammetta l'utilizzo di questo coefficiente di associazione).

Va anche nuovamente ricordata l'importanza degli strumenti grafici di presentazione dei risultati. Il ricercatore lavora spesso con tabelle perché presentano i dati con maggiore precisione e, inoltre, ne conosce bene l'utilizzo. Tale conoscenza può però non essere completamente condivisa da parte del pubblico a cui il ricercatore si rivolge.

> Ludwig Fleck (...) distingue quattro tipi di *audience* che contano per lo sviluppo del lavoro scientifico: a) il circolo più ristretto di specialisti - (...) l'*invisible college* - che capisce a fondo i problemi che uno studioso sta affrontando: una dozzina o una quindicina di persone, sparse in tutto il mondo; b) la cerchia più larga degli specialisti di quella disciplina o di discipline affini; c) il pubblico colto; d) il pubblico in generale (Toulmin 1984: 15).

Come è evidente, per lo meno per le ultime due categorie può essere molto più agevole la presentazione con l'immediatezza comunicativa della forma grafica al posto di aride tabelle di numeri, spesso afflitti dall'insopportabile precisione di una sfilza di decimali assolutamente ingiustificati e ingiustificabili nella maggior parte dei casi, dal momento che spesso si tratta di indagini campionarie con margini di errore che confliggono con la pretestuosa "precisione fuori luogo" a cui si è fatto riferimento in precedenza.

La presentazione dei risultati attraverso il rapporto di ricerca dovrebbe comunque essere più di tutto il momento della "trasparenza". Il ricercatore dovrebbe cioè sentirsi in dovere di accompagnare il lettore nell'esposizione di tutte le fasi della ricerca e ciò costituisce l'occasione per un riesame del lavoro effettuato da parte del ricercatore stesso. Egli, in pratica, dovrebbe porsi criticamente alcune domande così riassunte da Garson (1976): Le prove che si sono raccolte sono forse il risultato del caso? Le prove stesse portate in supporto alle relazioni tra le variabili sono veramente abbastanza forti da confermare ed essere coerenti con il modello ipotizzato a priori? Le prove sarebbero le stesse in un altro posto o tempo? (E cioè, qual è l'ambito di validità delle relazioni che abbiamo trovato?) Inoltre, abbiamo veramente misurato ciò che intendevamo misurare? Abbiamo controllato che le relazioni non siano spurie a causa di variabili di cui non abbiamo tenuto conto nel modello? Se i dati sono di natura quantitativa, ci sono test di significatività statistica e misure di associazione che possano essere usati e, se li abbiamo utilizzati, abbiamo utilizzato quelli più appropriati per il livello di "misurazione" delle variabili, e così via? Bisogna essere in grado di esplicitare questi problemi e di rispondere con onestà.

11. Riferimenti bibliografici

Agnoli M.S. (2003), *Il disegno della ricerca sociale*, Carocci, Roma.

Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca (2014), *Rapporto sullo stato del sistema universitario e della ricerca - 2013*, https://www.anvur.it/wp-content/uploads/2014/03/Rapporto%20ANVUR%202013__UNIVERSITA%20e%20RICERCA__integrale.pdf.

AlmaLaurea, *La condizione professionale dei laureati. Indagine 2020, Note metodologiche*, https://www.almalaurea.it/sites/almalaurea.it/files/docs/universita/occupazione/occupazione19/note-occupazione2020.pdf.

Amaturo E. (1988), *Analyse des données e analisi dei dati nelle scienze sociali*, Centro scientifico editore, Torino.

Arculeo A., Marradi A. (1985), *Relazione fra elezioni e referenda negli anni settanta*, «Rivista italiana di scienza politica», n.1: 99-141.

Benzécri J.P. *et al.* (1973), *L'analyse des données*, Dunod, Parigi.

Biolcati-Rinaldi F., Vezzoni C. (2012), *L'analisi secondaria nella ricerca sociale. Come rispondere a nuova domande con dati già raccolti*, Il Mulino, Bologna.

Biorcio R. (1993), *L'analisi dei gruppi*, FrancoAngeli, Milano.

Chiesi A.M. (1999), *L'analisi dei reticoli*, FrancoAngeli, Milano.

Cibois P. (1984), *L'analyse des données en sociologie*, Puf, Paris.

Cobalti A., Schizzerotto A. (1994), *La mobilità sociale in Italia*, Il Mulino, Bologna.

