5 IL CAMPIONAMENTO

di Stefano Nobile

I. Popolazione e campione: il problema della rappresentatività campionaria

Se potessero farlo – il che significa: se avessero il tempo, le risorse umane e il denaro necessari – i ricercatori sociali preferirebbero certamente ricorrere a un censimento piuttosto che a un campione. Perché un censimento rappresenta tutta la popolazione[1], mentre un campione ne rappresenta solo una parte. Si tratta di un problema che fino a un secolo e mezzo fa i ricercatori sostanzialmente non si ponevano nemmeno, giacché il concetto di campione è entrato in tempi relativamente recenti nel lessico della scienza[2], così come quello di statistica inferenziale, a esso collegato, ossia alla fine dell'Ottocento.

[1] Bisogna precisare che il termine *popolazione* non si riferisce necessariamente a delle persone, che sono le unità di analisi più frequenti nella ricerca quantitativa nonché della ricerca sociale in genere. Si possono avere popolazioni di documenti, di film, di capoluoghi di provincia, di automobili, eccetera. L'insieme di tutte le aziende del Veneto può essere considerata una popolazione, così come può esserlo l'insieme di tutte le scuole elementari di Viterbo o anche tutti gli articoli pubblicati sulla rivista *Sociologia e ricerca sociale* tra il 2000 e il 2020. Alcuni studiosi preferiscono usare il termine *universo* al posto di *campione*. Ma universo, pur non inciampando sul problema della somiglianza con una popolazione (la radice è la stessa di popolo, e, dunque, richiama gli umani), ha il difetto di suggerire un concetto dai confini estremamente porosi, dunque indefiniti.

[2] Il problema del campionamento riguarda tutte le scienze che abbiano come riferimento l'essere umano o gli esseri viventi in generale: oltre alla sociologia, alla psicologia, eccetera, anche la biologia, la medicina (con le sue branche come l'epidemiologia o l'infettivologia) e la botanica sono discipline che si servono della statistica inferenziale.

Ma cosa è un campione e perché si campiona? La parola ha un'origine antica e risale all'epoca medievale, in cui il miglior uomo d'armi lottava in rappresentanza di qualcun altro. Da lì il concetto di *miglior esemplare possibile*, che spiega perché anche a Maradona possa essere attribuito lo stesso sostantivo con funzione aggettivale. Si campiona, dunque, per evitare di dover considerare l'intera popolazione di riferimento, anche perché questa non sempre è definibile o, se lo è, non è possibile individuare a uno a uno i suoi membri. Dunque, se potessero, i ricercatori preferirebbero avere dati certi, ottenuti dall'intera popolazione, piuttosto che dati incerti ottenuti da una parte di essa. Il problema, allora, è quello di ridurre questo margine di incertezza, di scattare una "fotografia" della popolazione di riferimento più fedele possibile. Se il censimento è la fotografia ad altissima risoluzione, in cui si riescono a riconoscere i minimi dettagli, il campione è un'immagine a più bassa definizione, dove si vedono i pixel. Ma se la fotografia è fedele all'originale, ci basterà strizzare gli occhi per vedere l'insieme: magari vedremo qualche zona meno nitida, ma avremo comunque un'idea – per quanto approssimata[3] – dell'insieme. A questo punto verrebbe da dire, con un certo automatismo, che se il ricercatore aumenta progressivamente questa definizione (ossia aumenta la numerosità campionaria), quasi certamente aumenterà la capacità del campione di rappresentare la popolazione di riferimento. Ma questo assunto è vero fino a un certo punto: l'efficacia dei criteri di selezione dei casi conta assai di più della bruta quantità con cui si aumenta a dismisura la numerosità campionaria. Esiste un episodio notissimo che testimonia quanto appena detto: quello delle elezioni per la presidenza americana del 1936, in cui a contendersi la Casa Bianca c'erano il democratico Franklin D. Roosevelt e il repubblicano Alf Landon. Il *Literary Digest*, una rivista piuttosto blasonata, inviò un questionario a tutti gli americani iscritti al pubblico registro automobilistico e a quelli che disponevano di un'utenza telefonica. Dei dieci milioni di questionari inviati ne tornarono indietro circa due milioni e trecentomila: un'enormità. Con questi sontuosi dati, il *Literary Digest* predisse la larga vittoria di Landon. A contendere la riuscita della previsione al Golia dei sondaggi c'era un piccolo Davide, George Gallup, che intervistò appena duemila persone (un numero davvero lillipuziano rispetto a quello del *Literary Digest*). Per Gallup, le elezioni le avrebbe vinte

