Piergiorgio Corbetta

METODOLOGIA E TECNICHE DELLA RICERCA SOCIALE

LA LOGICA DELLA RICERCA SOCIALE

I PARADIGMI DELLA RICERCA SOCIALE

Kuhn e i paradigmi delle scienze

La parola *paradigma* è stata utilizzata da Platone (*modello*) e da Aristotele (*esempio*). In sociologia si usa molto, e ha diversi significati (*teoria*, *articolazione interna di una teoria*, *scuola o pensiero*).
Nel 1972 Thomas Kuhn scrive l'opera "La struttura della rivoluzioni scientifiche", in cui rifiuta la concezione tradizionale della scienza come accumulazione progressiva di nuove scoperte, affermando invece che in certi momenti (detti *rivoluzionari*) si interrompe il rapporto di continuità con il passato e si inizia un nuovo corso, in modo non completamente razionale. "Si tratta di un elemento arbitrario composto di accidentalità storiche e personali sempre presente come elemento costitutivo nelle convinzioni manifestate da una data comunità scientifica in un dato momento". Il passaggio da una teoria a un'altra è così globale e ha tali conseguenze che Kuhn lo chiama *rivoluzione scientifica*. C'è un cambiamento dei problemi da proporre all'indagine scientifica e dei criteri con cui si stabilisce cosa si considera come un problema ammissibile, cambia anche la struttura concettuale attraverso cui gli scienziati guardano il mondo (*paradigma*). Il paradigma è una prospettiva teorica che è condivisa e riconosciuta dagli scienziati, è fondata su acquisizioni precedenti e indirizza la ricerca riguardo alla scelta dei fatti rilevanti da studiare, alla formulazione delle ipotesi e ai metodi e tecniche di ricerca necessari.
Senza un paradigma una scienza non ha orientamenti né criteri di scelta, perché tutti i criteri, i problemi e le tecniche diventano ugualmente rilevanti. Il paradigma è una guida e fornisce agli scienziati un modello e le indicazioni per costruirlo. Con il paradigma lo scienziato acquisisce contemporaneamente teorie, metodi e criteri. Il paradigma è qualcosa di più ampio di una teoria, è una visione del mondo, una finestra mentale, una griglia di lettura che precede l'elaborazione teorica. La scienza normale corrisponde a quei periodi in cui esiste all'interno di una disciplina un paradigma condiviso dagli scienziati.
Nella storia della sociologia è difficile individuare un paradigma predominante, condiviso da tutti i sociologi. Solo tra gli anni '40 e '50 ha prevalso il concetto di sistema e la teoria funzionalista di T. Parsons. Egli rielabora il pensiero degli europei e crea una teoria basata sul sistema e sul consenso. A questo paradigma viene contrapposto quello di Marx, basato sul conflitto sociale. In questo modo possiamo parlare di disciplina multiparadigmatica.

Positivismo e interpretativismo

Da un punto di vista storico, possiamo individuare due paradigmi fondamentali che hanno indirizzato la ricerca sociale: il *positivismo* e l'*interpretativismo*. La profonda differenza tra i due paradigmi emerge dalle risposte che essi danno alle domande principali a cui si trova di fronte la scienza sociale: la realtà (sociale) esiste (*ontologia*)? È conoscibile (*epistemologia*)? Come può essere conosciuta (*metodologia*)?

Positivismo

Il paradigma positivista (il primo ad essere utilizzato nelle scienze sociali) studia la realtà sociale utilizzando gli apparati concettuali, le tecniche di osservazione e misurazione, gli strumenti di analisi matematica e i procedimenti di inferenza delle scienze naturali. Il primo vero sociologo positivista è Durkheim, la cui teoria impone di trattare i fatti sociali come cose effettivamente esistenti al di fuori delle coscienze individuali e studiabili oggettivamente.
L'ontologia del positivismo afferma quindi che la realtà sociale ha esistenza effettiva ed è conoscibile, come se si trattasse di una "cosa". Dal punto vista epistemologico, esso si basa sul dualismo tra ricercatore e oggetto di studio (che non si influenzano a vicenda in nessun modo), presume di ottenere risultati veri e certi, il suo obiettivo è quello di spiegare e di formulare leggi naturali e generali immutabili. La metodologia positivista prevede quindi esperimenti e manipolazioni della realtà, con osservazioni e distacco tra l'osservatore e l'osservato; il suo modo di procedere è prevalentemente induttivo (dal particolare al generale). Le tecniche utilizzate sono quantitative (esperimenti, statistica) e si utilizzano le variabili.

Neopositivismo

Il neopositivismo nasce per rispondere alle critiche che erano state avanzate al positivismo. Dal punto di vista ontologico, adotta il realismo critico, per cui afferma che esiste una realtà sociale esterna all'uomo, ma che essa è conoscibile solo imperfettamente, in modo probabilistico. L'epistemologia del neopositivismo prevede il riconoscimento del rapporto di interferenza tra studioso e studiato, che deve essere il più possibile evitato per poter formulare leggi non più assolute, ma limitate nel tempo e soggette alla continua falsificazione per poter arrivare sempre più vicini alla conoscenza assoluta. La metodologia resta sostanzialmente quella del positivismo, anche se c'è un'apertura ai metodi qualitativi.

Interpretativismo

L'interpretativismo, che vede in Weber il suo esponente principale, non si propone di spiegare la realtà bensì di comprenderla; in questo modo si pone all'opposto del positivismo per quanto riguarda i punti principali del paradigma. Infatti, la sua ontologia prevede il costruttivismo e il relativismo (realtà multiple), vale a dire che non esiste una realtà oggettiva (ogni individuo produce una sua realtà, e solo questa realtà è conoscibile); inoltre anche le singole realtà individuali o anche condivise tra i gruppi sociali, variano comunque tra le diverse culture e quindi non esiste una realtà sociale universale valida per tutti. L'epistemologia prevede una separazione tra studioso e oggetto dello studio, la ricerca sociale è vista come una scienza interpretativa alla ricerca di significato piuttosto che una scienza sperimentale in cerca di leggi. Nel perseguire il suo scopo (che è quello della comprensione del comportamento individuale), la ricerca sociale può servirsi di astrazioni e generalizzazioni: i tipi ideali e gli enunciati di possibilità. La metodologia prevede l'interazione tra studioso e studiato, perché solo in questo modo è possibile comprendere il significato attribuito dal soggetto alla propria azione. Le tecniche sono quindi qualitative e soggettive e il metodo usato è quello dell'induzione (dal particolare al generale).

Radicalizzazioni e critiche

Una radicalizzazione del positivismo consiste nel *riduzionismo*, cioè nel ridurre la ricerca sociale ad una mera raccolta di dati senza un'elaborazione teorica che li supporti. In questo modo la ricerca sociale diventa una massa sterminata di dati minuziosamente rilevati, misurati e classificati, ma non coordinati tra loro, privi di connessioni significative, incapaci di rendere una conoscenza adeguata dell'oggetto cui nominalmente si riferiscono.
In ogni caso, la critica maggiore mossa al positivismo è quella di separare troppo nettamente le categorie osservative da quelle teoriche; in altre parole non è possibile sostenere che le forme di conoscenza non siano storicamente e socialmente determinate e quindi dipendenti dalle teorie utilizzate.
Per quanto riguarda l'interpretativismo, le critiche maggiori sono rivolte ai filoni sviluppatisi dalla teoria originale di Weber (che pur affermando la centralità dell'intenzione soggettiva, non escludeva la possibilità di arrivare a delle forme di generalizzazione conoscitiva, i *tipi ideali*) che si spingevano verso un soggettivismo estremo. In questo modo si esclude la possibilità dell'esistenza della scienza sociale, perché se tutto è soggettivo e unico non possono esistere leggi sociali comuni a più individui che hanno assunto autonomia rispetto ai singoli, come le istituzioni. La seconda critica afferma che se la realtà è una pura costruzione soggettiva, non è possibile andare oltre alla persona, si nega l'acquisibilità di generalizzazioni sovrapersonali e quindi si nega l'oggettività della scienza. Se il ricercatore non può trascendere l'oggetto dell'indagine non può esistere la conoscenza oggettiva.

RICERCA QUANTITATIVA E RICERCA QUALITATIVA

Ricerca quantitativa e ricerca qualitativa: un confronto

Impostazione della ricerca

Nei due approcci è fondamentalmente diverso il rapporto instaurato tra *teoria* e *ricerca*.
Nel caso della ricerca quantitativa neopositivista, il rapporto è strutturato in fasi logicamente sequenziali , secondo un'impostazione sostanzialmente deduttiva (la teoria precede l'osservazione), che si muove nel contesto della *giustificazione*, cioè di sostegno, tramite i dati empirici, della teoria precedentemente formulata sulla base della letteratura.
Nel caso della ricerca qualitativa interpretativista elaborazione teorica e ricerca empirica procedono intrecciate, in quanto il ricercatore vede nella formulazione iniziale di una teoria una possibile condizionamento che potrebbe inibirgli la capacità di comprendere il soggetto studiato. In questo modo la letteratura ha una minore importanza.
Anche i *concetti* sono usati in modo diverso dai due approcci. I concetti sono gli elementi costitutivi della teoria, e tramite la loro *operativizzazione* (trasformazione in variabili empiricamente osservabili) permettono alla teoria di essere sottoposta a controllo empirico.
Nell'approccio neopositivista la chiarificazione dei concetti e la loro operativizzazione in variabili avvengono prima ancora di iniziare la ricerca. Questo metodo, se da un lato offre il vantaggio di poter rilevare empiricamente il concetto, dall'altro comporta anche lo svantaggio di una forte riduzione e impoverimento del concetto stesso, con il rischio ulteriore che la variabile sostituisca il concetto (*reificazione*).
Un ricercatore qualitativo avrebbe invece utilizzato il concetto come orientativo (*sensitizing concept*), che predispone alla percezione, ancora da definire non solo in termini operativi, ma anche teorici, nel corso della ricerca stessa. I concetti diventano quindi una guida di avvicinamento alla realtà empirica, non riduzioni della realtà stessa in variabili astratte.
Per quanto riguarda il *rapporto generale con l'ambiente studiato*, l'approccio neopositivista non ritiene che la reattività del soggetto possa rappresentare un ostacolo di base, e crede che un certo grado di *manipolazione controllata* sia ammissibile. Viceversa la ricerca qualitativa si basa sull'*approccio naturalistico*, vale a dire che il ricercatore non manipola in alcun modo la realtà in esame. I due modi di fare ricerca trovano illustrazioni tipiche e opposte nelle tecniche dell'esperimento e dell'osservazione partecipante.
Se passiamo alla specifica *interazione psicologica con i singoli soggetti studiati*, il ricercatore quantitativo assume un punto di vista esterno al soggetto studiato, in modo neutro e distaccato; inoltre studia solo ciò che egli ritiene importante. Il ricercatore qualitativo invece si immerge il più completamente possibile nella realtà del soggetto e quindi tende a sviluppare con i soggetti una relazione di immedesimazione empatica. Ma in questo modo sorge prepotentemente il problema dell'oggettività della ricerca.
Anche l'*interazione fisica con i singoli soggetti studiati* è differente per i due approcci. La ricerca quantitativa spesso non prevede alcun contatto fisico tra studioso e studiato, mentre nella ricerca qualitativa il contatto fisico è una precondizione essenziale per la comprensione.
Il soggetto studiato quindi risulta *passivo* nella ricerca quantitativa, mentre ha un ruolo *attivo* nella ricerca qualitativa.

Rilevazione (disegno della ricerca)

Nella ricerca quantitativa il disegno della ricerca (decisioni operative che sovrintendono all'organizzazione pratica della ricerca) è costruito a tavolino prima dell'inizio della rilevazione ed è rigidamente strutturato e chiuso. Nella ricerca qualitativa invece è destrutturato, aperto, idoneo a captare l'imprevisto, modellato nel corso della rilevazione. Da queste diverse impostazioni deriva la diversa concezione della *rappresentatività* dei soggetti studiati. Nella ricerca quantitativa il ricercatore è più preoccupato della rappresentatività del pezzo di società che sta studiando piuttosto che della sua capacità di comprendere, mentre l'opposto vale per la ricerca qualitativa, alla quale non interessa la rilevanza statistica bensì l'importanza che il singolo caso sembra esprimere.
Anche lo *strumento di rilevazione* è differente per i due tipi di ricerche. Nella ricerca quantitativa esso è uniforme o uniformante per garantire la validità statistica, mentre nella ricerca qualitativa le informazioni sono approfondite a livelli diversi a seconda della convenienza del momento.
Allo stesso modo, anche la *natura dei dati* è diversa. Nella ricerca quantitativa essi sono oggettivi e standardizzati (*hard*), mentre la ricerca qualitativa si preoccupa della loro ricchezza e profondità soggettive (*soft*).

Analisi dei dati

L'analisi dei dati è completamente differente per le due impostazioni della ricerca, a partire dall'*oggetto dell'analisi*. La ricerca quantitativa raccoglie le proprietà individuali di ogni soggetto che sembrano rilevanti per lo scopo della ricerca (*variabili*) e si limita ad analizzare statisticamente queste variabili. Il soggetto non viene quindi più ricomposto nella sua unitarietà di persona. L'obiettivo dell'analisi sarà *spiegare la varianza* delle variabili dipendenti, trovare cioè le cause che provocano la variazione delle variabili dipendenti.
La ricerca qualitativa invece non frammenta i soggetti in variabili, ma li considera nella loro interezza, sulla base del ragionamento che l'individuo è qualcosa in più della somma delle sue parti. L'obiettivo è quindi quello di *comprendere le persone*, interpretando il punto di vista dell'attore sociale.
Le tecniche matematiche e statistiche sono fondamentali per la ricerca quantitativa, mentre sono considerate inutili e dannose nella ricerca qualitativa.

