ROBERTO ALBANO, TANIA PARISI
Dipartimento Culture, Politica e Società
Università degli studi di Torino

IL CAMPIONAMENTO NELLA RICERCA QUANTITATIVA

Data di revisione: ottobre 2015

Molta parte della ricerca sociale è svolta su campioni di popolazione, poiché per varie ragioni risulterebbe non conveniente, se non impossibile, effettuare la rilevazione sull'intera popolazione che si vuole indagare[1]. Ad esempio, l'inchiesta campionaria, più nota con il termine inglese *survey*, consiste in una serie di interviste con questionario condotte su un sottoinsieme di una popolazione dai contorni definiti a priori, la popolazione bersaglio (o *target*). Ma l'estrazione di campioni può riguardare anche altre unità di rilevazione e analisi: si possono campionare comuni, sezioni elettorali, documenti amministrativi, spot commerciali, immagini tratte dal web, e molto altro ancora. È solo per semplicità che di seguito faremo principalmente riferimento nei nostri esempi al campionamento di singoli individui, certamente una situazione molto frequente nella ricerca sociale.

Per iniziare, supponiamo di voler studiare mediante interviste con questionario la condizione dei giovani cittadini italiani, residenti e compresi nella fascia di età 16-29 anni, estraendo a tal fine un campione di individui. I soggetti potenzialmente intervistabili della popolazione così definita (16-29enni in possesso della cittadinanza italiana, iscritti all'anagrafe) costituiscono le unità di rilevazione; i soggetti che vengono effettivamente selezionati, un numero evidentemente ristretto rispetto alla popolazione, costituiscono i casi. Questi, vengono selezionati con apposite procedure che si possono suddividere in due grandi famiglie: le procedure di campionamento probabilistico e quelle di campionamento non probabilistico.

Le prime sono indispensabili quando si voglia applicare l'inferenza statistica, ossia generalizzare i risultati ottenuti analizzando i dati quantitativi del campione alla popolazione da cui esso è stato estratto[2]. Nell'indagine campionaria si parla spesso di campione rappresentativo per indicare il campione ottenuto con metodi probabilistici. Va chiarito comunque che la rappresentatività non è una proprietà del singolo campione estratto casualmente: questo, in linea di principio, può anche essere molto diverso dalla popolazione da cui è estratto[3], anche se si cerca di evitare, con vari accorgimenti e correttivi, che il caso sia, per così dire, "maligno". La rappresentatività è invece una proprietà del disegno di campionamento: estratti casualmente molti campioni di uguale numerosità da una popolazione invariata, quelli che fotografano meglio la popolazione risulteranno più frequenti di quelli che deviano da essa in modo consistente[4].

---

[1] La popolazione, o universo, è il complesso delle unità statistiche le cui caratteristiche, o proprietà, sono oggetto di interesse della ricerca.

[2] Non è facile stabilire con esattezza quando nacque l'idea di studiare un campione casuale per ricavare informazioni su un insieme più ampio nell'ambito delle scienze umane e sociali. Quel che è certo è che la questione della rappresentatività di un campione, e quindi della generalizzabilità dei risultati conseguiti mediante indagine campionaria, fu trattata sistematicamente dall'Istituto Internazionale di Statistica (IIS) a partire dal 1895. Una prima teoria sistematica del campionamento probabilistico in indagini tipo social survey fu esposta dall'economista Arthur Bowley nel 1926 in un saggio di 62 pagine pubblicato nel bollettino dell'IIS.

[3] Per fare un esempio semplice, se da un sacchetto contenente 2 biglie nere e 2 biglie bianche, ne estraiamo casualmente 2 non ci aspettiamo di averne con certezza 1 bianca e 1 nera, ma questo è il risultato più probabile. Facciamo il caso del campionamento senza ripetizione: i distinti campioni estraibili di numerosità pari a 2, considerando anche l'ordine di estrazione, sono 12; otto di essi, senza riguardo all'ordine di estrazione, sono rappresentativi, in quanto composti da una pallina bianca e una nera. È comunque possibile che il campione estratto sia formato da due palline nere (due possibilità su 12) o da due palline bianche (idem): è dunque possibile estrarre un campione non rappresentativo, anche se è meno probabile (quattro possibilità su 12).

[4] Riprendendo l'esempio della nota precedente: i campioni formati da una pallina bianca e una nera, senza riguardo all'ordine di estrazione, si presenteranno in un grande numero di estrazioni con una frequenza prossima alla probabilità teorica che è di 8/12 (espressa in percentuale: circa il 67%); il grado di approssimazione della frequenza alla probabilità teorica sarà tanto maggiore quanti più campioni si

La ricerca che non impiega la statistica inferenziale non ha bisogno di campioni casuali e, di conseguenza, fa ricorso a procedure di selezione dei casi non probabilistiche. Tuttavia, va anche detto che nella ricerca quantitativa talvolta si fa un uso para-probabilistico di procedure che non prevedono l'estrazione casuale dei casi.

Le principali procedure di campionamento probabilistico

Cominciamo col dire che in un campione genuinamente probabilistico:

- ciascuna delle unità della popolazione ha una probabilità nota e diversa da zero di essere inclusa nel campione;
- i meccanismi di distorsione sono ridotti al minimo; per vari motivi che non sono controllabili dai ricercatori, la scelta non guidata dal caso può ricadere più facilmente su unità di un certo tipo che su quelle di altro tipo.

Una procedura di campionamento è detta probabilistica, o casuale (da non confondere con causale), se, fissata la numerosità, tutti i campioni che è possibile estrarre hanno ciascuno la stessa probabilità degli altri di essere scelti.

- *Campionamento casuale semplice; con ripetizione e senza ripetizione*

La più elementare delle tecniche di campionamento probabilistico è il campionamento casuale semplice, in cui ciascuna unità ha la stessa probabilità delle altre di essere inclusa nel campione[5].

