9. L'elaborazione e l'analisi dei dati

9.1 Introduzione

Le fasi conclusive di un processo di ricerca consistono nell'elaborazione e analisi dei dati. Con il primo termine ci si riferisce a qualunque procedimento (non solo statistico) di trattamento dei dati rilevati; con il secondo alle riflessioni che applichiamo ai (o che ci vengono suggerite dai) risultati dell'elaborazione dei dati.

In effetti, separare l'analisi dei dati dalla loro rilevazione riproduce il solito paradigma positivistico, secondo il quale i dati preesistono all'osservatore neutrale, quindi vengono *prima* rilevati, *poi* sottoposti ad elaborazione e analizzati. In realtà, come si è detto ampiamente nei capitoli precedenti, anche nella ricerca *standard* i 'dati' sono il frutto di un processo di costruzione che chiama in causa, con gradi diversi di autonomia, sia l'osservatore che il soggetto osservato e derivano comunque, anche nelle scienze della natura, dall'interazione fra gli schemi mentali dell'osservatore e le percezioni sensoriali che gli derivano dal mondo della natura (un oggetto *è* un tavolo solo perché le percezioni relative alla sua forma, dimensione, ecc. *vengono strutturate dall'osservatore in termini di* 'tavolo').

Iniziamo dunque a riflettere sull'analisi dei dati tenendo conto del fatto che separare questa fase dalla raccolta costituisce un artificio improprio, anche perché la 'raccolta', al pari del trattamento cui saranno sottoposti i dati rilevati e delle riflessioni che ne deriveranno, è innanzi tutto determinata dal quadro teorico di riferimento adottato dal ricercatore e dalle ipotesi formulate inizialmente. Ciò vale, di norma, per qualunque tipo di ricerca (descrittiva, esplicativa, applicativa, valutativa, *standard*, *non standard*), perché

> una buona descrizione sociale è per definizione intrisa di teoria, nel senso che le domande e lo sguardo del ricercatore discendono necessariamente da un orizzonte precedentemente costruito ed elaborato (Dal Lago e Quadrelli, 2003: 23).

Il quadro teorico di riferimento permette di definire, in primo luogo, i criteri di rilevanza assunti all'inizio di un percorso di ricerca: in una prospettiva descrittiva, ad esempio, si ipotizza inizialmente che esista un certo numero di aspetti, successivamente operativizzati in variabili, meritevoli di essere rilevati. In questo caso, nello stadio dell'elaborazione ed analisi dei dati, ci si preoccuperà principalmente di fornire una descrizione di tali aspetti, con l'ausilio della statistica e la produzione di tabelle e grafici anche elementari.

Molto più spesso le ipotesi iniziali non riguardano solo la rilevanza di alcune variabili rispetto al tema oggetto d'indagine, ma concernono anche l'esistenza di relazioni tra variabili; in questo caso, attraverso l'elaborazione dei dati ci si prefiggerà di accertare tali relazioni e, ove possibile, di misurarne l'intensità, mentre in fase di analisi verrà valutata la rilevanza di tali relazioni rispetto sia a ricerche analoghe, sia alle teorie sociologiche di riferimento.

Nelle ricerche esplicative le ipotesi saranno formulate fin dall'inizio con un maggior grado di precisione: in altre parole, sapremo con molta chiarezza cosa stiamo cercando, e molto spesso anche i tipi di elaborazione cui sottoporre i dati saranno definiti fin dall'inizio, in ragione delle ipotesi di partenza.

In una ricerca valutativa gli obiettivi iniziali influenzeranno in modo altrettanto cogente l'elaborazione e l'analisi. Gli obiettivi della valutazione svolgeranno un ruolo simile a quello delle ipotesi da confermare, in quanto predetermineranno le relazioni tra variabili da misurare e le valutazioni che saranno fondate su tali relazioni.

Un esempio per chiarire. Gli stessi dati rilevati da un questionario, relativi ad esempio a percorsi scolastici e comportamenti devianti dei giovani, potranno essere elaborati ed analizzati in modo diverso a seconda del tipo di ricerca che si sta svolgendo. In una ricerca descrittiva ci si prefiggerà di conoscere la distribuzione dei due caratteri; per esempio, di conoscere la percentuale di abbandoni scolastici nella fascia dell'obbligo e la percentuale di consumatori di droghe tra i giovani. L'analisi dei dati potrà limitarsi a valutare se l'abbandono scolastico è aumentato o diminuito nel tempo, se è maggiore in un quartiere o in un altro, ecc. In una ricerca esplicativa sarà stata inizialmente ipotizzata una

relazione tra i due fenomeni; con l'elaborazione ed analisi dei dati si cercherà quindi di misurare l'esistenza e l'intensità di tale relazione. Ad esempio, quindi, si cercherà di vedere se tra i giovani che hanno abbandonato anzitempo la scuola si registra una concentrazione più elevata di consumatori di droga: ovvero se l'assunzione di droghe preceda o segua gli abbandoni scolastici. Infine, in una ricerca valutativa, mirata ad esempio alla valutazione dell'efficacia di una campagna di prevenzione, si porranno in relazione i consumi di droga registrati prima e dopo la campagna (oppure tra un gruppo esposto alla campagna e un altro che non ne è stato oggetto), per valutare se vi siano differenze e se tali differenze corrispondano all'obiettivo della campagna stessa.

Come si è detto in precedenza, molto spesso una ricerca sociale si pone, miscelati in vario modo, tutti gli obiettivi sopra detti. Una ricerca sull'abbandono scolastico può prefiggersi sia di misurarne la portata in generale e nei diversi contesti territoriali e sociali di appartenenza dei giovani, sia di cercarne le relazioni con altre variabili (status socio economico e culturale della famiglia, tipo di quartiere di residenza, ecc.), sia infine di individuare variabili causali sulle quali intervenire a livello di politiche sociali, ovvero di misurare gli effetti di politiche già avviate.

In tutti i casi, l'elaborazione e l'analisi dei dati serviranno a descrivere i fenomeni rilevati e a cercare relazioni tra le variabili costruite per rilevarli.

A ben vedere, le fasi finali della ricerca sono caratterizzate da due processi ben diversi tra loro. Da un lato, infatti, con l'elaborazione e l'analisi dei dati si conclude un processo di progressiva riduzione della realtà alle nostre categorie concettuali (che presiedono sia alla rilevazione che all'elaborazione e all'analisi dei dati); dunque un processo di scomposizione e parcellizzazione che apparentemente *impoverisce* la ricchezza informativa del mondo che abbiamo deciso di conoscere per questa via. Dall'altro lato, i risultati della ricerca dovrebbero permettere una conoscenza più precisa o più approfondita di quella di senso comune, una conoscenza quindi che *arricchisce* la consapevolezza delle stesse persone dalle quali sono stati 'estratti' i dati e alle quali sono state restituite le analisi degli stessi.

La riduzione della realtà operata dal processo di ricerca passa attraverso le fasi seguenti:

a) astrazione dalla realtà sociale (composta da infiniti fenomeni caratterizzati da innumerevoli aspetti degni d'interesse) di alcuni elementi rilevanti per l'oggetto della ricerca, sui quali verrà condotta l'indagine empirica (a fini descrittivi o interpretativi/valutativi); si tratta della fase di *definizione dei criteri di rilevanza*, ovvero di *"definizione dei fabbisogni informativi dell'indagine"* (Agnoli, 1994);
b) precisazione dei concetti rilevanti per la ricerca e loro traduzione in *variabili o indicatori* (qui l'ulteriore riduzione deriva dal fatto che quasi mai gli indicatori colgono per intero l'estensione dei concetti); si tratta della fase di costruzione degli indicatori secondo il paradigma lazarsfeldiano di cui si è detto in precedenza;
c) *misurazione* delle variabili (in questo caso si ha un'ulteriore riduzione della realtà dovuta al fatto che vengono 'forzati' nelle classificazioni utilizzate aspetti molto più articolati e complessi di quelli riferibili alle classificazioni stesse); è la fase della discesa sul campo e dell'applicazione degli strumenti di rilevazione precedentemente costruiti;
d) *elaborazione* dei dati. In questo caso una nuova riduzione deriva dall'utilizzo di forme di elaborazione che sottolineano alcune relazioni a svantaggio di altre, come accade, ad esempio, per statistiche che, sintetizzando in un indice o in una distribuzione un numero elevato di informazioni elementari, mettono in forte evidenza un aspetto a scapito di altri.

Si badi bene che questo processo è più evidente e scandito in modo più preciso nella cosiddetta ricerca *standard*, cui è principalmente dedicato questo capitolo, ma è inevitabile anche nella ricerca *non standard*, in quanto proprio di ogni procedimento scientifico (cfr. Gobo, 2001).

Anche in questo caso un esempio può essere d'aiuto. In un'indagine sui giovani, viene in primo luogo ridotto il loro mondo vitale a quei concetti che altre ricerche o teorie sociologiche hanno definito come rilevanti. Ad esempio, si decide di rilevare, tra i vari aspetti, il rendimento scolastico e lo status familiare. Per inciso, la scelta di operare su di un campione comporta un'ulteriore riduzione, in questo caso del campo d'indagine e non del tema oggetto della ricerca.

In secondo luogo, per tradurre questi concetti in domande sul questionario, occorre ridurre ulteriormente i concetti alle loro componenti misurabili. Decidere, ad esempio, di misurare il rendimento scolastico in base alle bocciature subite esclude altri aspetti, quali i voti conseguiti, l'apprendimento non registrato dalle votazioni, ecc.

In terzo luogo, in sede di codifica dei dati vengono costruite categorie che riducono ulteriormente le differenze; magari aggregando il numero di bocciature a prescindere dall'anno in cui si sono verificate o dalla loro sequenza.

Infine, in sede di elaborazione ci si può limitare a costruire indici statistici, come la media, che, a fronte del vantaggio di sintetizzare in un solo numero una distribuzione di migliaia di casi, presenta tuttavia il grosso limite di

'appiattire' le differenze interne. Ad esempio, una media di una bocciatura per studente può derivare sia da un'equidistribuzione del fenomeno (tutti i rispondenti hanno subito una ed una sola bocciatura), sia da una sua distribuzione polare (un terzo ha conosciuto tre bocciature, due terzi nessuna)[1].