Risultati

I risultati dei due tipi di ricerca sono naturalmente diversi. Già nella *presentazione dei dati* notiamo che la ricerca quantitativa si serve di *tabelle*, mentre quella qualitativa di *narrazioni*. Le tabelle hanno il pregio della chiarezza e della sinteticità, ma presentano il difetto di presentare uno schema mentale proprio dei ricercatori che può non corrispondere alle reali categorie mentali dei soggetti; inoltre impoveriscono inevitabilmente la ricchezza delle affermazioni dei soggetti. Le narrazioni riescono ad ovviare a questi difetti, perché riportano le parole degli intervistati e quindi si pongono come una "fotografia" dei loro pensieri.
Per quanto riguarda la *generalizzazioni* dei dati, la ricerca quantitativa si pone l'obiettivo di enunciare rapporti causali tra le variabili che possano spiegare i risultati ottenuti. La ricerca qualitativa, invece, cerca di individuare *tipi ideali* (nel senso weberiano), cioè categorie concettuali che non esistono nella realtà, ma che liberano i casi reali dai dettagli e dagli accidenti della realtà per estrarne le caratteristiche essenziali ad un livello superiore di astrazione; lo scopo dei tipi ideali è quello di essere utilizzati come modelli con i quali illuminare e interpretare la realtà stessa.
La ricerca qualitativa non si preoccupa di spiegare i meccanismi causali che stanno alla base dei fenomeni sociali, cerca invece di descriverne le differenze interpretandole alla luce dei tipi ideali. All'opposto, il fine ultimo della ricerca quantitativa è proprio quello di individuare il meccanismo causale.
Un'ultima questione è quella della *portata dei risultati*. A questo proposito notiamo che la profondità dell'analisi e l'ampiezza della ricerca sono inversamente correlate, vale a dire che ad un maggior numero di casi esaminati corrisponde un minore approfondimento dei singoli casi. Data la maggiore quantità di casi necessariamente esaminati dalla ricerca quantitativa, risulta indubbiamente una maggiore generalizzabilità dei risultati rispetto a quelli della ricerca qualitativa.

Due diversi modi di conoscere la realtà sociale

A questo punto ci si potrebbe chiedere se uno dei due approcci è "scientificamente" migliore dell'altro. A questo proposito si possono individuare tre posizioni. La prima afferma che i due approcci sono incompatibili tra di loro, e quindi i rispettivi sostenitori dei due paradigmi dicono che il proprio è corretto mentre l'altro è sbagliato. La seconda si ritrova nei neopositivisti, che affermano l'utilità dell'approccio qualitativo, ma solo in una prospettiva preliminare di stimolazione intellettuale (ruolo ancillare). La terza posizione infine sostiene la pari dignità dei due metodi, e auspica lo sviluppo di una scienza sociale che, a seconda delle circostanze e delle opportunità, scelga per l'uno o per l'altro approccio. Infatti contributi importanti alle scienze sociali sono arrivati da entrambi i tipi di ricerca, che possono essere rispettivamente adatti per differenti situazioni. Entrambi gli approcci si possono considerare come due diversi modi di fare ricerca che possono contribuire insieme alla conoscenza dei fenomeni sociali, integrandosi vicendevolmente per una migliore comprensione della realtà da punti di vista differenti.

LA RILEVAZIONE DEI DATI: TECNICHE QUANTITATIVE

LA TRADUZIONE EMPIRICA DELLA TEORIA

Struttura "tipo" della ricerca quantitativa

La ricerca scientifica è un processo creativo di scoperta che si sviluppa secondo un itinerario prefissato e secondo procedure prestabilite che si sono consolidate all'interno della comunità scientifica. Questo significa che esiste un atto della scoperta che sfugge alle analisi logiche, ma allo stesso tempo la ricerca empirica deve essere *pubblica*, *controllabile* e *ripetibile* per poter essere definita scientifica. Per questo esiste un percorso "tipico" della ricerca sociale che parte dalla teoria, attraversa le fasi si raccolta e analisi dei dati e ritorna alla teoria. Più precisamente, si possono individuare cinque fasi e cinque processi che le legano.
La prima fase è quella della *teoria*, la seconda quella delle *ipotesi*, legate tra di loro attraverso il processo della *deduzione*. La teoria è generale mentre l'ipotesi ne rappresenta un'articolazione specifica. La terza fase è quella della *raccolta dei dati*, a cui si arriva attraverso il processo di *operativizzazione*, cioè la trasformazione delle ipotesi in affermazioni empiricamente osservabili. L'operativizzazione porta alla definizione del *disegno della ricerca*, cioè di un piano di lavoro che stabilisce le varie fasi dell'osservazione empirica. La quarta fase è quella dell'*analisi dei dati*, preceduta dall'*organizzazione dei dati* rilevati. Di solito questa fase nella ricerca quantitativa consiste nella creazione di una *matrice di dati*. La quinta fase è quella della rappresentazione dei *risultati*, a cui si arriva tramite un processo di *interpretazione* delle analisi statistiche condotte nella fase precedente. Infine il ricercatore ritorna alla teoria iniziale tramite un processo di *induzione*, che confronta i risultati ottenuti con la teoria precedente.

Dalla teoria alle ipotesi

Per elaborare un'ipotesi si parte da una *teoria*, cioè un insieme di preposizioni tra loro organicamente collegate che si pongono a un elevato livello di astrazione e generalizzazione rispetto alla realtà empirica, le quali sono derivate da regolarità empiriche e dalle quali possono essere derivate della previsioni empiriche. Si tratta di un tentativo di spiegare un fenomeno sociale e deve essere verificata.
Una teoria deve essere organizzata in *ipotesi* specifiche. L'ipotesi implica una relazione tra due o più concetti, si colloca a un livello inferiore di astrazione e generalità rispetto alla teoria, e ne permette una traduzione in termini empiricamente controllabili.
La validità di una teoria dipende dalla sua traduzione in ipotesi empiricamente verificabili, perché se una teoria è troppo vaga per dar luogo ad ipotesi, non può essere verificata nella realtà. Il criterio della *controllabilità empirica* è il criterio stesso della *scientificità*.
È importante la differenza tra *generalizzazioni empiriche* e *teorie*: le prima sono proposizioni isolate che riassumono uniformità relazionali osservate tra due o più variabili, mentre le seconde nascono quando queste proposizioni sono raccolte e sussunte in un sistema concettuale che si colloca ad un livello superiore di astrazione (ad esempio, permette di avanzare ipotesi in campi diversi e remoti da quelli originari).
Talvolta la pratica delle ricerca si sviluppa con ordini diversi rispetto a quello canonico: è possibile che le ipotesi vengano sviluppate dopo aver raccolto i dati, e con questi confrontati *a posteriori*. Oppure si ricorre alla teoria *dopo* aver analizzato i dati, per spiegare un fatto anomalo o un risultato inaspettato. Infine, una nuova teoria può essere scoperta nel corso della fase empirica. Talora la rilevazione viene prima della delle ipotesi per ragioni di forza maggiore, nel caso dell'*analisi secondaria*, quando cioè si applica una seconda analisi a dati raccolti da altri ricercatori in tempi precedenti.

Dai concetti alle variabili

Nel suo significato più ampio, il termine *concetto* si riferisce al *contenuto semantico* (significato) dei segni linguistici e delle immagini mentali. Proprio per questa sua generalità, il concetto può includere ogni specie di segno o di procedura semantica, astratto, concreto, universale, individuale, ecc.
Essendo l'ipotesi una interconnessione tra concetti, emerge il fatto che i concetti sono i "mattoni della teoria", e attraverso la loro operativizzazione si realizza la traduzione empirica di una teoria. Il concetto è il legame tra la teoria e il mondo empirico osservabile.
I concetti possono riferirsi ad astrazioni impossibili da verificare empiricamente (potere, felicità, ecc), oppure a entità concrete (oggetti, persone, ecc). Ma se i concetti formano una teoria, come si può verificarla empiricamente? Bisogna passare dai concetti astratti alla loro applicazione come *proprietà* degli specifici oggetti studiati (chiamati *unità di analisi*). Una proprietà misurabile di una unità di analisi si chiama *variabile*. Per esempio, il peso è un concetto, ma il peso di un oggetto è la sua proprietà. Il peso dell'oggetto misurato con la bilancia è una variabile. Oppure, il livello culturale è un concetto astratto, ma se applicato a un individuo diventa una proprietà, e se è misurabile una variabile.
In definitiva, una variabile è una proprietà di una unità di analisi a cui sono assegnati valori diversi.

Unità di analisi

L'unità di analisi rappresenta l'oggetto sociale al quale afferiscono, nella ricerca empirica, le proprietà studiate. Esse devono essere determinate con precisione nel momento in cui si vuole sottoporre a controllo empirico una teoria mediante una specifica ricerca di tipo quantitativo, in quanto sono un elemento importante del disegno della ricerca (il programma di lavoro empirico). Le unità di analisi possono essere concretamente rappresentate dall'*individuo* (la più comune), dall'*aggregato di individui* (di solito basate sulla *territorialità*), dal *gruppo-organizzazione-istituzione* (quando l'unità di rilevamento è rappresentata dal collettivo stesso), dagli *eventi sociali* (quando gli eventi stessi sono le unità di analisi) e dalle *rappresentazioni simboliche – prodotto culturale* (quanto l'unità di analisi consiste da messaggi di comunicazione di massa di ogni genere).
L'unità di analisi è singolare ed astratta, mentre chiamiamo *casi* gli esemplari specifici di quella data unità di analisi che vengono studiati, sui quali si rilevano i dati. Essi sono gli oggetti specifici della ricerca empirica.

Variabili

Una *variabile* è un *concetto operativizzato*, o meglio la *proprietà operativizzata* di un oggetto, in quanto il concetto, per poter essere operativizzato, ha dovuto essere applicato ad un oggetto diventandone proprietà. Un concetto può essere operativizzato in modi diversi. Le variabili possono variare tra diverse modalità; il caso limite è quello in cui risulta invariante nello specifico sottoinsieme degli oggetti studiati, nel qual caso prende il nome di *costante*. Le variabili possono variare nel *tempo*, su uno stesso caso (studio *longitudinale* o *diacronico*) oppure *fra i casi*, nello stesso tempo (studio *trasversale* o *sincronico*). Nelle scienze sociali il secondo metodo è il più utilizzato.
Le variabili possono esser classificate secondo la loro *manipolabilità*, la posizione nella relazione *causa/effetto*, l'*osservabilità*, il carattere *individuale o collettivo* e il *trattamento dei loro valori*.
La prima distinzione è quella tra variabili *manipolabili* e *non manipolabili*. Le variabili manipolabili sono quelle che possono essere modificate dal ricercatore, viceversa quelle non manipolabili non possono essere controllate. La maggior parte delle variabili sociali non sono manipolabili, anche se esistono dei casi in cui il ricercatore può controllarle.
La seconda distinzione è quella tra variabili *dipendenti* e variabili *indipendenti*. In una relazione asimmetrica tra due variabili, quando cioè una variabile influenza un'altra, la variabile indipendente è ciò che influenza (la causa), mentre la variabile dipendente è ciò che è influenzato (l'effetto). Nel caso in cui le variabili indipendenti siano più di una abbiamo una relazione *multivariata*.
La terza distinzione è quella tra variabili *latenti* e variabili *osservate*. La distinzione si basa sulla *osservabilità*, ossia sulla possibilità di rilevazione empirica. Le prime sono variabili non direttamente osservabili in quanto rappresentano concetti molto generali o complessi, mentre le seconde sono facilmente rilevabili. In ogni caso, entrambe possono essere operativizzate, per cui anche nel caso delle variabili latenti c'è una sostanziale differenza con i concetti.
L'ultima distinzione è quella tra variabili *individuali* e variabili *collettive*. Le variabili individuali sono specifiche di ogni individuo, mentre quelle collettive sono proprie di un gruppo sociale. Le variabili collettive si suddividono a loro volta in variabili *aggregate*, dove la proprietà del collettivo deriva dalle proprietà dei singoli componenti del gruppo, e variabili *globali*, quando le caratteristiche esclusive del gruppo non derivano da proprietà dei membri che lo compongono.
Le variabili sono assolutamente fondamentali nella ricerca empirica, anche se a ogni definizione operativa è lasciata all'arbitrio del ricercatore, che deve solo esplicitare e giustificare le sue scelte. Per questo una definizione operativa non è mai perfettamente adeguata ed esiste sempre uno scarto tra variabile e concetto. Un altro pericolo che porta l'operativizzazione è quello della *reificazione*, cioè di identificare la definizione operativa di un concetto (necessariamente arbitraria e impoverita) con il concetto stesso. Tuttavia, con tutti i suoi limiti, la definizione operativa è necessaria per fondare *scientificamente* e *oggettivamente* la ricerca sociale.

Variabili nominali, ordinali e cardinali

Un'altra classificazione molto importante è quella tra che riguarda le operazioni *logico-matematiche* che possono essere effettuate sulle variabili. A questo proposito abbiamo variabili *nominali*, *ordinali* e *cardinali*.
Le variabili nominali sono tali quando la proprietà da registrare assume *stati discreti non ordinabili*, cioè finiti e delimitati che non hanno alcun ordine o gerarchia tra di essi. Gli stati di una proprietà così descritta si chiamano *categorie*, le categorie operativizzate (cioè gli stati della variabile) *modalità* e i simboli assegnati alle modalità *valori*. La procedura di operativizzazione che permette di passare dalla proprietà alla variabile è la *classificazione*. Nel caso in cui ci siano solo due modalità si parla di variabili *dicotomiche*.
Le variabili ordinali sono tali quando la proprietà da registrare assume *stati discreti ordinabili*. In questo caso è possibile stabilire non solo relazioni di eguaglianza e disuguaglianza, ma anche relazioni d'ordine. In questo caso la procedura di operativizzazione è l'*ordinamento*, che tiene conto dell'ordinabilità degli stati della proprietà. Quindi l'attribuzione dei valori alle singole modalità dovrà utilizzare un criterio che presevi l'ordine degli stati. Tipicamente si utilizzano i numeri naturali, che comunque non godono delle loro proprietà cardinali (cioè la distanza che corre tra le varie modalità non può essere confrontata con le altre). Le variabili possono essere ordinali perché derivano da proprietà originariamente costituite da stati discreti oppure perché derivano da proprietà continue che sono state registrate su una sequenza sono ordinale perché non si dispone di una unità di misura.
Le variabili *cardinali* sono tali perché i numeri che ne identificano le modalità non sono delle semplici etichette, ma hanno un pieno significato numerico (hanno cioè proprietà sia ordinali che cardinali). Tra le modalità delle variabili di questo tipo, oltre a stabilire relazioni di eguaglianza e diversità e d'ordine, si possono effettuare operazioni di somma e sottrazione tra i valori e tutte le altre operazioni statistiche.
Si possono ottenere variabili cardinali attraverso due processi: la *misurazione* (quando la proprietà da misurare è *continua* e si possiede una *unità di misura* prestabilita che permetta di confrontare la grandezza da misurare con una grandezza di riferimento) e il *conteggio* (quando la proprietà da registrare è *discreta* ed esiste una *unità di conto*, cioè una unità elementare che è contenuta un certo numero di volte nelle proprietà dell'oggetto).
Nelle scienze sociali molte variabili cardinali derivano operazioni condotte su altre variabili cardinali.
Le variabili *quasi-cardinali* sono un sottoinsieme delle variabili cardinali. Le proprietà più caratteristiche delle scienze sociali possono essere tutte immaginate come proprietà continue, che però non riescono a passare dalla condizione di proprietà continua a quella di variabile cardinale per la difficoltà di applicare una unità di misura agli atteggiamenti umani. Un tentativo di superare questo limite è dato dalla *tecnica delle scale*, che cerca di avvicinarsi a misurazioni in senso proprio, cioè a variabili in cui la distanza tra due valori sia nota. Le variabili prodotte da questa tecnica sono dette *quasi-cardinali*.