Se i casi già estratti non vengono nuovamente messi in ballottaggio allora si parla di campionamento casuale semplice senza ripetizione; in caso contrario avremo un campionamento casuale semplice con ripetizione, detto anche campionamento bernoulliano.

Molti test di inferenza statistica impiegati nella ricerca sociale si basano sul fatto che gli eventi siano indipendenti, e che perciò nel calcolare le loro probabilità non si debba fare ricorso alle probabilità condizionate.

A rigore il campionamento senza ripetizione non soddisfa appieno l'ipotesi di indipendenza, cosa che invece vale per quello con ripetizione.

Tuttavia, se la numerosità della popolazione (N) è molto grande rispetto alle dimensioni del campione (n), cioè se la frazione di campionamento ($f = n/N$) è inferiore a una certa soglia - spesso individuata in 0,05 - è possibile trascurare le differenze esistenti tra i due tipi di campionamento e sfruttare la maggiore semplicità del primo metodo.

Se si fa ricorso al campionamento casuale semplice senza ripetizione quando il campione è relativamente grande rispetto alla popolazione, occorre applicare correttivi nelle formule degli stimatori di alcuni parametri (Herzel, 1991, p. 630) e nelle formule per la determinazione stessa della numerosità campionaria minima.

Per effettuare un campionamento casuale occorre partire da una lista di campionamento, ossia un elenco completo e non ridondante degli elementi che compongono la popolazione. Se

---

estrarranno. Naturalmente, questi concetti esposti qui in modo informale possono essere formalizzati, come avviene nei testi dedicati alla teoria della probabilità.

[5] Con tale espressione si intende che le unità singolarmente prese devono avere la stessa probabilità di estrazione, e non le unità di un tipo rispetto a quelle di un altro tipo; se in una scatola ci sono cinque sfere di cui 3 bianche e 2 rosse, la probabilità di estrarne una di tipo bianca è maggiore di quella di estrarne 1 di tipo rosso; ma ogni unità fisica deve avere la stessa opportunità di essere scelta.

manca la corrispondenza biunivoca tra popolazione di riferimento e lista di campionamento, si incorre nel cosiddetto "errore di copertura"[6].
Sono esempi di liste di campionamento: i registri demografici della popolazione residente (compilati e conservati dai Comuni); gli elenchi delle imprese legalmente operanti su un territorio compilati e conservati dalle Camere di commercio; il pubblico registro automobilistico; l'elenco di tutte le scuole pubbliche di ogni ordine e grado detenuto dai Provveditorati agli studi ecc.
L'elenco telefonico è stato usato in passato come lista di campionamento; tuttavia, molte famiglie, e in numero crescente, non desiderano comparire nell'elenco degli abbonati[7] e la stessa sottoscrizione di abbonamenti di telefonia fissa è in calo, nel nostro come in altri paesi. Per giunta queste famiglie appartengono più frequentemente a particolari categorie sociali che non ad altre: confrontando le persone raggiungibili via telefono fisso con quelle irraggiungibili si è dimostrata, almeno in Italia, l'esistenza di "pattern sistematici nell'esclusione di strati sociali specifici dalle inchieste telefoniche" (Callegaro Poggio, 2004, p. 498). In altre parole, la popolazione raggiungibile (famiglie negli elenchi della telefonia fissa) non è rappresentativa della popolazione obiettivo (famiglie residenti sul territorio In pratica, la completezza è un concetto limite: anche i registri anagrafici che a prima vista potrebbero sembrare liste complete della popolazione residente, non lo sono di fatto a causa dei ritardi nelle registrazioni, delle dichiarazioni false e di comodo ecc.
Il modo in cui è stato costruito il campione e le modalità di conduzione dell'indagine vanno comunque sempre esplicitati nei rapporti di ricerca, in modo da rendere consapevole il lettore di quale sia l'effettiva popolazione a cui si fa riferimento.
Per esempio, in un'inchiesta condotta sui 14-19enni di una certa regione, un campione può risultare rappresentativo per i giovani che frequentano la scuola, in quanto il campionamento è stato effettuato sui frequentanti di tutte le scuole di quel territorio, ma sarebbe scorretto generalizzare i risultati a tutti i giovani lì residenti, compresi cioè quelli che non frequentano più la scuola.
Va osservato che spesso è il comportamento del soggetto prescelto a far parte del campione a impedire la piena realizzazione di un campione probabilistico.
Se si desidera condurre un'indagine sull'accettazione della violenza come strumento risolutivo dei conflitti, è molto probabile che lo strumento della survey non riesca a cogliere gli atteggiamenti di coloro che per esempio fanno parte di gang o persino di gruppi violenti organizzati (per le loro condizioni di clandestinità, per il loro antagonismo verso la 'cultura dominante' e le sue agenzie, per la paura di essere schedati ecc.). Ma anche quando il tema indagato non è così sensibile, si rileva – come verrà mostrato più avanti - una differente disponibilità a farsi intervistare, che è legata oltre che alle propensioni individuali, anche a caratteristiche socio-demografiche e di posizione nella stratificazione sociale degli individui (Marradi 1997; Urigh 2008). Una volta stabiliti i confini della popolazione di riferimento, e identificata una lista di campionamento completa, occorre mettere in atto una procedura di estrazione che sia effettivamente casuale.

[6] L'errore di copertura può essere di due tipi: si parla di 'sotttocopertura' quando la lista di campionamento esclude parte della popolazione di riferimento; si parla invece di 'sovracopertura' quando la lista di campionamento include unità che non fanno parte della popolazione di riferimento. Per una rassegna degli errori non campionari, si rimanda a Groves (1989).
[7] Il problema del mancato inserimento negli elenchi telefonici delle persone in possesso di apparecchio fisso viene evitato adottando software che compongono in modo casuale numeri telefonici (RDD, *random digit dialing)*.