[3] L'approssimazione è imputabile al fatto che quanta più certezza richiediamo al nostro campione (vogliamo essere *certi* che la stima di un certo parametro sia esattamente pari a una determinata grandezza), tanta meno precisione avremo (vogliamo conoscere il valore *preciso* di quel parametro).

Roosevelt. E così fu. Perché le cose andarono in questo modo? È evidente: perché quelli che possedevano un'automobile o un'utenza telefonica appartenevano ai ceti più abbienti della popolazione, quelli dunque più inclini a votare repubblicano. Dall'altra parte, al contrario, vennero intercettate tutte le fasce della popolazione e, dunque, la "fotografia" di cui si è parlato era certamente più fedele alla realtà. Cosa ci insegna questo aneddoto? Che per ottenere un campione efficace non basta fare leva sui numeri (quantunque abbiano la loro importanza), ma cercare di riprodurre le caratteristiche della popolazione di riferimento. Attenzione, però: una certa vulgata scientista, consistente quanto tetragona, sostiene che la rappresentatività statistica – ossia la capacità del campione di riflettere le caratteristiche della popolazione di riferimento – sia un diktat irrinunciabile, l'unico in grado di garantire la presunta scientificità dell'analisi dei dati. Non sono pochi i ricercatori che pretendono una patente di credibilità e rigore metodologico a suon di coefficienti di significatività statistica. Ma il feticcio della cifra sembra far prevalere esclusivamente la dimensione teorica dell'inferenza, senza tenere nel debito conto quella metodologica. Un conto, infatti, è voler riprodurre, in piccolo, ciò che accade in grande, una sorta di sineddoche metodologica, la parte che sta per il tutto. Ma tutt'altro conto è voler approfondire o studiare le relazioni tra variabili. Nel primo scenario, che corrisponde ai sondaggi[4], è del tutto legittimo che il campione debba riflettere le caratteristiche della popolazione. Se si vuole sapere quale partito prenderà più voti in occasione della prossima competizione alle urne è necessario adottare delle tecniche di campionamento che garantiscano la capacità del campione di rispecchiare la popolazione nella maniera più fedele possibile. Ma se andiamo alla ricerca di un nesso tra variabili, per esempio per spiegare se esiste una relazione tra la propensione al consumo di sostanze sintetiche e il rendimento scolastico, il problema della rappresentatività passa in secondo piano (Calandi 2003). Esso, inoltre, non è un concetto binario – come, ancora, sembrano pretendere alcuni (a titolo esemplificativo, Barisione e Mannheimer 2005) – bensì, come ha chiarito Marradi (2007), un continuum che va da un minimo a un massimo. In più, si dimentica che la rappresentatività statistica è garantita soltanto rispetto alla variabile (o alle

[4] Il concetto di *sondaggio* va distinto da quello di *inchiesta*. Il primo rimanda all'idea di sonda, dunque a quella di intercettare gli umori di una popolazione, di sapere, per esempio, cosa voteranno gli italiani alla prossima tornata elettorale. Il secondo ha la stessa origine di "chiedere", dunque risponde a esigenze conoscitive diverse, legate alla spiegazione della relazione tra variabili: perché accade un certo fenomeno?

