15 L'ANALISI BIVARIATA

di Maria Paola Faggiano

14.1. Il nesso tra due variabili: pianificare e impostare l'analisi, valutare un risultato

Il passaggio dallo studio dell'andamento delle variabili prese singolarmente (cfr. Cap. 12) al vaglio delle *relazioni tra variabili* rappresenta l'ingresso nel vivo dell'analisi dei dati. Adottare una strategia di ricerca quantitativa porta concretamente ad esaminare contemporaneamente *numerose proprietà definite operativamente* rispetto ad altrettanto *numerosi esemplari*, opportunamente selezionati, dell'unità di analisi prescelta. La predisposizione della *matrice dei dati casi per variabili* (cfr. Cap. 11) mette, pertanto, l'analista nella condizione di andare ben oltre la rendicontazione puntuale delle singole informazioni disponibili e di guardare, con opportuni strumenti, ad un'ampia trama di relazioni. Avviare l'analisi bivariata e, ad uno stadio più avanzato, quella multivariata equivale al tentativo di rispondere, attraverso il prezioso supporto dei dati empirici, ai molteplici *perché? come? quando?* che, con riferimento ad uno specifico oggetto di studio, emergono in un classico iter di ricerca. La messa a punto di un *modello analitico-procedurale* (Merton, 1949-1968a; 1949-1968b; 1955), quale *sistema di organizzazione, formalizzazione e rappresentazione delle ipotesi di ricerca*, costituisce un faro teorico ed operativo di fondamentale importanza nel complessivo percorso d'indagine; esso si traduce in un *piano di analisi*, che, combinando funzioni esplorative, descrittive ed esplicative, rappresenta una guida essenziale per il ricercatore, nella cruciale fase in cui egli seleziona e imposta il sistema di *legami* cui dare particolare risalto entro un dataset. Il piano, pur includendo operazioni di controllo della qualità dei dati (cfr. Cap. 10) e di sintesi delle informazioni disponibili (cfr. Cap. 13), come

anche diverse indicazioni tecniche funzionali alla stesura del report di ricerca (si pensi ai molteplici stili di comunicazione dei risultati d'indagine assumibili a seconda delle opzioni grafiche e statistiche privilegiate), veicola le *istruzioni* da mettere in pratica per la realizzazione dell'analisi dei dati, dal livello monovariato al bivariato, compresi gli step successivi. È evidente che la produzione di un piano di analisi implichi la presenza di obiettivi e ipotesi chiari da tradurre in operazioni di analisi concrete; ciò non significa che un analista non possa, per così dire, anche *mettersi in ascolto dei dati*, operando, davanti alla matrice, veri e propri tentativi in una direzione più propriamente esplorativa, non predisposti in sede di stesura del progetto di analisi. La capacità di utilizzare in sinergia il tratto indispensabile della pianificazione con doti di creatività e di disponibilità alla scoperta non potrà che conferire ricchezza, profondità e fondatezza all'analisi della base empirica.

Il capitolo si concentra sul *livello bivariato di analisi*, step intermedio di fondamentale importanza entro piani articolati di elaborazione dati; l'analisi bivariata si concretizza nella scelta, guidata da specifici interessi ed ipotesi di ricerca, di coppie di variabili di cui vagliare l'andamento congiunto e rispetto a cui, sulla base di un opportuno corredo statistico e grafico-tabellare, produrre riflessioni e note di commento[1]. È bene precisare sin da ora che l'analisi bivariata può coinvolgere anche variabili complesse, che rappresentano l'esito di una specifica procedura di sintesi; un

[1] Incrociare variabili e/o riflettere sugli spazi logici che scaturiscono dall'abbinamento di modalità espressive di specifici caratteri torna utile anche in fasi dell'indagine non strettamente ascrivibili all'analisi dei dati (cfr. Faggiano, 2012). Si pensi 1. alla *messa a punto* della più opportuna *strategia di campionamento* e alla selezione delle variabili da valorizzare ai fini dell'individuazione degli esemplari su cui condurre uno studio empirico (il *campionamento stratificato* o *per quote* costituiscono un valido esempio in questa direzione – cfr. Cap. 5); 2. alle tante occasioni in cui, entro un dato iter di ricerca, si renda necessario operare un *controllo* ed individuare una *strategia correttiva*. Costituiscono esempi concreti in questa direzione a. i *confronti tra campione progettato* – sulla base di specifiche variabili-criterio – *e campione effettivamente raggiunto* (ai fini dell'integrazione o revisione del piano, come anche della ponderazione del campione disponibile); b. diversi *controlli di qualità del dato* (cfr. Cap. 10): b1. *di congruenza* (di fronte a coppie di variabili come *titolo di studio* e *professione svolta*; *età* e *sviluppo di patologie senili*; *età* e *stato civile*, *status di genitore* ed *età dei figli*, ecc. si rende necessario un accurato check, dati alla mano, di alcune combinazioni di modalità logicamente o formalmente impossibili; ciò comporta, in presenza di errori riscontrati in matrice, la messa a punto di opportune operazioni di pulizia del dato), b2. *di coerenza nel tempo* (in caso di rilevazioni replicate ciclicamente a parità di casi di studio), b3. *di confronto tra risposte riferibili a gruppi diversi* (*sperimentale e di controllo*) entro disegni di ricerca dall'impronta sperimentale (cfr. Cap. 9). Infine, è di fondamentale importanza impostare incroci tra variabili e osservare nel dettaglio tutte le celle di una tabella di contingenza (in cui siano state calcolate le percentuali sul totale dei casi) − puntando, evidentemente, alla più proficua sintesi dei dati −, quando il piano di analisi preveda la realizzazione di *indici tipologici* attraverso la *procedura di costruzione/riduzione di uno spazio di attributi* (cfr. Cap. 13).

indice disponibile in matrice, quale che sia stata la tecnica adottata per la sua realizzazione (si può, ad esempio, trattare di un indice additivo, di un indice tipologico, di un fattore, di gruppi emersi attraverso la *cluster analysis*, ecc. – cfr. Capp. 13, 16 e 17) è, difatti, una colonna, tra le altre, della matrice dei dati. Il dataset di partenza, per così dire, "cresce" progressivamente sulla base delle elaborazioni messe a punto e i livelli di analisi dei dati – monovariato, bivariato, ecc. – ricorrono, ciclicamente, piuttosto che assumere un carattere di linearità.

Prima dell'intervento del ricercatore, che comporta evidentemente delle *decisioni* con riferimento alla destinazione d'uso della base empirica, tecnicamente, la relazione tra due variabili appare *neutralmente bidirezionale e simmetrica.* Apparato teorico-concettuale e conoscenze previe, misti ad immaginazione sociologica, suggeriranno di propendere per una classificazione della relazione proprio come *bidirezionale e simmetrica* (le due variabili si influenzano reciprocamente con pari forza), o come *unidirezionale* (la variabile X, *variabile indipendente*, influenza la variabile Y, *variabile dipendente*, e non viceversa), o, ancora, come *bi-direzionale asimmetrica* (X condiziona Y più di quanto sia da essa influenzata) (cfr. Marradi, 1997).

L'analisi della distribuzione congiunta di due variabili consente di affermare se tra esse esista una relazione («una qualche forma di sistematicità nel modo in cui le modalità di tali variabili sono associate» - cfr. Di Franco, 2001, p. 123), che forma assuma, che forza evidenzi. Possiamo distinguere, anzitutto, i due casi estremi – rispetto a cui gli esempi concreti di ricerca rappresentano generalmente specifiche sfumature collocate tra tali poli del continuum – dell'*indipendenza statistica* (la variabile X assume i propri valori indipendentemente da quelli presentati dalla variabile Y) e della *massima associazione* (o *dipendenza perfetta*) tra due variabili (a ciascuna delle modalità di una variabile è sistematicamente associata una sola modalità dell'altra variabile).

Un esempio chiarirà quanto espresso (Tabb. 14.1. e 14.2.). La prima tabella rappresenta una distribuzione congiunta entro la quale, per ogni singola cella, è riportato lo stesso numero di casi; in altri termini, ciascuna variabile presenta la medesima distribuzione rispetto a tutte le modalità dell'altra variabile.

Tab. 14.1. - Tendenza alla condivisione di stati d'animo sui Social Network in base alla Generazione di appartenenza (v.a.) – Esempio di indipendenza statistica

	Giovani	Adulti	Anziani	Totale
Nulla o Bassa	50	50	50	150
Media	50	50	50	150
Alta	50	50	50	150
Totale	150	150	150	450

Al contrario, nella seconda tabella, le frequenze si concentrano, in modo perfettamente bilanciato, in 3 delle 9 celle disponibili (nelle altre 6 non figurano casi empirici),

in corrispondenza di ben precise combinazioni di modalità (nell'esempio si tratta delle coppie: "anziani-nulla o bassa", "adulti-media", "giovani-alta").

Tab. 14.2. - Tendenza alla condivisione di stati d'animo sui Social Network in base alla Generazione di appartenenza (v.a.) – Esempio di massima associazione o dipendenza perfetta

	Giovani	Adulti	Anziani	Totale
Nulla o Bassa	0	0	150	150
Media	0	150	0	150
Alta	150	0	0	150
Totale	150	150	150	450

Prima di giungere ad evidenziare e commentare nel report di ricerca il nesso tra due variabili, l'analista predispone un accurato controllo delle distribuzioni congiunte. L'*esistenza della relazione* (e l'eventuale *generalizzazione* alla popolazione statistica di riferimento dei risultati ottenuti a livello campionario[2]) è accertata attraverso il calcolo di opportuni *coefficienti di significatività statistica*; è, inoltre, possibile dar conto della *forza della relazione* attraverso il calcolo dei *coefficienti di associazione statistica*; la *progettazione* e *realizzazione di grafici* consente di operare il vaglio della *forma della relazione*; la predisposizione, a fronte di incroci bivariati significativi emersi, di *controlli trivariati e multivariati* consente, infine, di classificare la relazione bivariata come *genuina* o meno (cfr. Statera, 1997; Di Franco, 2001), come anche di approfondire e articolare il quadro dei risultati di ricerca emersi (cfr. Capp. 15 e 16).