A ogni elemento, che deve comparire una sola volta nella lista (condizione di non ridondanza), va attribuito un numero (puntatore); un dispositivo generatore di numeri casuali (che si tratti di tavole di numeri casuali, di dispositivi meccanici o di procedure stocastiche informatizzate), che operi ovviamente nel *range* di numeri usati come puntatori, fornirà i numeri per scegliere gli elementi che vanno a costituire il campione.

- *Campionamento Sistematico*

Una approssimazione pratica del campionamento casuale può essere ottenuta con il cosiddetto campionamento sistematico.
Data una lista di N unità numerate progressivamente e scelta la dimensione n del campione, si sceglie un caso ogni k (ad es. ogni 10, ogni 20, ogni 35 ecc.).
k = N/n è detto intervallo di campionamento. È importante che il punto di partenza sia selezionato in modo casuale: se per esempio iniziassimo sempre dalla prima unità della lista, questa avrebbe probabilità di essere estratta pari a 1, e l'unità seguente avrebbe probabilità nulla di essere estratta, per cui verrebbe meno il principio di equipossibilità. Il punto di inizio r, è un numero minore o uguale a k.
Il campione conterrà così le seguenti unità:
{r, r+k, r+2·k....., r+(n-1)·k}.

Ad esempio, se la lista è di 150 unità (N), il campione desiderato è di 30 (n), l'intervallo di campionamento (k) è 5; se il punto di inizio (r) è il caso numero 4, si inseriranno nel campione i casi: {4 , 9 , 14, 19 , …, 149}.

Questo metodo non dà comunque uguale probabilità di estrazione a ogni campione. Per esempio le combinazioni che comprendono una unità e la successiva non hanno alcuna possibilità di essere estratte.

- *Campionamento Stratificato, proporzionale e non proporzionale*

Nel campionamento casuale stratificato, la popolazione viene prima di tutto partizionata (suddivisa) in strati, ossia sotto-popolazioni omogenee rispetto a una o più variabili. Per esempio, se la variabile stratificante è il genere, la popolazione è preliminarmente suddivisa in due strati, corrispondenti alle due modalità della variabile, lo strato dei maschi e quello delle femmine; se vi è più di una variabile stratificante, gli strati vengono individuati da tutte le combinazioni possibili tra le modalità delle diverse variabili. In una seconda fase si procede a scegliere un campione casuale in ciascuno degli strati.
La numerosità campionaria di norma è proporzionale alla numerosità di ogni strato: si parla in tal caso di campione autoponderato. Talvolta, invece, ciò non è opportuno e/o possibile, per svariate ragioni: ad esempio, può essere necessario sovrarappresentare nel campione particolari strati minoritari al fine di avere un numero adeguato di casi sui cui fare analisi approfondite. In un caso come questo si ha allora un campione stratificato non proporzionale[8].

[8] Un caso speciale di quest'ultimo è dato dal 'piano di ripartizione ottimale': esso consiste nella scelta di una quota per ogni strato direttamente proporzionale agli scarti quadratici medi della variabile in ciascuno strato (ovviamente ciò comporta una stima in anticipo della variabilità interna a ogni strato). La ratio è che se una variabile in uno strato si presenta con minor dispersione tra le varie modalità non occorre campionare tanti casi come invece è necessario in quegli strati in cui il fenomeno si presenta con maggior variabilità.

Se il campione non è autoponderato, in sede di analisi dei dati occorrerà effettuare un'operazione di ponderazione mediante opportuni pesi attribuiti ai casi a seconda dello strato a cui appartengono per ristabilire la rappresentatività del campione rispetto alla popolazione nel suo complesso.
Il campionamento stratificato è in genere più efficiente rispetto a quello casuale semplice, ossia permette, a parità di precisione richiesta nella stima dei parametri e di grado di fiducia nella stima desiderato, di costruire campioni più piccoli di quelli ottenuti con il campionamento casuale semplice. Ciò è vero solo se le variabili stratificanti sono in qualche misura correlate con le variabili oggetto di indagine.
Un altro obiettivo per cui si può decidere di ricorrere alla stratificazione, è quello di garantire che siano presenti in numero sufficiente nel campione rappresentanti di ogni strato – anche di quelli minoritari – per effettuare stime per ciascun strato, senza per questo dover aumentare la dimensione totale del campione estratto.

- *Campionamento a grappoli*

Nel campionamento a grappoli, le unità di campionamento si presentano al ricercatore riunite in gruppi detti appunto grappoli (in inglese sono detti *cluster*). Contrariamente alle procedure esaminate in precedenza, con questo tipo di campionamento non si selezionano singoli casi dalla lista di campionamento, bensì gruppi di casi. Scelti i grappoli con una procedura casuale, tutti i casi compresi al loro interno vengono inseriti nel campione, a meno che non siano previsti ulteriori stadi di campionamento. Di solito questo metodo è usato quando è impossibile o è troppo costoso costruire una lista di individui, oppure nell'analisi ecologica dei dati.
Affinché questo tipo di campionamento funzioni bene, producendo stime efficienti, è necessario che i grappoli siano internamente il più possibile eterogenei e che tra i grappoli non vi siano troppe differenze. Tale assunto però non sempre è rispettato. Ad esempio, in alcuni casi i grappoli sono rappresentati da unità amministrative, come le sezioni di censimento individuate dall'Istat (tipicamente costituite da isolati di caseggiati in un Comune ma anche altre perimetrazioni). Si parla in tal caso di campionamento per aree. Questa però è una procedura che potrebbe portare alla violazione dell'assunto precedente: infatti alcune categorie di soggetti potrebbero essere concentrate in alcune aree della città, cioè i grappoli potrebbero non rispondere al carattere dell'eterogeneità. Se si hanno delle spie di violazione di tale assunto si possono adottare dei correttivi, ad esempio stratificando preventivamente i grappoli e estraendo un congruo numero per ciascuno strato (v. punto successivo).