Concetti, indicatori e indici

Nelle scienze sociali esistono concetti che hanno un elevato grado di generalità, e si pongono lontani dall'esperienza. Per poterli definire in modo empiricamente controllabile è necessario darne una definizione operativa (tradurli in termini osservativi) tramite gli *indicatori*. Gli indicatori sono concetti più semplici, traducibili in termini osservativi, che sono legati ai concetti generali da un *rapporto di indicazione*, o rappresentanza semantica. Gli indicatori sono quindi dei ancora dei concetti, ma più facilmente operativizzabili. Tuttavia il rapporto tra concetto e indicatore è parziale: da una parte un concetto generale non può essere esaurito da un solo indicatore specifico, dall'altra un indicatore può sovrapporsi solo parzialmente al concetto per il quale è stato scelto, e dipendere per il resto da un altro concetto, anche profondamente diverso. Inoltre la scelta di un indicatore è lasciata unicamente all'arbitrio del ricercatore, il cui unico obbligo è quello di argomentare la sua scelta, non di dimostrarne la correttezza.
La rilevazione empirica di un concetto non direttamente osservabile passa attraverso quattro fasi: l'articolazione del concetto in *dimensioni* (i diversi aspetti e significati del concetto), la scelta degli indicatori, la loro operativizzazione, la formazione degli *indici*. L'indice è la sintesi globale della pluralità delle variabili che sono state prodotte dai diversi indicatori.

Errore di rilevazione

L'*errore di rilevazione* rappresenta lo scarto tra il concetto teorico e la variabile empirica. L'errore di rilevazione viene di solito suddiviso in *errore sistematico* ed *errore accidentale*.
L'errore sistematico è un errore costante, che si presenta in tutte i singoli casi di rilevazione e tende a sovrastimare o sottostimare il valore vero.
L'errore accidentale è invece un errore variabile, che varia da rilevazione a rilevazione, per cui si tratta di un'oscillazione che, ripetuta su tutti i soggetti, tende a zero.
L'errore sistematico è la parte di errore comune a tutte le applicazioni di una determinata rilevazione; l'errore accidentale è la parte di errore specifica di ogni singola rilevazione.
Gli errori possono verificarsi sia nella fase teorica, o di *indicazione* (in cui si scelgono gli indicatori), che in quella empirica, o di *operativizzazione* (in cui si rilevano gli indicatori stessi).
L'errore nella fase di indicazione è sempre sistematico, perché l'indicatore non è del tutto adatto al concetto e quindi si ha un difetto nel rapporto di indicazione.
L'errore nella fase di operativizzazione può esser sia sistematico sia accidentale, in quanto esistono tre fasi nell'operativizzazione (*selezione* delle unità studiate, rilevazione dei dati o *osservazione* e *trattamento dei dati*) in ognuna delle quali si possono compiere degli errori.
Gli errori di selezione sono quelli dovuti al fatto che si operi solo su un campione di soggetti e non sull'intera popolazione. Essi sono: l'*errore di copertura* (dovuta al fatto che la lista della popolazione da cui si estrae il campione non è completa), l'*errore di campionamento* (il fatto di condurre la ricerca su una frazione della popolazione) e l'*errore di non risposta* (quando i soggetti del campione non possono o non vogliono rispondere).
Gli errori di osservazione possono essere addebitati a quattro fonti: errori dovuti all'*intervistatore*, errori dovuti all'*intervistato*, errori dovuti allo *strumento* ed errori dovuti al *modo di somministrazione*.
Gli errori di trattamento dei dati sono errori di codifica, trascrizione, memorizzazione, elaborazione, ecc.
L'unico errore quantificabile è quello di campionamento, per questo spesso viene riportato come errore globale della rilevazione.

Attendibilità e validità

L'attendibilità ha a che fare con la riproducibilità del risultato, e segnala il grado con il quale una certa procedura di traduzione di un concetto in variabile produce gli stessi risultati in prove ripetute con lo stesso strumento di rilevazione (*stabilità*) oppure con strumenti equivalenti (*equivalenza*).
La validità fa invece riferimento al grado con il quale una certa procedura di traduzione di un concetto in variabile effettivamente rileva il concetto che si intende rilevare.
In genere si associa l'attendibilità all'errore accidentale e la validità all'errore sistematico.
Per determinare l'attendibilità si utilizza il concetto di *equivalenza*, deve l'attendibilità è misurata attraverso la correlazione tra due procedure diverse ma molto simili tra loro. Questa tecnica è però utile solo nel caso in cui la procedura di operativizzazione consiste in una batteria di domande.
La validità invece è meno facilmente controllabile, in quanto l'errore di validità in genere nasce dall'errore di indicazione. Per determinare la validità di un indicatore si utilizzano sue procedure di convalida: la *validità di contenuto* e la *validità per* criterio. La validità di contenuto indica il fatto che l'indicatore prescelto per un concetto copre effettivamente l'intero dominio di significato del concetto; una convalida di questo tipo può avvenire soltanto su un piano puramente logico. La validità per criterio consiste nella corrispondenza tra l'indicatore e un criterio esterno che per qualche motivo si ritiene correlato con il concetto. Questo criterio può essere rappresentato da un altro indicatore già accettato come valido oppure da un fatto oggettivo. La validità per criterio è stata distinta in *validità predittiva* (quando si correla il dato dell'indicatore con un evento successivo ad esso connesso), in *validità concomitante* (quando l'indicatore è correlato con un altro indicatore rilevato nello stesso momento) e in *validità per gruppi noti* (quando l'indicatore viene applicato a soggetti dei quali sia nota la posizione sulla proprietà da rilevare.
Esiste infine anche la *validità di costrutto*, che consiste nella rispondenza di un indicatore alle attese teoriche in termini di relazioni con altre variabili.

INCHIESTA CAMPIONARIA

Problemi di fondo della rilevazione tramite interrogazione

La standardizzazione ovvero l'invarianza dello stimolo

I problemi che si pongono nella rilevazione tramite interrogazione si possono ricondurre ai due paradigmi fondamentali della ricerca sociale. In particolare, si contrappongono coloro che ritengono che la realtà sociale sia esterna all'individuo e pienamente conoscibile (posizione *oggettivista*) e coloro che sostengono che il dato sociale viene generato dall'interazione tra i soggetti studiante e studiato (posizione *costruttivista)*.
In secondo luogo, esiste una diatriba tra chi ritiene che esistano uniformità empiriche nei fenomeni sociali, che quindi possono esser classificati e standardizzati (posizione *uniformista*) e chi sottolinea la fondamentale irriducibilità del soggetto umano a qualsiasi forma di generalizzazione e standardizzazione (posizione *individualista*).
Queste due diversità di vedute implicano due specifiche questioni: la prima riguarda il rapporto tra intervistato e intervistatore, la seconda concerne la standardizzazione dello strumento di informazione e dell'informazione rilevata.
Per quanto riguarda il rapporto tra intervistato e intervistatore, l'approccio oggettivista ritiene che esso debba essere il più possibile spersonalizzato per non alterare lo stato dell'oggetto studiato. Tuttavia non è possibile instaurare un rapporto neutro tra intervistato e intervistatore, esiste sempre un certo grado di interazione.
Per quanto riguarda invece la standardizzazione della rilevazione, l'approccio uniformista prevede l'uniformità dello strumento della rilevazione-interrogazione (questionario con domande e risposte prefissate). I limiti del questionario sono due: non tiene conto della disuguaglianza sociale e uniforma l'individuo al livello dell'uomo medio.
L'obiettivo della posizione oggettivista-uniformista è quindi quello di ottenere la neutralità dello strumento di rilevazione, cioè ottenere l'*invarianza dello stimolo*. Ma non è sicuro che all'invarianza dello stimolo corrisponda l'uniformità dei significati, infatti una stessa domanda o parola possono avere diversi significati per lo stesso individuo, sia per motivi culturali che per le circostanze stesse in cui si svolge l'intervista.
A questo punto il ricercatore deve scegliere se appoggiarsi ad una tecnica che massimizza la ricerca di uniformità (questionario) e una che predilige l'individualità del soggetto studiato (intervista strutturata). Se si sceglie il questionario, bisogna essere consapevoli che studiando solo le uniformità del comportamento delle persone (ciò che esse hanno in comune) si limita inevitabilmente la piena comprensione dei fatti sociali.

L'affidabilità del comportamento verbale

Molti scienziati sociali hanno espresso dubbi sulla possibilità che la realtà sociale possa essere compresa attraverso i resoconti verbali. Le risposte alle domande (standardizzate) degli intervistatori possono non essere attendibili per due motivi: la *desiderabilità sociale* delle risposte e la *mancanza di opinioni*.
La desiderabilità sociale è la valutazione, socialmente condivisa, che in una certa cultura viene data ad un certo atteggiamento o comportamento individuale. Se uno di questi è valutato positivamente o negativamente, una domanda che abbia questo come oggetto può dare risposte distorte, perché l'intervistato può essere riluttante a rivelare opinioni o comportamenti che ritiene indesiderabili e può essere tentato di dare di sé la migliore immagine possibile, anche se non veritiera, in modo volontario o anche involontario.
La mancanza di opinioni concerne domande su tematiche complesse, sulle quali è plausibile che un certo numero di intervistati non abbia mai riflettuto, e quindi molti rispondono a caso oppure formulano sul momento un'opinione che può essere solo passeggera. Questo fenomeno è accentuato anche dal fatto che spesso la risposta "non so" viene percepita come un'ammissione di incapacità mentale.
Un altro problema delle domande standardizzate è che esse misurano l'opinione, ma non la sua intesità né il suo radicamento.

Sostanza e forma delle domande

Dati sociografici, atteggiamenti e comportamenti

Le domande di un questionario possono essere riconducibili alla tripartizione tra proprietà sociografiche di base, atteggiamenti e comportamenti.
Domande relative alle *proprietà sociografiche di base*: riguardano le caratteristiche sociali di base di un individuo (genere, età, luogo di nascita), quelle ereditate dalla famiglia (classe sociale di origine, titolo di studio), quelle temporanee (professione, stato civile, comune di residenza). Queste domande seguono delle formulazioni standard.
Domande relative agli *atteggiamenti* (opinioni, motivazioni, sentimenti, giudizi, valor): interrogare direttamente gli individui è l'unica via per ottenere queste informazioni, ma questo è anche il campo più difficile da esplorare, e le risposte sono influenzate dal modo in cui sono poste le domande.
Domande relative ai *comportamenti*, che rilevano quello che il soggetto dice di fare o di aver fatto. Questo è un aspetto più facile da indagare rispetto agli altri.

Domande aperte e domande chiuse

Le *domande aperte* sono quelle in cui si lascia piena libertà all'intervistato nella formulazione della risposta; si rivolgono di solito ad un campione ridotto. Il vantaggio della domanda aperta è quello di concedere una maggiore libertà di espressione e spontaneità, ma la risposta deve essere trascritta per intero. Lo svantaggio consiste nel fatto che la risposta è difficile da classificare successivamente in categorie predeterminate. Questo crea dei problemi di codifica, perché le risposte possono essere generiche o imprecise. Solo un buon intervistatore può sollecitare a precisare meglio il significato delle risposte, ma questo comporta un maggiore impegno dell'intervistato e quindi un maggior rischio di rifiuti, senza contare l'aggravio dei costi.
Le *domande chiuse* offrono la possibilità di scegliere tra risposte prefissate, quindi la risposta sarà standard. Sono il solo tipo di domande che si possono utilizzare con un campione di grandi dimensioni. I vantaggi delle domande chiuse consistono nella maggiore facilità di codifica, nello stimolo dell'analisi e della riflessione e nella maggiore economicità (in un campione ampio). La domande sono poste a tutti con lo stesso schema di risposte e chiariscono all'intervistato qual è il piano di riferimento della ricerca, evitando così risposte vaghe. Gli svantaggi sono il rischio di non considerare tutte le altre possibili alternative di risposta non previste e di influenzare la risposta con le alternative proposte. A volte l'intervistato sceglie una delle alternative anche se non è convinto. Le risposte, inoltre, non hanno significato uguale per tutti, e tutte le alternative possono essere troppe per essere ricordate.

Formulazione delle domande

La formulazione delle domande è importantissima perché può influenzare pesantemente la risposta; bisogna quindi porre molta attenzione al linguaggio, alla sintassi e al contenuto stesso delle domande.
Semplicità di linguaggio: il linguaggio del questionario deve essere adatto alle caratteristiche del campione studiato, il questionario autocompilato deve essere più semplice rispetto a quello con intervistatore e in ogni caso non bisogna far conto sulle sue spiegazioni, perché di solito gli intervistati si vergognano di ammettere di non capire le domande.
Lunghezza delle domande: di solito le domande devono essere concise, ma nel caso di tematiche complesse sono preferibili le domande lunghe perché facilitano il ricordo, danno più tempo per pensare e agevolano una risposta più articolata.
Numero delle alternative di risposta: non devono essere troppo numerose; se presentate a voce non devono superare il numero di cinque.
Espressioni in gergo: è preferibile non utilizzare espressioni gergali perché potrebbero irritare l'intervistato.
Definizioni ambigue: occorre fare molta attenzione a non utilizzare termini dal significato non ben definito.