Come si vede dall'esempio, i mille aspetti rilevanti della condizione giovanile sono stati via via ridotti fino ad essere espressi in un numero (la media delle bocciature). Anche se si tratta di un caso estremo, l'esempio illustra come la conoscenza scientifica derivi da un processo continuo di semplificazione o astrazione dalla realtà sociale: si tratta del prezzo da pagare per conseguire una conoscenza più puntuale di quella posseduta da ogni membro competente della società, ovvero della cosiddetta conoscenza del senso comune.

Ovviamente non ci sono solo prezzi da pagare, ma anche vantaggi. Il più evidente deriva dalla possibilità di fornire una descrizione puntuale di fenomeni che a livello di senso comune sono conosciuti solo in modo approssimato, se non errato. Un secondo importante vantaggio è connesso alla possibilità di cogliere aspetti e relazioni tra fenomeni che sfuggono allo sguardo degli attori sociali. Il terzo, connesso soprattutto alle ricerche *standard*, deriva dalla possibilità di misurare con precisione relazioni tra variabili che a livello di senso comune possono essere solo ipotizzate: un conto è dire che una famiglia di alto status sociale aiuta nella riuscita scolastica, un altro è poter misurare l'intensità di questa relazione, ovvero trasformarla in un percorso causale.

Qui emerge il secondo processo, opposto al primo, che caratterizza la ricerca: quello che consente di *arricchire* la conoscenza del mondo sociale grazie ai dati rilevati e alle elaborazioni effettuate. Sotto questo aspetto, il ricercatore deve anche sapersi fare 'trovatore', ossia deve essere disposto a riconoscere nessi causali o relazioni rilevanti che non era stato capace di ipotizzare fin dall'inizio della ricerca.

Si ricorda infatti che il processo di ricerca realmente praticato è costituito da un continuo dialogo tra gli schemi cognitivi con i quali il ricercatore affronta il suo oggetto e le sollecitazioni, che provengono dalla realtà sociale, a modificare, aggiornare, integrare, gli schemi iniziali, attraverso il costante adeguamento delle ipotesi iniziali, necessario per contenere una realtà sempre più ricca dei nostri schemi mentali (sempre che, ovviamente, abbiamo l'onestà, umiltà e competenza per accorgercene).

9.2 L'analisi dei dati nella ricerca *standard*

Passando a trattare dell'analisi dei dati nelle ricerche che si prefiggono di costruire una matrice dei dati alle quali applicare strumenti più o meno complicati di analisi, si può dire, in termini generali, che in una ricerca descrittiva l'elaborazione consisterà principalmente in una restituzione dei dati in forma grafica o di tabella, tale da evidenziare la distribuzione delle diverse caratteristiche del campione (o dell'universo) in ragione delle varie proprietà considerate rilevanti.

Nel caso di una ricerca esplicativa, i dati verranno organizzati in modo da consentire il controllo delle ipotesi. Si ricorda peraltro che la distinzione è relativa, in quanto nelle ricerche descrittive si tende comunque a mettere in rilievo alcune relazioni significative tra variabili, o elaborando tabelle a doppia entrata, oppure ricorrendo a procedimenti statistici più sofisticati, mentre ogni ricerca esplicativa contiene comunque una più o meno ampia parte descrittiva.

In coerenza con il carattere di circolarità del processo di ricerca più volte citato, è sicuramente necessario mantenere la disponibilità alla revisione delle ipotesi iniziali alla luce dei risultati delle elaborazioni effettuate. Ma nessuna elaborazione ha realmente un significato, se non è sorretta da una riflessione sociologicamente fondata che sappia leggere ogni relazione individuata in termini di connessioni di senso al livello degli attori sociali e di riferimento a teorie sociologiche più generali. Weber affermò del resto che

> se manca l'adeguazione di senso... ha luogo soltanto la probabilità statistica *non intelligibile*, o intelligibile solo in maniera imperfetta. ... Solamente le uniformità statistiche che corrispondono al senso *intelligibile* di un agire sociale, costituiscono tipi di azione intelligibili... e quindi 'regole sociologiche' (1922/1974, I, 11, corsivo nel testo).

Inoltre, qualunque persona di buon senso è in grado di proporre una spiegazione plausibile per ogni relazione rilevata, come si è detto nel capitolo iniziale, richiamando Merton. Il sociologo deve saper andare

[1] Questo accade anche nelle ricerche *non standard*. Nel caso di una ricerca qualitativa, ad esempio, una lunga intervista viene via via scomposta in parti, analizzate in comparazione con altre interviste piuttosto che secondo la sequenza delle affermazioni dell'intervistato, fino ad apparire magari nel *report* finale solo con un paio di citazioni.

al di là della plausibilità, formulando le proprie ipotesi in modo tale da poterle confermare o smentire con i dati di ricerca e non solo con un *cocktail* casereccio di teorie, evidenze empiriche e senso comune.

Proprio i nessi che legano le conoscenze acquisite mediante la ricerca empirica a modelli interpretativi di più ampia portata rendono la ricerca sul campo una fonte di continuo ripensamento, revisione e progresso della teoria. Se un'analisi condotta sui giovani presuppone una serie di modelli, teorie ed ipotesi generali sulla struttura delle società, sui processi di socializzazione, sui percorsi di trasmissione culturale, i suoi risultati costringono quasi sempre a perfezionare o ripensare tali modelli e teorie, in un processo senza fine di approfondimento della conoscenza della realtà sociale.

Anche l'impiego pratico della ricerca è impossibile senza adeguati riferimenti al livello teorico: è difficile condurre una buona ricerca valutativa senza disporre di un modello di funzionamento della realtà sociale sul quale il pianificatore compie simulazioni (ex ante) o costruisce percorsi causali (ex post) che gli permettano di organizzare in modo efficace gli interventi o valutarne gli esiti.

In linea generale, l'elaborazione dei dati serve dunque sia per descrivere in modo più compatto e sintetico i dati raccolti, anche a fini comparativi, sia per ricercare relazioni tra variabili. Al di là delle competenze specifiche possedute dal ricercatore, è bene ricordare che ogni statistica risponde ad un interrogativo e mette in luce alcuni aspetti a scapito di altri. Quindi va utilizzata nella consapevolezza dei suoi limiti e delle sue potenzialità. Inoltre, anche l'utilizzo di grafici e tabelle per presentare i dati deve essere coerente con l'intendimento illustrativo del ricercatore. I dati non parlano da soli, e neppure si presentano da soli. Ogni elaborazione corrisponde ad una scelta e come tale va meditata e motivata.

In tema di elaborazione dei dati le principali distinzioni corrono tra variabili qualitative e quantitative e tra elaborazioni riferite a una, due o più variabili. Ci avvaliamo di quest'ultima strategia espositiva, essendo la più diffusa, facendo riferimento al tipo di variabili all'interno di ogni paragrafo.

9.2.1 L'analisi monovariata

Richiamiamo brevemente, per chiarezza, la matrice dei dati, che costituisce il punto di partenza di ogni elaborazione: sappiamo che essa è costituita da un numero di colonne corrispondente al numero di variabili utilizzato e da tante righe quanti sono i casi rilevati. La prima e più semplice elaborazione consiste nel conteggio, per ogni variabile, di quanti casi sono stati rilevati per ogni classe in cui la variabile stessa è stata articolata.

Occorre infatti ricordare che ogni variabile (qualitativa o quantitativa che sia) è costituita da due elementi: il primo è rappresentato dalle classi in cui è articolata la proprietà o il fenomeno che è stato rilevato (colore degli occhi, anno di nascita, classe di reddito, ecc.); il secondo elemento corrisponde invece al numero di casi che rientrano in una delle classi costruite in sede di rilevazione (o, in taluni casi, aggregate in sede di codifica dei dati). Su questa base si fonda ogni elaborazione statistica (o presentazione in forma tabellare o grafica) dei dati. Se ad esempio in un questionario rivolto a 1.000 studenti è stato chiesto quanti sono stati promossi, rimandati o bocciati nell'anno precedente, la variabile così costruita risulta la seguente:

*Tab. 1 - Distribuzione degli studenti secondo l'esito degli scrutini**

Modalità	promossi	rimandati	bocciati	totale
Numero casi	750	150	100	1.000
Frequenze	75%	15%	10%	100%

* Si noti che il titolo della tabella evidenzia innanzi tutto chi sono i soggetti dei quali viene rilevata una proprietà (gli studenti) e, in subordine, la proprietà di cui viene riportata la distribuzione.

Nella sua forma *standard* la variabile è composta dalle prime due righe (Modalità e numerosità; nella terminologia adottata dalla statistica, valori argomentali e frequenze assolute). La terza riga (Frequenze) costituisce un'elaborazione della seconda, e viene ottenuta dividendo la numerosità di ogni singola cella per il totale, moltiplicando poi per 100 il risultato. L'elaborazione di percentuali è sempre possibile, indipendentemente dalla forma, alfabetica o numerica, in cui sono espresse le modalità della variabile, in quanto le operazioni aritmetiche vengono svolte solo sul numero di casi.

Abbiamo introdotto subito le percentuali perché rappresentano un mezzo molto efficace per comparare diverse distribuzioni della stessa variabile. Prendiamo ad esempio tre diverse scuole, in cui è stata rilevata la distribuzione riportata nella Tab. 2: dalla lettura dei valori assoluti non è di immediata evidenza quale sia la scuola con la maggior incidenza di promossi, di rimandati o di bocciati.

Tab. 2 - Distribuzione degli studenti delle scuole A, B, C secondo l'esito degli scrutini

Modalità	promossi	rimandati	bocciati	totale
Scuola A	750	150	100	1.000
Scuola B	600	117	92	829
Scuola C	847	141	126	1.114

Le comparazioni sono invece immediatamente effettuabili se i dati vengono percentualizzati, come in Tab 3.

Tab. 3 - Distribuzione percentuale degli studenti delle scuole A, B, C secondo l'esito degli scrutini

Modalità	promossi	rimandati	bocciati	totale
Scuola A	75,0	15,0	10,0	100,0
Scuola B	72,4	16,5	11,1	100,0
Scuola C	76,0	12,7	11,3	100,0

Anche un esame frettoloso della Tab. 3 consente di vedere che la più alta percentuale di promossi è stata registrata nella Scuola C e la più bassa nella B, e di procedere in modo analogo per un confronto delle altre due modalità. Si noti che nella Tab. 3 la percentuale è stata calcolata fino al primo decimale. La scelta del numero ottimale di decimali dipende dal numero di modalità della variabile e, soprattutto, dal numero di casi. Se la base di calcolo è molto ampia può avere un senso scendere a due decimali, soprattutto se anche le modalità sono numerose, in quanto scostamenti anche modesti nei valori percentuali possono corrispondere a differenze consistenti in valore assoluto (ad es., una differenza di 0,1% corrisponde, in un campione di mille unità, a un individuo; in un universo di 10 milioni di unità a ben 10.000 soggetti); di contro, se la base di calcolo è modesta (ad es., 200 unità) e ridotto il numero di modalità (al limite, 'sì' e 'no') può essere opportuno evitare addirittura i decimali.