- *Campionamento a stadi*

Soprattutto quando si applica il campionamento a grappoli, ma non solo, può accadere che il numero degli individui appartenenti ad ogni grappolo sia troppo elevato. Da questo insieme di unità primarie occorre cioè estrarre un ulteriore campione di unità secondarie, per esempio con l'estrazione casuale semplice. Nel campionamento a stadi, questi ultimi possono essere anche più di due, e si possono combinare metodi di campionamento diversi.
Consideriamo per esempio il campionamento che le società di ricerca compiono per il sondaggio delle intenzioni di voto. A un primo stadio vengono estratti i Comuni in cui effettuare le interviste. Questi rappresentano i punti di campionamento primario. Tale scelta riduce lo spazio geografico su cui mobilitare gli intervistatori. Di solito si estraggono i grappoli da strati costituiti dai comuni di diversa ampiezza e di diversa macro-area geografica.
In un secondo stadio si procede all'estrazione di sezioni elettorali che rappresentano i punti di campionamento secondario. Infine si procede, in un terzo stadio, alla selezione casuale semplice di singoli intervistati dalle diverse sezioni.

Determinazione dell'ampiezza dei campioni casuali (*)[9]

Quanto deve essere grande il campione estratto secondo modalità aleatorie? A tale domanda si possono dare risposte diverse a seconda degli obiettivi conoscitivi dell'indagine, del tipo di analisi che si intende applicare, nonché del disegno di campionamento utilizzato. In generale vale però il principio di economicità: poiché la ricerca ha dei costi non indifferenti, alcuni dei quali fissi ma altri variabili in funzione del numero dei casi esaminati, è opportuno che la numerosità campionaria sia quella minima necessaria.

Il problema di individuare delle formulazioni matematiche di determinazione esatta della numerosità campionaria minima è stato studiato soprattutto in riferimento alla statistica inferenziale. Nei corsi introduttivi di statistica e metodologia della ricerca l'attenzione va quasi sempre esclusivamente alle formule utilizzabili nel caso della stima intervallare di un parametro relativo a una singola variabile. Anche noi vi faremo riferimento, avvertendo però che chi fa ricerca empirica nelle scienze sociali si trova spesso di fronte a situazioni più complesse; tanto per fare alcuni esempi: la stima intervallare della correlazione tra due o più variabili; l'applicazione di tecniche multivariate che richiedono un numero minimo di casi, di solito dipendente dal numero di variabili trattate; la presenza di dati incompleti. Purtroppo in casi come questi la trattazione del tema si complica notevolmente, oppure non trova nella letteratura una base teorica e sono risolvibili solo sulla base di una certa esperienza di ricerca.

Consideriamo solo un caso elementare, quello della stima della media di una variabile nella popolazione attraverso un campione casuale semplice con ripetizione.

Se il campionamento non è affetto da errori sistematici, problema su cui torneremo successivamente, otterremo una statistica campionaria $\bar{x}$ tale che:

$$\bar{x} = \mu \pm e$$

dove $\bar{x}$ è la media campionaria, *e* è l'errore casuale o **errore di campionamento**, μ è il vero valore (*true score*), cioè il parametro della popolazione. Sappiamo che l'errore standard $(\sigma/\sqrt{n})$ della distribuzione campionaria delle medie diminuisce all'aumentare della numerosità campionaria: pertanto a parità di condizioni, un campione più grande di un altro fornisce rispetto a quest'ultimo una stima del parametro più accurata.

D'altro canto occorre tenere conto del fatto che un campione più grande comporta costi più elevati: pertanto le dimensioni del campione vengono di solito concretamente determinate tenendo conto sia del livello di accuratezza di stima desiderato, sia in funzione del costo complessivo dell'indagine.

Fissiamo a priori il livello di accuratezza desiderato e chiediamoci quale numerosità campionaria minima occorra per ottenere tale accuratezza.

E' necessario fissare a priori il livello di fiducia dell'errore che si è disposti ad accettare.

Poniamo di volere costruire un campione casuale semplice che ci fornisca una stima della media della popolazione con un errore $e = \pm 0{,}1$ rispetto al parametro reale, con un livello di fiducia prescelto, per esempio il 95% dei casi.

Ciò significa che se estraessimo infiniti campioni, 95 volte su 100 le stime ottenute varierebbero in un intervallo non superiore a 0,2 attorno al parametro vero.

Sappiamo che in una distribuzione normale il 95% dei casi è compreso in un intervallo di $\pm 1{,}96\sigma$ (la stima della media cioè, avrà una probabilità 0,95 di differire al massimo di $\pm 1{,}96\sigma$

[9] (*) La lettura di questo paragrafo può eventualmente essere omessa perché presuppone la conoscenza di concetti della statistica inferenziale.

rispetto al valore corretto); perciò l'errore standard della distribuzione campionaria, $\sigma/\sqrt{n}$, moltiplicata per 1,96 deve essere uguale alla precisione voluta (nel nostro caso 0,2). Svolgendo i passaggi algebrici calcoliamo un *n* che soddisfa questo vincolo:

$$e = 1.96\sigma / \sqrt{n};$$
$$\sqrt{n} = 1{,}96\sigma / e;$$
$$n = (1{,}96\sigma / e)^2$$