Parole dal forte connotato negativo: è bene evitare anche i termini carichi di significato emotivo, soprattutto se questo è negativo.
Domande sintatticamente complesse: la domanda deve avere una sintassi chiara e semplice, evitando ad esempio la doppia negazione.
Domande con risposta non univoca: bisogna evitare le domande esplicitamente multiple (domande in cui ne sia inclusa un'altra) e quelle dalla problematica non sufficientemente articolata.
Domande non discriminanti: le domande devono esser costruite in modo tale da operare delle discriminazioni significative nel campione degli intervistati.
Domande tendenziose (viziate o a risposta pilotata): è necessario presentare le domande in modo equilibrato, senza orientare l'intervistato verso una possibile risposta.
Comportamenti presunti: è indispensabile evitare di dare per scontati comportamenti che non lo sono.
Focalizzazione nel tempo: occorre sempre definire con precisione l'arco temporale al quale si riferisce la domanda.
Concretezza – astrazione: la domanda astratta può dare facilmente luogo a risposte generiche o normative, mentre la domanda concreta facilita la riflessione e rende più difficile il fraintendimento.
Comportamenti e atteggiamenti: data la difficoltà di determinare gli atteggiamenti, è buona regola, quando possibile, limitarsi ai comportamenti piuttosto che restare nell'ambito dell'opinione.
Desiderabilità sociale delle risposte: per evitare risposte normative bisogna formulare domande il più possibile concrete. Altre indicazioni sono quelle di giustificare anche la risposta meno accettabile; considerare normale e diffuso anche il comportamento negativo; equilibrare la desiderabilità delle risposte ("Alcuni dicono che… altri pensano che…"); attribuire all'intervistato il comportamento condannato, lasciandogli il compito dell'eventuale smentita; formulare le domande in terza persona; e così via. In ogni caso è impossibile eliminare del tutto gli effetti della desiderabilità sociale.
Domande imbarazzanti: andrebbero studiate attraverso domande aperte e con interviste non-strutturate, con le quali si può conquistare la fiducia degli intervistati.
Mancanza di opinione e non so: bisogna far presente all'intervistato che "non so" è una risposta legittima come le altre, per esempio includendola espressamente tra le alternative possibili. Bisogna inoltre evitare di indirizzarlo, anche in maniera indiretta o inconsapevole.
Intensità degli atteggiamenti: è importante cogliere anche l'intensità degli atteggiamenti, perché è quest'ultima che determina i comportamenti. La rilevazione dell'intensità degli atteggiamenti necessita di solito di domante ulteriori.
Acquiescenza: si riferisce alla tendenza di scegliere risposte che esprimono accordo piuttosto che negative. Un problema simile è quello dell'*uniformità delle risposte*, quando si tende a scegliere la stessa risposta per una serie di domande che contemplano lo stesso tipo di alternativa.
Effetto memoria: per ovviare alla inevitabile distorsione causata dalla memoria si possono stabilire limiti temporali al ricordo; utilizzare punti di riferimento temporali relativi ad eventi più salienti rispetto a quello studiato; presentare all'intervistato liste di possibili risposte; utilizzare diari o strumenti analoghi; ecc.
Sequenza delle domande: è meglio mettere all'inizio domande facili, che abbiano lo scopo di rassicurare l'intervistato e di metterlo a proprio agio. Le domande imbarazzanti si posizioneranno quindi a metà questionario, in modo che l'intervistatore abbia avuto un po' di tempo per conquistare la fiducia dell'intervistato. Anche le domande impegnative dovranno essere collocate a metà dell'intervista, in modo tale da assecondare la curva di interesse dell'intervistato. Alla fine si potranno porre le domande più noiose ma che non richiedono riflessione come quelle sociometriche.
È bene seguire anche il passaggio da domande generali a domande particolari, stringendo progressivamente sugli aspetti più specifici.
Bisogna tenere conto infine dell'*effetto contaminazione*, cioè del fatto che in certi casi la risposta ad una domanda può essere influenzata dalle domande che l'hanno preceduta.

Batterie di domande

Le *batterie di domande* sono domande che essendo, tutte formulate nello stesso modo (stessa domanda introduttiva e stesse alternative di risposta, varia solo l'oggetto al quale si riferiscono), vengono presentate all'intervistato in un unico blocco. Le batterie di domande hanno gli obiettivi di risparmiare spazio sul questionario e tempo dell'intervista, facilitare la comprensione del meccanismo di risposta, migliorare la validità della risposta e permettere al ricercatore di costruire indici sintetici che riassumono in un unico punteggio le diverse domande della batteria. Gli svantaggi delle batterie di domande consistono nel pericolo che le riposte siano date a caso e che le risposte siano meccanicamente tutte uguali tra di loro.

Modalità di rilevazione

Interviste faccia a faccia

Nel caso che stiamo trattando, vale a dire quello dell'intervista con questionario standardizzato, l'obiettivo è quello di limitare l'effetto dell'intervistatore, standardizzandone il comportamento e limitandone i margini di discrezionalità attraverso una fase di addestramento. In altre parole, l'intervistatore deve inibirsi qualsiasi comportamento che può influenzare l'intervistato; per questo motivo gli intervistatori devono presentare alcuni tratti particolari per raggiungere questo scopo.
Le loro caratteristiche: l'intervistatore ideale è donna, sposata, di mezza età, diplomata, casalinga, di ceto medio, con un abbigliamento neutrale.
Le loro aspettative: le aspettative degli intervistatori possono essere trasmesse inconsciamente agli intervistati, influenzandone le risposte soprattutto per quanto riguarda intervistati insicuri.
La loro preparazione: l'intervistatore deve essere consapevole dell'influenza che ha nella formulazione delle risposte, e per questo deve essere istruito per limitare al massimo questi effetti.
La loro motivazione: l'intervistatore deve essere convinto dell'importanza del proprio lavoro, perché un atteggiamento contrario potrebbe riverberarsi in modo negativo sull'intervistato.

Interviste telefoniche

L'intervista telefonica presenta numerosi vantaggi: permette una grande velocità di rilevazione; ha costi ridotti; presenta minori resistenze alla concessione dell'intervista e maggiore garanzia di anonimato; permettere di raggiungere a parità di costo anche gli intervistati della periferia del paese; facilita enormemente il lavoro di preparazione degli intervistatori e la loro supervisione; consente di utilizzare direttamente il computer in fase di rilevazione.
I suoi svantaggi sono: il minore coinvolgimento dell'intervistato che porta a una maggiore incidenza di risposte superficiali; il più rapido logoramento del rapporto con l'intervistato; l'impossibilità di utilizzare materiale visivo; l'impossibilità di raccogliere dati non verbali; l'impossibilità di raggiungere tutti gli strati sociali; il fatto che anziani e persone poco istruite risultano sottorappresentate; il fatto che le domande sono spesso elementari a causa della ristrettezza del tempo a disposizione. I limiti più gravi sono comunque l'assenza di contatto e la mancanza di tempo, che non rendono adatta l'intervista telefonica quando si vogliono analizzare tematiche complesse.

Questionati autocompilati

I questionari autocompilati sono quelli che il soggetto compila da solo, senza l'intervento dell'intervistatore. Il vantaggio principale di questa tecnica è l'enorme risparmio dei tempi di rilevazione. I limiti invece consistono nel fatto che deve essere breve, conciso e il più semplice possibile per venire incontro al maggior numero possibile di persone; inoltre di solito coloro che restituiscono il questionario sono un segmento particolare della popolazione in esame (*autoselezione*), cosa che limita l'estensibilità dei risultati.
Esistono due casi principali di autocompilazione: la *rilevazione di gruppo* e la *rilevazione individuale*.
La rilevazione di gruppo avviene in presenza di un operatore che distribuisce i questionari, assiste alla compilazione e ritira i questionari; in questo caso i rischi di questa ricerca vengono molto ridotti.
La rilevazione individuale si può dividere a sua volta in rilevazione con *restituzione vincolata* e con *restituzione non vincolata*. Nel primo caso, l'operatore deposita il questionario presso il soggetto in questione e passa a ritirarlo in seguito (un tipico caso è il censimento); anche questa tecnica limita i rischi prima ricordati. Nel secondo caso, tipicamente quello del questionario postale, il questionario viene inviato per posta con una lettera di presentazione e con una busta di ritorno già affrancata. I vantaggi di questa tecnica sono i risparmi altissimi dei costi, la possibilità di essere compilati in qualsiasi momento, la maggiore garanzia di anonimato, l'assenza di distorsioni dovute all'intervistatore, l'accessibilità a soggetti residenti in zone poco raggiungibili. Gli svantaggi sono la bassa percentuale di risposte, l'autoselezione del campione, la necessità che il livello di istruzione della popolazione studiata sia medio-alto, la mancanza di controllo sulla compilazione, l'impossibilità di questionari complessi e la sua lunghezza non eccessiva.
Sia nel caso dei questionari telefonici che di quelli autocompilati sono da escludere le domande aperte.

Interviste computerizzate (elettroniche)

Il computer può essere utilizzato sia nelle interviste telefoniche (CATI, *Computer Assisted Telephone Interviewing*) che nelle interviste faccia a faccia (CAPI, *Computer Assisted Personal Interviewing*), che si differenzia da una normale intervista faccia a faccia per il fatto che le risposte vengono riportate direttamente su un PC portatile con risparmio notevole di tempo.
Una altro impiego del computer è quello della *teleintervista*, in cui l'intervistato risponde direttamente con il suo PC al questionario trasmesso per via elettronica. I vantaggi di questa tecnica consistono nell'elimizazione dell'intervistatore ma soprattutto nel fatto che si possono condurre inchieste longitudinali, cioè rilevazioni ripetute nel tempo sugli stessi soggetti. I suoi limiti consistono nel fatto che non è possibile accertarsi di chi effettivamente compili il questionario e inoltre la consapevolezza di esser studiati può alterare il comportamento.

Organizzazione della rilevazione

Queste fasi normalmente precedono la rilevazione vera e propria.
Lo studio esplorativo: consiste in interviste preliminari, che partono da una massima destrutturazione e da strumenti qualitativi fino a strumenti sempre più strutturati, che hanno lo scopo di analizzare perfettamente il problema di oggetto e di formulare con la massima precisione possibile le risposte alternative per le domande chiuse.
Il pre-test: consiste in una sorta di "prova generale" del questionario, sottoposto a poche decine di soggetti, che ha lo scopo di evidenziare eventuali problemi o cattive formulazioni del questionario stesso.
La preparazione e la supervisione degli intervistatori: gli intervistatori devono essere informati su tutto ciò che riguarda la ricerca, "collaudati" sul campo con il pre-test e controllati durante la ricerca vera e propria da appositi supervisori.
Il contatto iniziale con i soggetti intervistati: il problema dei rifiuti. Il momento più delicato dell'intervista è quello iniziale, in cui i soggetti devono decidere se collaborare o meno. Per rendere massima la collaborazione è importante insistere sull'anonimità delle risposte, sul prestigio dell'istituzione committente e sulla figura dell'intervistatore. È anche opportuno inviare una lettera di presentazione che spieghi chiaramente chi è il committente della ricerca, quali ne sono gli obiettivi, perché ci si rivolge proprio a lui, sottolineare l'importanza delle sue risposte e rassicurarlo sull'anonimato.
La forma grafica del questionario: è opportuno distinguere chiaramente il testo che riguarda l'intervistato e quello che è di pertinenza dell'intervistatore, i passaggi tra le domande devono essere indicati chiaramente, il questionario deve essere graficamente compatto e non estendersi su troppe pagine. I questionari autocompilati devono inoltre essere autoesplicativi, le domande devono essere semplici e brevi (meglio se dello stesso formato), l'impostazione grafica deve esser compatta e chiara.

Analisi secondaria e inchieste ripetute nel tempo

L'*analisi secondaria* è una ricerca che viene condotta su dati di inchiesta campionaria già precedentemente raccolti e disponibili nella forma di matrice-dati originale. L'analisi secondaria nasce nell'ambito di un maggiore approfondimento dei dati già raccolti sulla base di successive scoperte o di nuove teorie avanzate nelle scienze sociali; a questo punto i dati già raccolti possono essere suscettibili di nuove elaborazioni e approfondimenti. Inoltre, a causa dell'estrema onerosità della fase di raccolta dei fati, sono nate apposite agenzie che riescono risorse comuni mettendo poi i dati a disposizione di tutti i ricercatori; ovviamente queste agenzie non raccolgono i dati per un unico tema, ma si rivolgono ad un ampio spettro di problematiche sociali.
I vantaggi di questi sviluppi si possono riassumere in un generale risparmio economico, nella garanzia del rigore della rilevazione stessa e nella possibilità anche per i ricercatori con poche risorse di effettuare ricerche di ampio respiro. Gli svantaggi sono legati alla qualità dei dati, in quanto i dati raccolti nel passato possono non essere stati trattati in modo corretto; altri svantaggi sono la limitazione degli interrogativi e il fatto che possano nascere ricerche a partire dai dati disponibili piuttosto che dalle ipotesi teoriche.
La *meta-analisi* si differenzia dall'analisi secondaria perché non riesamina i vecchi dati ma si propone di "analizzare le analisi", applicando metodi statistici per giungere a delle sintesi dei risultati delle ricerche considerate.
Le *inchieste replicate nel tempo* (*diacroniche*) si possono dividere in *longitudinali* e *trasversali replicate*.
Le inchieste longitudinali consistono nell'intervistare ripetutamente gli stessi soggetti in un ampio arco di tempo (*panel*). I problemi di questa tecnica consistono nella "mortalità" del campione (molti soggetti saranno irraggiungibili per diversi motivi a ogni nuova rilevazione), nell'effetto memoria e nel fatto che il soggetto sapendosi osservato può modificare il suo comportamento normale.
L'*inchiesta retrospettiva* è invece una normale inchiesta trasversale nella quale si pongono agli intervistati domande sul loro passato, con gli evidenti limiti che l'affidamento alla memoria può comportare.
Le *inchieste trasversali replicate* consistono nell'intervistare diversi campioni di soggetti, sullo stesso argomento, ma in momenti diversi.
Il maggiore difetto degli studi che comportano il fattore "tempo" è naturalmente il costo molto elevato.

LA TECNICA DELLE SCALE

L'operativizzazione dei concetti complessi

La *tecnica delle scale* (*scaling*) consiste in un insieme di procedure messe a punto per misurare concetti complessi e non direttamente osservabili. L'unico modo per poterli registrare è quello di usare un insieme coerente ed organico di indicatori, mettendo anche a punto criteri intersoggettivi per controllare l'effettiva sovrapposizione fra indicatori e concetto e la completezza della procedura. Possiamo quindi dire che una scala è un insieme coerente di elementi che sono considerati indicatori di un concetto più generale.
La tecnica delle scale è usata soprattutto nella misura degli *atteggiamenti*, dove l'unità d'analisi è l'individuo, il concetto generale è un atteggiamento (credenze di fondo non rilevabili direttamente) e i concetti specifici sono le opinioni (espressione empiricamente rilevabile di un atteggiamento).
Le variabili prodotte dalla tecnica delle scale non possono essere considerate pienamente cardinali, perché scaturiscono da dimensioni sottostanti immaginate come proprietà continue non misurabili, anche se la teoria delle scale tenta di dare una risposta a questo problema. Per questo le variabili della teoria delle scale vengono chiamate *quasi-cardinali*.