Va da sé che, quale che sia il numero di decimali utilizzato, è quasi sempre indispensabile 'arrotondare' le cifre, in modo che la loro somma corrisponda al totale (100). Si usa arrotondare al valore superiore quando il primo decimale successivo a quello che verrà indicato in tabella è superiore o uguale a 5 ed evitare l'arrotondamento quando è inferiore a 5. Ad esempio, 7,5723 diventa 7,6, mentre 7,5486 rimane 7,5. Nei non infrequenti casi in cui, anche usando l'arrotondamento, si ottiene come somma il poco gradevole valore di 100,1 o di 99,9, si è soliti contravvenire alla regola sopra detta, evitando l'arrotondamento del più basso tra i valori con scarto superiore a 5 ovvero, nel caso opposto, arrotondando al decimale superiore il più elevato dei valori tra quelli con scarto inferiore a 5. Nella Tab. 4 è riportato un esempio.

La distribuzione percentuale dei casi consente anche, se la variabile è espressa almeno su di una scala ordinale, di calcolare le cosiddette frequenze cumulate. Una volta prescelto un ordine di organizzazione delle categorie (nell'esempio fin qui utilizzato, il grado decrescente di successo scolastico, secondo il quale la modalità 'rimandato' sta prima di 'bocciato' ma dopo 'promosso'), si riporta per ognuna di queste la somma delle percentuali relative alle modalità maggiori o uguali. Un esempio è riportato nella tab. 5, che riprende i dati di Tab. 3 al fine di evidenziare l'utilità comparativa delle frequenze cumulate.

Tab. 4 - Esempio di arrotondamenti di percentuali

Modalità	promossi	rimandati	bocciati	totale
Numero casi	753	150	100	1.003
Frequenza a 5 decimali	75,07477	14,95513	9,97008	99,99998
Arrotondamento	75,1	15,0	10,0	100,1
Arrotond. corretto	75,1	14,9	10,0	100,0
Numero casi	747	150	100	997
Frequenza a 5 decimali	74,92477	15,04513	10,03009	99,99999
Arrotondamento	74,9	15,0	10,0	99,9
Arrotondamento corretto	74,9	15,1	10,0	100,0

Tab. 5 - Distribuzione degli studenti delle scuole A, B, C secondo le frequenze cumulate dell'esito degli scrutini

Modalità	Promossi	Promossi o rimandati	Promossi, rimandati o bocciati

Scuola A	75,0	90,0	100,0
Scuola B	72,4	88,9	100,0
Scuola C	76,0	88,7	100,0

Si noti che nella tabella non viene riportata la colonna del totale perché questa costituirebbe una duplicazione dell'ultima modalità, che già contiene la somma di tutte le modalità precedenti più quella propria della colonna. Mentre nella Tab. 3 si notava che la Scuola C era caratterizzata dalla più elevata percentuale sia di promossi che di bocciati, dalla Tab. 5 emerge che la Scuola C è quella caratterizzata dal più basso tasso di successo, se si considerano unitamente le due modalità 'promossi o rimandati'.

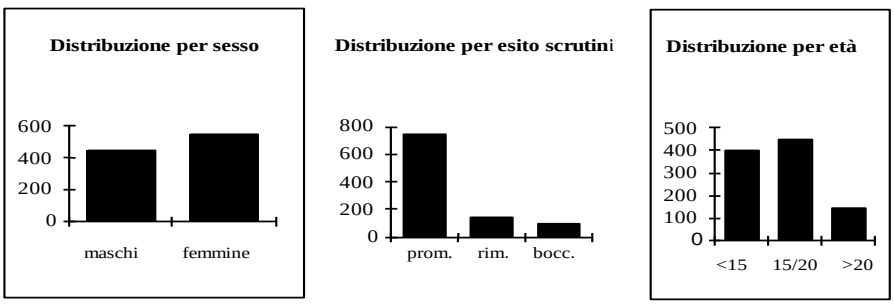
Naturalmente ognuna delle tabelle riportate come esempio è suscettibile di essere tradotta in un grafico. Le rappresentazioni grafiche possibili sono assai numerose; ciascuna evidenzia alcuni aspetti mettendone in ombra altri. Data l'efficacia e l'immediatezza della comunicazione iconica, è consigliabile scegliere attentamente il tipo di rappresentazione che meglio si adatta al messaggio che si intende trasmettere o all'aspetto che si vuole evidenziare.

Per brevità consideriamo qui solo il caso degli istogrammi, ossia i rettangoli costruiti su di un piano cartesiano in cui di norma si riportano sull'asse X le modalità e sull'asse Y le frequenze (o il numero assoluto di casi). È bene ricordare tuttavia che i grafici, essendo caratterizzati dalla continuità dei valori disposti sugli assi, dovrebbero essere correttamente utilizzati solo per rappresentare distribuzioni di variabili misurate su scale metriche. Per un esempio si vedano i tre grafici sotto riportati, relativi l'uno alla distribuzione per sesso, l'altra per esito scolastico, l'altra ancora per età di un ipotetico campione di 1.000 persone.

In effetti, il primo grafico è scorretto in quanto suggerisce un ordine tra i due caratteri che non è ipotizzabile per una variabile nominale, mentre il secondo suggerisce l'equidistanza fra le tre modalità che non può essere postulata per una variabile ordinale. Entrambi riportano poi una base sull'asse X relativa alle modalità (maschi/femmine; promossi/rimandati/bocciati), mentre non è possibile ipotizzarne alcuna (la modalità dovrebbe essere indicata da un punto e l'istogramma ridursi a un segmento). Perfino il terzo grafico, in cui è corretto l'ordine crescente delle tre classi, è in realtà sbagliato per due ragioni: in primo luogo perché assegna uguale estensione alle tre classi, mentre la prima ha un'origine indefinita, fissabile arbitrariamente a 6 anni, e quindi un'estensione di 10 anni (dall'età scolare a 15 anni) e la terza un termine indefinito (di fatto l'età dello studente più anziano del campione); in secondo luogo perché posiziona arbitrariamente i punti di origine e termine delle classi, con una soluzione di continuità tra '15' e '>15' e tra '20' e '>20' che non esiste in realtà (la distanza è infinitesima).

Di contro, l'impiego rigoroso delle rappresentazioni grafiche deve obbedire a requisiti specifici. Nel caso in esame, degli istogrammi di frequenza, vanno rispettate le condizioni secondo le quali le aree dei singoli rettangoli sono proporzionali alla frequenza e le basi sono proporzionali all'estensione delle singole classi (calcolabile solo per variabili metriche).

Grafico 1 - Esempio di distribuzione di una popolazione per tre diversi caratteri



Come si è anticipato, non esistono solo rappresentazioni grafiche o tabellari delle variabili. È disponibile un'ampia gamma di statistiche utilizzabili nell'analisi monovariata (per una trattazione esaustiva vedi Marradi, 1993), al cui interno le più utilizzate sono, in sintesi, le seguenti:

a) *moda:* evidenzia la classe che è caratterizzata dal maggior numero di casi, ovvero il valore o i valori che ricorrono più di frequente nella distribuzione. Nella variabile di Tab. 3, ad esempio, la moda è costituita dalla modalità 'promosso' in tutte e tre le scuole considerate. Come si evince dall'esempio, la moda può essere calcolata anche per le variabili nominali, in quanto chiama in causa solo le frequenze e non le categorie in cui si articola la variabile. Ovviamente, una distribuzione può avere più d'una moda se più d'una modalità presenta un massimo relativo di casi (ad esempio: promossi 500, rimandati 500, bocciati 200 è una distribuzione detta bimodale, in quanto presenta due mode in corrispondenza delle modalità 'promosso' e 'rimandato');
b) *mediana:* è il valore assunto dal caso che sta a metà di una distribuzione ordinata (non importa se in senso crescente o decrescente). La mediana può dunque essere calcolata anche per variabili misurate su scale ordinali, attraverso la distribuzione delle frequenze cumulate (in tal caso la modalità mediana è la prima che supera il 50% dei casi delle frequenze cumulate);
c) *decili, quartili, percentili:* in una distribuzione ordinata, sono i valori assunti articolando i casi in blocchi pari al dieci, al venticinque o all'un per cento dei casi. Sono utilizzati quando si intende sottolineare quale parte della distribuzione sia al di sotto o al di sopra di certi valori di soglia. Affermare ad esempio che il primo decile della distribuzione del reddito degli italiani ha un valore di 500 Euro al mese significa affermare che il 10% degli italiani ha un reddito inferiore o uguale a questa cifra. Dalla distribuzione di frequenza deriva la distribuzione delle frequenze cumulate, possibile solo per variabili ordinali o metriche, che consente di calcolare il *coefficiente di concentrazione* (una statistica che ha senso se si è in presenza di una proprietà redistribuibile[2]);
d) *media aritmetica:* è data dalla somma dei valori registrati da ogni caso fratto il numero totale dei casi. Poiché richiede un'operazione di somma delle modalità in cui è articolata una variabile, è consentita solo per quelle misurate su scale a intervalli o di rapporti. A partire dal noto aforisma di Trilussa[3], la media è una delle statistiche più usate e criticate. In questa sede basta ricordare che rappresenta un valore di sintesi di un'intera distribuzione, ed è quindi utile a fini comparativi, a condizione che le distribuzioni stesse non siano troppo disomogenee. Ad esempio, calcolare l'età media dei residenti nelle regioni italiane è utile per confrontarne in modo sintetico la distribuzione per età in quanto non ci sono ragioni per pensare che le distribuzioni siano troppo diverse al loro interno. Analogamente, il reddito pro capite dei vari Paesi avanzati è utile per ordinarli secondo il benessere economico della popolazione, ma a condizione che la variabilità interna non sia troppo diversa.
   Va ricordato che esistono altre medie, dalle formule più complesse (media geometrica, media armonica, in generale medie di potenza), che sono tuttavia in casi particolari assai più utili della media aritmetica per sintetizzare in un indice statistico una distribuzione;
e) *scarto quadratico medio:* viene utilizzato per rappresentare con un solo indice numerico la variabilità di una distribuzione, ossia il grado di dispersione attorno alla media dei diversi valori. Può essere calcolato solo per variabili misurate su scale a intervalli o di rapporti, in quanto chiama in causa nella formula anche le categorie in cui una variabile è articolata. È costituito dalla radice quadrata della somma dei quadrati degli scarti dalla media fratto il numero totale dei casi, ovvero dalla radice quadrata della media aritmetica dei quadrati degli scarti dei valori dalla media. Spesso usato come 'correttivo' della media, consente di capire quanto quest'ultima rappresenti bene la distribuzione. Uno scarto quadratico medio molto basso si ha infatti quando i valori rilevati sono distribuiti in un'area molto ridotta attorno alla media, che dunque rappresenta bene la realtà (nell'aforisma di Trilussa, abbiamo mangiato circa mezzo pollo a testa). Di contro, il valore dello scarto quadratico medio sale rapidamente se i valori sono distribuiti in modo più ampio (ossia se ci sono parecchi soggetti che hanno digiunato e altri che si sono mangiati un pollo intero). Elevato al quadrato, assume il nome di *varianza*. Una prima, grezza misura della dispersione è consentita già dall'esame del *campo di variazion*e (differenza tra il valore più elevato