variabili) che è stata usata per definire il campione. Ma ciò non implica affatto che le stesse proporzioni trovate tra popolazione e campione, per esempio, rispetto al titolo di studio, si ritrovino anche in quelle relative alla predisposizione che popolazione e campione hanno rispetto alla popolazione immigrata. Vale a dire che la variabile criterio *non* si trascina dietro tutte le altre variabili che saranno utilizzate nella ricerca (Di Franco 2010): se ci fosse questo rapporto biunivoco tra variabili, non ci sarebbe neppure bisogno di fare ricerca. In quel caso, infatti, sapremmo già che a un certo titolo di studio corrisponde, per esempio, un determinato atteggiamento o un certo comportamento. Ne discende che, tra l'altro, l'idea che un campione probabilistico sia più affidabile di un campione non probabilistico è, in molti casi, una forzatura. Allo stesso modo, è infondata l'idea secondo cui i campioni probabilistici sono quelli utilizzati nella ricerca quantitativa e quelli non probabilistici in quella qualitativa. Vediamo, dunque, cosa sono i campioni dell'uno e dell'altro tipo.

2. I campioni probabilistici

I campioni probabilistici sono quelli rispetto ai quali tutte le unità della popolazione hanno (virtualmente) la stessa probabilità di essere estratte, come accade nel gioco della tombola, in cui – pescando a casaccio nel sacchetto – tutti e novanta i numeri hanno la stessa probabilità di estrazione; nei campioni non probabilistici questa chance non è data. E allora perché non scegliere sempre i primi, visto che garantiscono la già richiamata rappresentatività statistica, ossia la capacità – al netto dell'errore di campionamento – di rappresentare adeguatamente la popolazione di riferimento? Una prima ragione sta nel fatto che, per potervi fare ricorso, bisogna disporre di informazioni certe sulle dimensioni della popolazione. Ma se la ricerca dovesse riguardare i fan di Achille Lauro, gli immigrati irregolari della Campania o i tweet scritti in occasione dell'assegnazione dei premi Oscar, come possiamo fare, visto che non potremmo stabilirne il numero, ossia la grandezza della popolazione di riferimento? Non ci rimane altra soluzione che ricorrere ai campioni non probabilistici.

Cerchiamo dunque di capire quali opportunità ci offrono le tecniche di campionamento dell'uno e dell'altro tipo.
I campioni probabilistici sono rappresentati, per eccellenza, dal campione casuale semplice[5]. Per esso vale la logica dell'estrazione dall'urna: le palline numerate all'interno di essa sono i casi. Si pesca in modo accidentale. Nonostante le sue ottime credenziali, il campione casuale semplice non è immune da aspetti paradossali, che diventano ancora più evidenti quando si ricorre ad altri tipi di campioni. Poniamo, infatti, di trovarci in una situazione semplicissima, ben lontana dalla realtà, ma utile a capire come funzionano realmente le tecniche di campionamento. Supponiamo di voler stimare il voto medio di una popolazione composta da appena sei studenti all'esame di metodologia della ricerca sociale. I loro voti sono riportati in Tabella 5.1. Immaginiamo adesso di voler costruire un campione di due elementi (ben il 33% della popolazione, che è come se l'ISTAT, nelle sue indagini, intervistasse venti milioni e mezzo di persone).

Tabella 5.1 Voti presi all'esame di Metodologia della ricerca sociale da una popolazione di studenti

Studente	**Voto**
Alessia	18
Beatrice	24
Christian	30
Dario	22
Elisa	29
Francesco	21
Media	*24*

Sappiamo che la media del voto preso da questa popolazione di studenti è di 24 e vogliamo stimare lo stesso voto campionando due casi. Abbiamo le seguenti combinazioni possibili:

[5] Per definizione «un campione casuale è scevro da errori di selezione» (Stuart 1996, p. 15) e, dunque, «ha credenziali esemplari» (*ibidem*). Il problema principale della teoria dei campioni, infatti, è quello dell'*errore di selezione*, che comporta una distorsione del campione rispetto alla popolazione di riferimento.