Nei paragrafi successivi, con l'ausilio di esempi mirati e riducendo al massimo la complessità di passaggi ed implicazioni di ordine matematico e statistico, il tema della relazione tra due variabili sarà illustrato approfonditamente, mirando all'indicazione puntuale di regole da seguire ed errori da evitare in caso di incroci tra variabili categoriali, di tipo misto, o cardinali/quasi-cardinali. L'obiettivo principale è quello di prospettare soluzioni concrete – dalla costruzione di grafici e tabelle, al calcolo di coefficienti di significatività e associazione – tenendo conto dei diversi tipi di variabile e della loro struttura, come anche delle unità di analisi coinvolte (individui, aggregati territoriali, ecc. – cfr. Cap. 11). Uno spazio adeguato è, inoltre, dedicato all'impostazione dell'analisi e alla lettura ed interpretazione dei risultati ottenuti a livello bivariato, anche in vista di successive e più complesse analisi. Ai fini della produzione di un sistema di esemplificazioni sufficientemente vasto ed eterogeneo, utile a mettere il lettore nella condizione di cogliere la versatilità dell'analisi bivariata, comprese le molteplici forme di comunicazione di dati incrociati, sono presentati output

[2] Operazione possibile, con un margine di errore calcolabile, solo nel caso di campioni probabilistici (cfr. Cap. 5).

a livello bivariato a partire da basi di dati differenziate (un dataset include casi individuali, un altro i Paesi dell'Unione Europea), sistematicamente connesse a recenti e significative occasioni di ricerca empirica.

14.2. Analizzare e rappresentare graficamente relazioni tra variabili categoriali

L'analisi bivariata tra due variabili categoriali viene condotta costruendo *tabelle di contingenza* o *a doppia entrata*. La *distribuzione congiunta* altresì detta corrisponde ad una tabella in cui le variabili selezionate ai fini dell'analisi sono disposte l'una in *riga*, l'altra in *colonna* (cfr. Tab. 14.3.). Le *distribuzioni condizionate* consistono nel sistema di celle in cui è riportata la singola *frequenza di associazione* (ad esempio n_{ij}), equivalente al numero di casi collocati nel punto di intersezione di ciascuna coppia di modalità della X e della Y.

Tab. 14.3. – La struttura di una tabella di contingenza: formalizzazione

		X						Totale
		x_1	x_2	…	x_j	…	x_K	
Y	y_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1K}	$n_{1\cdot}$
	y_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2K}	$n_{2\cdot}$
	:	:	:	…	:	…	:	:
	y_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{iK}	$n_{i\cdot}$
	:	:	:	…	:	…	:	:
	y_H	n_{H1}	n_{H2}	…	n_{Hj}	…	n_{HK}	$n_{H\cdot}$
Totale		$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot j}$	…	$n_{\cdot K}$	N

Distribuzioni condizionate

Frequenze assolute della variabile X (*Distribuzione marginale di colonna*)

Frequenze assolute della variabile Y (*Distribuzione marginale di riga*)

Ad esempio, laddove la X sia rappresentata dalla *generazione d'appartenenza* – articolata nelle 4 modalità "giovani", "giovani adulti", "adulti" e "anziani" – e la Y sia il *possesso di un pc portatile* – con 2 modalità, "sì" e "no" –, il numero di soggetti intervistati caratterizzati dalla duplice caratteristica dell'essere giovani e del possedere un

pc portatile rappresenta una combinazione di modalità, tra le 8 possibili[3], cui si associa un certo numero di casi empirici, porzione, più o meno ampia, del totale dei casi, N. Le *distribuzioni marginali di riga e colonna* coincidono con le *frequenze assolute* di ciascuna variabile, ovvero con le singole *distribuzioni di frequenza* (cfr. Cap. 12).

La previa analisi monovariata delle variabili da incrociare è di fondamentale importanza ai fini del contenimento di effetti distorsivi nelle fasi successive di analisi. Solo *variabili non eccessivamente analitiche* − che, inevitabilmente, moltiplicano il numero di celle di una tavola di contingenza[4] − e, possibilmente, *bilanciate* mettono al riparo da inconvenienti (come celle vuote o con un numero di casi troppo esiguo, coefficienti distorti, ecc.), anche vistosi, e rendono *efficace* la lettura dei dati.

Le variabili categoriali, specie quelle nominali, si articolano in modalità corrispondenti a parole o espressioni linguistiche, dotate di un'elevata *autonomia semantica* (Cfr. Cap. 7). Questa caratteristica mette l'analista nella condizione di leggere *cella per cella* la tabella, anche ai fini dell'individuazione di *specifiche associazioni locali* (ad esempio, l'attrazione tra le due modalità "predilezione di studi universitari ad indirizzo umanistico" e "genere femminile"), come anche di compiere una *lettura globale* dei dati (restando sullo stesso esempio, si potrebbe affermare che, complessivamente, la *scelta universitaria* è influenzata dal *genere*).

Oltre ad inserire le frequenze di associazione, in una tabella di contingenza (se i casi osservati sono almeno 100) vanno riportate le *percentuali*; queste ultime si possono calcolare per *colonna*, per *riga* o *sul totale dei casi* (*% di cella*). Le percentuali di colonna normalizzano i dati rispetto ai marginali di colonna ed abbattono le differenze tra le frequenze delle modalità inserita in colonna; in tal caso, la lettura dei dati si attua valutando gli scarti percentuali di riga, utilizzando quali termini di confronto, i valori posti nella distribuzione marginale di riga, equivalente, come sopra accennato, alla distribuzione di frequenza della variabile Y. Le percentuali di riga normalizzano i dati rispetto ai marginali di riga; quelle di cella (la cui somma è pari a 100%) sono calcolate sul totale dei casi (nell'esempio riportato, 13.473) e permettono di vagliare l'intero sistema delle distribuzioni condizionate (cfr. Tab. 14.4.[5]).

[3] Il numero delle combinazioni possibili equivale al *prodotto* tra *numero delle modalità di X* e *numero delle modalità di Y*.

[4] Con un effetto di "frantumazione" dei dati, che, nei casi più estremi si traduce nell'illeggibilità della tabella e nell'impossibilità stessa di individuare una qualche tendenza a livello bivariato. Si pensi, a titolo esemplificativo, a variabili, nella loro foggia originaria e grezza, come la *provincia italiana di residenza*, traducibile in una nuova variabile più compatta come l'*area geografica di provenienza*, o il *numero di figli*, prima articolata in valori progressivi da 0 in su, poi semplificata in poche modalità ("nessuno", "uno", "più di uno") ai fini dell'analisi incrociata dei dati.

[5] Questo esempio e i seguenti (per tabelle e grafici riferiti ad elaborazioni su dati individuali) sono estratti a partire dal dataset dell'indagine *La vita ai tempi del Coronvirus*, svolta durante il primo lockdown

Tab 14.4. – Uso di Facebook in base alle Classi d'età (frequenze, % di colonna, % di riga, % di cella)

		Meno di 25	25-34	35-54	55-64	65 e oltre	*Totale*
No	*Freq.*	1.438	372	1.162	552	217	3.741
	% colonna	53,9	15,7	20,9	27,1	25,7	27,8
	% riga	38,4	9,9	31,1	14,8	5,8	100,0
	% sul totale	10,7	2,8	8,6	4,1	1,6	27,8
Sì	*Freq.*	1.232	2.001	4.390	1.482	627	9.732
	% colonna	46,1	84,3	79,1	72,9	74,3	72,2
	% riga	12,7	20,6	45,1	15,2	6,4	100,0
	% sul totale	9,1	14,9	32,5	11,0	4,7	72,2
Totale	*Freq.*	2.670	2.373	5.552	2.034	844	13.473
	% colonna	100,0	100,0	100,0	100,0	100,0	100,0
	% riga	19,8	17,6	41,2	15,1%	6,3	100,0
	% sul totale	19,8	17,6	41,2	15,1	6,3	100,0

Normalmente, valorizzando un *principio di parsimonia*, si seleziona un tipo specifico di percentuale da esibire in tabella, esplicitando, peraltro, in tal modo il *verso della* sua *lettura*; ad esempio, si opta per la percentuale di colonna quando si intende analizzare l'influenza della variabile inserita in colonna (in ipotesi *indipendente*) rispetto a quella posta in riga. Possiamo definire *efficiente*, *completa* e *adeguata* la presentazione tabellare prescelta laddove essa 1. sia adeguatamente numerata e contempli un'intestazione chiara e, per quanto sintetica, completa di tutti gli elementi utili al lettore per la sua comprensione (ad es., l'indicazione della fonte, se esterna, il tipo di percentualizzazione impiegato, il numero di casi mancanti, ecc.); 2. riporti i totali in percentuale al fine di cogliere la direzione della lettura; 3. contempli (magari in parentesi) la base, in valore assoluto, su cui sono state calcolate le percentuali, oltre che il totale complessivo dei casi di riferimento (in tal modo, evitando scorretti camuffamenti – il fruitore di una qualunque statistica deve sapere se il report riguarda poche decine, centi-

nazionale, nella primavera 2020 (cfr. Lombardo e Mauceri, a c. di, 2020, scaricabile gratuitamente al seguente link: https://ojs.francoangeli.it/_omp/index.php/oa/catalog/book/566).

naia o migliaia di casi! –, pur non riportando, per esigenze di sintesi, le singole frequenze di associazione, sarà comunque possibile ed agevole per il lettore interessato calcolarle); 4. esibisca: percentuali che *quadrano* (i diversi totali devono corrispondere a 100); un numero di decimali, mai eccessivo per semplicità, rispondente ad un principio di sensatezza; arrotondamenti compiuti nel rispetto di noti principi (compresa l'esigenza di far quadrare i totali a 100 – cfr. Corbetta, 1999).