[2] Si tratta di un indice statistico di estremo interesse per le scienze sociali, in quanto mette a confronto la distruzione osservata di una variabile, ad es. il reddito, con i due limiti teorici di massima disuguaglianza, che si ha quando un solo soggetto possiede per intero la variabile (caso-limite, che a rigori non permetterebbe neppure di parlare di 'distribuzione', ma solo di concentrazione della grandezza in una sola persona) e di totale equidistribuzione, nei casi in cui tutti i soggetti posseggono la stessa quantità della variabile (caso-limite perché quando una variabile assume una sola modalità diventa una costante). Ovviamente se invece del reddito (variabile che è possibile sia accumulare che ridistribuire) si analizza l'età o la statura l'indice di concentrazione non ha senso.

[3] Secondo il quale la statistica è la scienza in forza della quale, se io mangio un pollo e tu digiuni, abbiamo mangiato mezzo pollo ciascuno.

e quello più basso assunto dalla variabile) e dalla differenza interquartile (differenza tra i valori posti al 25° e al 75° percentile di una distribuzione ordinata);

f) *coefficiente di variabilità:* è dato dal rapporto tra scarto quadratico medio e media. Molto usato a scopo comparativo, in quanto anche lo scarto quadratico medio è influenzato dal valore della media, mentre tale influenza viene depurata da questo procedimento[4];

g) *indici di asimmetria o di curtosi:* consentono di evidenziare, sintetizzandolo in un indice numerico, il grado di scostamento della distribuzione dalla cosiddetta 'normale' o gaussiana, che costituisce una distribuzione di riferimento (La curva normale è caratterizzata dalla simmetricità della distribuzione rispetto alla media, che coincide con la mediana e la moda).

Le statistiche relative ad una variabile sono anche usualmente distinte in misure della tendenza centrale o indici di posizione (moda, mediana, media), della dispersione (scarto quadratico medio, varianza), della forma della distribuzione (indici di asimmetria e di curtosi).

A scopo esemplificativo riportiamo una serie di statistiche relative alla distribuzione di una variabile in tre ipotetici campioni, riportati in Tab. 6. Nella Tabella, per motivi di semplicità, si utilizza una distribuzione per classi chiuse, ovvero caratterizzate tutte da un valore iniziale e finale predefinito. Di norma nella distribuzione del reddito si utilizza invece una classe finale 'aperta', che tuttavia comporta problemi nel calcolo della media e delle statistiche a questa collegate.

Tab. 6 - Distribuzione del reddito mensile in quattro popolazioni. Valori in migliaia di Euro

	0-1	>1-2	>2-4	>4-7	>7-10	Totale
Popolazione A	20,0	35,0	20,0	15,0	10,0	100,0
Popolazione B	25,0	20,0	25,0	15,0	10,0	100,0
Popolazione C	45,0	4,0	2,0	4,0	45,0	100,0
Popolazione D	5,0	10,0	70,0	10,0	5,0	100,0

Le principali statistiche proprie di ognuna delle quattro popolazioni sono le seguenti

	Moda	Mediana	Media	Scarto quadratico medio	Coefficiente di variazione
Popolazione A	classe >1-2	1,857143	2,9	2,44949	0,844652
Popolazione B	classi 0-1 e >2-4	2,4	2,85	2,44875	0,85921
Popolazione C	classi 0-1 e >7-10	3	3	4,094509	1,364836
Popolazione D	classe >2-4	3	3	1,635543	0,545181

Si nota agevolmente che media e mediana sono uguali nella terza e quarta distribuzione, volutamente costruite in modo simmetrico, ossia con uguale andamento ai due lati della media, non ostante raffigurino due situazioni opposte: la C di grossa dispersione e la D di forte concentrazione dei casi attorno alla media. Lo scarto quadratico medio è infatti più basso nella popolazione D, per la quale la media costituisce un indice statistico un po' più attendibile che negli altri tre casi. Soprattutto nella popolazione C, che presenta infatti lo scarto quadratico medio più alto, la media rappresenta molto male la distribuzione, dal momento che solo il 2% della popolazione fa parte della classe che comprende la media. Solo la moda fa capire che la popolazione C presenta una distribuzione di tipo polare, è caratterizzata cioè da un consistente numero di individui a basso reddito e da uno altrettanto elevato ad alto reddito. In tutti i casi, la disuguaglianza nella

[4] Il coefficiente di variabilità è molto utile per comparare distribuzioni diverse tra loro per media e scarto quadratico medio. È utilizzato ad esempio negli studi elettorali, perché permette di cogliere il radicamento territoriale dei partiti. Analizzando la distribuzione di voto nei vari quartieri di due partiti che raccolgono percentuali assai diverse di voti, ad esempio il 5% e il 15%, se hanno lo stesso *cv* (coefficiente di variazione) si deduce che hanno lo stesso radicamento territoriale (in media in ogni quartiere l'uno prende il triplo di voti dell'altro), mentre se il primo ha un *cv* più alto si deve ritenere che esso sia fortemente rappresentato in alcuni quartieri e poco presente in altri. Negli anni Ottanta, ad es., il Partito Liberale era presente in forze solo in alcuni quartieri genovesi, mentre il Partito Repubblicano raccoglieva consensi in modo più omogeneo sul territorio; anche se raggiungevano percentuali di voto simili a livello cittadino, il *cv* del primo era molto più elevato del secondo (cfr. Palumbo 1985).

distribuzione del reddito, molto alta nella popolazione C e modesta nella D, si coglie solo grazie allo scarto quadratico medio e al coefficiente di variazione (entrambi elevati nella popolazione C e bassi nella D). Utile nelle due distribuzioni non simmetriche (la A e la B) il confronto tra mediana e media, dal quale si legge che nella popolazione A la media rappresenta la distribuzione meno fedelmente che nella B, dal momento che la distanza tra media e mediana è in un caso di oltre un milione, nell'altro di meno della metà.

Esula dagli scopi di questo libro illustrare nel dettaglio la differenza di calcolo delle diverse statistiche nel caso di una distribuzione di dati grezzi rispetto ad una con i dati raggruppati in classi. Basti qui dire che in quest'ultima si ipotizza l'equidistribuzione dei casi entro la classe. Se ad es. la classe di reddito '>1.000 fino a 2.000'[5] contiene 11 casi, si ipotizza che il primo sia posto a 1.001; il secondo a 1.100; il terzo a 1.200 e così via fino a 2.000. Assumendo questa ipotesi, la media di ogni distribuzione interna a una classe è, per definizione, uguale al valore centrale della classe stessa, che ne costituisce la media e la mediana. Nei calcoli ci si regola quindi come se tutti i casi fossero concentrati nel punto medio della classe. Questa convenzione può essere fonte di distorsioni, quando accade che i valori reali sono distribuiti in modo diverso entro la classe. Se ad es. degli undici casi della distribuzione illustrata in precedenza sei assumessero il valore di 1.200 e cinque di 1.500, la media reale della classe non sarebbe 1.500 Euro, bensì (1.200*6 +1.500*5)/11 = (7.200 + 7.500)/11 = 14.700/11 = 1.340 Euro.

Si noti per inciso che questa possibilità di distorsione sta alla base delle scelte del ricercatore di raccogliere, per certi fenomeni, dati puntuali piuttosto che aggregati. Ad esempio, di chiedere in un questionario di indicare l'età piuttosto che di barrare una casella in corrispondenza di una classe d'età. È infatti sempre possibile aggregare successivamente in classi i dati, soprattutto per la loro restituzione in forma di tabella, mentre non si può risalire ai dati puntuali se questi sono rilevati in forma aggregata. Analogamente, è rilevante anche il numero di classi che il ricercatore decide di costruire e la loro ampiezza: è abbastanza ovvio che un numero eccessivo di classi può rendere difficile apprezzare la distribuzione, soprattutto se riguarda pochi casi, mentre per altro verso poche classi possono rendere meno leggibili differenze importanti nella distribuzione. La scelta delle classi è poi spesso influenzata dal tipo di aggregazione di eventuali dati di confronto. L'Istat, ad esempio, di solito raggruppa le età giovanili in base alle classi scolastiche e prevede quindi una classe da 6 a 10 anni; sarebbe sciocco, in assenza di forti ragioni contrarie, che il ricercatore ne costruisse una da 6 a 9, rendendo impossibili i confronti.

Le elaborazioni presentate, e altre che esulano dalle finalità di questo testo, rientrano nella cosiddetta statistica descrittiva, che mira a rappresentare in modo sintetico, attraverso indici numerici, la distribuzione di una variabile. Una crescente diffusione sta avendo tuttavia la cosiddetta statistica esplorativa, che consente di selezionare le parti della distribuzione che meglio si prestano alla più puntuale definizione del fenomeno. Ad esempio, una distribuzione caratterizzata da alcuni casi limite (che nel caso di un questionario potrebbero corrispondere anche ad errori di codifica) rischierebbe di generare indici statistici poco affidabili. Si pensi ad esempio ad una comunità di poveri al cui interno vengano registrati due miliardari: la media e lo scarto quadratico medio schizzerebbero verso l'alto, mentre basterebbe escludere i due nababbi dalla distribuzione per ottenere delle statistiche assai affidabili. Senza procedere oltre nell'esempio, occorre dire che, grazie anche alla diffusione di programmi quali l'SPSS e il SAS, è possibile con poca fatica 'interrogare' i dati in modo ben più approfondito che in passato. Corrispondentemente, cresce tuttavia la necessità di inserire nel bagaglio culturale di ogni ricercatore una robusta dose di statistica, se non si vuole rischiare di porre ai dati meno domande di quelle a cui potrebbero rispondere (o, peggio, di far porre dal computer domande sbagliate ovvero di ottenere risposte inutili o incomprensibili).