Tabella 5.2 Medie dei possibili campioni

Combinazioni	Medie
AB	21,0
AC	24,0
AD	20,0
AE	23,5
AF	19,5
BC	27,0
BD	23,0
BE	26,5
BF	22,5
CD	26,0
CE	29,5
CF	25,5
DE	25,5
DF	21,5
EF	25,0

Come possiamo osservare, delle 15 combinazioni possibili tra coppie di studenti (dei quali sono riportate soltanto le iniziali dei nomi), soltanto una (Alice e Christian) restituisce effettivamente il valore riscontrato nella popolazione. Il che ci dimostra, in modo lapalissiano, che il campione casuale semplice, la tecnica più impermeabile a possibili errori di ogni altra tra tutte le tecniche di campionamento, non garantisce affatto la rappresentatività del campione[6]. Non solo: se ci volessimo accontentare di stime approssimative (per esempio, ritenendoci soddisfatti di avere un solo voto di scarto dalla media effettiva), avremmo sì un range più ampio di campioni (in questo caso andrebbero bene anche AD, AE, BD, EF), ma a una condizione: quanta più certezza richiediamo, tanta meno precisione avremo (e viceversa). Tanto è vero che se accettassimo uno scarto non più di uno, ma di due voti dalla media, i campioni che risponderebbero a questo requisito diventerebbero ben dieci, dai soli due iniziali che avevamo a disposizione.

Più in generale, ossia al di là dello specifico caso del campione casuale semplice, possiamo osservare che si pone, in filigrana, un problema di

[6] Su questo punto, Bruschi (1999, p. 380) è icastico: «l'estrazione casuale non garantisce la rappresentatività del campione; quest'ultima ne è solo una conseguenza probabile».

scarto dalla media. In altre parole, i campioni estraibili hanno tutti un certo scarto dalla media effettiva, che è ciò che intendiamo stimare. L'insieme di tutti questi scarti è lo scostamento semplice medio, una misura che si trova alla base della deviazione standard e della varianza. Come vedremo meglio più avanti, è proprio la deviazione standard uno dei parametri di cui il ricercatore deve servirsi quando vuole stimare l'ampiezza di un campione.

Tempo addietro, quando non c'erano i personal computer a disposizione, il campione casuale semplice si otteneva dalle cosiddette tavole dei numeri aleatori. Si trattava di tavole che riportavano sequenze di numeri che fungevano da riferimento per l'estrazione dei casi. Ciò presupponeva la necessità di disporre di una lista di campionamento, ossia di poter attribuire un qualche riferimento a ciascuno dei casi che fa parte della popolazione dalla quale si andrà a estrarre il campione. È evidente che questo possa costituire un limite, perché non sempre il ricercatore dispone di questa lista. Il problema non cambia da quando si è passati dall'uso delle tavole aleatorie a quello dei software per l'estrazione randomizzata dei casi. Il ricercatore fornisce al programma la lista di tutti i casi che fanno parte della popolazione (per esempio, tutti i comuni italiani, oppure tutte le aziende iscritte alla camera di commercio, o tutti gli studenti immatricolati alla Sapienza di Roma nell'anno accademico 2022-2023 o tutti gli attori che fanno parte dell'annuario del cinema italiano) e, stabilita la numerosità campionaria, chiede l'estrazione di n casi.