Gli esempi riportati di seguito (cfr. Tabb. 14.5. - 14.11.) costituiscono degli utili spunti per il lettore, anche in virtù della loro varietà.

Nella Tab. 14.5. la variabile *frequenza con cui si cucina da quando sono in vigore le restrizioni governative* è utilizzata nella sua vesta analitica originaria (riflette perfettamente la domanda riportata nel questionario) ed è analizzata in base al *genere*.

Tab 14.5. – Frequenza con cui si cucina da quando sono in vigore le restrizioni governative in base al Genere (% di colonna)

	Uomini	Donne	*Totale*
È aumentata	48,2	**65,6**	61,3
È rimasta invariata	**33,9**	26,6	28,4
Si è ridotta	4,3	3,8	3,9
Si è interrotta	1,1	0,8	0,9
Non ho mai svolto questa attività	**12,5**	3,2	5,5
Totale	100,0 (3.400)	100,0 (10.073)	100,0 (13.473)

La distribuzione della Y (come, peraltro, quella della X) si presenta fortemente sbilanciata (le modalità "si è ridotta" e "si è interrotta" presentano percentuali risibili), per quanto sia possibile osservare delle interessanti associazioni locali. Focalizzando l'attenzione sulle distribuzioni marginali di colonna è possibile cogliere come sia per gli uomini che per le donne il valore modale sia "è aumentata"; tuttavia, per le donne tale pratica è divenuta particolarmente assidua durante il lockdown (si tratta del 65,6% delle intervistate a fronte del 48,2% degli intervistati – a livello monovariato, tale azione si riferisce al 61,3% dei soggetti complessivamente raggiunti dall'indagine e ciò permette di comprendere che la percentuale connessa agli uomini è anche sensibilmente inferiore al dato campionario), mentre il genere maschile risulta associato alle modalità "è rimasta invariata" e "non ho mai svolto questa attività" (scarti percentuali significativi, ≥ 5%, rispetto agli analoghi valori riportati in corrispondenza del campione femminile).

Lo stesso incrocio è presentato di seguito (cfr. Tab. 14.6.) utilizzando la variabile *frequenza con cui si cucina* in una sua veste fortemente ridotta in seguito alla procedura di *ricodifica* (cfr. Cap. 13). In tal caso, alla modalità "è aumentata" è contrapposto il blocco, che accorpa il resto delle originarie modalità analitiche, "non è aumentata". Le differenze di genere, in tal caso, emergono in modo ancora più schiacciante, per quanto la veste sintetica dei dati non dia più, inevitabilmente, conto di alcune informazioni puntuali riportate nella tabella precedente.

Tab 14.6. – Frequenza con cui si cucina (versione ridotta) da quando sono in vigore le restrizioni governative in base al Genere (% di colonna)

	Uomini	Donne	*Totale*
È aumentata	48,2	**65,6**	61,3
Non è aumentata	**51,8**	34,4	38,7
Totale	100,0 (3.400)	100,0 (10.073)	100,0 (13.473)

Nella tabella 14.7. una tipologia (che, nella sua veste finale, corrisponde ad una variabile nominale) ottenuta in seguito all'applicazione combinata dell'Analisi delle Corrispondenze Multiple e della Cluster Analysis (cfr. Tab. 14.7.), riferibile alle *rappresentazioni del futuro in fase pandemica*, risulta incrociata con le *classi d'età*. La tipologia si articola in 3 profili individuali distinti, di cui è bene dar conto, data la natura ultra-sintetica della variabile:

1. gli *Inclusi*: sono soggetti residenti al Sud Italia e in contesti solo sfiorati dalla pandemia; pur consapevoli della crisi italiana e internazionale in atto e dell'acuirsi dei problemi occupazionali a livello globale, guardano positivamente al futuro e giudicano con fiducia gli interventi a livello politico e medico-scientifico. "Un radicato senso di sicurezza", una "collocazione stabile e garantita nel settore pubblico", una "solida e preziosa rete sociale e familiare", "un forte bagaglio culturale e informativo" rappresentano le caratteristiche su cui si innestano le prospettive future di questo gruppo.

2. Gli *Esclusi*: sono residenti al Nord Italia e in contesti ad alta diffusione del virus; risultano essere caratterizzati da una totale sfiducia verso il futuro. I componenti di questo gruppo, convinti che le prospettive per sé e per la propria famiglia siano negative su tutti i fronti, ritengono che la fine dell'emergenza sia lontana e che la pandemia inasprirà le diseguaglianze sociali preesistenti. Si tratta in particolare di soggetti in condizione di "precarietà lavorativa", ma anche di "lavoratori autonomi" o "dipendenti del settore privato", il cui livello di istruzione è tendenzialmente basso e le cui opportunità culturali sono pressoché modeste.

3. I *Sospesi*, disorientati e incapaci di fare qualunque previsione (se non quella dell'irrisolvibilità nel breve periodo dell'emergenza Covid-19), sono calati in uno stato di insicurezza acuta, che non lascia spazio ad alcuna percezione di un avvenire. Sono prevalentemente "donne", del "Sud Italia", soggetti "non occupati" o "precari", "inseriti nel nucleo familiare d'origine", con "basso livello di status socioculturale".

La tabella 14.7. evidenzia che tra gli *Inclusi* spicchino i soggetti in età avanzata, tra gli *Esclusi* quelli in età adulta, tra i *Sospesi* i giovani.

Tab 14.7. – Profili individuali in base alla rappresentazione del futuro e Classi d'età (%)

	Meno di 25 anni	24-34	35-54	55-64	65 e oltre	*Totale*
Inclusi	33,6	32,3	34,8	**41,8**	**44,0**	36,4
Sospesi	**26,5**	18,5	16,1	17,2	17,1	18,3
Esclusi	39,9	**49,2**	**49,1**	41,0	38,9	45,3
Totale	100,0 (2.020)	100,0 (2.156)	100,0 (5.659)	100,0 (2.156)	100,0 (1.482)	100,0 (13.473)

La tabella 14.8. riporta l'incrocio tra una tipologia ottenuta attraverso l'applicazione della procedura di costruzione/riduzione dello spazio di attributi[6], *Preoccupazione per gli effetti diretti e indiretti del Covid-19*, e il *Genere*. Le informazioni originarie riassunte da tale variabile sintetica sono state rilevate attraverso una batteria di domande (10 item e una comune scala di preoccupazione, con punteggi da 0 = nessuna preoccupazione a 5 = massima preoccupazione), volta a rilevare l'intensità della preoccupazione degli intervistati rispetto agli effetti del Covid-19 e il tipo di preoccupazione (per sé stessi – ad es. nel caso di *contrazione personale e con sintomi del Coronavirus* e per gli altri – ad es. nel caso del *contagio di un familiare stretto*). Per quanto allineati sulle posizioni intermedie, uomini e donne appaiono associati a modalità dell'indice tipologico agli antipodi: gli uomini sono in generale meno preoccupati delle donne, che, al contrario, risultano essere decisamente preoccupate sia per sé stesse che per gli altri.

Tab 14.8. – Preoccupazione per gli effetti diretti e indiretti del Covid-19 (Indice tipologico) in base al Genere (% di colonna)

	Uomini	Donne	*Totale*

[6] A partire dalla combinazione di due indici additivi parziali, uno centrato sulla dimensione della preoccupazione per sé stessi, l'altro sulla preoccupazione per gli altri, semplificati e dicotomizzati entrambi nelle modalità "medio-alta preoccupazione" e "bassa preoccupazione".

Poco preoccupati per sé e per gli altri	**30,4**	20,4	23,0
Molto preoccupati solo per sé	7,6	7,5	7,5
Molto preoccupati solo per gli altri	18,8	14,9	15,9
Molto preoccupati per sé e per gli altri	43,2	**57,2**	53,6
Totale	100,0 (3.085)	100,0 (8.656)	100,0 (11.741)

Valori mancanti: 12,9%

La tabella 14.9. riporta l'incrocio tra un indice additivo semplificato di *Valutazione della comunicazione pubblica in tema di Covid-19 da parte delle istituzioni sanitarie* e le *Classi d'età*.

Tab 14.9. – Valutazione della comunicazione pubblica sul Covid-19 da parte delle istituzioni sanitarie in base alle Classi d'età (%)

	Meno di 25 anni	24-34	35-54	55-64	65 e oltre	*Totale*
Poco affidabile (*valori inferiori alla media*)	10,7	15,6	22,6	22,0	**27,2**	19,2
Molto affidabile (*valori uguali o superiori alla media*)	**89,3**	84,4	77,4	78,0	72,8	80,8
Totale	100,0 (2.353)	100,0 (2.173)	100,0 (5.103)	100,0 (1.863)	100,0 (753)	100,0 (12.245)

Valori mancanti: 9,1%

L'indice (Cfr. Cap. 13), costruito sommando per tutti i soggetti intervistati il punteggio di scala da 0 a 5 (0 = "per nulla affidabile" e 5 = "del tutto affidabile") su 4 item (affidabilità accordata a: *Ministero della Salute*, *Istituto Superiore di Sanità*, *Ordine dei Medici*, *Organizzazione Mondiale della Sanità*) degli 11 inseriti in una batteria di domande (in cui trovano posto istituzioni di diverso tipo che si sono pubblicamente espresse sul Covid-19 in fase emergenziale), si articola in due sole modalità: "poco affidabile" (valori inferiori alla media sull'indice nella sua veste quasi-cardinale originaria) e "molto affidabile" (valori uguali o superiori alla media[7]). Come si può osservare, in uno scenario in cui per tutti i target d'età risulta essere preponderante un'elevata fiducia nei confronti delle istituzioni sanitarie deputate alla comunicazione sul Covid-19, coloro che spiccano per la maggiore affidabilità accordata sono i giovani; agli antipodi figurano gli anziani, con la percentuale di sfiducia più elevata nei sub-campioni messi a confronto.