Nello schema che segue si riportano le statistiche di uso corrente consentite a seconda del tipo di variabile, o meglio del tipo di scala su cui essa è misurata.

Tipo di statistica	Tipi di variabili cui è applicabile
Moda	Tutte
Mediana	Ordinali o metriche
Frequenze accumulate e indice di concentrazione	Ordinali o metriche
Campo di variazione, differenza interquartile	Metriche

[5] Si noti che le classi debbono rispettare il criterio della mutua esclusività (un caso deve essere assegnato ad una ed una sola classe); pertanto non si possono costruire classi del tipo 'da 1.000 a 2.000 Euro' e 'da 2.000 a 3.000', perché un caso che registrasse il valore di 2.000 Euro apparterrebbe ad entrambe le classi. Di solito si usano simboli convenzionali per costruire classi 'chiuse' a destra o a sinistra: ad es., 0-┤ 1.000; 1.000-┤ 2.000 significa che la prima classe (aperta a destra e chiusa a sinistra) comprende il valore zero e arriva fino al valore 999,99 Euro, mentre la seconda include il valore 1.000 e arriva fino al valore 1.999,99. Un altro modo di rappresentare questo accorgimento, forse più comprensibile, è quello di scrivere classi aperte aventi come confine i limiti reali della classe; ad es., 0-999 Euro, 1.000-1.999. In questo caso si tratta di classi aperte sia a destra che a sinistra. Si noti che in questo secondo caso un valore rilevato di 999,99 Euro, per effetto dell'arrotondamento, va nella seconda classe (da 1.000 a 1.999), mentre nel primo caso lo stesso valore va nella prima (0-┤ 1.000).

Media	Metriche
Scarto quadratico medio e varianza	Metriche
Indici di asimmetria e di curtosi	Metriche

Per concludere, una breve riflessione sull'utilizzo delle analisi monovariate. In generale, la distribuzione di una variabile secondo le diverse modalità assume un interesse modesto, se non viene impiegata all'interno di una comparazione, nel tempo o nello spazio. Sapere che il 26,2% dei residenti a Genova, alla data del Censimento 2001, ha 65 anni o più è ben poco significativo, se non si compara questo dato a quello relativo ad altre grandi città o alla regione nel suo complesso, ovvero al valore che ha assunto a Genova nei Censimenti precedenti.

La comparazione nello spazio avviene di solito costruendo tabelle a doppia entrata in cui si pongono in riga i singoli casi (ad esempio le diverse città che si intende comparare) e in colonna le diverse modalità della variabile (nell'esempio, i residenti per classi di età). Va da sé che una percentualizzazione dei dati 'in orizzontale' (cioè fatto uguale a 100 il totale di riga) consente una comparazione immediata.

Un poco più complessa la comparazione nel tempo. Il modo più semplice di realizzarla è quello di calcolare le variazioni, assolute e percentuali. Le prime si ottengono come differenza tra il valore a fine periodo e quello a inizio periodo, le seconde come quoziente (per cento) tra la differenza assoluta e il valore di inizio periodo. Se tuttavia si dispone di una serie storica di dati è più agevole utilizzare i numeri indice. In questo caso si pone pari a cento il valore di un anno (di solito quello di inizio periodo, ma si può scegliere un anno diverso, ad esempio quello in cui è stato registrato il valore massimo o minimo della serie[6]) e si indica per ogni anno il quoziente (per cento) ottenuto dividendo il valore assoluto dell'anno per quello dell'anno assunto come base di calcolo. In questo modo si può evidenziare l'andamento nel tempo di serie storiche che risulterebbero illeggibili se si riportassero le sole variazioni percentuali tra un anno e l'altro. Per inciso, il rapporto tra due numeri indice (meno 100) corrisponde alla variazione percentuale tra i due anni cui si riferisce l'indice, mentre la differenza tra il numero indice di un anno e 100 corrisponde alla variazione percentuale tra l'anno di riferimento e quello considerato. La tab. 7 fornisce un esempio di quanto detto. I due decimali consentono di evidenziare che ogni numero indice è composto da 100 più la differenza percentuale tra l'anno di riferimento e quello cui si riferisce il numero indice. Con semplici calcoli il lettore potrà constatare che il rapporto (per cento) tra due numeri indice è pari alla variazione percentuale del periodo meno cento.

Tab 7 - Serie temporale dei residenti a Genova. Censimenti e relative statistiche.

anni	1951	1961	1971	1981	1991	2001
Residenti	688.447	784.194	816.872	762.895	678.771	610.307
Var. ass.*		95.747	32.678	-53.977	-84.124	-68.464
var. %*		13,91	4,17	-6,61	-11,03	-10,09
N. Indice**	100	113,91	118,65	110,81	98,59	88,65

* Le variazioni sono calcolate rispetto al dato censuario precedente

** I numeri indice sono calcolati assumendo come base il 1951

9.2.2. L'analisi bivariata e multivariata

Si è detto più volte che l'interesse del ricercatore è rivolto prevalentemente alla ricerca delle relazioni tra variabili, siano esse ipotizzate fin dall'inizio della ricerca, ovvero emergenti dall'analisi dei dati, richiedendo in tal caso una revisione o un arricchimento delle ipotesi iniziali.

In buona parte le tecniche di elaborazione dei dati sono quindi rivolte all'esplorazione delle relazioni tra variabili, ovvero alla cosiddetta analisi bivariata o multivariata.

Vale la pena innanzi tutto di notare che l'esame delle relazioni tra variabili fornisce informazioni assai più ricche della semplice analisi delle loro distribuzioni. Se ad esempio in una ricerca si registrasse, da un lato, che il 50% del campione è composto da ricchi, l'altro 50% da poveri, e d'altro canto che il 50% è composto da appartenenti alla fede religiosa A, l'altro 50% da fedeli della religione B, l'aspetto più interessante

[6] Si noti che scegliendo come base di calcolo un anno particolarmente sfortunato la serie mostra valori superiori a 100, mentre scegliendo un anno fortunato si ottiene l'effetto opposto. Dunque nei numeri indice è importante motivare adeguatamente la scelta della base di calcolo; in casi difficili, è meglio usare come base la media del periodo piuttosto che costruire serie troppo ottimistiche o pessimistiche ad effetto di una base di calcolo poco difendibile (ad es., il primo anno per il quale si dispone di dati).

sarebbe costituito non già dalle due distribuzioni isolate, ma dalla relazione tra le due distribuzioni: se tutti i poveri professassero la religione A, tutti i ricchi la B, saremmo in presenza di una società a forte discriminazione su base religiosa. Del pari, se in una società scoprissimo che i figli riproducono lo status sociale dei padri parleremmo di società priva di mobilità sociale. Nel converso, se la povertà fosse equidistribuita rispetto alla religione, o lo status sociale dei figli non riproducesse quello dei padri, riterremmo di essere in presenza di una società non discriminatoria, quantomeno sotto questi due importanti punti di vista.

In linea generale, per evidenziare l'esistenza di una relazione tra due o più variabili occorre considerare la loro distribuzione congiunta (infatti in statistica si parla di variabilità a due o più dimensioni). Il modo più semplice per farlo è quello di costruire una tabella, detta a doppia entrata, che riproduca tale distribuzione. Si tratta anche in questo caso di una matrice, diversa tuttavia dalla matrice dei dati, dalla quale peraltro discende. È infatti una matrice 'variabile per variabile', in quanto porta nella prima riga e nella prima colonna (dette anche 'tessuto metrico') le modalità assunte dalle due variabili; nelle caselle che risultano dagli incroci delle diverse modalità (dette 'tessuto connettivo') il numero di casi che presentano congiuntamente i valori delle due variabili descritti dall'intestazione di riga e di colonna; nell'ultima riga e colonna (dette 'marginali') si hanno i totali dei casi che presentano una modalità specifica della variabile di riga o di colonna. Per convenzione, si pone usualmente nelle righe la variabile dipendente e nelle colonne quella indipendente (anche se non sempre è possibile ipotizzare a priori una relazione di dipendenza tra le due variabili).

L'esempio di Tab. 8 illustra quanto fin qui detto.

*Tab. 8 - Distribuzione di una popolazione campionaria adulta per livello d'istruzione ed età**

età liv. istr.	18-30 anni	31-50 anni	51 e oltre	Totale
Basso	45	75	175	295
Medio	131	231	143	505
Alto	107	70	23	200
Totale	283	376	341	1.000

***** Si noti che nell'indicazione delle due variabili si pone al primo posto quella riportata nelle righe e al secondo quella riportata nelle colonne.

La Tab. 8 rappresenta una matrice quadrata, in quanto il numero delle righe è pari al numero delle colonne, o una matrice '3 x 3' (si contano, ovviamente, solo le modalità assunte dalle due variabili, non le intestazioni di riga o colonna e i totali finali, ossia le righe o colonne marginali). Si tratta altresì di una tabella in cui sicuramente una variabile è indipendente (l'età), mentre l'altra potrebbe essere dipendente (qualora si ipotizzasse, com'è plausibile, che la recente esplosione della scolarizzazione superiore abbia interessato più i giovani che gli adulti, ormai usciti dal circuito scolastico). In altri termini, è possibile che l'età influenzi il titolo di studio, ma certamente è impossibile che il titolo di studio influenzi l'età[7].

Analogamente a quanto si è visto nell'analisi monovariata, i valori assoluti non aiutano a capire se vi sia o no una relazione tra le due variabili. In questo caso, tuttavia, abbiamo la scelta fra ben tre diverse modalità di costruire le percentuali, riportate nelle tabelle seguenti.