Supponiamo che uno psicologo voglia stimare il tempo di reazione medio a uno stimolo; da precedenti ricerche si sa che esso potrebbe essere intorno ai 3 secondi, con una deviazione standard pari a 0,8.
Si desidera effettuare una stima con un errore pari a 0,2 a un livello di confidenza del 95%. Applicando ai dati la formula precedente si ottiene:

$$n = (1{,}96 \cdot 0{,}8 / 0{,}2)^2$$
$$n = 61$$

Si sarà probabilmente notato che nella formula di calcolo di n non compare come parametro l'ampiezza della popolazione (N). Ciò ha come utile conseguenza che, anche nel peggiore dei casi (cioè con una deviazione standard molto elevata), con poche migliaia di casi – diciamo due o tremila – si possono fare stime molto accurate, a un livello di confidenza elevato, per popolazioni di qualsiasi dimensione.
Tuttavia, se il rapporto tra n campionario e N, numerosità della popolazione, è superiore ad un quinto e il campionamento è senza ripetizione, l'errore campionario va calcolato tenendo conto della frazione di campionamento. La formula di calcolo di n allora si ottiene con i passaggi seguenti:

$$e = 1{,}96 \frac{\sigma}{\sqrt{n}} \left(\sqrt{\frac{N-n}{N-1}} \right);$$
$$e^2 (N-1) = 1{,}96^2 \frac{\sigma^2}{n} (N-n);$$
$$e^2 (N-1) + 1{,}96^2 \sigma^2 = \frac{1{,}96^2 \sigma^2 N}{n}$$
$$n = \frac{1{,}96^2 N \sigma^2}{e^2 (N-1) + 1{,}96^2 \sigma^2}$$

Si noterà che in entrambi i casi le equazioni non sono ancora risolvibili in quanto noi non conosciamo il valore di σ.
Sarà perciò necessario procedere a una sua stima, basandosi su ricerche precedenti (sulla stessa popolazione o su popolazioni ritenute simili), oppure su uno studio pilota, o ancora basandosi sul parere di testimoni privilegiati e esperti. In questo secondo caso, qualora gli esperti non sappiano indicare tale grandezza, possiamo chiedere quali sono secondo loro i valori minimi e massimi della variabile presa in questione; al posto di σ^2 utilizzeremo allora come misura di variabilità il campo di variazione, cioè la differenza tra il valore massimo e quello minimo.

In generale converrà aumentare la deviazione standard stimata per sicurezza, anche se ciò avrà lo svantaggio di aumentare la numerosità campionaria (e quindi il costo del campionamento).
Se si ha a che fare con una stima di proporzioni, la scelta a priori di σ è più semplice; sappiamo che s in questo caso è pari a $\sqrt{p \cdot q}$; questo valore è massimo (0,25) quando p = q. In assenza di altre informazioni fissiamo proprio questo valore di σ.
Poiché spesso intendiamo studiare più variabili, è opportuno che la scelta del numero di casi da campionare sia effettuata in base alla variabile che presenta una maggiore eterogeneità. In una *survey* con 50 domande o anche più ciò può risultare alquanto laborioso.
Ma una complicazione maggiore è rappresentata dal fatto che nella ricerca sociale si desidera di solito operare incrociando la variabile oggetto di stima con altre variabili di interesse. Tornando all'esempio precedente, potremmo essere interessati a stimare, con lo stesso livello di precisione e lo stesso intervallo di confidenza, il tempo di reazione medio a uno stimolo nel sottogruppo dei maschi e delle femmine. Ma ciò evidentemente non è possibile se non raddoppiando la numerosità del campione, in modo da avere una sessantina di casi per ciascuno dei due sottogruppi.
Questo peraltro non è che un esempio delle complicazioni a cui va incontro nella ricerca survey la determinazione della numerosità del campione; più in generale, possiamo dire che la formula di determinazione della numerosità campionaria deve essere adattata quando l'analisi statistica è di tipo multivariato.
Terminiamo questo paragrafo con un esempio che può dare al lettore un'idea approssimativa delle numerosità campionarie tipiche della ricerca che ha come obiettivo la generalizzazione dei risultati dal campione alla popolazione.
Un'importante survey cross-country ripetuta per lo studio del mutamento dei valori degli europei è la European Values Survey. Avviata nel 1981, è arrivata nel 2008/9 alla quarta edizione (nel 2017/8 si terrà la quinta wave).
Nella tabella 1 abbiamo selezionato 5 dei paesi in cui l'indagine EVS è stata condotta in tutte le edizioni (wave); nelle celle interne è indicata la numerosità campionaria per ciascun paese e per ciascuna edizione.

Tabella 1 – Numerosità campionarie per wave e paesi considerati

	1981-1984	1990-1993	1999-2001	2008-2010
Danimarca	1.182	1.030	1.023	1.507
Francia	1.200	1.002	1.615	1.501
Gran Bretagna	1.167	1.484	1.000	1.561
Italia	1.348	2.018	2.000	1.519
Spagna	2.303	2.637	1.200	1.500

L'errore sistematico

Abbiamo visto in precedenza che da un campione probabilistico otteniamo una statistica campionaria affetta da errore stocastico; nel caso preso in esame, quello della stima del parametro μ:

$$\bar{x} = \mu \pm e$$

Tuttavia nell'inchiesta campionaria è facile che i dati siano anche affetti da disturbi non casuali; nella formula precedente dovremo quindi aggiungere una componete additiva che definiamo errore sistematico e indichiamo con la lettera *b* (dall'inglese *bias*):

$$\bar{x} = \mu \pm e + b$$

Purtroppo non c'è una teoria su questo tipo di errore che ci informi sul suo valore atteso.
È detto 'profilo dell'errore' l'analisi delle fonti di errore sistematico. Il ricercatore dovrebbe fare sempre questo tipo di studio e esplicitare in sede di presentazione dei risultati di ricerca i disturbi sistematici di cui possono essere affetti i dati impiegati. Tale studio richiede una descrizione completa e ordinata delle operazioni e delle potenziali fonti di errore, nonché dell'effetto dell'errore di ciascuna operazione sull'errore complessivo (Frosini, Montinaro, Nicolini, 1994, p. 6).
Gli errori sistematici si manifestano anche, e con maggiore intensità, nelle rilevazioni complete. Alcuni di essi sono relativi a ogni singola operazione di rilevazione dello stato sulla proprietà: questi pertanto non decresceranno all'aumentare della dimensione del campione, ma al contrario aumenteranno.
I tipi di errore sistematico più comuni nell'inchiesta campionaria sono i seguenti:

- errori di copertura: causati, come detto, da difettosità delle liste di campionamento o da aspetti operativi; sono dovuti a incompletezza, elementi inesistenti, ridondanza;
- errori da mancate risposte totali (*missing respondent*) dovuti a mancato contatto del soggetto selezionato dalla lista di campionamento oppure a rifiuto a rispondere da parte di un soggetto contattato; le mancate risposte, in gergo definite anche *cadute*, portano a una discrepanza tra popolazione bersaglio e popolazione raggiunta; si possono avere anche errori da mancate risposte parziali (*missing values*), quando i soggetti contattati si lasciano intervistare ma per qualche ragione (rifiuto dell'intervistato, errore di rilevazione ecc.) non vengono registrati gli stati su alcune delle proprietà rilevate;
- infedeltà delle risposte: discrepanze tra dati rilevati e gli effettivi stati sulle proprietà dei soggetti. Riguardano in special modo l'indagine che ha per oggetto atteggiamenti, preferenze, opinioni, rappresentazioni sociali, dichiarazioni di comportamento ecc. Sono a loro volta distinguibili in:
- errori dovuti allo strumento di rilevazione: domande doppie, domande non comprese dall'intervistato, formati di risposta incompleti, ecc.;
- errori di risposta: volontà di fornire informazioni false, incapacità di introspezione, acquiescenza, desiderabilità sociale ecc.;
- errori di codifica dei dati dal questionario alla matrice dati (*wild code*).

Soffermiamoci ancora sugli errori da mancate risposte, in quanto problema specifico dell'inchiesta campionaria quando si intendano fare operazioni di inferenza statistica.
Le cadute nelle interviste faccia a faccia o telefoniche oscillano in genere tra il 20% e il 30% (sono più elevate di solito in quelle telefoniche: è più facile abbassare la cornetta del telefono che chiudere la porta in faccia). I questionari postali hanno in genere dei ritorni ancora più bassi.
Proseguiamo ora la riflessione sulla base di un esempio ipotetico. Supponiamo di partire da un campione perfettamente casuale di 1000 soggetti estratti (SE=1000); di questi si è riusciti a contattarne 850 (SC=850; SNC=150); 700 rispondono al questionario (SR=700) e 150 si rifiutano (SNR=150). Abbiamo in totale 300 cadute (SNC+SNR).
Dopo aver tentato attraverso nuovi contatti di convincere almeno alcuni dei soggetti caduti, non resterà altra strada che quella di estrarre casualmente un nuovo campione tale per cui si riesca infine a ottenere SR=1000.
Tuttavia, questa operazione non è 'indolore'. Il fenomeno non sarebbe grave se i soggetti SNR e SNC costituissero dei sottocampioni casuali di SE.
Il problema è che non lo sono affatto: come è stato dimostrato mediante indagini specifiche (cfr. Marradi 1997, p. 37), i SNC e in misura ancora maggiore i SNR hanno distribuzioni si-

stematicamente diverse dai SR su molte delle tipiche variabili socio-demografiche (per esempio tra i non rispondenti sono sovra-rappresentate donne anziane, sole, con basso titolo di studio) ma anche per quanto riguarda abitudini di vita, situazione economica, opinioni politiche ecc.
Poiché gran parte delle variabili socio-demografiche sono note al censimento, il rimedio è quello di ponderare il campione perché rispecchi l'effettiva struttura della popolazione.
Se però c'è una distorsione sistematica che riguarda variabili oggetto di ricerca e di cui non si hanno dati di censimento il problema non è facilmente risolvibile.
Poiché né i SNC né ancor meno i SNR possono essere considerati campioni casuali di SE, e a fortiori della popolazione obiettivo, abbiamo come conseguenza negativa che anche un campione perfettamente casuale al momento dell'estrazione può trasformarsi in un campione non casuale a seguito delle cadute.

Il campionamento non probabilistico

Si è già fatto notare che non necessariamente nell'analisi dei dati rilevati in un'indagine campionaria si intende applicare l'inferenza statistica: in tal caso, si può anche ricorrere a disegni di campionamento non probabilistico.
Questi ultimi sono per esempio utilizzati negli studi pilota: prima di affrontare i costi di un campionamento probabilistico, in assenza di ricerche precedenti o di altre informazioni, un piccolo campione di comodo, composto dai primi casi che si trovano a disposizione, servirà ad accertare se un certo fenomeno è effettivamente presente in modo rilevante, secondo quali modalità si manifesta e se lo strumento di rilevazione è appropriato.
Un altro esempio in cui la scelta dei casi non è di tipo probabilistico è dato dai disegni sperimentali; per questi di solito va bene un campione di comodo, anche se è bene evitare che il campione sia troppo omogeneo su altre caratteristiche non in esame ma che potrebbero avere relazioni con le variabili di cui si vogliono studiare le relazioni causa-effetto. I soggetti selezionati per l'esperimento vengono poi assegnati, questa volta in modo casuale, ad un gruppo sperimentale e ad un gruppo di controllo (per approfondimenti rimandiamo a un manuale di metodologia della ricerca; fra i tanti: Corbetta, 1999).
Anche negli studi di covariazione talvolta si rinuncia all'obiettivo della rappresentatività del campione rispetto alla popolazione da cui è tratto per concentrarsi sul rapporto che intercorre tra le variabili. Anche in questi casi, come nei disegni sperimentali, è meglio che i campioni siano numerosi e soprattutto eterogenei per quanto concerne le variabili di controllo.
Infine, in molte survey si fa uso prevalente di statistiche descrittive, con l'obiettivo di descrivere un fenomeno settoriale o localmente situato, e non si ha alcun interesse a effettuare generalizzazioni di carattere statistico – inferenziale.
Vediamo, infine, quali sono i più diffusi disegni di campionamento non probabilistico e come trovano impiego nell'indagine quantitativa.