Colombis A. (1989), *Ricerca sociale, problemi gnoseologici e "rivoluzione" informatica alla fine degli anni '80: the day before*, in Crespi F. (cur.), *Sociologia e cultura: nuovi paradigmi teorici e metodi di ricerca nello studio dei processi culturali*, FrancoAngeli, Milano: 53-78.

Coppola L. (2011), *NVivo: un programma per l'analisi qualitativa*, FrancoAngeli, Milano.

De Gregorio E., Lattanzi P.F. (2011), *Programmi per la ricerca qualitativa. Guida pratica all'uso di ATLAS.ti e MAXQDA*, FrancoAngeli, Milano.

Delli Zotti G. (2021), *Metodi e tecniche della ricerca sociale. Vol. I La rilevazione dei dati*, Eut, Trieste, https://www.openstarts.units.it/handle/10077/32180.

Delli Zotti G. (cur.) (2014), *Children's Voices. Etnicità e bullismo nella scuola*, Bonanno, Acireale-Catania.

Delli Zotti G. (2010), *Tecniche grafiche di analisi e rappresentazione dei dati*, FrancoAngeli, Milano.

Delli Zotti G. (2004), *Introduzione alla ricerca sociale. Problemi e qualche soluzione*, FrancoAngeli, Milano.

Delli Zotti G. (1999), *L'analisi esplorativa delle tabelle di contingenza. Nuova edizione - esempi realizzati con Spss per Windows 7.5*, Dip. Est, Udine, Quaderni del Dipartimento Est, 99-15.

Delli Zotti G. (1992), *Il problema più importante per noi*, in Marradi A. e Gasperoni G. (cur.), *Costruire il dato 2. Vizi e virtù di alcune tecniche di raccolta delle informazioni*, FrancoAngeli, Milano: 130-141.

Delli Zotti G. (1985), *Tipologia delle matrici utilizzate nella ricerca sociale*, «Rassegna Italiana di Sociologia», XXVI, 2.

Delli Zotti G., Pellizzoni L. (1996), *Giovani verso l'Europa. Popoli, paesi, istituzioni*, Isig, Gorizia.

Di Franco G. (2017), *Tecniche e modelli di analisi multivariata*, FrancoAngeli, Milano.

Di Franco G. (2010), *L'analisi dei dati con Spss. Guida alla programmazione e alla sintassi dei comandi*, FrancoAngeli, Milano.

Di Franco G. (2006), *Corrispondenze multiple e altre tecniche multivariate per variabili categoriali*, FrancoAngeli, Milano.

Di Franco G. (2001), *EDS: Esplorare, descrivere e sintetizzare i dati. Guida pratica all'analisi dei dati nella ricerca sociale*, Angeli, Milano.

Di Franco G., Marradi A. (2020), *L'analisi bivariata*, FrancoAngeli, Milano.

Duverger M. (1969), *I metodi delle scienze sociali*, Etas-Kompass, Milano [ed. orig.: 1961, *Mèthodes des sciences sociales*, Puf, Paris].

Enciclopedia Treccani, *Analisi fattoriale*, https://www.treccani.it/enciclopedia/ricerca/analisi-fattoriale/.

Erikson R. (1984), *Social Class of Men, Women and Families*, «Sociology», 4, pp. 500-514.

Everitt B. (1978), *Graphical Techniques for Multivariate Data*, Heinemann, London.

Fabbris G.P. (1990), *L'analisi esplorativa dei dati multidimensionali*, Cleup, Padova.

Galtung J. (1970), *Theory and Methods of Social Research*, Allen & Unwin, London.

Gangemi G. (1985), *La logica della comparazione: controlli statistici e controlli nel disegno della ricerca*, Edas, Messina.

Garson G.B. (1976), *Handbook of Political Science Methods*, sec. ed., Holkbrook, Boston.

Gubert R. (1987), *Analisi fattoriale*, in Demarchi F., Ellena A., Cattarinussi B. (cur.), *Nuovo dizionario di sociologia*, Paoline, Milano: 124-132.

Helliwell J.F. *et al.* (2020), *World Happiness Report 2020*, https://worldhappiness.report/ed/2020/.

Inglehart R. (1996), *Modernization and Postmodernization. Cultural, economic and Political Change in 43 societies*, Princeton U.P., Princeton, New Jersey [trad. it.: 1998, *La società postmoderna. Mutamento, ideologie e valori in 43 paesi*, Editori Riuniti, Roma].