Le differenze tra le tabelle derivano dal fatto che rispondono a domande differenti. La 9, infatti, consente di evidenziare se i tre livelli d'istruzione siano differenziati rispetto alla distribuzione per età, ovvero di rispondere alla domanda "qual è la composizione per età di ogni gruppo di individui con pari titolo di studio?". La risposta è che esiste una composizione differenziata: infatti gli individui in possesso di elevato titolo di studio sono per oltre il 50% giovani, non ostante questi rappresentino meno del 30% del campione; di contro, solo per l'11,5% sono anziani, benché gli anziani abbiano un peso addirittura maggiore dei giovani sul totale generale. Una situazione quasi speculare si registra per gli individui con basso livello d'istruzione.

*Tab. 9 - Distribuzione percentuale per livello d'istruzione secondo l'età**

età livello istruzione	18-30 anni	31-50 anni	51 e oltre	Totale

[7] Questo vale, naturalmente, se si considera l'età in termini strettamente anagrafici. Se si parlasse di età percepita (dai singoli soggetti e dai loro interlocutori) si potrebbe invece trovare una relazione tra titolo di studio ed età di tipo diverso, perché di solito una maggiore istruzione (in via diretta, o perché legata ad un reddito maggiore) aiuta le persone in età matura a gestire la propria immagine fisica in modo da mostrare un'età minore e a percepirsi come più giovani socialmente, mentre nel converso un titolo di studio più elevato potrebbe 'invecchiare' socialmente dei giovani.

Basso	15,25	25,42	59,32	100,00
Medio	25,94	45,74	28,32	100,00
Alto	53,50	35,00	11,50	100,00
Totale	28,30	37,60	34,10	100,00

* Il titolo della tabella chiarisce che è l'età la variabile secondo la quale viene effettuata la percentualizzazione di ogni singolo livello d'istruzione.

Questo breve commento mostra che i dati sono stati analizzati mettendo a confronto i valori delle singole caselle con la media e tra di loro, sottolineando gli scostamenti positivi o negativi più rilevanti. Le conclusioni che trarremmo da questa modalità di percentualizzazione sono che l'età influenza il livello d'istruzione in modo abbastanza sensibile.

Nella Tab. 10 leggiamo invece la distribuzione di ogni gruppo d'età secondo il livello d'istruzione. In questo caso si ci si chiede come sono distribuiti i diversi gruppi d'età rispetto all'istruzione e si noterà che i giovani fanno registrare una consistente presenza di livelli d'istruzione elevati e, soprattutto, che gli anziani sono caratterizzati da una notevole presenza di livelli d'istruzione bassi. Questa elaborazione potrebbe rivelarsi utile nel caso, ad esempio, intendessimo programmare una campagna informativa rivolta a giovani o ad anziani, per calibrare il linguaggio utilizzato o le conoscenze di carattere generale possedute dalla popolazione *target*.

*Tab. 10 - Distribuzione percentuale per età secondo il livello d'istruzione**

età livello istruzione	18-30 anni	31-50 anni	51 e oltre	Totale
Basso	15,90	19,95	51,32	29,50
Medio	46,29	61,44	41,94	50,50
Alto	37,81	18,62	6,74	20,00
Totale	100,00	100,00	100,00	100,00

* Il titolo chiarisce che si tratta della situazione opposta rispetto alla tabella precedente

Solo la Tab. 11 evidenzia tuttavia in modo immediato le *relazioni* che corrono tra le due variabili. Nel caso specifico, si nota un addensamento di valori lungo la diagonale secondaria della matrice[8], che conferma l'esistenza di quella relazione di proporzionalità inversa che già avevamo ipotizzato analizzando le tabelle precedenti. In questo caso i valori di ogni casella riproducono il 'peso' di ognuna di queste sul totale generale, mentre le marginali indicano la quota di casi appartenenti a quella modalità sul totale.

Tab. 11 - Distribuzione percentuale per livello d'istruzione ed età

età livello istruzione	18-30 anni	31-50 anni	51 e oltre	Totale
Basso	4,50	7,50	17,50	29,50
Medio	13,10	23,10	14,30	50,50
Alto	10,70	7,00	2,30	20,00
Totale	28,30	37,60	34,10	100,00

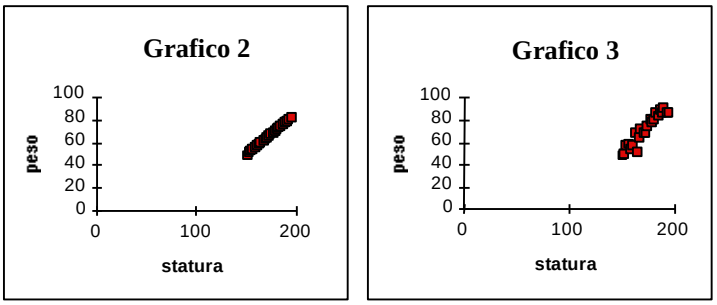
Alcune regole pratiche in tema di tabelle: qualora per brevità si riportino solo le distribuzioni percentuali che si considerano importanti per la presentazione dei dati, non si mancherà di evidenziare, sotto il totale percentuale di 100 (che va sempre indicato, per consentire di cogliere subito la direzione di calcolo delle percentuali), il numero totale dei casi di riga o di colonna. Inoltre, si eviterà di calcolare delle percentuali se i valori assoluti sono modesti: su di un totale di 50 casi ha già poco senso elaborare le frequenze relative, ma non ne ha alcuno elaborare le frequenze di una tabella a doppia entrata, che presenterebbe sicuramente caselle vuote e righe o colonne con meno di dieci casi!

[8] In una matrice quadrata, la diagonale principale è quella che unisce le caselle di intersezione delle classi aventi lo stesso numero d'ordine (le prime, le seconde, le terze, delle due variabili; nell'esempio, 'Basso' e '18-30 anni'; 'Medio' e '31-50 anni'; 'Alto' e '51 e oltre'); la diagonale secondaria è quella ortogonale alla prima, di intersezione tra la prima e l'ultima classe, la seconda e la penultima, ecc.; nell'esempio, unisce 'Alto' e '18-30 anni'; 'Medio' e '31-50 anni'; 'Basso' e '51 e oltre'.

Una corretta presentazione dei dati in forma tabellare richiede inoltre un titolo corretto ed esaustivo (nelle Tabb. 9, 10, 11, ad es., non si dice di quali soggetti venga riportata la distribuzione), nonché l'indicazione della fonte (rilevazione diretta, dati Istat, ecc.) e degli anni cui si riferiscono i dati o la rilevazione.

Molto utili sono anche le rappresentazioni grafiche delle relazioni tra variabili. Una tra le più comuni, utilizzata soprattutto usando i valori ordinati secondo le due variabili, più raramente con i dati accorpati in classi, è costituita dal cosiddetto *scatter diagram*, che si ottiene disegnando un piano definito da due assi ortogonali che corrispondono alle due variabili e rappresentando ogni caso con un punto in corrispondenza dei valori assunti sulle due variabili. In questo modo si nota agevolmente se esistono addensamenti dei casi in certe zone del piano, ovvero se i casi tendono a disporsi secondo una curva nota.

I due grafici che seguono riportano il diagramma di dispersione di due distribuzioni di un ipotetico campione di adulti. Nella prima è stata ipotizzata una perfetta relazione tra peso e statura (Grafico 2): in questo caso tutti i valori sono distribuiti lungo una retta. Nel secondo caso (Grafico 3), pur rimanendo leggibile la relazione lineare tra le due variabili, si nota una certa dispersione rispetto alla retta del Grafico 2.



L'interesse del ricercatore si focalizza, ovviamente, sulla ricerca di relazioni tra le due variabili. Ci si chiede, innanzi tutto, se esista una qualche relazione; in caso affermativo, di che tipo sia e, infine, quale sia la sua intensità.

Per rispondere a queste domande si è soliti confrontare la distribuzione in esame con due tipi di situazioni-limite, che possono essere sempre costruite a partire da tale distribuzione.

Il primo caso-limite è costituito dall'assenza di qualsiasi tipo di relazioni tra le due variabili. Questa situazione è nota in statistica come 'indipendenza stocastica', e si verifica quando le frequenze di ogni singola cella (calcolate in diagonale, ossia secondo le modalità di cui alla Tab. 11) sono uguali al prodotto delle frequenze marginali. Si usa dire in questi casi che $p_{ij}=p_i*q_j$, dove p_{ij} è la frequenza di una casella all'incrocio tra la colonna i e la riga j, p_i è la frequenza di colonna e q_j è la frequenza di riga. La Tab. 12 riproduce la situazione di indipendenza stocastica relativamente ai dati di Tab. 8, a partire dai valori delle marginali ivi contenuti.

Il modo più semplice di misurare la distanza che intercorre tra la situazione teorica di indipendenza stocastica e la distribuzione reale in esame è quello di calcolare le differenze che si registrano, cella per cella, tra le due distribuzioni. Tali differenze, ottenute sottraendo per ogni casella le frequenze osservate da quelle teoriche, prendono il nome di contingenze; la loro formula è $c_{ij}=p_{ij}-p_i*q_j$.

Tab. 12 - Distribuzione di una popolazione campionaria adulta per livello d'istruzione ed età in condizione di indipendenza stocastica

*Valori assoluti**

età livello istruzione	18-30 anni	31-50 anni	51 e oltre	Totale
Basso	83	111	101	295
Medio	143	190	172	505
Alto	57	75	68	200
Totale	283	376	341	1000

Valori percentuali

età livello istruzione	18-30 anni	31-50 anni	51 e oltre	Totale
Basso	8,35	11,09	10,06	29,50
Medio	14,29	18,99	17,22	50,50

Alto	5,66	7,52	6,82	20,00
Totale	28,30	37,60	34,10	100,00

* I valori assoluti si ottengono moltiplicando i totali di riga e di colonna e dividendo il prodotto per il numero totale dei casi; ad esempio: 83 (casella 1.1) è = 295 x 283 ÷ 1.000

La Tab. 13 riporta i valori delle contingenze sia in valore assoluto che in percentuale, ottenute nel primo caso sottraendo ai valori della Tab. 8 quelli della prima parte della Tab. 12; nel secondo sottraendo ai valori della Tab. 11 quelli della seconda parte della Tab. 12. Si ricordi che gli indici di connessione utilizzano i valori assoluti o le frequenze pure, ossia non percentualizzate.

Si noti che i totali delle contingenze di riga e di colonna sono pari a zero, dal momento che la tecnica di calcolo è tale da evidenziare gli scostamenti interni alla riga o alla colonna registrati tra la distribuzione osservata e quella teorica di riferimento. Di conseguenza, gli indici statistici costruiti a partire da questa tabella, come il chi quadro e altri che da questo derivano, utilizzano sempre i quadrati delle contingenze, per la stessa ragione per cui lo scarto quadratico medio utilizza i quadrati degli scarti dalla media.