- *Campionamento di comodo.*

Nel campionamento di comodo, o di convenienza, i casi vengono selezionati sulla base della loro immediata disponibilità per la ricerca. È ampiamente impiegato negli studi pilota e nei disegni sperimentali. Eventuali generalizzazioni a una popolazione più ampia risultano alquanto azzardate, salvo che si possa argomentare che il campione non presenta delle atipicità rispetto alla popolazione target per le variabili in esame.

- *Campionamento a valanga*

Il campionamento a valanga è una tecnica di campionamento a più fasi proposta originariamente da J. S. Coleman per la *social network analysis* nel 1958 e formalizzata da L. Goodman nel 1961; è una tecnica oggi utilizzata in vari tipi di ricerche anche quantitative. Questo tipo di campionamento, come quello di comodo, può avere come finalità la mera costituzione di un campione di adeguata numerosità senza alcune pretese di rappresentatività; oppure è da taluni impiegato come surrogato del campionamento probabilistico quando è impraticabile la costruzione della lista di campionamento. Nella prima fase si individuano soggetti con determinate caratteristiche, per esempio appartenenti a particolari cerchie sociali, di solito poco visibili socialmente; oltre a essere essi stessi persone su cui viene svolta l'inchiesta, questi soggetti fungono da informatori per individuare altri soggetti appartenenti a tali cerchie o gruppi. Il processo può procedere in ulteriori stadi. È utile per indagare popolazioni con caratteristiche rare o nascoste, quando queste caratteristiche sono all'origine di un legame sociale (per esempio i consumatori di droghe illegali hanno spesso contatti con altri consumatori; così i membri di una setta, le persone che frequentano mense per i poveri ecc.).

Rispetto al campionamento probabilistico, presenta il limite di portare alla selezione dei soggetti più visibili, più attivi, che non sono necessariamente quelli più rappresentativi della popolazione obiettivo. Inoltre c'è il rischio che la catena di contatti prenda una strada troppo specifica. Per esempio in una ricerca su persone ex tossicodipendenti, il punto di avvio, per ragioni pratiche, è spesso rappresentato da individui che hanno frequentato una comunità o un SERT; questi tenderanno a segnalare altre persone che a loro volta sono passate negli stessi servizi: il rischio è quindi di avere un campione molto specifico, rappresentativo della popolazione degli utenti di quei servizi ma non della popolazione complessiva delle persone ex tossicodipendenti.

Prima di proseguire, può essere utile soffermarsi a riflettere sul fatto che, talvolta, proprio le popolazioni più interessanti dal punto di vista sociologico o (1) non sono censite in alcun registro e, di conseguenza non è disponibile alcuna lista di campionamento dalla quale partire, oppure (2) hanno dimensioni talmente ridotte da non consentirne l'inclusione in un campione utilizzando le strategie di campionamento classiche ("popolazioni rare"). Se l'accento è posto sull'assenza di una lista di campionamento, si parla solitamente di popolazioni "sfuggenti", "nascoste" o "difficilmente raggiungibili" (Kish 1991). Se invece si pone l'accento sulla loro incidenza numerica sul totale della popolazione, talmente ridotta da rendere vana la speranza di intercettarne un numero sufficiente con le tecniche di campionamento classico, si parla di popolazioni "rare", comprendenti meno di un decimo della popolazione totale. Un esempio di popolazione "sfuggente" è dato dalle persone presenti irregolarmente in un paese. Un esempio di popolazione "rara" è rappresentato dalle madri sole, dalle famiglie molto numerose (con più di quattro figli) o, ancora, dalle famiglie in grave disagio abitativo. Poiché è vano sperare di intercettare un numero sufficiente di membri di queste popolazioni ricorrendo alle tecniche di campionamento classico, spesso ci si avvale delle tecniche che sfruttano i legami sociali, come ad esempio il campionamento e valanga. Ma il costo di questa scelta, come detto, è l'impossibilità di generalizzare i risultati cui si perviene con la ricerca.

Verso la metà degli anni Novanta, sono state specificate alcune tecniche di campionamento basate sui legami sociali tra i membri della popolazione oggetto di studio che, sfruttando le proprietà di particolari strutture matematiche dette "grafi", consentono di calcolare, le probabilità di inclusione nel campione dei soggetti[10]. Una volta affrontato con successo questo

[10] La formalizzazione matematica dei modelli matematici sottostanti a questa proposta esula gli obiettivi di questa introduzione al campionamento nella ricerca sociale. Si rimanda chi volesse approfondire questo

problema, si possono ottenere stime non distorte dei parametri nella popolazione nel complesso, ed è quindi in teoria possibile fare correttamente inferenza come avviene utilizzando i campioni probabilistici. Le precauzioni metodologiche da adottare nello specificare piani di campionamento di questi tipo sono tali, tuttavia, da consigliarne l'impiego, per ora, solo da parte di specialisti e in casi limitati.

- *Campionamento per quote*

Il campionamento per quote è molto usato nei sondaggi di opinione, nel marketing, nelle survey relative a comportamenti, atteggiamenti e valori diffusi tra la popolazione o sue particolari fasce.
Nel campionamento per quote la popolazione viene suddivisa secondo variabili criterio note al censimento (età, genere, comune di residenza ecc.). Il totale dei casi da campionare viene suddiviso tra le celle generate dall'incrocio delle modalità di ciascuna variabile criterio, in modo da rispecchiare le proporzioni esistenti nella popolazione. Immaginiamo per esempio di voler fare una ricerca sui giovani torinesi e che come variabili criterio siano state individuate l'età e il genere secondo le modalità indicate nella tabella seguente:

Tabella 2 – Giovani residenti a Torino al 31/12/2011

	M	F
15 – 19enni	17.894 (14,63%)	16.664 (13,56%)
20 – 24enni	20.363 (16,57%)	19.435 (15,81%)
25 – 29 enni	24.376 (19,83%)	24.094 (19,60%)
14- 29enni	62.723	60.193

fonte: Anagrafe del Comune di Torino

Se il campione complessivo sarà di numerosità n = 1000, 146 dovranno essere maschi 15-19enni, 135 ragazze 15-19enni, 165 maschi 20-24enni ecc.