Domanda e risposta graduata: l'autonomia semantica delle risposte

Gli elementi di una scala sono tipicamente domande, possiamo quindi affermare che una scale è costituita da una batteria di domande (raramente da una domanda singola).
Le domande (sempre chiuse) possono essere proposte in tre modi diversi. Il primo consiste nel presentare risposte *semanticamente autonome*, cioè ciascuna ha un suo intrinseco significato compiuto che non necessita, per essere compreso, di essere messo in relazione con il significato delle altre alternative presenti nella scala. Il secondo caso è quello in cui le categorie di risposta sono a *parziale autonomia semantica*, quando il significato di ogni categoria è parzialmente autonomo dalle altre ("molto", "abbastanza", "poco", "per nulla"). Infine ci sono le *scale auto-ancoranti*, dove solo le due categorie estreme sono dotate di significato, mentre tra di esse si colloca un continuum entro il quale il soggetto colloca la sua posizione.
Le variabili prodotte dalla prima situazione sono senza dubbio ordinali, mentre nella seconda è probabile che scatti un processo di comparazione quantitativa. Per quanto riguarda il caso delle risposte auto-ancoranti è ancora più probabile che si metta in moto una procedura mentale di suddivisione graduata dello spazio tra i due estremi, suddivisione che è però soggettiva e non valida per tutti. Per questo si parla di variabili quasi-cardinali.
Nel caso delle variabili a parziale autonomia semantica è preferibile offrire la possibilità di un punto neutro e dell'opzione "non saprei". Il numero delle opzioni disponibili di solito è 5 o 7, tranne nell'intervista telefonica, dove si usano domande con risposte binarie per motivi di semplicità.
Nel caso delle graduatorie auto-ancoranti si possono usare diverse soluzioni come quella delle caselle vuotre, della sequenza di cifre oppure della linea continua.
Le preferenze possono essere espresse in *termini assoluti* (quando ogni domanda riguarda isolatamente una singola questione) oppure in *termini relativi* (nella forma di confronti e scelte tra diversi oggetti).
È preferibile scegliere scale con più domande rispetto a scale con una domanda sola per tre motivi: la complessità dei concetti rende improbabile la loro copertura con un singolo indicatore; una rilevazione singola manca di precizione, in quanto non riesce a discriminare in maniera fine tra le diverse posizioni dei soggetti sulla proprietà considerata; infine le singole domande sono più esposte agli errori accidentali. Le domande ad un solo elementi sono quindi meno valide, meno precise e meno attendibili.

Scala di Likert

La procedura che sta alla base delle scale di Likert consiste nella somma dei punti attribuiti ad ogni singola domanda. Il formato delle singole domande della scala di Likert è rappresentato da una serie di affermazioni per ognuna delle quali l'intervistato deve dire se e in che misura è d'accordo. Di solito le alternative di risposta sono cinque, da "molto d'accordo" a "fortemente contrario".
La costruzione della scala avviene in quattro fasi. Nella prima, la *formulazione delle domande*, si individuano le dimensioni dell'atteggiamento studiato e si formulano delle affermazioni che coprano i vari aspetti del concetto generale che si vuole rilevare.
Nella seconda fase, la *somministrazione delle domande*, la scala viene sottoposta ad un campione limitato di intervistati con un certo livello di istruzione.
In seguito, nella terza fase (*analisi degli elementi*), si selezionano le domande e si valuta il grado di coerenza interna della scala, cioè se la scala misura effettivamente il concetto in esame. È infatti possibile che alcuni elementi non risultino in linea con gli altri e vadano quindi eliminati. Gli strumenti utilizzati nella terza fase sono la *correlazione elemento-scala* e il *coefficiente alfa*. Per la correlazione elemento-scala, si calcola per ogni soggetto il punteggio su tutta la scala e si calcola il coefficiente di correlazione tra questo punteggio e il punteggio di ogni singolo elemento. Il coefficiente di correlazione è una misura che quantifica il grado di relazione tra due variabili cardinali e indica se il punteggio di ogni singolo elemento si muove nella stessa direzione del punteggio globale che tiene conto di tutti gli altri elementi. Se ciò non avviene la domanda non è congruente con la scala e va eliminata. Il coefficiente alfa serve invece a valutare la coerenza interna complessiva della scala. Esso si basa sulla matrice di correlazione tra tutti gli elementi della scala e il loro numero; più alti sono i valori (da 0 a 1) maggiore è la coerenza interna alla scala.
Infine si apre la quarta fase, quella dei controlli sulla *validità* e l'*unidimensionalità* della scala. Tralasciando i controlli di validità, la tecnica più efficace per il controllo di unidimensionalità è quella dell'*analisi fattoriale*. Il suo scopo è quello di ridurre una serie di variabili tra loro collegate ad un numero inferiore di variabili ipotetiche tra loro indipendenti, in modo da controllare se dietro agli elementi di una scala che si presume unifattoriale, vi sia un solo fattore o più fattori.
I vantaggi della scala Likert consistono nella sua semplicità e applicabilità, mentre i suoi svantaggi sono il fatto che i suoi elementi vengono trattati come scale cardinali pur essendo ordinali (a parziale autonomia semantica), la mancata riproducibilità (dal punteggio della scala non è possibile risalire alle risposte delle singole domande) e il fatto che il punteggio finale non rappresenta una variabile cardinale.

Scalogramma di Guttman

La scala di Guttman nasce con l'obiettivo di fornire una soluzione al problema dell'unidimensionalità della scala di Likert e consiste in una sequenza di gradini, una successione di elementi aventi difficoltà crescente, in modo che chi ha risposto affermativamente ad una certa domanda deve aver risposto affermativamente anche a quelle che la precedono nella scala di difficoltà. In questo modo, se gli elementi della scala sono perfettamente scalati, solo alcune sequenze di risposte sono possibili; inoltre dal risultato finale è possibile risalire alle risposte date dal soggetto ai singoli elementi della scala (*riproducibilità*). Questa tecnica prevede solo elementi dicotomici, cioè ogni domanda può avere solo due risposte opposte e distinte. Le due risposte possibili vengono di solito contrassegnate con i numeri 0 e 1.
Anche la scala di Guttman segue tre-quattro fasi nella sua costruzione. La prima è quella della *formulazione delle domande*, con considerazioni analoghe a quelle relative alla scala di Likert tranne che le domande devono essere dicotomiche e disposte secondo un ordine crescente di forza. Anche la seconda fase (*somministrazione*) è simile a quella della scala di Likert, con il vantaggio che la forma binaria agevola le risposte e rende più veloce la compilazione (anche se talvolta la forte semplificazione indotta dal carattere binario delle scelte può creare problemi all'intervistato).
La specificità della scala di Guttman sta nell'analisi dei risultati, quando si valuta la scalabilità degli elementi, si scartano quelli meno coerenti col modello, si stabilisce un indice di scalabilità della scale e se accettarla o meno. In primo luogo si devono individuare gli errori della scala, cioè le risposte che non si inseriscono nelle sequenze previste nel modello. Per questo si utilizza un indice (*coefficiente di riproducibilità*) che misura il grado di scostamento della scala osservata dalla scala perfetta. Questo indice può variare da 0 a 1; per poter essere accettabile, il valore dell'indice deve essere maggiore o uguale a 0,90 (cioè errori pari o inferiori al 10% delle risposte). Esiste anche un altro indice, detto *di minima riproducibilità marginale*, che segnala il valore minimo al di sotto del quale il coefficiente di riproducibilità non può scendere, quali che siano le sequenze delle risposte. Esso deve essere confrontato con il coefficiente di riproducibilità: solo se il secondo, oltre ad essere maggiore di 0,90, è anche nettamente superiore al primo, si può affermare che la buona riproducibilità della scala è dovuta ad un'effettiva scalabilità dei suoi elementi e non alla distribuzione marginale delle risposte.
L'ultima fase è quella di attribuire i punteggi ai soggetti; per far questo si sommano i punteggi 0/1 ottenuti nelle varie risposte.
I problemi della scala di Guttman consistono nel fatto che il punteggio finale è ancora una variabile ordinale; si tratta di una tecnica applicabile solo ad atteggiamenti ben definiti e scalabili; il modello risulta rigidamente deterministico di fronte ad una realtà sociale interpretabile solo attraverso modelli probabilistici.

Modelli probabilistici (la scala di Rasch)

Nell'approccio probabilistico la probabilità di dare una certa risposta ad un dato elemento non è solo 0 o 1, ma si colloca tra questi due estremi. Questa impostazione presuppone un modello di relazione tra posizione del soggetto sul continuum e probabilità di risposta ad un determinato elemento della scala che viene chiamata *traccia*. La traccia è quindi una curva che descrive la probabilità di rispondere affermativamente ad un certo elemento a seconda della posizione dell'individuo sul continuum sottostante. La traccia non assume la forma lineare, ma quella di una curva ad "S" detta *curva logistica*. La posizione di ciascun soggetto è data dal valore *v*. La "difficoltà" di un elemento della scala (vale a dire la probabilità di una risposta "Sì") è data dal parametro *b* che corrisponde al valore della variabile latente per il quale la probabilità di risposta affermativa è il 50%. Maggiore è il valore di *b*, maggiore è la "difficoltà" della domanda. La probabilità di risposta positiva dipende quindi dalla differenza $v - b$: se essi coincidono, la probabilità è del 50%; se $v > b$ la probabilità di risposta affermativa è superiore a quella della risposta negativa; viceversa se $v < b$.
I vantaggi di questo modello sono due: esso è una descrizione molto più adeguata ai reali meccanismi che generano le risposte rispetto al modello deterministico e le variabili prodotte da questo modello sono variabili cardinali. In questo modo può dirsi realizzato l'obiettivo della misurazione nelle scienze sociali.

Unfolding *di Coombs*

L'*unfolding* di Coombs (dall'inglese *unfold*, "aprire", "(di)spiegare") è una tecnica specificamente pensata per trattare dati derivanti da preferenze relative. Essa permette di individuare se dietro le preferenze espresse dai soggetti esiste un unico continuum comune a tutti i soggetti sul quale gli elementi (e gli intervistati stessi) sono ordinabili. Se esiste un'unica dimensione sottostante e questa è utilizzata come criterio di valutazione da parte degli intervistati, allora solo determinate sequenze di risposta sono possibili. L'analisi delle preferenze espresse dagli intervistati permette in questo caso di risalire sia all'ordine degli elementi sul continuum sia alla posizione dei soggetti sullo stesso. Questa tecnica permette di arrivare ad una scala cardinale, nella quale cioè anche le distanze tra gli intervalli sono note. Il suo principale difetto consiste nel suo rigido determinismo, anche se recenti sviluppi hanno permesso la formulazione di modelli probabilistici.

Differenziale semantico

La tecnica del differenziale semantico si propone di rilevare con il massimo della standardizzazione il *significato* che i concetti assumono per gli individui. Il modo più semplice per scoprire cosa significa una certa cosa per una certa persona è quello di chiederglielo direttamente, ma solo persone intelligenti, istruite e con eccellenti capacità verbali possono fornire dati utilizzabili.
La tecnica del differenziale semantico supera questi limiti in quanto si basa sulle associazioni che l'intervistato instaura tra il concetto in esame ed altri concetti proposti in maniera standardizzata a tutti gli intervistati. In concreto si utilizzano una serie di scale auto-ancoranti nelle quali solo le categorie estreme hanno significato autonomo, mentre il significato graduato delle categorie intermedie viene stabilito a giudizio dell'intervistato. La lista di questi attributi bipolari non deve avere necessariamente relazione con l'oggetto valutato, e quindi deve essere sempre la stessa. Il numero delle domande di solito va dalle 12 alle 50, in base al disegno della ricerca.
Il modo più importante di utilizzare il differenziale semantico è rappresentato dall'esplorazione delle dimensioni dei significati. Si ritiene cioè che attraverso l'analisi fattoriale sia possibile determinare quali sono le dimensioni fondamentali che stanno dietro ai giudizio di un certo campione di soggetti intervistati. In linea generale, si possono trovare tre dimensioni fondamentali: la *valutazione*, la *potenza* e l'*attività*, in ordine di importanza. La valutazione sembra rappresentare l'atteggiamento verso un certo oggetto.
Il contributo più importante della tecnica del differenziale semantico è proprio quello di aver introdotto la multidimensionalità dei significati nella struttura degli atteggiamenti.

Test sociometrico

Il test sociometrico nasce al fine di rilevare le relazioni interpersonali esistenti all'interno di un gruppo di individui; infatti il suo campo di applicazione ideale è rappresentato da una classe scolastica. Nella sua forma più semplice il test sociometrico consiste in un questionario con poche domande, che ruotano intorno al tema della preferenza e rifiuto nei confronti degli altri appartenenti al gruppo. In questo modo si può conoscere lo *status sociometrico personale*, cioè il grado di prestigio di cui gode un elemento rispetto agli altri componenti del gruppo, e la *struttura sociometrico del gruppo*, cioè l'organizzazione interna del gruppo (con sottogruppi, persone isolate, relazioni tra i sottogruppi, ecc.). La tecnica è quindi utile sia come strumento di diagnosi individuale che per cogliere la struttura relazionale del gruppo. Tuttavia esso è adatto solo per gruppi strutturati perché è necessario che sia delimitato nettamente il raggio di scelta del soggetto. In questi ultimi tempi questa tecnica ha ripreso piede all'interno della *network analysis* (analisi di rete).

IL CAMPIONAMENTO

Popolazione e campione

Anche se può sembrare strano, la scelta casuale (tipica del campionamento) deve segue ire regole ben precise. Un *campionamento* infatti può essere definito come un procedimento attraverso il quale si estrae, da un insieme di unità (*popolazione*) costituenti l'oggetto di studio, un numero ridotto di casi (*campione*) scelti con criteri tali da consentire la generalizzazione all'intera popolazione dei risultati ottenuti studiando il campione. La rilevazione campionaria presenta i vantaggi per quanto riguarda il costo di rilevazione, i tempi di raccolta dati ed elaborazione, l'organizzazione, l'approfondimento e l'accuratezza.

Errore di campionamento

La tecnica del campionamento presenta tuttavia anche degli svantaggi. Infatti, se l'indagine totale fornisce il valore esatto del parametro che si vuole conoscere, l'indagine campionaria ne fornisce solo una *stima*, cioè un valore approssimato. Ciò significa che il valore in questione non è certo, ma solo probabile, e inoltre questa probabilità può variare entro un certo intervallo (detto *intervallo di fiducia*). La stima del campione sarà quindi sempre affetta da un errore, che si chiama *errore di campionamento*. Se però il campione è *probabilistico* (cioè scelto secondo una procedura rigorosamente casuale), la statistica ci permette di calcolare l'entità di tale errore.