Il chi quadro può assumere un valore minimo di zero, in condizione di identità tra distribuzione osservata e distribuzione teorica (ossia quando esiste indipendenza stocastica tra le due variabili). Il valore massimo è invece indeterminato, in quanto dipende dalla numerosità dei casi contenuti nella tabella (per un approfondimento sul tema, si veda Bailey, 1991: 473 e segg.).

Il vantaggio che presentano il chi quadro e gli indici statistici da questo derivati consiste nel fatto che sono calcolabili per qualunque tipo di coppia di variabili, anche se nominali. Il chi quadro nulla dice, tuttavia, sulla forma assunta dall'eventuale relazione riscontrata.

Tab. 13 - Tabella delle contingenze

Valori assoluti

età livello istruzione	18-30 anni	31-50 anni	51 e oltre	Totale
Basso	-38,49	-35,92	74,41	0,00
Medio	-11,91	41,12	-29,21	0,00
Alto	50,40	-5,20	-45,20	0,00
Totale	0,00	0,00	0,00	0,00

Valori percentuali

età livello istruzione	18-30 anni	31-50 anni	51 e oltre	Totale
Basso	-3,85	-3,59	7,44	0,00
Medio	-1,19	4,11	-2,92	0,00
Alto	5,04	-0,52	-4,52	0,00
Totale	0,00	0,00	0,00	0,00

Di contro, per valutare anche il tipo di legame tra due variabili è necessario che queste siano quantomeno ordinali, meglio ancora metriche. In questi casi si utilizzano altri casi-limite di riferimento. Il più diffuso è quello della relazione perfetta tra due variabili. Il caso più semplice di relazione perfetta tra variabili è dato da una relazione lineare, che sta alla base del calcolo degli indici di correlazione e di regressione. Si tratta della situazione opposta a quella presa a riferimento dal chi quadro: non più l'assenza di qualunque relazione, ma la presenza di una specifica relazione. In questo caso si assume che una variabile sia espressa da una funzione (lineare) dell'altra. La Tab. 14 riproduce una matrice quadrata in cui si ha perfetta relazione lineare positiva, ossia un rapporto di proporzionalità diretta tra le due variabili, mentre la 15 riproduce con gli stessi dati una situazione di perfetta relazione negativa, ossia di proporzionalità inversa.

Tab. 14 - Esempio di relazione lineare positiva perfetta

età livello istruzione	18-30 anni	31-50 anni	51 e oltre	Totale
Basso	300	0	0	300
Medio	0	400	0	400
Alto	0	0	300	300
Totale	300	400	300	1.000

Tab. 15 - Esempio di relazione lineare negativa perfetta

età livello istruzione	18-30 anni	31-50 anni	51 e oltre	Totale
Basso	0	0	300	300
Medio	0	400	0	400
Alto	300	0	0	300
Totale	300	400	300	1.000

Nelle Tabelle 14 e 15 si ha una corrispondenza biunivoca tra righe e colonne: gli individui con livello d'istruzione basso sono tutti di età superiore a 50 anni, quelli con livello alto al di sotto dei 31 anni, ecc. Si noti che questa situazione è possibile, quando si opera su dati aggregati, solo se vengono rispettate due condizioni: la matrice è quadrata e le marginali di riga e colonna sono uguali a due a due.

Nel caso della Tab. 14 i valori sono allineati lungo la diagonale principale della matrice e, poiché entrambe le variabili sono articolate in classi ordinate in senso crescente, si ha una relazione lineare positiva o proporzionalità diretta (ossia, al crescere dell'età cresce anche il livello d'istruzione). Nella Tab. 15 si ha invece una relazione lineare negativa perfetta, ovvero una proporzionalità inversa, in quanto al crescere dell'età decresce il titolo di studio. Occorre anche aggiungere che per utilizzare i coefficienti di correlazione lineare occorre che entrambe le variabili siano metriche. Dunque i tre livelli di istruzione dovranno essere espressi in termini numerici, magari in termini di numero medio di anni di studio compiuti.

Il lettore attento noterà inoltre che, quando ci si discosta dalla situazione tipo illustrata (matrici quadrate con corrispondenza tra le marginali), l'eventuale relazione tra variabili può essere messa in ombra. Consideriamo ad esempio i dati di Tab. 16.

Tab. 16 Distribuzione di una popolazione campionaria per livello d'istruzione ed età

età livello istruzione	18-24	25-30	31-40	45-50	51-60	61 e oltre	Totale
alfabeta	0	0	0	0	50	100	150
licenza elementare	0	0	0	0	100	50	150
media inferiore	0	0	0	200	0	0	200
media superiore	0	0	200	0	0	0	200
laurea	150	50	0	0	0	0	200
post laurea	0	100	0	0	0	0	100
Totale	150	150	200	200	150	150	1000

La Tabella contiene, con maggior dettaglio, gli stessi dati delle Tabb. 14 e 15. Tuttavia, si nota innanzi tutto che non c'è corrispondenza biunivoca tra ogni riga e colonna: la classe di età 25-30 e le due finali presentano infatti due caselle piene e non una, e lo stesso accade per le righe 'alfabeta', 'licenza elementare' e 'laurea'. Pur trattandosi degli stessi dati di partenza, quindi, la loro distribuzione congiunta non evidenzia la stessa relazione lineare perfetta che emerge dalla Tab. 15 o dalla 14. Analogamente, se si decidesse di accorpare alla voce 'livello d'istruzione basso' l'insieme dei soggetti compresi tra alfabeti e in possesso di licenza media inferiore, lasciando nella categoria 'medio' il solo diploma, si otterrebbe di nuovo una tabella da cui non emerge la relazione perfetta evidenziata dalle tabelle. 14 e 15.

Senza procedere oltre nell'esempio, va quindi ribadito che la ricerca di relazioni tra variabili è influenzata anche dalle operazioni di accorpamento delle classi in cui si articola una variabile, oltre che, ovviamente, dalla sua struttura iniziale. Va tuttavia distinto il caso in cui l'articolazione in classi è costitutiva della variabile, essendo operata prima della rilevazione come prerequisito essenziale per effettuarle (come accade, ad esempio, se si chiede di barrare una casella in corrispondenza di un livello d'istruzione predefinito) da quello in cui viene rilevato un dato più fine (ad es. il numero di anni scolastici frequentati con successo o il reddito mensile in Euro), successivamente accorpato in classi. Le formule per il calcolo delle relazioni tra variabili operano infatti sia su dati aggregati in classi che su dati individuali: in questo secondo caso l'accorpamento effettuato in sede di presentazione dei dati non influenza, ovviamente, il calcolo, effettuato sui dati rilevati e non su quelli accorpati successivamente.

La relazione lineare tra variabili è misurata dal coefficiente di correlazione lineare, il cui valore varia tra 1 (perfetta relazione positiva), 0 (assenza di relazione lineare), -1 (perfetta relazione negativa). Tuttavia

l'assenza di relazione lineare non comporta l'assenza di qualunque tipo di relazione. Nel mondo sociale abbondano anzi i casi di relazioni curvilinee (si pensi ad esempio alla soddisfazione per unità di prodotto: sale fino al livello di saturazione, dopo di che scende progressivamente; in questo caso potrebbe esistere una relazione parabolica perfetta, a fronte della quale la relazione lineare sarebbe ovviamente nulla). In altre parole, poiché i dati rispondono alle nostre domande, dobbiamo avere l'accortezza di porre quelle giuste (magari aiutati da una prima lettura del grafico di distribuzione dei casi, che già suggerisce la forma delle eventuali relazioni tra le variabili). Ciò significa, da un lato, disporre di un *set* di ipotesi iniziali dalle quali far discendere le elaborazioni cui sottoporre i dati; d'altro lato la capacità di passare dai risultati delle elaborazioni all'analisi dei dati e viceversa, ossia di saper usare la statistica per interrogare i dati alla luce delle conoscenze sostantive sul fenomeno che stiamo analizzando.

Le conoscenze sostantive sono fondamentali anche per un'altra ragione, che qui citiamo solo brevemente: la statistica fornisce informazioni sulla relazione tra due variabili, sulla loro intensità e sul 'verso' (relazione diretta o inversa), non sul senso di questa relazione. In altre parole, se si registra una correlazione perfetta tra livello d'istruzione e reddito, è in base a conoscenze sostantive che si definisce quale sia la variabile dipendente e quella indipendente. In aggiunta, l'assegnazione preliminare del carattere dipendente o indipendente a una variabile influenza anche il tipo di elaborazione statistica da effettuare. Se è ipotizzata fin dall'inizio una relazione di dipendenza, si utilizza la regressione; se si ipotizza l'interdipendenza, ci si avvale della correlazione. La formula di calcolo del coefficiente di correlazione presuppone infatti che gli scostamenti dal caso-limite ipotizzato (nella correlazione lineare, una retta) siano attribuibili ad entrambe le variabili. Quella del coefficiente di regressione presuppone invece che solo la variabile dipendente sia responsabile degli scostamenti (nella regressione lineare, ancora una retta).

Il coefficiente di correlazione lineare merita tuttavia un ulteriore cenno, in quanto costituisce la base di molte elaborazioni statistiche più complesse, che rientrano nell'analisi multivariata, quali l'analisi dei fattori e l'analisi dei gruppi (*cluster analysis*; per una trattazione approfondita si rinvia a Micheli e Manfredi, 1995). Il suo calcolo presuppone la standardizzazione delle variabili di partenza, ossia un'operazione (consentita solo su variabili metriche) in forza della quale ogni valore di una variabile viene trasformato in un nuovo valore, dato dalla differenza tra il valore originario e la media della distribuzione, diviso per lo scarto quadratico medio. In questo modo si ottiene la cosiddetta 'variabile standardizzata', costituita anche nei valori argomentali da numeri puri (in quanto quoziente di grandezze dello stesso ordine e tipo). Questa operazione è fondamentale perché consente di ottenere una nuova distribuzione avente media zero e scarto quadratico medio uguale a uno. Le due variabili così trasformate sono agevolmente rappresentabili in uno spazio cartesiano in cui l'origine (zero) corrisponde alla media delle due variabili e la posizione dei singoli punti è correttamente raffrontabile con le bisettrici del primo e terzo o secondo e quarto quadrante, che identificano, rispettivamente, i casi di perfetta relazione positiva o negativa. Già l'esame del diagramma di dispersione di due variabili standardizzate consente di ipotizzare l'esistenza di una relazione lineare o curvilinea e di capire quanto un'eventuale curva nota (retta, parabola, iperbole, ecc.) interpolata tra i diversi punti li rappresenti correttamente (detto anche grado di *fitness* di una curva).