Il campionamento per quote non va confuso con il campionamento stratificato.
Nel campionamento per quote si danno istruzioni agli intervistatori sui criteri con cui essi devono attivarsi nella scelta dei soggetti da inserire nel campione: per esempio scegliere 30 soggetti di età compresa tra i 14 e i 20 anni, metà maschi e metà femmine, abitanti in un certo quartiere. Per il resto si lascia all'intervistatore libertà nella scelta dei soggetti.
Questa procedura ha però un limite evidente: essendo l'intervistatore, al pari di chiunque altro, inserito in determinate reti di relazioni sociali, propenderà (anche involontariamente) a selezionare soggetti con particolari caratteristiche piuttosto che altre: per esempio persone con un livello culturale simile, che condividono un certo habitus o stile di vita, ecc. In tal modo non tutti i soggetti all'interno di un certo strato definito dai criteri dati all'intervistatore hanno eguali probabilità di essere scelti, come accade invece nel campionamento casuale stratificato.
La fallibilità del campionamento per quote è documentata; un caso molto citato è il fallimento della previsione elettorale nelle presidenziali americane del 1948, in cui i sondaggi davano per vincente Dewey su Truman, cosa che invece non avvenne. La colpa venne attribuita proprio al campionamento per quote.

---

tipo di tecniche di campionamento alla vasta letteratura metodologica ormai disponibile (ad esempio, Heckathorn 1997, 2002, 2007).

In base a esperienze come queste sono state messe a punto varianti del campionamento per quote che riducono, pur senza eliminarle, le distorsioni connesse alla libertà di scelta degli intervistatori.

-Campionamento del caso tipico

Con il campionamento del caso tipico (*typical case sampling*) il ricercatore seleziona un certo numero di unità ecologiche che a suo parere, o a parere di altri esperti, presentano un certo grado di normalità/tipicità, su alcune caratteristiche, rispetto a una popolazione più ampia; all'interno di queste unità ecologiche provvederà poi a individuare, con metodi probabilistici o non, le unità di rilevazione. Per esempio: la selezione di alcuni casi tipici di impresa industriale nell'Italia del Nord-ovest e la selezione di un certo numero complessivo di addetti da intervistare.

Una pratica diffusa è quella di selezionare le unità ecologiche minimizzando la differenza tra le medie di alcune importanti caratteristiche di ogni unità e le corrispondenti medie nella popolazione.

Caso particolare è il *critical case sampling* o campionamento da aree barometro: si sceglie un aggregato (*cluster*), che in passato si è rivelato essere rappresentativo di una popolazione più ampia o, se non rappresentativo, utile in qualche modo per fare delle previsioni. Per esempio, si può analizzare la serie storica delle elezioni amministrative di una Regione e individuare un Comune in cui i risultati elettorali riproducono in piccolo quelli dell'intera Regione; una volta individuato, il Comune potrà in futuro costituire un campione utile per fare dei sondaggi elettorali in prossimità delle elezioni e per prevedere il risultato nell'intera Regione.

Bibliografia citata

CALLEGARO M., POGGIO T. (2004), *Espansione della telefonia mobile ed errore di copertura nelle inchieste telefoniche*, "Polis", 2, pp. 477-508.

CHIARI G., CORBETTA P. (1973), *Il problema del campionamento nella ricerca sociologica*, "Rassegna Italiana di Sociologia", 3-4, pp. 473-513 e 643-667.

CICCHITELLI G., HERZEL A., MONTANARI G. E. (1997), *Il campionamento statistico*, il Mulino, Bologna, seconda edizione.

COLEMAN J.S. (1958), *Relational Analysis: the Study of Social Organizations with survey Methods*, "Human Organization", 17, pp. 28-36.

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

FROSINI B.V., MONTINARO M., NICOLINI G. (1994), *Il campionamento da popolazioni finite: metodi e applicazioni*, Utet libreria, Torino.

GOODMAN L. (1961), *Snowball Sampling*, "Annals of Mathematical Statistics", 32, pp. 148-170.

GROVES R.M. (1989), *Survey Errors and Survey Costs*, Wiley, New York.

HECKATHORNE D.D. (1997), *Respondent-Driven Sampling: A New Approach to the Study of Hidden Population*, "Social Problems", 44, 2, pp. 174-199.

HECKATHORNE D.D. (2002), *Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Population*, "Social Problems", 49, 1, pp. 11-34.

HECKATHORNE D.D. (2007), *Extensions of Respondent-Driven Sampling: Analyzing Continuous Variables and Controlling for Differential Recruitment*, "Sociological Methodology", 37, 1, pp. 151-207.

HERZEL A. (1991), *Teoria e tecniche dei campioni*, in AA.VV., *Enciclopedia delle Scienze Sociali*, Istituto dell'Enciclopedia Italiana, Roma.
KISH L. (1991), *Taxonomy of Elusive Populations*, "Journal of Official Statistics",7, pp. 340-7.
MARRADI A. (1997), *Casuale e rappresentativo: ma cosa vuol dire* ?, in CERI P. (a cura di), *Politica e sondaggi*, Rosenberg & Sellier, Torino.
NEYMAN J. (1934), *On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection*, "Journal of the Royal Statistical Society", 97, pp. 558-606.
URIGH N. (2008), *The Nature and Causes of Attrition in the British Household Panel Study,* 5, Working paper, Institute Social and Economic Research, University of Essex.