Campioni probabilistici: il campione casuale semplice

Nei campioni probabilistici l'unità d'analisi è estratta con una probabilità nota e diversa da zero. È necessario conoscere la popolazione. Il caso più semplice del campione probabilistico è quello del *campionamento casuale semplice*, in cui ogni individuo della popolazione ha uguali possibilità di essere scelto per il campione. Si devono estrarre gli individui senza riferimento a caratteristiche individuali; si assegna un numero a ciascuna persona e si sceglie a caso.
L'errore di campionamento del campione casuale semplice è direttamente proporzionale al livello di fiducia che vogliamo avere nella stima (cioè il grado di certezza) ed alla variabilità del fenomeno studiato (cioè la dispersione della distribuzione della variabile), mentre è inversamente proporzionale all'ampiezza del campione.
Un altro importante passo è quello della determinazione dell'ampiezza del campione. L'ampiezza del campione è direttamente proporzionale al livello di fiducia desiderato per la stima ed alla variabilità del fenomeno studiato, ed inversamente proporzionale all'errore che il ricercatore è disposto ad accettare. Questo significa che la dimensione della popolazione non ha grande importanza per determinare l'ampiezza del campione, infatti ad esempio un campione di 1.000 casi può essere sufficiente per arrivare a stime della stessa precisione per popolazioni di 10.000 o 100.000 elementi. Al limite, se si desidera avere stime della precisione di due punti percentuali, sono sufficienti 2.500 casi per qualunque dimensione della popolazione, anche a livello mondiale.

Altri campioni probabilistici

Campionamento sistematico: è simile al casuale semplice, ma con diversa tecnica di estrazione. I soggetti si scelgono secondo un intervallo stabilito (uno su *k*). Si usa quando non c'è periodicità e quando la lista non è completa (ad esempio nei controlli di qualità sui prodotti oppure negli *exit polls*). In ogni caso deve essere rispettato il requisito che tutte le unità abbiano la stessa probabilità di essere incluse nel campione e inoltre deve essere evitata ogni forma di scelta diversa da quella predeterminata dall'intervallo di campionamento.
Campionamento stratificato: la popolazione è divisa in strati omogenei rispetto alla variabile e si estrae un campione casuale semplice da ciascuno strato; in seguito si uniscono i campioni dei singoli strati per ottenere il campione finale. Questa procedura richiede che per tutte le unità della popolazione sia nota la variabile posta alla base della stratificazione. Il campione ottenuto può essere *stratificato proporzionale* (se si decide di riprodurre la stessa composizione degli strati nella popolazione) oppure *stratificato non proporzionale* (se si decide di sovrarappresentare alcuni strati e sottorappresentare altri).
Campionamento a stadi: la popolazione è suddivisa su più livelli gerarchicamente ordinati, i quali vengono estratti in successione con un procedimento ad "imbuto". Se presumiamo di avere due stasi, il campionamento si effettua in due momenti: prima si estraggono le *unità primarie* (gruppi di soggetti che costituiscono le unità di analisi vere e proprie) e successivamente si estrae casualmente un campione di *unità secondarie* (le unità di analisi) in ognuna delle unità primarie selezionate dalla prima estrazione. I vantaggi di questa tecnica consistono nel fatto che non è necessario avere la lista di tutta la popolazione, ma solo delle unità primarie; inoltre la rilevazione viene concentrata sulle unità estratte, con notevole riduzione dei costi.
Campionamento per aree: è molto simile al campionamento a stadi e si utilizza quando mancano del tutto i dati sulla popolazione oppure quando le liste sono incomplete.
Campionamento a grappoli: si usa quando la popolazione risulta naturalmente suddivisa in gruppi di unità spazialmente contigue (*grappoli*). Al posto delle unità elementari vengono estratti i grappoli, e poi tutte le unità elementari appartenenti ai grappoli vengono incluse nel campione. Questa tecnica semplifica di molto la rilevazione ed è molto utile quando manca la lista delle unità elementari mentre esiste la possibilità di estrarre con procedura probabilistica i grappoli.
Campioni complessi: sono quelli in cui si utilizzano congiuntamente le tecniche ora presentate.

Il campionamento nella ricerca sociale

L'errore nella ricerca sociale può essere distinto in tre parti: *errore di selezione*, *errore di osservazione* ed *errore di trattamento dati*. La procedura di campionamento produce un errore del primo tipo, che a sua volta può essere distinto in ulteriori tre componenti: *errore di copertura*, *errore di campionamento* ed *errore di trattamento dati*. Finora ci siamo occupati del solo errore di campionamento; tratteremo ora anche gli altri.

Errore di copertura. Lista della popolazione

Nel caso in cui si conosce la lista della popolazione, è possibile procedere con campionamenti probabilistici. Questo accade di solito quando l'oggetto di studio è l'intera popolazione (anche nazionale), perché esistono anagrafi e liste elettorali che forniscono l'elenco completo della popolazione. Il problema si pone per i sottoinsiemi della popolazione, perché di solito non si è in possesso di una lista completa della popolazione. Quando invece l'unità di analisi non è un individuo ma un collettivo, la situazione è migliore perché in genere un aggregato di individui esiste in forma istituzionalizzata e registrata.
Se non c'è la possibilità di conoscere la lista della popolazione bisogna rinunciare a tecniche di campionamento probabilistico, perché in questi casi non è possibile assegnare a tutte le unità della popolazione una certa probabilità di estrazione.
Ma non è sufficiente che le liste esistano, bisogna anche che siano aggiornate, complete ed esenti da duplicazioni. Il problema della completezza è il più grave; in questo caso il ricercatore può ridefinire la popolazione, trascurare gli esclusi oppure procedere ad un'integrazione del campione.

Errore di campionamento. Ampiezza del campione

Se consideriamo il caso di una ricerca *monovariata* (quando si stimano le variabili ad una ad una) la dimensione del campione può essere adeguata, ma se nella stessa ricerca studiamo le *relazioni* tra le variabili (analisi *bivariata* o *multivariata*) l'errore cresce subito fino a livelli inaccettabili. La dimensione ideale del campione dipende dalla distribuzione delle variabili studiate e dal tipo di analisi che si intende fare. In generale l'ampiezza del campione dovrà essere tanto maggiore quanto più il fenomeno da studiare è minoritario.

Errore di non-risposta. Mancati contatti e rifiuti

L'*errore di non-risposta* consiste nel fatto che i soggetti selezionati dal campionamento non sono contattabili o si rifiutano di rispondere. Il problema del mancato contatto con i soggetti può essere causato dalla difficoltà di raggiungerli oppure dalla loro irreperibilità; in ogni caso si tratta di problemi fastidiosi ma risolvibili.
Molto più grave è il problema dei rifiuti a rispondere, in quanto spesso coloro che non vogliono rispondere sono diversi dagli altri e quindi non rappresentano una selezione casuale del campione originario. In questo modo si compromette la validità del campione stesso, che sovrarappresenterà alcune categorie di persone a scapito di altre. La percentuale di mancate risposte in Italia varia dal 20% al 50%, a seconda della diversa forma di contatto utilizzata (ad esempio di solito le interviste faccia a faccia hanno un tasso di risposta superiore a quelle telefoniche).
Una soluzione per rimediare alle mancate risposte può essere quella di sostituire i soggetti con altri scelti a caso, ma questa tecnica spesso non è efficace perché i sostituti assomigliano più ai rispondenti che non ai non rispondenti.
Per contrastare efficacemente il problema delle mancate risposte ci sono due metodi: il primo è quello di tornare il più possibile dalla persone che non rispondono per incontrarle o convincerle; il secondo consiste nella *ponderazione*, cioè nell'attribuire alle persone non raggiunte dall'intervista le risposte medie date dal gruppo sociale al quale esse appartengono.

Campioni non probabilistici

Quando il disegno probabilistico non può essere impostato si ricorre sin dall'inizio ai *campioni non probabilistici*.

Campionamento per quote: si divide la popolazione in strati rilevanti e il ricercatore sceglie *a sua discrezione* i soggetti all'interno degli strati rispettando la proporzione (non c'è casualità). I limiti di questa procedura consistono nel fatto che il ricercatore cerchi i soggetti più facilmente raggiungibili, enfatizzando in questo modo l'errore di non-risposta.

Disegno fattoriale: il disegno fattoriale si colloca a mezza strada tra una tecnica di campionamento e un esperimento. Il suo scopo è quello di cogliere le relazioni che vigono all'interno della popolazione; per far questo i gruppi che si creano dalle combinazioni delle variabili (es.: istruzione, età e genere) hanno tutti dimensione uguale e non proporzionale alla popolazione. Il disegno fattoriale non arreca alcun vantaggio allo studio della relazione tra variabile dipendente e indipendente.

Campionamento a scelta ragionata: in questo caso le unità campionarie non sono scelte in maniera probabilistica, ma sulla base di alcune loro caratteristiche.

Campionamento bilanciato: è una forma di campionamento ragionato, nel quale si selezionano la unità di modo che la media del campione, per determinate variabili, sia prossima alla media della popolazione (deve trattarsi quindi di variabili delle quali sia nota la distribuzione nella popolazione). Esso viene usato soprattutto in caso di campioni molto piccoli.

Campionamento a valanga: è caratterizzato da fasi successive: prima si intervistano le persone che hanno le giuste caratteristiche, da queste si ricevono indicazioni per rintracciare altre persone con le stesse caratteristiche, e così via. Per questo è particolarmente utile in caso di popolazioni clandestine.

Campionamento telefonico: la particolarità di questo campionamento consiste nel fatto che la selezione è fatta automaticamente tramite computer, a partire da elenchi telefonici oppure da numeri generati direttamente dal computer (*random digit dialing*). Questa tecnica presenta il vantaggio che il computer registra i motivi dei mancati contatti e gestisce l'esclusione del numero o la ripetizione della chiamata. Questo tipo di campionamento presenta il difetto che chi vive da solo ha maggiore possibilità di essere estratto di chi vive in una famiglia numerosa.

Campionamento di convenienza: l'unico criterio di questa tecnica è che si scelgono le persone più facilmente accessibili; naturalmente va il più possibile evitato.

Ponderazione

La *ponderazione* è quella procedura con la quale modifichiamo artificialmente la composizione del campione onde renderla più prossima alla distribuzione della popolazione. Essa si realizza attribuendo un "peso" alle unità campionarie che varia a seconda delle loro caratteristiche.
Le procedure di ponderazione sono essenzialmente tre e si basano sulle probabilità di inclusione delle unità nel campione, sulle conoscenze che si hanno sulla popolazione e sulle conoscenze che si hanno sulle non-risposte.
Il caso che si basa sulle probabilità di inclusione delle unità nel campione consiste nel campionare negli strati in modo deliberatamente non proporzionale alla loro presenza nella popolazione per avere un numero di soggetti sufficiente per l'analisi statistica. In questo caso la probabilità di inclusione non è uguale per tutti i soggetti, ma è nota; si resta quindi tra i campioni probabilistici.
Il caso più comune che si basa sulle conoscenze che si hanno sulla popolazione è detto della *post-stratificazione* e consiste nel correggere la distribuzione nella popolazione del campione di alcune variabili in modo da farla corrispondere alla distribuzione della popolazione totale, assegnando a ogni caso un coefficiente di ponderazione (*peso*) pari al rapporto *quota teorica* / *quota rilevata* della categoria di appartenenza. Esso copre l'errore di copertura. In questo caso non siamo più in presenza di campioni probabilistici.
Il caso che si basa sulle conoscenze che si hanno sulle non-risposte copre invece l'errore di non-risposta e consiste nel classificare le persone che si rifiutano di rispondere sulla base di un certo numero di variabili e quindi le risposte raccolte vengono ponderate attribuendo loro un peso che tiene conto dei rifiuti. Lo scopo di questa procedura è quello di attribuire ai non rispondenti il comportamento medio delle persone appartenenti al loro stesso gruppo sociale. Anche in questo caso non si tratta di una tecnica probabilistica.
Un ulteriore intervento, che si usa per attenuare la distorsione prodotta dalla mancata risposta solo a qualche domanda del questionario, consiste nel procedere ad una stima delle mancate risposte a partire dalle informazioni che si hanno sugli intervistati parzialmente reticenti.

Bontà di un campione

Alla validità di un campione concorrono due fattori: la sua *rappresentanza* e la sua *ampiezza*.
Un campione è *rappresentativo* quando fornisce un'immagine in piccolo ma senza distorsioni della popolazione; la rappresentatività dipende dalla *casualità* della procedura con la quale è stato costruito. Questa rappresentatività è valida per tutte le variabili della popolazione. Possiamo infine dire che se le stime del campione sono sufficientemente piccole, il campione è rappresentativo.
Ma è praticamente impossibile realizzare una procedura completamente casuale, per cui la rappresentatività del campione è un obiettivo limite al quale ci si può solo avvicinare minimizzando gli errori di copertura e di non-risposta (*accuratezza*).
In parte l'*ampiezza* del campione è condizione della rappresentatività: se il campione è troppo piccolo, allora l'errore di campionamento è troppo elevato e il campione non può essere definito rappresentativo. In parte invece l'ampiezza del campione è un requisito autonomo dalla rappresentatività, e dipende dal tipo di analisi che vogliamo fare sui dati (monovariata, bivariata o multivariata).
Tra i due requisiti dovrebbe essere privilegiata l'ampiezza per la sua maggiore importanza.
È importante anche la *finalità della ricerca*: se si tratta di uno studio *descrittivo*, il campione deve essere il più possibile rappresentativo, se invece l'obiettivo è di tipo relazionale, il campione può anche non essere perfettamente rappresentativo. In ogni caso il ricercatore può trascurare l'accuratezza della rilevazione, applicando il più possibile il campionamento casuale nonostante la sua difficoltà.

L’ANALISI DEI DATI

L'ANALISI MONOVARIATA

Tipi di variabili e analisi statistica

Le caratteristiche logico-matematiche delle variabili (nominali, ordinali e cardinali) definiscono le procedure da seguire nella fase di analisi dei dati. Le diverse variabili sono quindi analizzate in modo diverso sin dai livelli più elementari. La maggior parte delle tecniche sono state elaborate per le variabili nominali o cardinali, mentre le variabili ordinali dovrebbero essere trattate come nominali perché non è corretto assegnare loro le proprietà delle variabili cardinali. Un caso particolare delle variabili nominali è quello delle cosiddette variabili *dicotomiche*, che hanno la proprietà di poter essere trattate statisticamente come variabili cardinali; per questo talvolta il ricercatore "dicotomizza" variabili a più categorie (*politomiche*).