[7] Molto elevata nel campione in analisi.

Nella tabella 14.10 il sistema di 21.143 risposte collezionate attraverso una domanda a risposta multipla del questionario (volta ad esplorare le *Attività svolte sui Social Network in fase pandemica*), variamente allocate rispetto alle modalità riportate nel prospetto, sono lette in base al *Genere*. Rapportando il totale delle risposte al totale dei casi validi (13.173 dei 13.473) è possibile quantificare il numero medio di attività svolte ad opera del campione nel periodo delle restrizioni governative (1,6).

Tab 14.10. – Analisi delle risposte multiple: Attività svolte sui Social Network nel periodo delle restrizioni governative in base al Genere (% di colonna)

	Uomini	Donne	*Totale*
Ho partecipato ad un video-aperitivo	15,5	15,5	15,5
Ho partecipato ad un flashmob	3,6	4,6	4,3
Ho utilizzato l'hashtag #iorestoacasa e/o #andratuttobene per la condivisione di stati/ foto/video	11,3	15,6	14,7
Ho raccontato me stesso, le mie emozioni ed opinioni	9,7	11,3	10,9
Ho promosso il mio lavoro (creazione di pagine e profili, attività di posting, etc.)	8,3	6,5	6,9
Ho conosciuto persone nuove, instaurato nuove relazioni	7,1	4,9	5,4
Ho condiviso news sullo stato di emergenza da Covid-19	22,4	22,4	22,4
Nessuna di queste attività	22,1	19,2	19,9
Totale	100,0 (5.177)	100,0 (15.966)	100,0 (21.143)

Casi validi: 13.173

Come è possibile osservare, le piccole differenze tra uomini e donne riscontrabili con riferimento a ciascuna attività considerata non superano in alcun caso la soglia statistica del 5%; pertanto, possiamo concludere che esse siano dovute al caso e che il genere non abbia alcuna significativa influenza sulle modalità di utilizzo delle piattaforme social in pandemia.

Un ultimo esempio, reso attraverso la predisposizione di una tavola compatta, ha a che fare con una batteria di domande focalizzata sull'*Adozione di comportamenti preventivi durante la fase più acuta della pandemia*. L'organizzazione dei dati in tabella prevede l'indicazione, per ciascun item, della percentuale legata alla risposta affermativa (quella connessa con la risposta negativa è agevolmente calcolabile per differenza). Ogni variabile risulta essere incrociata col *Genere* e le differenze significative sono evidenziate in grassetto.

Tab 14.11. – Adozione di comportamenti preventivi (9 item) durante la fase più acuta della pandemia in base al Genere (% di colonna)

	Uomini	Donne	*Totale*
Indossare la mascherina quando esco	78,9	80,7	80,3 (10.814)
Detergere le mani con il gel antibatterico	82,1	**87,0**	85,8 (11.560)
Disinfettare oggetti e superfici lavabili	80,5	**90,0**	87,6 (11.802)
Usare guanti monouso quando esco	66,7	**73,9**	72,1 (9.714)
Lavarsi le mani dopo essere rientrati a casa	88,0	86,7	87,0 (11.728)
Mettere a lavare i vestiti dopo essere rientrati a casa	56,2	**64,6**	62,5 (8.415)
Togliersi le scarpe subito dopo essere rientrati a casa	77,3	81,2	80,2 (10.811)
Rispettare il metro di distanza tra me e gli altri quando esco	85,1	84,1	84,3 (11.363)
Evitare di prendere i mezzi pubblici	94,5	96,0	95,6 (12.885)

Casi validi (su tutti gli item): 13.473

Tra le alternative possibili, il *grafico a barre raggruppate* costituisce un'efficace rappresentazione di incroci bivariati tra variabili categoriali. Sono stati ripresi due degli esempi precedenti di cui è stata già commentata la relativa tabella di contingenza per poter operare un utile confronto (cfr. Figg. 14.1. e 14.2.).

Fig. 14.1. – Frequenza con cui si cucina da quando sono in vigore le restrizioni governative in base al genere (%)

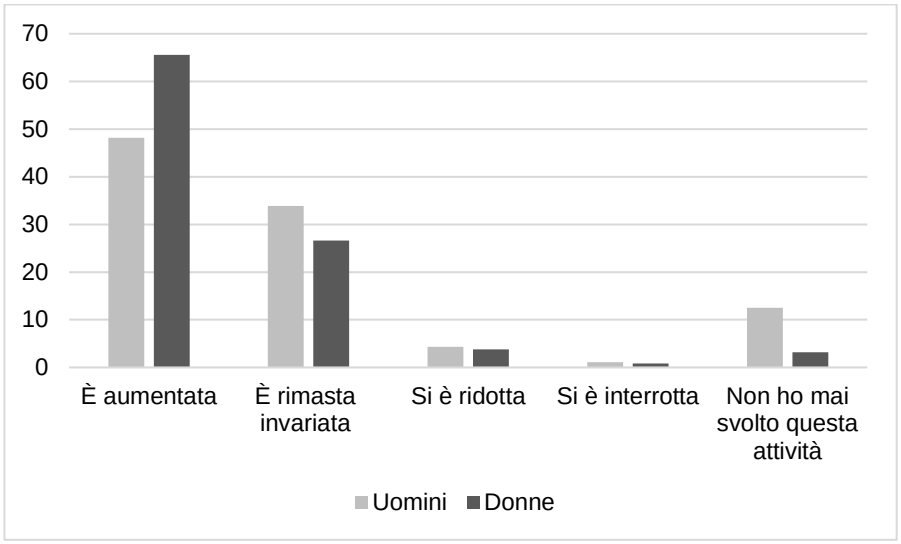
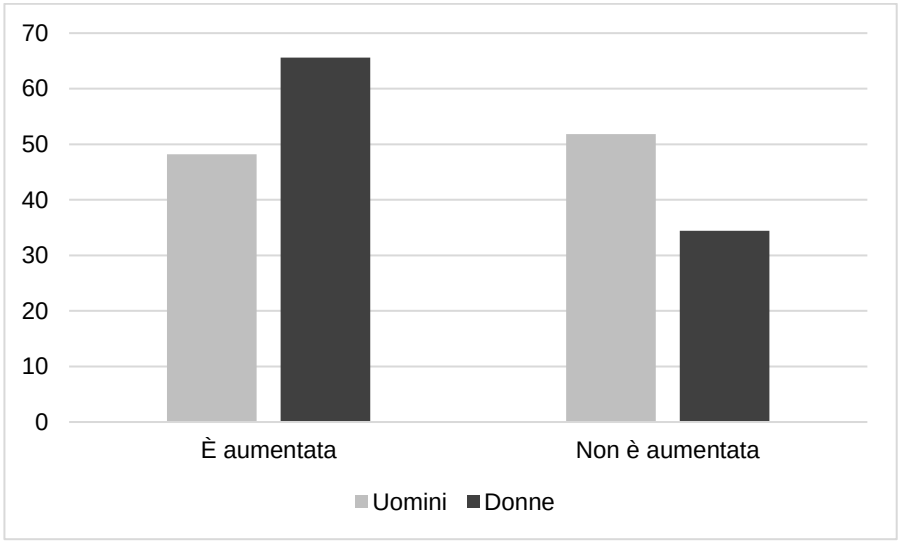


Fig. 14.2. – Frequenza con cui si cucina da quando sono in vigore le restrizioni governative (versione ridotta) in base al Genere (%)



Filtrando i dati in base alla variabile utilizzata come chiave di lettura (il *Genere*), quindi costruendo un grafico (*a barre semplici* o *a torta*) per il subcampione delle donne ed uno per il subcampione degli uomini, si otterrebbe un risultato affine, efficace laddove i due grafici, posti l'uno accanto all'altro, si prestino ad immediati confronti (per esplorare la vasta gamma di possibilità grafiche e per approfondimenti sugli accorgimenti da adottare e sugli errori da evitare quando si progettano i grafici, cfr. Delli Zotti, 2010; Cairo, 2020; Jones, 2020).

La costruzione di tabelle di contingenza si rivela un utile strumento anche per lo studio della relazione tra due variabili con categorie ordinate[8] (fr. Tab. 14.12.). Laddove la *tabella* sia *quadrata*, come nell'esempio riportato, ovvero nel caso in cui le due variabili presentino lo stesso numero di modalità, è possibile individuare *due diagonali*: 1. quella della *cograduazione* (celle evidenziate in grigio chiaro), che procede in senso discendente dalla prima cella in alto a sinistra (combinazione "zona a bassa diffusione del contagio" e "sistema sanitario nazionale giudicato inefficace") all'ultima cella in basso a destra (combinazione "zona ad alta diffusione del contagio" e "sistema sanitario nazionale giudicato del tutto efficace") e 2. quella della *contro-graduazione*, che segue l'andamento opposto (celle evidenziate in grigio scuro).