Occorre aggiungere che le analisi multivariate non mirano solo a rilevare l'esistenza di relazioni tra variabili, ma possono essere anche rivolte all'analisi della distribuzione dei casi in base al possesso congiunto di due o più caratteristiche, come accade nella *cluster analysis.* Inoltre, tecniche di analisi dei dati assai sofisticate permettono di lavorare anche sulle distribuzioni congiunte di variabili non metriche, come accade per i modelli *log-lineari*, per l'*analisi delle corrispondenze*, ecc.

Si è fino a questo punto parlato di analisi statistica della relazione tra due variabili senza distinguere tra una distribuzione campionaria e quella dell'intera popolazione; limitandoci cioè alla statistica descrittiva senza chiamare in causa la statistica inferenziale. In realtà, poiché di solito si lavora su campioni, si pone innanzi tutto il problema di stabilire quanto la relazione rilevata nel campione corrisponda ad un fenomeno proprio anche dell'universo, ovvero vada attribuita all'errore di campionamento. Non a caso, ad esempio, il chi quadro viene presentato dai manuali di statistica come un *test*, per decidere tra queste due alternative.

La scelta viene effettuata ovviamente su basi probabilistiche, ossia stabilendo un livello di probabilità al di sopra del quale si può ritenere che la relazione registrata non sia attribuibile ad errore di campionamento. Solo successivamente si procede, ove possibile, a misurare l'intensità e la forma di tale relazione.

Si è detto 'ove possibile', in quanto il tipo di elaborazioni cui sottoporre i dati dipende, anche nel caso dell'analisi bivariata, dal modo in cui queste sono state misurate. Mentre le variabili metriche ammettono la quasi totalità delle elaborazioni possibili (fatta salva la differenza, segnalata da Ricolfi 1995, tra scale assolute e scale posizionali), quelle ordinali e, soprattutto, le nominali subiscono molte limitazioni. Modelli di elaborazione di crescente diffusione, come la già citata *analyse des données* e i modelli log lineari (cfr. Chiari e Peri 1987; Corbetta 1994; Lanzetti, 1995), hanno peraltro arricchito in modo significativo le possibilità di lavoro del sociologo, le cui variabili più significative sono spesso di carattere non metrico.

Un altro modo per superare le limitazioni che derivano dall'impiego di variabili non metriche è quello di trasformarle in una variabile dicotomica, assegnando i valori convenzionali di zero (assenza di proprietà) e uno (presenza della proprietà) alle due categorie così ottenute. Nel caso del livello d'istruzione, sopra riportato, è possibile 'dicotomizzare' la variabile trasformandola in vario modo. Ad esempio distinguendo tra 'livello inferiore o pari all'obbligo' e 'oltre l'obbligo', oppure tra 'alfabeti senza titolo di studio' e 'in possesso di un titolo di studio'. Va da sé che queste trasformazioni corrono spesso il rischio di deformare la distribuzione, rendendo in realtà invisibili relazioni che pure esistono, o peggio di invalidare l'operazione di classificazione da cui la variabile dicotomica ha origine (cfr. Marradi, 1987). Se ad es. si dicotomizza tra 'nessun titolo di studio' (= alfabeti) e 'con titolo di studio' (= tutte le altre categorie), si assegna, di fatto, lo stesso peso al possesso di una specializzazione post laurea e alla licenza elementare. È pur vero che, per prudenza, di solito si ripete l'operazione per ogni categoria (considerando residuali tutte le altre), oppure si cerca di effettuare accorpamenti sensati (ad esempio tra obbligo scolastico e post obbligo), ma in tutti i casi i rischi di distorsione sono consistenti.

Ci si è soffermati su questo esempio perché in sede di elaborazione dei dati si pone, in modo rovesciato, un problema analogo a quello segnalato da Ricolfi (1995) in ordine all'ispezionabilità della base empirica. In altri termini, se in sede di elaborazione dei dati vengono compiute delle scelte non trasparenti per il lettore, i principi della ripetibilità, pubblicità e controllabilità del processo di ricerca, che sono costitutivi del suo carattere scientifico, vengono messi in discussione altrettanto seriamente. È peraltro vero che la disponibilità dei dati elementari consentirebbe a qualunque ricercatore di sottoporre gli stessi a procedure di elaborazione diverse, mentre la ripetibilità è assai più difficile da realizzare in gran parte delle ricerche qualitative.

Senza alcuna pretesa di esaustività, ci limitiamo a ricordare nel seguito le principali tecniche statistiche utilizzate nell'analisi bivariata in ragione del tipo di variabile in esame.

Quando la variabile dipendente è di tipo qualitativo si utilizzano tabelle di contingenza e misure derivate (come il Chi quadro), ivi incluse le misure di associazione consentite da variabili ordinali; l'analisi logistica; l'analisi delle corrispondenze.

Quando la variabile dipendente è di tipo quantitativo si ricorre:

- se l'indipendente è qualitativa all'analisi della varianza;
- se l'indipendente è quantitativa a correlazione e regressione;
- indipendentemente dal carattere della variabile indipendente è possibile impiegare l'analisi della covarianza e la regressione con variabili fittizie (ossia ottenute dicotomizzando le quantitative).

Per quanto un sociologo tenti di semplificare i propri modelli di lettura della realtà, difficilmente è in grado di trarre conclusioni importanti sulla base della sola analisi bivariata. Fenomeni quali la devianza, lo status sociale, il grado di socializzazione, non possono essere spiegati da una sola variabile indipendente, ma dipendono in diversa misura da famiglia di origine, tipo e livello d'istruzione posseduta, età, condizione occupazionale, ecc. È evidente allora che molto spesso è necessario utilizzare tecniche di elaborazione che consentano di tener conto simultaneamente di più variabili, per misurare le relazioni reciproche. Per una trattazione approfondita del caso più semplice, quello relativo a tre variabili, si rinvia a Ricolfi (1993); per una panoramica più generale, si veda Corbetta (1994).

La base di partenza è spesso costituita dalla matrice dei coefficienti di correlazione lineare tra le diverse variabili: quindi un nuovo tipo di matrice, variabili per variabili, che all'interno delle celle riporta le correlazioni tra quelle indicate in riga o in colonna. Il problema più rilevante che si pone nell'elaborare questa matrice deriva dal fatto che quasi mai ci si trova dinanzi ad una situazione in cui una variabile può essere definita dipendente e tutte le altre indipendenti. In altre parole, non è mai di immediata evidenza, conoscendo i coefficienti di correlazione di una variabile dipendente Y da altre due variabili X e Z considerate indipendenti, quanto davvero Y dipenda da X e da Z, poiché occorre (cfr. Ricolfi, 1993) sia scontare l'effetto delle relazioni tra queste due, sia tener conto dei limiti propri del coefficiente di correlazione lineare (che misura solo le relazioni lineari dell'intera distribuzione, mentre potrebbero esistere sia relazioni n*online*ari, sia relazioni lineari parziali). Inoltre, assume particolare importanza nell'analisi multivariata la disponibilità, da parte del ricercatore, di un modello plausibile di relazioni tra le variabili. Un modello, ad esempio, in grado di definire quali siano le variabili rilevanti per l'analisi e quali relazioni possano essere spurie, (ossia dovute all'effetto di una terza variabile interveniente, o apparenti, ossia annullate dall'inserimento di una variabile di controllo (cfr. per alcuni esempi Bailey, 1991: 513 e segg; Boudon, 1970: 67-77). In questa sede ci limitiamo a citare due modalità di elaborazione assai diffuse nella ricerca sociale. La prima è costituita dall'analisi dei fattori, che si è ormai differenziata in numerosi filoni, a seconda che prevalgano gli interessi interpretativi o quelli classificatori (consente infatti, come si è accennato in precedenza, di assegnare pesi ai diversi indicatori per costruire un indice sintetico del concetto misurato). Nell'ambito dell'impiego interpretativo, l'analisi dei fattori nasce dalla constatazione che le variabili che

compongono la matrice di correlazione possono essere considerate indicatori di un minor numero sottostante di fattori, ai quali le variabili sono in vario modo correlate[9].

La seconda è costituita dalla *cluster analysis*, che opera invece un accorpamento in gruppi dei soggetti analizzati sulla base degli stati che presentano sulle proprietà considerate.

Recenti sviluppi hanno portato alla *fuzzy cluster*, che presuppone la non esclusione reciproca dell'appartenenza dei casi a diverse categorie e che promette interessanti sviluppi in una società caratterizzata sempre più dalle appartenenze multiple e temporanee (cfr. Veronesi e Visioli, 2000).

[9] È stato pubblicato recentemente un volume che non solo illustra ampiamente la famiglia di tecniche statistiche raggruppate sotto il nome di 'analisi fattoriale', ma che evidenzia molto bene anche la pluralità di scopi che possono essere perseguiti con queste tecniche (Di Franco e Marradi, 20013).

Per saperne di più

Si rinvia ai testi pubblicati da Franco Angeli nella collana di Metodologia delle scienze umane e in particolare a:

Biorcio R. (1993), *L'analisi dei gruppi*, Franco Angeli, Milano.

Marradi A. (1993), *L'analisi monoviariata*, Franco Angeli, Milano.

Micheli G. A. e Manfredi P. (1995), *Correlazione e regressione*, Franco Angeli, Milano.

Ricolfi L. (1993), *Tre variabili. Un'introduzione all'analisi multivariata*, Franco Angeli, Milano.

Si veda inoltre:

Amaturo E. (1989), *Analyse del données e analisi dei dati nelle scienze sociali*, Centro Scientifico Editore, Torino.

Di Franco G. (2003), *L'analisi multivariata nelle scienze sociali. Modelli log lineari e variabili categoriali*, Carocci, Roma.

Di Franco G. e Marradi A. (2003), *Analisi fattoriale e analisi in componenti principali*, Bonanno Editore, Acireale-Roma.

Lanzetti C. (1995), *Elaborazioni di dati qualitativi*, Franco Angeli, Milano.

Veronesi M. e Visioli A. (2000), *Logica fuzzy: teoria e applicazioni*, Franco Angeli, Milano.