Matrice dei dati

La matrice dei dati consiste in un insieme rettangolare di numeri, dove in riga abbiamo i *casi* e in colonna le *variabili*; in ogni cella derivante dall'incrocio tra una riga e una colonna abbiamo un *dato*, cioè il valore assunto da una particolare variabile su un particolare caso. Per potere essere organizzate in una matrice, le informazioni devono avere due caratteristiche: l'unità d'analisi deve essere sempre la stessa e su tutti i casi studiati devono essere rilevate le stesse informazioni.
L'operazione di traduzione del materiale empirico grezzo in matrice viene chiamata *codifica* ed avviene con due strumenti, il *tracciato record* (la posizione di ogni variabile nella riga della matrice) e il *codice* (che assegna ad ogni modalità della variabile un valore numerico).
Ogni riga della matrice corrisponde ad un caso (leggendo ogni riga possiamo ottenere il *profilo* di un caso), mentre ogni colonna corrisponde ad una variabile (leggendo una colonna conosciamo le risposte date a quella domanda da tutti gli intervistati).

Distribuzione di frequenza

Distribuzioni assolute e relative

Per dare una rappresentazione sintetica di una colonna della matrice si una la *distribuzione di frequenza*, che è una rappresentazione nella quale ad ogni valore della variabile viene associata la frequenza con la quale esso si presenta nei dati analizzati. La distribuzione di frequenza può essere *assoluta*, quando ci si limita semplicemente a contare i casi che presentano quel valore, oppure *relativa*, quando sono rapportate ad un totale comune. Un modo per operare questa relativizzazione è la *proporzione*, che consiste nel dividere ogni singola frequenza assoluta per il numero totale di casi; più spesso si usa la *percentuale*, che si ottiene dalle proporzioni moltiplicandole per 100. Il fatto di relativizzare le frequenze permette di effettuare dei confronti fra distribuzioni di frequenza della stessa variabile ma ottenute da popolazioni di diversa numerosità.
Una forma particolare di distribuzione di frequenza è costituita dalla *distribuzione cumulativa di frequenza*, nella quale in corrispondenza di ogni valore dalla variabile viene riportata la somma delle frequenze corrispondenti a quel valore e a tutti quelli inferiori.
Se le variabili da sintetizzare sono ordinali, si tende a raggrupparli in *classi* di valori adiacenti, perché spesso sono in numero elevato e altrimenti si otterrebbe una distribuzione troppo dispersa.

La presentazione delle tabelle

Distribuzione di frequenza in forma compatta: il ricercatore deve attenersi al massimo della parsimoniosità nella presentazione dei dati per non confondere il lettore, per cui si limiterà a presentare le percentuali e il totale in valore assoluto (*base* del calcolo delle percentuali)
Cifre decimali: di solito le percentuali si riportano con una cifra decimale oppure senza decimali se la base delle percentuali è minore di 100; questo perché esiste sempre un errore che può essere di diversi punti.
Arrotondamenti: se il decimale da eliminare si colloca tra 0 e 4, si arrotonda per difetto, se si colloca tra 0 e 5 si arrotonda per eccesso.
Il decimale zero: se si decide di riportare i decimali deve essere presente anche lo zero (es. 22,0%)
Quadratura: a causa degli arrotondamenti può succedere che la somma delle percentuali sia diversa da 100; in questo caso è opportuno alterare le cifre per avere delle percentuali la cui somma sia 100.

"Pulizia" dei dati e preparazione del file *di lavoro*

Controlli di plausibilità: si tratta di controllare che tutti i valori delle variabili siano plausibili, appartengano cioè al ventagli di valori previsti dal codice.
Controlli di congruenza: si possono confrontare le distribuzioni di due variabili per far emergere eventuali incongruenze tra le variabili stesse.
*Valori mancanti (*missing values): ad un certo caso in una certa variabile viene assegnato "valore mancante" se quel caso è privo di informazioni su quella variabile. Esistono quattro casi di valore mancante: "non sa", "non applicabile", "non risponde", "valore implausibile". Di solito si tende ad esporre i "non risponde" nell'analisi monovariata e ad escluderli nell'analisi a più variabili.

Analisi monovariata

L'*analisi monovariata* è un'analisi puramente *descrittiva* dei fenomeni studiati, che si limita ad esporre come ogni variabile è distribuita fra i casi rilevati, senza porsi problemi sulle relazioni tra le variabili. Essa rappresenta un passaggio inevitabile e necessario di ogni analisi multivariata, perché solo con questa analisi il ricercatore perviene a quella conoscenza diretta dei dati che gli permetterà di analizzarli con piena consapevolezza. Essa inoltre rappresenta una prima descrizione dei fenomeni analizzati e contribuisce alla comprensione della struttura del campione e della sua rappresentatività.

Misure di tendenza centrale

Le misure di tendenza centrale dicono qual è, in una distribuzione di frequenza, il valore che meglio di qualsiasi altro esprime la distribuzione quando si decidesse di sintetizzarla in un unico numero.
Variabili nominali: la moda. Se la variabile è nominale, l'unica misura di tendenza centrale calcolabile è la *moda*. La moda è la modalità di una variabile che si presenta nella distribuzione con maggior frequenza.
Variabili ordinali: la mediana. Nel caso delle variabili ordinali, oltre alla moda si può calcolare la *mediana*. La mediana è la modalità del caso che si trova al centro della distribuzione dei casi che va dal minore al maggiore (distribuzione *ordinata* dei casi secondo quella variabile).
Variabili cardinali: la media aritmetica. La *media aritmetica* è la misura di tendenza più nota e comune, ed è data dalla somma dei valori assunti dalla variabile su tutti i casi divisa per il numero dei casi. Se nella distribuzione di frequenza i dati sono raggruppati in classi, per il calcolo della media si assume il valore centrale della classe.
La media si può calcolare solo se la variabile è cardinale, in quanto richiede operazioni che possono essere effettuate solo se i valori hanno pieno significato numerico. Tuttavia ci sono dei casi in cui è preferibile usare la mediana anche nel caso di variabili cardinali, tipicamente quando si desidera una misura meno sensibile ai casi estremi (come il reddito medio della popolazione).

Misure di variabilità

Variabili nominali: indici di omogeneità/eterogeneità. Una variabile nominale ha una distribuzione massimamente *omogenea* quando tutti i casi si presentano con la stessa modalità; viceversa è massimamente *eterogenea* quando i casi sono equidistribuiti tra le modalità. Il più semplice indice di omogeneità (*assoluta*) è dato dalla somma dei quadrati delle proporzioni (cioè delle frequenze relativizzate al totale 1). L'indice di omogeneità *relativa* invece neutralizza l'influenza del numero delle modalità.
Variabili ordinali: la differenza interquartile. I *quartili* sono i valori che segnano i confini tra i quattro quarti di una distribuzione ordinata divisa in quattro parti di eguale numerosità. La *differenza interquartile* è la differenza tra il terzo ed il primo quartile; si usa per eliminare il 25% dei valori più alti e il 25% dei valori più bassi. Questa differenza si usa anche per le variabili cardinali.
Variabili cardinali: deviazione standard e varianza. La *deviazione standard* (o *scarto quadratico medio*) consiste nella somma degli scarti dei singoli valori dalla media elevati al quadrato (per annullare il loro segno) sotto radice. Se togliamo la radice otteniamo la *varianza* della distribuzione. Essa costituisce l'oggetto primario di tutta l'analisi dei dati.
Se si vogliono confrontare tra di loro le variabilità di distribuzioni aventi medie fortemente diverse, conviene utilizzare un indice di variabilità che tenga conto del valore della media (*coefficiente di variazione*).
La concentrazione. Quando la variabile è cardinale e consiste in *quantità possedute* dalle unità d'analisi si può calcolare la *concentrazione* di questa variabile nelle unità studiate. La variabile è *equidistribuita* se il suo ammontare complessivo è distribuito in parti uguali tra le unità, mentre è *concentrata* se l'ammontare complessivo è tutto attribuito ad una sola unità. Tipicamente gli indici di concentrazione sono utilizzati per studiare le disuguaglianze nella distribuzione della ricchezza.

Indici di distanza e di dissimilarità

Indici di distanza fra casi

Oltre all'analisi della matrice per variabili è possibile condurre delle analisi sulle righe della matrice, cioè a partire dai casi. La distanza tra i profili di due soggetti (*indice di similarità*) misura quanto i due soggetti hanno dato giudizi complessivamente simili o dissimili. La distanza tra due casi può essere tuttavia calcolata solo su variabili cardinali, oppure su variabili nominali con opportuni artifizi.

Indici di dissimilarità tra distribuzioni

L'obiettivo degli indici di dissimilarità è quello di sintetizzare attraverso un unico numero la distanza che esiste tra due distribuzioni di frequenza della stessa variabile. La distanza o la dissimilarità tra due distribuzioni può essere calcolata solo se esse presentano le stesse modalità.

Classificazioni, tipologie e tassonomie

La *classificazione* è quel processo secondo il quale i casi studiati vengono raggruppati in sottoinsiemi (*classi*) sulla base di loro similarità. Le classi devono essere *esaustive* (tutti i casi devono trovare collocazione in una classe) e *mutualmente esclusive* (un caso può appartenere ad una sola classe).

Classificazione unidimensionale: aggregazione delle modalità in classi

La classificazione più semplice è quella in cui i casi sono classificati per la somiglianza relativamente ad una sola variabile. In questi termini la determinazione delle classi corrisponde con quella delle modalità. Tuttavia, nel caso delle variabili nominali, in fase di analisi dei dati talvolta è necessario *aggregare* alcune modalità dal significato affine per poter procedere all'analisi bivariata. Questo processo spesso costringe il ricercatore a scelte insoddisfacenti e sforzate a causa della difficoltà di aggregare variabili troppo diverse tra loro; di solito quindi nell'analisi bivariata si tende a scartare le componenti troppo esigue e non aggregabili.
Nel caso delle variabili cardinali il problema dell'aggregazione è più semplice, basta raggruppare le modalità in classi di maggiore ampiezza. Esistono tre criteri di aggregazione: il primo consiste nel raggruppare i valori della variabile in intervalli di uguale ampiezza; il secondo nel raggruppare i valori assumendo a riferimento il loro significato; il secondo si basa sulla configurazione della distribuzione di frequenza, prendendo come soglia di divisione i quantili.

Classificazione multidimensionale: tipologie e tassonomie

Le classificazioni multidimensionale (sulla base di più variabili) si possono dividere in *tassonomie* e *tipologie*. La tassonomia è una classificazione nella quale le variabili che la definiscono sono considerate *in successione*, in una struttura gerarchica che procede per variabili di generalità decrescente.
La tipologia è invece una classificazione nella quale le variabili che la definiscono sono considerate *simultaneamente*. Le classi di una tipologia sono dette *tipi*. Un tipo è un concetto il cui significato si colloca all'intersezione dei significati delle modalità delle variabili che costituiscono la tipologia e quindi il suo significato è superiore alla semplice combinazione dei significati delle due variabili. In questo modo la tipologia ha uno scopo *euristico*, cioè non è solo una mera classificazione, ma ha finalità di interpretazione e spiegazione; è un punto cruciale nel collegamento tra dato empirico e teoria.
Quando ci sono in gioco più variabili, il numero dei tipi risulta molto elevato (perché esso è dato dal prodotto delle numero delle modalità che formano le variabili), quindi è necessario procedere ad una riduzione del loro numero mediante unificazione di alcuni tipi in un unico tipo (*riduzione dello spazio degli attributi*).

Trasformazioni delle variabili

La standardizzazione delle variabili

La standardizzazione delle variabili consiste nella trasformazione del valore originario in un valore standard che non risenta della unità di misura della variabile e della dispersione della distribuzione. La standardizzazione consiste nell'ottenere una nuova variabile sottraendo a quella originaria la media della distribuzione e dividendo questa differenza per la deviazione standard della distribuzione. La nuova variabile ha quindi media = 0 e deviazione standard = 1. In questo modo le variabili standardizzate sono tra loro perfettamente confrontabili avendo eliminato le differenze di scala e di dispersione. La standardizzazione consente di anche di confrontare variabili provenienti da diverse distribuzioni.

La costruzione di indici

Un *indice* è una variabile funzione di altre variabili, che sintetizza le informazioni contenute nelle singole variabili operativizzando un concetto complesso del quale le singole variabili sono espressioni parziali. Gli indici si possono costruire in diversi modi, ad esempio l'indice *additivo* è quello ottenuto sommando i punteggi delle singole variabili. Esistono anche altri indici ottenuti attraverso procedure più complesse, con punteggi diversi a domande diverse e anche con punteggi negativi. Le operazioni attraverso le quali si costituiscono le nuove variabili (indici) possono essere di tipo *algebrico* o di tipo *logico*. Spesso alle spalle di questa differenza ci sono diversi tipi di variabile: ad esempio sulle variabili nominali non è possibile effettuare operazioni algebriche.

Dati aggregati

Dati individuali e dati aggregati

Se consideriamo una variabile nominale riferita ad un livello individuale, notiamo che quando l'unità di analisi è un aggregato di individui questa variabile dà luogo a tante variabili cardinali quante sono le sue modalità. Quindi quando l'unità di analisi è un aggregato, la variabile è nella grande maggioranza dei casi cardinale. Naturalmente la variabile così ottenuta deve essere messa in rapporto con la dimensione della popolazione.

Rapporti statistici

Quando ci si trova nella situazione di confrontare fenomeni che fanno riferimento a realtà diverse, nelle quali le quantità assolute dei fenomeni risentono della diversa dimensione degli aggregati o del diverso ammontare dei fenomeni considerati, si pone la necessità di relativizzare le quantità assolute alle rispettive basi di riferimento del fenomeno mediante un rapporto.
Rapporto di composizione: consiste nel rapportare una parte del fenomeno al fenomeno stesso nella sua totalità.
Rapporto di coesistenza: è il rapporto tra due parti, cioè tra la frequenza di una modalità e la frequenza di un'altra.
Rapporto di derivazione: è il rapporto tra la misura del fenomeno e quella di un altro che può essere considerato un suo presupposto necessario.
Rapporti medi: sono diffusissimi e si hanno ogni volta che il fenomeno posto al numeratore si può associare mediamente ad ogni unità posta al denominatore. Sono una sorta di categoria residua nella quale si collocano i rapporti che non ricadono nei casi precedenti. In genere il ricercatore ha ampia possibilità di scegliere cosa mettere al denominatore per rendere i numeratori confrontabili; per questo deve cercare le scelte più ragionevoli per non ottenere risultati fuorvianti.