Tab 14.12. – Giudizio sul sistema sanitario nazionale in quanto a capacità di far fronte all'emergenza pandemica (primo lockdown) in base alla Zona di residenza classificata per tasso di diffusione del contagio (% di colonna)

	Bassa	Media	Alta	*Totale*
Inefficace	22,1	25,1	27,5	24,4
Parzialmente efficace	5,2	4,0	4,0	4,6
Del tutto efficace	72,7	70,9	68,5	71,0
Totale	100,0 (2.020)	100,0 (2.156)	100,0 (5.659)	100,0 (13.473)

Le celle non evidenziate in grigio, compresa quella nel riquadro al centro (in comune tra le due diagonali), si riferiscono alle *coppie legate*. Parliamo di cograduazione (*relazione diretta* – 4.335 casi nell'esempio selezionato) o contro-graduazione (*relazione inversa* – 5.926 casi) se le frequenze della tabella di contingenza si concentrano lungo la diagonale corrispondente e le frequenze delle coppie legate sono della minima entità possibile[9] (3.213). Nel caso presentato, per quanto di lieve entità (e nonostante il carattere residuale della modalità intermedia della Y), si può parlare di relazione inversa tra le due variabili[10].

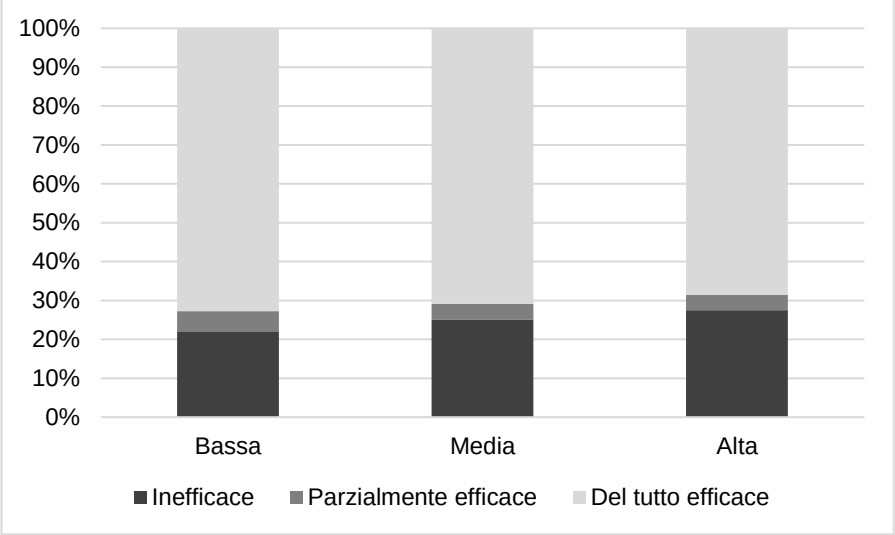
[8] Per approfondimenti sui *rapporti di probabilità* e sull'*analisi bivariata* condotta a partire da *due dicotomie*, cfr. Marradi, 1997; Corbetta, 1999; Di Franco, 2001.

[9] Più precisamente, quando le frequenze non si addensano in alcuna delle due diagonali, il risultato dell'analisi è l'assenza di relazione tra le due variabili o la presenza di associazioni locali che svelano una relazione non monotonica fra le due variabili.

[10] I coefficienti di associazione (finalizzati a valutare la *forza della relazione bivariata*) per coppie di variabili nominali e ordinali (entrambe nominali o una nominale e una ordinale – *coefficiente di contingenza*, *phi* e *v* di Cramèr, *lambda*, *q* di Yule, *coefficiente di incertezza*), non sono particolarmente utilizzati nella ricerca sociale (per approfondimenti, cfr. Marradi, 1997; Corbetta, 1999; Di Franco, 2001; Di Franco e Marradi, 2020). Ciò soprattutto in ragione della completa ispezionabilità di una tabella di contingenza

Il *diagramma a barre sovrapposte* rappresenta un'efficace resa grafica di quanto visto in precedenza nella tabella 14.12[11].

Fig. 14.3. – Giudizio sul sistema sanitario nazionale in quanto a capacità di far fronte all'emergenza pandemica (primo lockdown) in base alla Zona di residenza classificata per tasso di diffusione del contagio (%)



In seguito alla predisposizione ed accurata lettura di tabelle di contingenza e grafici, per accertare l'*esistenza della relazione* tra due variabili categoriali si può ricorrere al calcolo del *coefficiente di significatività statistica* del *chi-quadrato*. Il valore del chi-quadrato si interpreta come una misura della distanza tra la tabella di contingenza osservata (in cui sono riportate le *frequenze empiriche* o *osservate*, f_o) ed un'altra tabella costruita in base all'assunto di indipendenza tra le due variabili (in cui sono riportate le *frequenze teoriche* o *attese*, f_e). Ad ogni singola frequenza di associazione osservata, corrisponderà una frequenza attesa equivalente al prodotto dei due marginali corrispondenti, diviso il totale dei casi (N). Quanto più è elevato il valore del chi-quadrato, tanto più è bassa la probabilità di commettere un errore nel rifiutare l'ipotesi di indipendenza statistica tra le due variabili nella popolazione di riferimento. Se frequenze

(grazie al numero contenuto di celle e all'autonomia semantica delle modalità delle variabili incrociate), attraverso la quale il ricercatore scopre non di rado come, a fronte di un'associazione statistica significativa (che rappresenta una misura sintetica della relazione tra due variabili), emergano una o poche associazioni locali rilevanti (quindi tra specifiche coppie di modalità tra le varie osservabili).

[11] Per approfondimenti sui coefficienti bidirezionali e unidirezionali di associazione idonei per le variabili ordinali (*gamma* di Goodman e Kruskal, *d* di Somers, *tau-b* e *tau-c* di Kendall, tutti con campo di variazione tra -1 e +1), cfr. Marradi, 1997; Corbetta, 1999; Di Franco, 2001; Di Franco e Marradi, 2020.

empiriche ed attese coincidono, tra le due variabili considerate vi è indipendenza statistica e il chi-quadrato assume valore 0 (esso non ha, tuttavia, un limite superiore e può assumere qualsiasi valore numerico positivo diverso da 0).

La formula del chi-quadrato è la seguente:

$$X^2 = \sum \frac{(fo-fe)2}{fe}$$

Proviamo con un esempio, già visto nelle pagine precedenti, a chiarire i passaggi necessari per il suo calcolo:

Tab 14.13. – Tabella di contingenza osservata: Frequenza con cui si cucina (versione ridotta) da quando sono in vigore le restrizioni governative in base al Genere (f_o)

	Uomini	Donne	*Totale*
È aumentata	1.640	6.604	8.244
Non è aumentata	1.760	3.469	5.229
Totale	3.400	10.073	13.473

Tab 14.14. – Tabella di contingenza teorica (impostazione del calcolo): Frequenza con cui si cucina (versione ridotta) da quando sono in vigore le restrizioni governative in base al Genere (f_e)

	Uomini	Donne	*Totale*
È aumentata	(3.400x8.244)/13.473	(10.073x8.244)/13.473	8.244
Non è aumentata	(3.400x5.229)/13.473	(10.073x5.229)/13.473	5.229
Totale	3.400	10.073	13.473

Tab 14.15. – Tabella di contingenza teorica (frequenze attese): Frequenza con cui si cucina (versione ridotta) da quando sono in vigore le restrizioni governative in base al Genere (f_e)

	Uomini	Donne	*Totale*
È aumentata	2.080,4	6.163,6	8.244
Non è aumentata	1.319,6	3.909,4	5.229

Totale	3.400	10.073	13.473

Tab 14.16. – Tabella delle contingenze (f_o-f_e): Frequenza con cui si cucina (versione ridotta) da quando sono in vigore le restrizioni governative in base al Genere (es. 1.640 - 2.080,4 = -440,4*)*

	Uomini	Donne
È aumentata	-440,4	440,4
Non è aumentata	440,4	-440,4

Abbiamo tutti gli elementi per il calcolo del chi-quadrato:

$X^2 = (-440,4)^2/2.080,4 + (440,4)^2/6.163,6 + (440,4)^2/1.319,6 + (-440,4)^2/3.909,4 =$
$= 93,2 + 31,5 + 147 + 49,6 = 321,3$

Per convenzione, in fase di analisi dei dati, viene respinta l'ipotesi nulla di indipendenza statistica tra le variabili se il chi-quadrato è così elevato da avere una *probabilità di errore*[12] del 5% o meno ($p \leq .05$ – *soglia statistica*) ed una del 95% o più che la relazione riscontrata tra le variabili sia presente anche nella popolazione da cui deriva il campione a partire dal quale vengono prodotti i risultati dell'indagine.

È bene precisare che, a rigore, laddove si analizzino i dati dell'intera popolazione o il campione non sia stato estratto casualmente, è inappropriato calcolare il chi-quadrato (esso rappresenta un test dell'*inferenza statistica* e si rivela di grande utilità nel caso si sia nelle condizioni di generalizzare alla popolazione di riferimento il risultato ottenuto su base campionaria); difatti, solo in presenza di campioni probabilistici, è possibile utilizzare la *tavola della distribuzione teorica del chi-quadrato*[13], in cui è indicato il *livello di fiducia* con cui si può respingere l'ipotesi di indipendenza tra le due variabili[14].

[12] Per probabilità di errore si intende la probabilità che le differenze riscontrate empiricamente siano dovute al caso.

[13] Nella tavola suddetta troviamo tante righe (o distribuzioni del X^2) quanti sono i gradi di libertà della tabella, ovvero il numero di valori *liberi di variare*. I gradi di libertà di una tabella di contingenza si calcolano nel modo seguente: (n. righe -1) (n. colonne – 1), dove righe e colonne corrispondono al numero di modalità di ciascuna variabile inclusa nell'incrocio bivariato. Una tabella di contingenza 2 x 2, come nell'esempio sopra riportato, ha un grado di libertà: se, difatti, si fissa la frequenza di una cella, anche le frequenze delle altre celle risultato automaticamente fissate (marginali vincolanti).