Inglehart R. (1977), *The Silent Revolution. Changing Values and Political Styles among Western Publics*, Princeton University, Princeton [trad. it.: 1983, *La rivoluzione silenziosa*, Rizzoli, Milano].

Istat (2020), *Censimento della popolazione 2011*, http://dati.istat.it/Index.aspx.

Klecka W.R. (1976), *Discriminant Analysis*, Sage, London.

Kovács, A. (1999), *Antisemitic Prejudices in Contemporary Hungary*, «Analysis of Current Trends in Antisemitism», Acta n. 16, sicsa.huji.ac.il/16 Kovacs.html.

Lancia F. (2004), *Strumenti per l'analisi dei testi. Introduzione all'uso di T-LAB*, FrancoAngeli, Milano.

Marradi A. (1997), *Linee guida per l'analisi bivariata dei dati nelle scienze sociali*, FrancoAngeli, Milano.

Marradi A. (1993), *L'analisi monovariata*, FrancoAngeli, Milano.

Merton R.K., Barber E.G. (2002), *Viaggi e avventure della Serendipity*, Il Mulino, Bologna.

Metastasio R., Cini F. (2009), *L'analisi del contenuto. Procedure di analisi dei dati con il programma SPAD*, FrancoAngeli, Milano.

Mussino A. (1983), *L'analisi statistico-sociale di grandi matrici di dati nell'applicazione dell'analisi delle corrispondenze*, «Sociologia e ricerca sociale», n.10: 69-99.

OECD (2020), *Better Life Index*, https://www.oecdbetterlifeindex.org/it/countries/italy-it/.

Pacifico M., Coppola L. (2010), *Nvivo: una risorsa metodologica. Procedure per l'analisi dei dati qualitativi*, FrancoAngeli, Milano.

Palumbo (2012), *Ricerca*, in Bezzi. C. (cur.), *Glossario del gruppo tematico "Metodi e tecniche" dell'AIV*, https://lavorosocialemeridiano.files.wordpress.com/ 2013/07/ bezzi_glossariometoditecnichevalutazione_2012. pdf).

Parra Saiani P. (2009), *Gli indicatori sociali*, FrancoAngeli, Milano.

Pavan E. (2016), *La Network analysis con NodeXL*, FrancoAngeli, Milano.

Pirsig R.M. (1981), *Lo Zen e l'arte della manutenzione della motocicletta*, Bompiani, Milano.

Ricolfi R. (2000), *Tre variabili. Un'introduzione all'analisi multivariata*, FrancoAngeli, Milano.

Ricolfi L. (1995), *La ricerca empirica nelle scienze sociali. Una tassonomia*, «Rassegna italiana di sociologia»; n.3: 389-418.

Ricolfi L. (1982), *Modelli di analisi e software per le scienze sociali*, «Rassegna italiana di sociologia», n.3: 331-372.

Rizzi S. (2009), *What-if analysis*, in: *Encyclopedia of Database Systems*, Springer-Verlag, Berlin.

Rosenberg M. (2003), *La logica dell'analisi trivariata*, FrancoAngeli, Milano.

Schwartz H., Jacobs J. (1987), *Sociologia qualitativa. Un metodo nella follia*, Il Mulino, Bologna [ed. orig.: 1979, *Qualitative Sociology. A Method to the Madness*, Free Press, New York].

Sedmack M., Medaric Z, Walker S. (eds.) (2014), *Children's Voices: Studies of interethnic conflict and violence in European schools*, Routledge, London.

Stanzione M. (1981), *Introduzione*, in Campbell D.T., *Epistemologia evoluzionistica*, Armando, Roma: 7-62.

Strassoldo R., G. Delli Zotti (1990), *Le matricole dell'Ateneo friulano*, Inserto speciale del Notiziario dell'Università degli Studi di Udine, 2.

Tommaselli V. (1993), *Analisi multidimensionale e inferenza sociologica. Validità ed attendibilità delle tecniche statistiche: fattori, componenti e* "scaling", FrancoAngeli, Milano.

Toulmin S.E. (1984), *Filosofia e teoria della scienza, storia e sociologia della scienza*, «Sociologia e ricerca sociale», numero monografico su "Metascienza", n.13: 9-19.

Tukey J.W. (1977), *Exploratory Data Analysis*, Addison-Wesley, Reading.