Serie temporali e serie territoriali: numeri indice

Serie temporali e territoriali

La *serie temporale* (o *serie storica*) è la sequenza dei valori assunti da una variabile nello stesso aggregato territoriale in tempi diversi; la *serie territoriale* è la sequenza dei valori assunti da una variabile nello stesso momento in diversi aggregati territoriali. Pur non essendo distribuzioni di frequenza, a queste serie è possibile applicare molte delle operazioni che si applicano alle distribuzioni di frequenza (tendenza centrale, variabilità, ecc.); di esse si possono dare anche rappresentazioni grafiche. Una rappresentazione grafica molto efficace per le serie territoriali è il *cartogramma*, che raffigura la distribuzione geografica del fenomeno studiato.

Lo studio della variazione: i numeri indice

Differenza assoluta e relativa: la differenza assoluta tra due grandezze omogenee ha un significato diverso a seconda dell'entità delle grandezze stesse. Se invece vogliamo calcolare la variazione relativa tra le due grandezze, dobbiamo fare la differenza tra le due e successivamente dividere per quella che si assume per riferimento. Naturalmente la variazione relativa risente fortemente della base di partenza.
Numeri indice: il numero indice è una proporzione che serve a mettere in luce le variazioni di una serie temporale o territoriale rispetto ad un tempo o a un luogo assunti come base di riferimento. Esso non dipende dall'unità di misura o di conto in cui è espresso (sono cioè *numeri puri*) e permettono quindi di fare confronti con variabili più disparate.

L'ANALISI BIVARIATA

Relazioni tra variabili

Affermare che c'è una relazione tra due o più variabili significa dire che c'è una variazione concomitante tra i loro valori (una *covariazione*). Si tratta di relazioni *statistiche*, ovvero *probabilistiche*; ma la statistica non può dire se esiste effettivamente una relazione *causale* tra le variabili esaminate (covariazione non significa causazione). Sarà il ricercatore a conferire a tale relazione il significato di nesso causale, sulla base di una teoria preesistente che non ha alcun legame con l'analisi statistica.
Esamineremo solo l'analisi *bivariata*, in cui vengono considerate solo le relazioni tra due variabili, dette rispettivamente *dipendente* e *indipendente* in quanto il ricercatore di solito interpreta le relazioni in termini di nessi causali.
Le tecniche di analisi bivariata dipendono in maniera determinante del tipo di variabili considerate. Se entrambe le variabili sono nominali, la tecnica usata sarà quella delle *tavole di contingenza*; se entrambe le variabili sono cardinali la tecnica sarà quella della *regressione-correlazione*; se la variabile indipendente è nominale e quella dipendente cardinale si userà la tecnica dell'*analisi della varianza*.

Tavole di contingenza

Direzione delle percentuali (percentuali di riga e percentuali di colonna)

La *tavola di contingenza* consiste in una tabella a doppia entrata in cui è collocata in riga una variabile (*variabile di riga*) e l'altra in colonna (*variabile di colonna*), mentre nelle *celle* definite dall'incrocio fra le righe e le colonne troviamo il numero di casi che presentano le corrispondenti modalità delle due variabili (*frequenza*). L'*ordine* di una tavola di contingenza è il prodotto delle righe per le colonne, mentre la *dimensione* è il numero di variabili in essa implicate. L'analisi bivariata tratta quindi solo tabelle bidimensionali.
Dalla tabella con i valori assoluti è possibile ricavare tre diverse tabelle percentuali: le *percentuali di riga* (che si ottiene ponendo uguale a 100 la variabile di colonna e registrando quindi i corrispondenti valori percentuali della variabile di riga), le *percentuali di colonna* (che si ottiene ponendo uguale a 100 la variabile di riga e registrando quindi i corrispondenti valori percentuali della variabile di colonna) e le *percentuali sul totale* (che si ottengono percentualizzando tutte le frequenze di cella sul totale generale). Se la tabella è stata costruita per analizzare la relazione tra le due variabili quest'ultima percentualizzazione è inutile. Lo scopo della percentuale è infatti quello di "pareggiare" basi diverse.
È necessario porre molta attenzione nella scelta delle due percentuali rimanenti perché una è corretta mentre l'altra è errata; per compiere la scelta giusta bisogna ricordare che si sceglie la percentuale di colonna quando si vuole analizzare l'influenza che la variabile di colonna ha su quella di riga e viceversa. In altri termini, si definisce qual è la variabile indipendente e si percentualizza all'interno delle sue modalità. Talvolta, quando gli obiettivi sono diversi, può essere utile calcolare invece l'altra percentualizzazione oppure calcolarle entrambe.

Presentazione delle tavole

Gli elementi caratterizzanti di una buona presentazione delle tavole sono cinque.
Parsimoniosità: la tabella deve riportare solo le percentuali che servono all'analisi (es. solo quelle di riga).
Totali: ogni riga o colonna finisce con il totale 100 per far capire immediatamente al lettore in che direzione sono state calcolate le percentuali.
Basi delle percentuali: deve essere sempre riportata la base percentuale, cioè il numero assoluto di casi sui quali è stata operata la percentualizzazione.
Cifre decimali, decimale zero, arrotondamenti, quadratura: valgono le considerazioni già sviluppate riguardo alla presentazione delle distribuzioni di frequenza.
Intestazione: le tabelle devono sempre essere intestate per poter essere autoesplicative.
Somme di percentuali: la somma di percentuali è legittima se i valori sommati appartengono alla stessa distribuzione, ma è errata se le percentuali sommate appartengono a due diverse distribuzioni.

Interpretazione delle tavole

Nell'interpretazione e commento delle tabelle è opportuno selezionare le modalità più significative della variabile dipendente e centrare su queste l'analisi; inoltre è preferibile trascurare differenze percentuali esigue (inferiori ai 5 punti percentuali). Per fare un commento efficace si prende una modalità significativa della variabile dipendente e si vede come essa varia al variare della variabile indipendente. La scelta della modalità da commentare dipende dalla linea argomentativa del ricercatore. Nel caso di variabili ordinali risulta utile aggregare le modalità estreme e contigue della variabile dipendente per una maggiore chiarezza. Un sistema spesso utilizzato per interpretare le tabelle consiste nell'*indice di differenza percentuale*, cioè nella differenza tra due modalità di risposta o tra le risposte positive e negative; esso permette di leggere i dati tenendo conto simultaneamente dell'andamento di più modalità della variabile dipendente.

Presentazione compatta delle tavole

Spesso, per economizzare lo spazio o per facilitare il confronto tra domande aventi la stessa struttura, si compattano più tavole semplici a doppia entrata in un'unica tavola, presentando un'unica modalità. Si possono incrociare diverse variabili dipendenti con la stessa variabile indipendente oppure viceversa.

Tavole di mobilità sociale

Nelle tavole di mobilità sociale su una dimensione si colloca la classe sociale dei soggetti studiati e sull'altra quella dei loro padri. Essa è di particolare importanza perché offre molteplici linee di lettura. Iniziando dalle celle, poiché le due variabili (classe sociale padre e classe sociale figlio) hanno le stesse modalità, sulla diagonale si trovano i soggetti immobili, mentre nel triangolo superiore alla diagonale ci sono i soggetti che hanno sperimentato un processo di mobilità ascendente e nel triangolo sotto alla diagonale ci sono invece i soggetti che hanno sperimentato un processo di mobilità discendente.
In questo caso inoltre tutte e tre le forme di percentualizzazione assumono un significato: le percentuali entro le modalità della variabile indipendente ci dicono qual è l'influenza della classe sociale di partenza su quella di arrivo, le percentuali per riga ci danno informazioni sull'origine sociale dei ceti attuali e infine la percentualizzazione sul totale ci dà informazione sul processo generale di mobilità sociale. Lo stesso approccio viene impiegato nelle *tavole di movimento elettorale*.

Significatività della relazione tra due variabili nominali: il test del chi-quadrato

Il *test del chi-quadrato* è un criterio oggettivo sulla base del quale è possibile dire che tra due variabili esiste o meno una relazione. Il test del chi-quadrato si basa sulla *falsificazione*, cioè si assume che non esista alcuna relazione tra le due variabili e si cerca di dimostrare che questa affermazione è falsa. Se la dimostrazione riesce, resta un'unica alternativa disponibile, cioè che tra le due variabili *esista* effettivamente una relazione. In particolare, con il test del chi-quadrato si costruisce una tavola ipotetica che rappresenta le frequenze che ci si aspetterebbe in caso di assenza della relazione. In seguito si calcolano le frequenze effettivamente trovate nei dati e si trova la differenza tra le frequenze attese e quelle osservate. Se la differenza è sufficientemente grande si accetta l'ipotesi di esistenza di una relazione. Il chi-quadrato è un indice che misura la distanza tra le tabelle delle frequenze osservate e quelle delle frequenze attese: più grande è il suo valore, maggiore è la differenza. Per convenzione si respinge l'ipotesi di indipendenza se il valore del chi-quadrato è così grande da avere solo il 5% o meno di possibilità di essere dovuto al caso, mentre ha il 95% di possibilità di essere causato da una relazione tra le variabili.
Se il campione è costituito da pochi casi, si può respingere l'ipotesi sottoposta a verifica solo se i risultati sono fortemente indicativi; mentre se il campione è molto ampio, anche piccole differenze possono essere considerate significative. Questo fatto è evidente nel chi-quadrato: il valore del chi-quadrato dipende fortemente della numerosità del campione.
Il test del chi-quadrato, inoltre, essendo basato sullo scarto tra frequenze attese e frequenze osservate può risultare significativo anche solo per l'anomalia di un'unica cella, che presenta valori fortemente devianti rispetto al valore atteso: per questo è necessario sempre ispezionare attentamente la tabella.

Misura della forza della relazione tra variabili nominali e ordinali

Misure di associazione tra variabili nominali

Le misure di associazione servono a valutare la *forza* (o *intensità*) di una relazione fra variabili nominali. Esse possono essere basate sul *chi-quadrato* oppure sulla *riduzione proporzionale dell'errore*.
Misure di associazione basate sul chi-quadrato: si può prendere direttamente come misura della forza di una relazione il chi-quadrato (maggiore è il chi-quadrato, maggiore è la forza) a condizione che le due tabelle presentino lo stesso numero di casi. Se le tabelle hanno numeri di casi differenti si utilizza l'indice Φ (*phi*), che si ottiene dividendo il chi-quadrato per il numero dei casi. Siccome questo valore non varia da 0 o 1 (non è cioè normalizzato), è stato introdotto l'indice *V* (*V di Cramèr*) che invece varia da 0 (indipendenza) a 1 (relazione perfetta).
Caso particolare della tabella 2x2: in questo caso gli indici Φ e V coincidono, e coincidono anche con il *coefficiente di correlazione r di Pearson*, che è l'indice di forza della relazione da utilizzare quando entrambe le variabili sono cardinali e si può utilizzare anche quando le variabili sono dicotomiche.
Misure di associazione basate sulla riduzione proporzionale dell'errore: queste misure si basano sulla riduzione di errore che si fa nel predire una variabile conoscendo il valore dell'altra, rispetto agli errori che isi farebbero se si fosse privi di tale conoscenza. La misura di associazione corrisponde alla proporzione di riduzione degli errori di previsione nel calcolare per le unità di analisi il valore di *Y* conoscendo il valore che esse hanno su *X*.

Misure di cograduazione tra variabili ordinali

Nel caso delle variabili ordinali diventa importante anche il *segno* della relazione, vale a dire se a valori alti di una variabile tendono a corrispondere valori alti anche dell'altra variabile la relazione si dice *positiva* (o *diretta*), mentre se a valori alti dell'una corrispondono valori bassi dell'altra si dice *negativa* (o *inversa*).
Le misure di *cograduazione* (specifiche delle variabili ordinali) si basano sul confronto tra i valori assunti dalle variabili *X* e *Y* su tutte le possibili coppie di casi. Una coppia di casi è detta *concordante* se su un caso i valori di *X* e *Y* sono entrambi maggiori (o minori) dei valori delle stesse variabili sull'altro caso, mentre è detta *discordante* se una variabile assume su un caso un valore maggiore mentre l'altra un valore minore rispetto ai valori assunti sull'altro caso. Se la maggioranza delle coppie sono concordanti (o discordanti) c'è una relazione (cograduazione) tra le due variabili (positiva se concordanti, negativa se discordanti); se ci sono tante coppie concordanti quante discordanti allora non c'è cograduazione.

Conclusioni su misure di associazione e di cograduazione

Le misure di forza delle relazioni tra variabili nominali e ordinali non sono molto utilizzate per tre motivi. In primo luogo, non esiste un unico indice standard; secondariamente, la forza della relazione può essere dovuta al comportamento specifico di alcune modalità; infine queste misure sono di difficile interpretazione. È quindi molto importante leggere attentamente la tabella anche per interpretarne la forza.

Rapporti di probabilità

Il *rapporto di probabilità* (*odds*) è il rapporto tra la frequenza di una categoria e la frequenza della categoria alternativa (nel caso delle variabili dicotomiche) e si indica con la lettera ω. Esso è anche definibile come il rapporto tra la probabilità che un individuo, estratto a caso dell'universo, appartenga ad una categoria della variabile considerata e la probabilità che non vi appartenga. Il rapporto di probabilità assume il valore 1 quando le due categorie della variabile hanno lo stesso peso (equivalente alla proporzione di 0,5 per entrambe); ha come valore minimo 0 e come valore superiore ha +∞.
Il rapporto di probabilità può essere esteso anche al caso di due variabili; in questo caso si usa il *rapporto si associazione*, che varia da 0 a +∞, passando per il valore 1 che si verifica nel caso di indipendenza tra le due variabili. Maggiore è la distanza da 1, maggiore è la forza della relazione. Valori superiori a 1 indicano una associazione *positiva*, mentre valori inferiori a 1 una associazione *negativa*.
Il rapporto di associazione non risente delle dimensioni del campione è cambia se entrambe le frequenze di una riga o di una colonna sono moltiplicate per una costante. Questa stabilità è utile per poter cogliere la struttura della relazione tra due variabili senza risentire delle variazioni campionarie.

Analisi della varianza

Principi e calcoli

L'analisi della varianza serve a studiare la relazione tra una variabile nominale e una cardinale. Se abbiamo la variabile cardinale dipendente, possiamo notare che essa varia tra i casi (*varianza*). La varianza può essere suddivisa in due componenti: quella parte dovuta alla variabilità del fenomeno *entro i gruppi* (*devianza non spiegata* o *interna*) e quella parte dovuta alla variabilità del fenomeno *tra i gruppi* (*devianza spiegata* o *esterna*). Si dice *spiegata* perché la spiegazione di quella parte della variabilità della variabile dipendente che è attribuibile alla variabile indipendente. Il diverso peso relativo di devianza interna ed esterna può essere utilizzato per valutare la significatività e la forza della relazione.