[14] Tra i limiti del chi-quadrato se ne menzionano due: a. è inaffidabile in caso di frequenze attese per una o più celle inferiori a 5; b. cresce all'aumentare del numero dei casi, pertanto, in caso di campioni consistenti risulta essere di norma significativo. Per approfondimenti, cfr. Marradi, 1997; Corbetta, 1999; Di Franco, 2001; Lucchini, 2018; Bocci e Mingo, 2020; Di Franco e Marradi, 2020.

Nel paragrafo appena chiuso si è stabilito di estrapolare gli esempi da un unico dataset, al fine di andare incontro al lettore nella comprensione dei temi trattati e dei passaggi esplicitati. Cionondimeno, con la cassetta degli attrezzi allestita, si invita il lettore stesso a "trasferire" e mettere in pratica le competenze acquisite in altre occasioni di ricerca, rispetto a temi diversi e ad altre unità di analisi[15].

14.3. L'analisi bivariata con variabili di tipo misto: cenni all'analisi della varianza

Quando l'obiettivo è quello di incrociare variabili di tipo misto, al fine di comprendere se esista e che intensità abbia la relazione tra queste ultime, si può fare ricorso all'*analisi della varianza*[16] (ANOVA – *Analysis of Variance*), un modello di analisi dei dati attraverso cui verificare ipotesi relative a differenze tra le medie di due o più popolazioni, campioni, sub-campioni rispetto a determinati fenomeni[17]. Nella sua forma più semplice[18] (il livello bivariato di analisi), sono coinvolte nell'analisi una *variabile categoriale* (attraverso le cui modalità individuare classi/gruppi dal valore strategico entro

[15] Si pensi, a titolo esemplificativo, ad una matrice in cui i casi osservati siano rappresentati da istituti secondari di secondo grado operanti sul territorio italiano e rispetto a cui si sia intenzionati ad incrociare le variabili *entità della dotazione tecnologica disponibile*, articolata in 3 classi, e *competenze digitali del personale docente*, graduata su cinque livelli; o ancora, si immagini un dataset in cui siano archiviati tutti i post pubblicati su Facebook da parte delle principali forze politiche italiane nell'occasione di una specifica campagna elettorale e rispetto a cui si intenda indagare l'*entità dell'attività di posting*, graduata su tre livelli, in base alla *settimana di campagna di riferimento*.

[16] È noto l'impiego dell'analisi della varianza in disegni di ricerca di taglio sperimentale (Cfr. Cap. 9). Si pensi ad un'occasione d'indagine in cui si proceda a rilevare con opportune tecniche di *scaling*, prima e dopo l'esposizione ad una campagna di sensibilizzazione contro il fumo, il grado di competenza acquisita sui rischi per la salute e/o l'atteggiamento verso il fumo da parte dei soggetti intervistati. Ci si concentri, inoltre, sull'obiettivo di valutare gli effetti di tale *trattamento* rispettivamente su gruppo sperimentale e di controllo, segmenti individuabili e confrontabili a partire da una variabile dicotomica disponibile in matrice che distingua gli intervistati a seconda che abbiano o meno partecipato alla campagna suddetta.

[17] Per approfondimenti e dettagli sui passaggi matematici e gli *assunti* dell'analisi della varianza, cfr. Barbaranelli, 2007.

[18] Per approfondimenti su modelli più complessi dell'analisi della varianza − ad es., i *disegni fattoriali* (che prevedono due o più variabili indipendenti ed in cui si procede con l'analisi e il controllo degli *effetti*

il dataset – si pensi, a titolo esemplificativo, alla suddivisione di un campione di intervistati per classi d'età o all'articolazione in aree geografiche delle province italiane) – in ipotesi *indipendente* – ed una *cardinale* – in ipotesi *dipendente*.

L'analisi, concretamente, consiste nell'esaminare la varianza di ciascun *gruppo* rispetto alla *media del gruppo* stesso (*interna*) e, comparativamente, la varianza fra i gruppi rispetto alla *media del campione* complessivo (*esterna*). Una volta impostati gli incroci tra variabili ritenuti interessanti ai fini della ricerca, scopo dell'analisi è quello di misurare la significatività statistica della differenza fra le medie delle categorie.

Il calcolo di *media aritmetica* e *varianza* (affrontati dettagliatamente nel Cap. 12 – qui, in particolare, il riferimento è alla *devianza*, numeratore della varianza) si rivela essenziale nel caso si proceda con l'applicazione di tale modello procedurale:

$$\bar{X} = \frac{X_1 n_1 + X_2 n_2 + X_{\dots} n_{\dots} + X_k n_k}{N} = \frac{\sum_{i=1}^{k} X_i n_i}{N}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{N} = \frac{\sum_{k=1}^{n} (X_k - \bar{X})^2 \, n_k}{N}$$

L'ANOVA si fonda sulla scomposizione della variabilità totale dei dati in analisi (*devianza totale*, corrispondente alla somma dei quadrati degli scarti tra i singoli punteggi e la media generale) in *due fonti di variazione*:

- una dovuta alla *differenza tra i gruppi* (la *varianza between*, *spiegata* o *esterna*, corrispondente alla somma dei quadrati degli scarti tra i punteggi medi di gruppo e la media generale), ovvero tra segmenti di casi (individui, aggregati territoriali, ecc.), individuabili a partire dalle diverse modalità della X. Quest'ultima, denominata nel gergo tecnico *fattore*, come accennato, riflette l'effetto del trattamento sperimentale o di un raggruppamento preesistente, disponibile, sotto forma di variabile categoriale, in matrice;
- una dovuta alla *diversità dei casi entro i gruppi* (la *varianza within*, *intra-gruppo* o *residua*, corrispondente alla somma dei quadrati degli scarti tra i punteggi di ogni caso e la relativa media di gruppo), ovvero alle differenze tra un caso e l'altro, naturalmente riscontrabili e casualmente distribuite anche all'interno di un gruppo tendenzialmente omogeneo.

In estrema sintesi, la devianza totale risulta essere scomposta in una parte dovuta alla deviazione delle medie di ogni gruppo dalla media generale e in un'altra dovuta alla deviazione dei punteggi dei casi dalla media del gruppo d'appartenenza. Se la *variabilità* della variabile dipendente è *minima entro* le categorie in cui si articola la

principali e degli *effetti di interazione* tra le variabili indipendenti) o *multivariati* (che includono due o più variabili dipendenti), cfr. Barbaranelli, 2007.

variabile indipendente e, al contempo, *massima fra* tali categorie, possiamo sostenere l'esistenza di una relazione fra le due variabili considerate. Quale *misura di significatività statistica*, volta ad esaminare l'esistenza di una differenza apprezzabile tra variabilità tra gruppi ed entro i gruppi, viene utilizzato il *rapporto F di Fisher*. Attraverso *F* vengono vagliate le seguenti ipotesi: 1. H_0 (ipotesi nulla di indipedenza tra le variabili): i gruppi individuabili a partire dalle modalità di X hanno medie uguali sulla variabile dipendente; 2. H_1: almeno due gruppi (solo due nel caso in cui la X sia dicotomica) presentano medie significativamente diverse tra loro. Se la X non influenza l'andamento di Y (in tal caso non si può rifiutare H_0), le devianze tra i gruppi ed entro i gruppi saranno molto simili tra loro e, conseguentemente, il rapporto F assumerà valori molto bassi (prossimi a 0). Se, al contrario, la X produce effetti (tanto da poter rifiutare H_0), la devianza tra i gruppi sarà maggiore della devianza entro i gruppi ed il rapporto F assumerà valori elevati.

Il coefficiente di associazione (che misura la forza dell'associazione) utilizzato per l'ANOVA è *eta quadrato*; in tal caso, più è alta la devianza spiegata (coincidente con quella totale nel caso di differenze nulle intra-gruppo) più è forte l'associazione tra variabile cardinale (o quasi-cardinale) e categoriale. Tale coefficiente varia tra 0 e 1 e non può assumere valori negativi (si tratta di una proporzione ed è elevato al quadrato)[19].

Si riporta di seguito un esempio per rendere più agevole la comprensione di quando esposto sinteticamente finora. L'incrocio tra variabili considerato è quello tra *suddivisione territoriale dei Paesi UE* (distinzione delle 27 nazioni UE[20] in quattro modalità/gruppi: *settentrionali*, *orientali*, *occidentali*, *meridionali*), in qualità di variabile indipendente, e *Incidenza dell'istruzione terziaria nella popolazione femminile*[21] (Fonte Eurostat, 2020; si tratta della percentuale di donne che hanno conseguito il livello terziario di educazione – formazione accademica e/o professionale avanzata – sul totale della popolazione femminile tra i 15 e i 74 anni).

[19] Su misure di significatività e associazione, controlli statistici da effettuare prima di procedere con l'ANOVA, *controlli post hoc* (utili a controllare la significatività di ogni singola differenza riscontrata tra un gruppo e l'altro nel caso la X si configuri come una variabile politomica), cfr. Barbaranelli, 2007.

[20] I casi dell'esempio selezionato sono rappresentati da aggregati territoriali (Paesi facenti parte dell'UE) e il dataset a 27 righe, entro cui trovano posto le variabili selezionate ai fini dell'incrocio bivariato, è riferito ad una popolazione statistica.

[21] Si tratta di un *rapporto di composizione*, pertanto di una misura normalizzata che si presta a confronti.