Oltre all'impossibilità, in alcune circostanze, di accedere alle liste di campionamento, il campione casuale semplice pone un secondo problema: quello di avere due varianti. Una prima variante presuppone la reimmissione, l'altra no. Per quanto possa apparire paradossale, il campione casuale semplice con reimmissione implica che un caso, una volta estratto, possa – per così dire – tornare all'interno dell'urna ed essere estratto una seconda volta. Sembra un controsenso, se si pensa che, ipoteticamente, questo significherebbe, per esempio, intervistare una persona che è giù stata intervistata. Nella pratica, infatti, si estrae senza reimmissione. Ma nella teoria, la reimmissione è rilevante per due motivi: il primo è che – senza reimmissione – tutti i casi che non sono stati selezionati alla prima estrazione aumentano automaticamente la loro probabilità di estrazione. Facciamo un esempio banale e decisamente parossistico: se la popolazione è composta da quattro casi e vogliamo selezionarne due, una volta estratto il primo (che aveva ¼ di probabilità di estrazione), ai

rimanenti tre rimane 1/3 di probabilità ciascuno. Il secondo motivo è che il campione con reimmissione si riconnette con il teorema del limite centrale. Secondo questo teorema, «se si estraggono ripetuti campioni casuali di dimensione n da una qualsiasi popolazione che abbia media μ e varianza σ^2, all'aumentare della dimensione n del campione le distribuzione campionaria delle media dei campioni tenderà ad avvicinarsi alla normalità e avrà come media μ e come varianza σ^2/n» (Blalock, 1969; tr. it. 1984, pp. 226-227). Per cui, in sintesi, tanto più si rimpicciolisce il numero di casi estraibili dalla popolazione (perché si è già estratto, senza reimmetterlo, il primo, e poi il secondo e così via), tanto più i campioni tenderanno a deviare tra di loro e a non rispettare il teorema del limite centrale[7].

Alla luce di tutto ciò, per quanti galloni di credibilità possa essersi guadagnato sulla carta, il campione casuale semplice non è poi di così semplice applicazione[8] e, in alcune circostanze, può risultare persino una opzione sconveniente.

Ecco allora che ci vengono in soccorso altre tecniche che, anziché riposare sul più salomonico dei criteri (pesco a caso nel mucchio, disponendo della lista), rispondono meglio a determinate circostanze, necessità e limiti in cui può venirsi a trovare il ricercatore. Pertanto, le altre tecniche di campionamento probabilistico che completano il quadro sono quello sistematico, quello a stadi, quello a grappoli e quello stratificato. Ognuno di essi ha i suoi pro e i suoi contro. Vediamoli aduno ad uno.

Il campione sistematico riesce ad ovviare alla necessità di dover disporre della lista di campionamento. È il tipico campione che si usa nel caso tipico degli exit poll, fuori dai seggi elettorali. Non potendo intervistare il signor Fabrizio B. o la signora Cristina A. – perché non sappiamo se si sono recati o meno al seggio elettorale – si intervista una persona ogni *k*, dove *k* è la cosiddetta frazione di campionamento, ossia il rapporto tra la numerosità della popolazione e quella del campione. Se il campione è un cinquantesimo della popolazione (in questo caso costituita da tutti coloro che sono iscritti a una certa sede elettorale), *n* sarà uguale a 50: si

[7] A questo proposito, Radini (2007, p. 246) fa notare come nel «campionamento senza reimmissione l'applicazione del Teorema del Limite Centrale produca il risultato contraddittorio secondo cui la validità del campionamento dipende tanto dall'aumento quanto dalla diminuzione dell'ampiezza campionaria.

[8] Per avere un'idea delle difficoltà reali di estrazione di casi, soprattutto quando questi sono degli individui, basterebbe pensare agli ostacoli che il ricercatore può incontrare per ottenere una lista dalla quale estrarre i nominativi. Per un approfondimento su questa e altre difficoltà, si veda Di Gioia (2009, p.101).