Tab 14.17. – Incidenza dell'istruzione terziaria nella popolazione femminile in base alla Suddivisione territoriale dei Paesi UE (Eurostat, 2020) – Media, N, Deviazione Standard

	Media intra-gruppo	*N*	*Deviazione std. intra-gruppo*
Paesi settentrionali	38,6143 (+)	7	3,55782
Paesi occidentali	28,8833 (=)	6	5,20208
Paesi orientali	21,8500 (-)	6	5,18180
Paesi meridionali	25,1000 (-)	8	6,87667
Totale	28,7222	27	8,20410

A partire dal ricco e articolato output ottenibile a seguito dell'applicazione di tale procedura, le tabelle e i grafici riportati contengono una selezione degli elementi considerati più significativi ed utili (nella Tab. 14.17. i simboli +, - e = semplificano al massimo i confronti tra medie nei singoli gruppi e media nella popolazione di riferimento). L'incrocio risulta significativo e l'associazione statistica tra le variabili forte (*F* elevato, *probabilità di errore* prossima a 0, *E quadrato elevato*) (cfr. Cap. 17).

Tab 14.18. – Incidenza dell'istruzione terziaria nella popolazione femminile in base alla Suddivisione territoriale dei Paesi UE (Eurostat, 2020): ANOVA

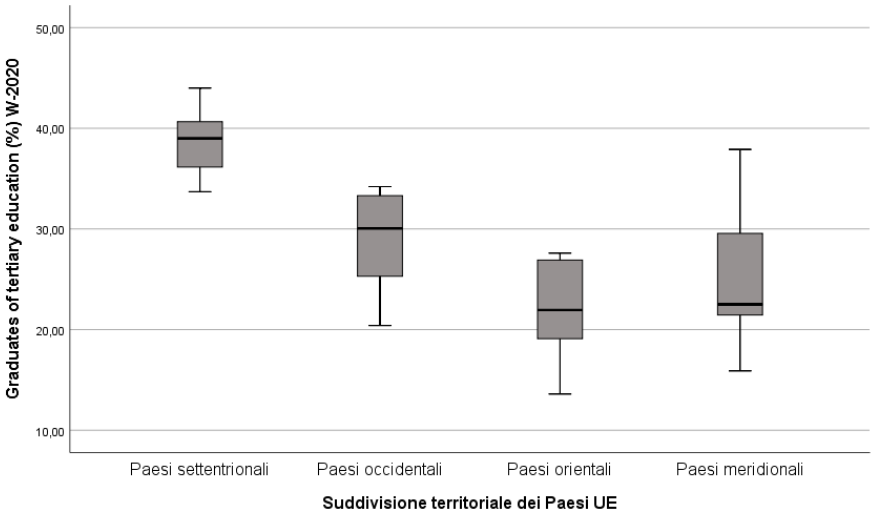
	Somma dei quadrati	*F*	*Sign. (p)*	*Eta quadrato*
Devianza tra gruppi	1073,455	12,165	,000	.613
Devianza entro i gruppi	676,532			
Devianza totale	1749,987			

Sarà interessante, nelle fasi successive di analisi, cercare di capire a quali altri fattori sia dovuta la quota di variabilità non intercettata dalla suddivisione dei Paesi UE in aree territoriali, ma anche provare ad inserire nel modello più di una variabile indipendente di cui analizzare anche gli effetti congiunti sulla Y.

Il *diagramma a scatola* (cfr. Fig. 14.4.) costituisce una rappresentazione grafica efficace di quanto più analiticamente riportato nelle tabelle (in tal caso apparato tabellare e grafico sono complementari e scegliere, eventualmente, di inserire entrambi gli elementi in un report di ricerca non comporta delle ridondanze). Ciascuna scatola (il rettangolo grigio) è delimitato dal primo e dal terzo quartile; essa appare divisa al suo interno dalla mediana (linea spessa). I segmenti in alto e in basso (i cosiddetti "baffi") indicano il valore minino e massimo osservati entro ciascun gruppo. Come è evidente,

i Paesi settentrionali sono quelli che si caratterizzano, complessivamente, per la maggiore incidenza dell'istruzione terziaria tra le donne; peraltro, il gruppo dei Paesi del Nord appare anche quello caratterizzato dalla maggiore omogeneità interna. I Paesi meridionali, pur evidenziando un valore mediano simile a quello dei Paesi orientali, sono caratterizzati da una variabilità interna particolarmente vistosa (elemento questo che varrebbe la pena di approfondire e, peraltro, perfettamente ispezionabile caso per caso in una matrice così piccola).

Fig. 14.4. – Incidenza dell'istruzione terziaria nella popolazione femminile in base alla Suddivisione territoriale dei Paesi UE (Eurostat, 2020): Diagramma a scatola



L'esempio selezionato consentirà al lettore attento di immaginare (l'output è identico e l'analista avrà cura di impostare incroci tra variabili di tipo misto secondo i criteri esplicitati all'inizio del paragrafo) l'applicazione dell'ANOVA anche su una matrice le cui righe siano rappresentate da casi individuali. Si pensi alla possibilità di analizzare il *Reddito mensile percepito in euro* in base al *Genere*; gli *item di una scala d'atteggiamento* (ad es. la *Paura di contrarre il Coronavirus con sintomi gravi* rilevata con una scala 0-5) in base alle *Classi d'età*; un *indice additivo* (ad es. la versione originale quasi-cardinale dell'*Indice di valutazione della comunicazione pubblica sul Covid-19 da parte delle istituzioni sanitarie*) in base al *Livello di istruzione*; i *fattori* estratti attraverso l'applicazione dell'Analisi delle Corrispondenze Multiple (ACM) o le *componenti* individuate attraverso la tecnica dell'Analisi in Componenti Principali (ACP) in base ad altre significative variabili categoriali presenti nel nostro dataset (cfr. Capp. 7, 13 e 16).

14.4. La relazione tra variabili cardinali e quasi-cardinali: soluzioni grafiche e coefficienti in uso

Nel caso di incroci tra variabili cardinali o quasi-cardinali, normalmente molto *sensibili*[22] e le cui modalità, rappresentate da valori numerici, hanno una scarsa *autonomia semantica*, l'analisi bivariata si attua attraverso la predisposizione di un opportuno apparato grafico e il calcolo di specifici coefficienti.

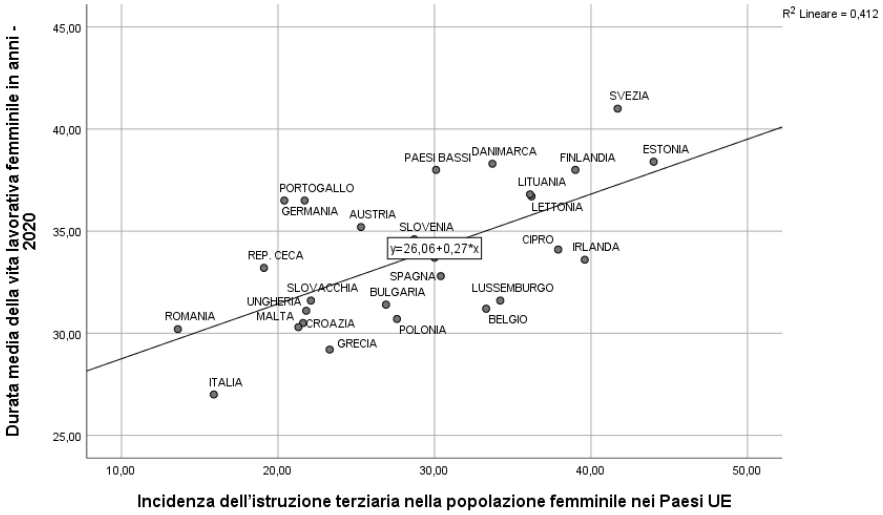
Il *grafico a dispersione* (*scatter diagram* o *scattergram*) costituisce, per variabili previamente normalizzate o standardizzate, una fruttuosa soluzione ai fini del vaglio della forma della relazione e dell'individuazione di eventuali *casi outliers* (casi devianti, molto distanti dalla media campionaria o della popolazione statistica di riferimento), la cui presenza può esercitare effetti distorsivi, anche importanti, sui coefficienti di associazione di cui sia stato effettuato il calcolo, sia facendo emergere fittiziamente una relazione altrimenti blanda o inesistente, sia, al contrario, oscurando o ridimensionando la forza di un'associazione.

Quando tra due variabili si ipotizza una relazione asimmetrica, si inserisce la variabile indipendente sull'asse delle X e quella dipendente sull'asse delle Y. Quanto più i *punti* della *nuvola* rappresentata nel grafico sono vicini tra loro assumendo un *andamento rettilineo*, tanto è maggiore l'*associazione lineare* (*diretta* se la nuvola segue un *andamento crescente*, *inversa* se essa segue un *andamento decrescente*[23]). Il grafico presentato di seguito (cfr. Fig. 15.5.) evidenzia l'esistenza di una relazione tendenzialmente lineare tra le variabili, riferite ai Paesi UE, *Incidenza dell'istruzione terziaria* e *Durata media della vita lavorativa*, entrambe riferite alla popolazione femminile residente nei contesti in analisi.

[22] La *sensibilità di una variabile*, ovvero di uno strumento di classificazione/ordinamento/conteggio/misurazione consiste nella sua capacità, in seguito alla definizione operativa accolta in matrice, di rappresentare in modo fedele la gamma degli stati possibili su una proprietà; essa corrisponde al rapporto tra numero di forme che può assumere il fenomeno in analisi e il numero di modalità presenti nel dataset.

[23] Nel caso di *assenza di relazione* tra le variabili selezionate ai fini dell'analisi, la nuvola dei punti si dispone parallelamente rispetto all'asse delle ascisse o delle ordinate.

Fig. 14.5. – Durata media della vita lavorativa femminile in anni e Incidenza dell'istruzione terziaria nella popolazione femminile nei Paesi UE (Eurostat, 2020): Diagramma a dispersione



La misura utile per vagliare la *forza* e il *verso* della relazione (diretta o inversa) tra variabili cardinali o quasi-cardinali[24] è rappresentata dal *coefficiente di correlazione lineare di Person* (r). Esso esprime l'*associazione lineare simmetrica* tra tali variabili e varia tra -1 (*relazione negativa perfetta*) e +1 (*relazione positiva perfetta*); quando assume valore 0 la relazione lineare risulta assente, anche se questo non significa necessariamente che non esista affatto una relazione tra le due variabili, la quale potrebbe assumere una forma diversa dalla retta. Il coefficiente corrisponde al rapporto tra la *codevianza*[25] (somma dei prodotti degli scarti della X e della Y dalle rispettive medie) e il prodotto degli *scarti-tipo* delle due variabili[26]; in altri termini, esso rapporta l'entità della *covariazione* (*varianza in comune*) tra le due variabili in analisi al prodotto delle rispettive quantità di variazione, espresse in unità di scarto-tipo. Sono riportate di seguito le formule per il calcolo della *covarianza* e del *coefficiente r*[27].

$$cov\ (x, y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) * (Y_i - \bar{Y})}{N}$$

[24] Idoneo anche per valutare la relazione tra due *variabili dicotomiche*.

[25] Numeratore della *covarianza*.

[26] Cfr. Cap. 12 per approfondimenti sulle *misure di variabilità* e *dispersione*.

[27] In considerazione del fatto che *r di Pearson* sovrastima la quantità di varianza comune tra le variabili prese in esame, una misura più precisa della varianza comune è rappresentata da r al quadrato (R^2), definito *coefficiente di determinazione*.

$$r = \frac{cov(x,y)}{s_x * s_y}$$

Tornando all'esempio riportato nella figura 14.5. il coefficiente di correlazione lineare è pari a +0,642: la relazione è significativa, piuttosto forte e diretta, a dimostrazione del fatto che nei Paesi UE, con un focus sulla popolazione femminile residente, al crescere dell'*Incidenza dell'istruzione terziaria* cresce anche la *Durata media della vita lavorativa* (si guardi, per curiosità, alla situazione dell'Italia, con valori così bassi su entrambi gli indicatori da risultare "fanalino di coda" entro il più ampio paniere dei Paesi UE)[28].

Anche in un dataset organizzato su *dati individuali* è utile realizzare grafici a dispersione (tenendo presente, tuttavia, che in tal caso i singoli individui proiettati sugli assi costituiscono elementi "anonimi" della nuvola dei punti, in quanto non etichettabili come nel caso di *aggregati territoriali*) e calcolare il coefficiente di correlazione (includendo negli incroci variabili cardinali e quasi-cardinali semplici e sintetizzate in indici – si pensi alla relazione tra due *Indice additivi*, uno *di valutazione*, più o meno positiva ed espressa in punteggi di scala, *della comunicazione pubblica sul Covid-19 ad opera dei media* e l'altro, *ad opera delle istituzioni politiche*; o, ancora, all'incrocio tra *Età in anni compiuti* e *Numero di libri letti l'anno*).

Il *coefficiente di regressione lineare* (*b*), diversamente da *r*, rappresenta una *misura asimmetrica* e *unidirezionale* ed esprime la *variazione che subisce la variabile dipendente* quando la *variabile indipendente varia di un'unità*. La sua formula è la seguente:

$$b = \frac{cov(x,y)}{s_x}$$

Se torniamo a osservare la figura 14.5., sulla nuvola dei punti osservati riportata nel diagramma a dispersione è sovra-impressa la cosiddetta *retta interpolante*. Per cogliere fino in fondo la valenza del calcolo di *b* e del ricorso al *modello*[29] *di regressione lineare bivariato*, è bene riprendere la funzione della retta: $Y = a + bX$, in cui *a* rappresenta l'*intercetta* della retta sull'asse delle ordinate Y, ovvero il punto in cui la retta interseca tale asse (*a* corrisponde a Y, se X = 0) e *b* (il nostro coefficiente di regressione)

[28] Per approfondimenti sulla *matrice delle correlazioni*, in cui è riportato il sistema completo dei coefficienti *r* calcolabili, per ogni coppia di variabili, rispetto al sistema di indicatori selezionati ai fini dell'applicazione dell'*Analisi in Componenti Principali* (ACP), cfr. il Cap. 16.

[29] Un *modello di analisi* rappresenta un procedimento di elaborazione dei dati il cui risultato corrisponde ad una *stima della bontà di adattamento di un modello teorico alla struttura dei dati*. Il ricorso ad un modello implica l'esplicitazione a monte di *assunti* circa le *relazioni* esistenti *tra* le *variabili* e la sua funzione consiste nel *testare ipotesi*, ovvero nel *verificare la tenuta e la compatibilità di uno schema teorico rispetto al piano osservativo* (cioè ai dati contenuti in matrice).

rappresenta il *coefficiente angolare della retta*, indicando la sua inclinazione. Laddove *b* sia maggiore di 0, la retta assume un andamento ascendente (inclinazione dal basso a sinistra verso l'alto a destra), se *b* è minore di 0, la retta si caratterizza invece per una forma discendente (con un'inclinazione dal basso a destra verso l'alto a sinistra). Obiettivo di uno studio empirico incentrato sulla relazione di dipendenza lineare tra due variabili cardinali o quasi-cardinali è quello di individuare per ciascun punto osservato un punto teorico, posto alla minore distanza possibile e che giaccia sulla retta di regressione (la quale attraversa la nuvola dei punti). La differenza tra ciascun valore osservato di *y* e il suo corrispettivo teorico si definisce *residuo*. La migliore retta individuabile è quella che minimizza tali residui. Un coefficiente di regressione di segno positivo indica un'*influenza lineare diretta* di X su Y; un *b* negativo indica un'*influenza lineare inversa* di X su Y; infine, un *b* pari a 0 evidenzia l'assenza di dipendenza lineare tra X e Y (in tal caso, la retta di regressione è parallela all'asse delle X e Y si rivela indipendente da X).

Tornando al nostro esempio, *b* è pari a +0,27 (*p=.000*), come anche riportato nel grafico a dispersione (cfr. Fig. 14.5.). Peraltro, R^2 (il cui campo di variazione oscilla tra 0 e 1) è interpretabile anche come misura complessiva di adattamento del modello teorico al piano empirico (per approfondimenti, cfr. Corbetta, 1999; Lucchini, 2018; Bocci e Mingo, 2020).

Si potrebbe continuare a parlare a lungo dell'analisi bivariata; tuttavia, per esigenze di sintesi e di semplificazione, si chiude qui, ricordando, ancora una volta, al lettore come i diversi livelli di analisi (a partire dallo stadio del pretrattamento dei dati) siano strettamente connessi tra loro. L'analisi bivariata, che consente al ricercatore sociale di avviare lo studio della trama delle relazioni tra variabili, rappresenta, pertanto, un essenziale *incipit* in tale direzione, una sorta di "apripista" ai fini dell'applicazione di tecniche e modelli più complessi, di tipo multivariato.

Riferimenti bibliografici

Barbaranelli C. (2007), *Analisi dei dati. Tecniche multivariate per la ricerca psicologica e sociale*, Milano, LED.

Bocci, L., Mingo, I. (2020), *Statistiche: tra produzione e fruizione. Fonti e strumenti per l'analisi dei dati*, Roma, Edizioni Nuova Cultura.

Cairo A. (2020), *Come i grafici mentono. Capire meglio le informazioni visive*, Milano, Raffaello Cortina.

Corbetta P.G. (1999), *Metodologia e tecniche della ricerca sociale*, Bologna, il Mulino.

Delli Zotti G. (2010), *Tecniche grafiche di analisi e rappresentazione dei dati*, Milano, Franco Angeli.
Di Franco G. (2005), *EDS: esplorare, descrivere e sintetizzare i dati. Guida pratica all'analisi dei dati nella ricerca sociale*, Milano, FrancoAngeli.
Di Franco G., Marradi A. (2020), *L'analisi bivariata*, Milano, FrancoAngeli.
Faggiano M.P. (2012), *Gli usi della tipologia nella ricerca sociale empirica*, Milano, FrancoAngeli.
Jones B. (2020), *Data analysis & visualization. Sette insidie da evitare per analizzare e rappresentare dati*, Adria (RO), Apogeo Editore.
Lombardo, C., Mauceri, S. (2020), *La società catastrofica. Vita e relazioni sociali ai tempi dell'emergenza Covid-19*, Milano, FrancoAngeli.
Lucchini M. (2018), *Metodologia della ricerca sociale*, Milano, Pearson.
Marradi A. (2002), *Linee guida per l'analisi bivariata dei dati nelle scienze sociali*, Milano, FrancoAngeli.
Merton, R.K. (1949-1968a), *The Bearing of Sociological Theory upon the Development of Empirical Research*, in Merton, 1949-1968a; tr. it., *L'influenza della teoria sociologica sulla ricerca empirica*, in *Teoria e struttura sociale*, vol. I, Bologna, il Mulino, 1970.
Merton, R.K. (1949-1968b), *The Middle Range Theories*, in Merton, 1949-1968a; tr. it., *Sulle teorie sociologiche di medio raggio*, in *Teoria e struttura* sociale, vol. I, Bologna, il Mulino, 1970.
Merton, R.K. (1955), *A Paradigm for the Study of the Sociology of Knowledge*, in Lazarsfeld e Rosenberg, 1955; estratto da *The Sociology of Knowledge*, in Gurvitch e Moore (*eds.*), 1945, *Twentieth Century Sociology*, New York, The Philosophical Library Inc.; tr. it., in *Teoria e struttura sociale*, vol. III, Bologna, il Mulino, 1970.
Statera G. (1997). *La ricerca sociale. Logica, strategie, tecniche*, Roma, SEAM.