intervista una persona ogni 50 che esce dal seggio[9]. Analogamente, se si volesse condurre un'indagine sugli acquisti fatti dai romani nei supermercati, si potrebbe optare per il campione sistematico, intervistando una persona ogni k. Al lettore più accorto non sfuggirà il fatto che questa tecnica di campionamento, pur rispettando i crismi dell'equiprobabilità di estrazione, difficilmente potrà essere impiegata da sola, se non in presenza di popolazioni molto piccole. Infatti, nei due esempi riportati – quello degli exit poll e quello degli acquisti al supermercato – è evidente che il campionamento sistematico può essere impiegato soltanto in seconda battuta, dopo aver provveduto a stratificare la popolazione di riferimento, ossia dopo averla ripartita rispetto ad altri criteri classificatori. Nel primo dei nostri due esempi, il criterio potrebbe essere quello di tutte le scuole adattate a seggio elettorale; nel secondo, il criterio potrebbe prevedere una prima ripartizione del territorio capitolino per quartieri e poi estrarre casualmente dei supermercati da ciascun quartiere e solo a quel punto introdurre il criterio della sistematicità, ossia dell'estrazione – come suggerisce lo stesso nome – di un soggetto ogni k. Soltanto nel caso in cui si volesse campionare una popolazione molto piccola (per esempio, le persone che escono dal parco giochi di Disneyland per avere informazioni sul grado di divertimento e sulla capacità di offerta della struttura), il campionamento sistematico potrebbe essere impiegato autonomamente. Attenzione però: anche in questo caso bisognerebbe avere delle accortezze. Per esempio, potrebbe essere necessario ripetere il campionamento in orari diversi della giornata, giorni diversi della settimana e stagioni diverse, poiché ciascuno di questi fattori temporali potrebbe influire sulle risposte ricevute.

Il campione sistematico, dunque, ben si adatta alle necessità del ricercatore in assenza di liste di campionamento e ancorato a un processo di stratificazione. Può tuttavia manifestare, in alcune circostanze, un difetto insidioso: quello per cui la frazione di campionamento coincide con una qualche struttura organizzativa interna della popolazione, determinando una distorsione altrettanto sistematica. A questo proposito, Babbie (1973, p. 93) riporta il caso di una ricerca realizzata durante la II Guerra Mondiale, in occasione della quale veniva selezionato un soldato ogni dieci. I

[9] Nella realtà le cose sono ben più complesse di così, perché i campioni usati negli exit poll possono sì essere sistematici, ma solo in seconda battuta. In prima istanza, vengono individuati degli strati sui quali campionare. D'altronde, se davvero intervistassimo una persona ogni cinquanta, otterremo – sulla popolazione italiana avente diritto al voto – un campione esageratamente grande di circa 800.000 casi.

ricercatori ignoravano il fatto che le liste dei militari di ciascuna squadra erano costruite tutte secondo la gerarchia militare (dai sergenti ai soldati semplici), sicché – estraendo un soldato ogni dieci – vennero estratti unicamente sergenti. È un caso davvero raro, che tradisce anche una certa ingenuità metodologica, ma può capitare.

Si è detto del campionamento stratificato. Vediamo cos'è e come funziona. Si tratta innanzitutto di una tecnica assai virtuosa, capace di miscelare efficacia ed efficienza. Essa consiste nel suddividere la popolazione in un certo numero di "strati", ossia di gruppi, rispetto a una o più variabili criterio (per esempio, il sesso, il titolo di studio o l'età, ripartita per classi; le regioni o i capoluoghi di provincia, se invece parliamo di unità ecologiche), a condizione che si disponga delle informazioni su come si distribuisce la (o le, se ne usano di più) variabile criterio all'interno della popolazione di riferimento. Per esempio, immaginando di voler condurre una ricerca sulla popolazione universitaria di tutti gli immatricolati alla Sapienza di Roma, si può decidere di ripatire questa popolazione in ragione del tipo di diploma conseguito alle scuole medie superiori, campionando poi (casualmente) all'interno di ciascuno strato così ottenuto, in proporzione a quanti possiedono le diverse caratteristiche con cui si presenta la variabile criterio all'interno della popolazione (istituto tecnico industriale, liceo artistico, istituto magistrale e così via). Se la variabile criterio fosse appunto il tipo di diploma conseguito, i soggetti campionati dovrebbero mantenere le stesse proporzioni che hanno nella popolazione di riferimento[10]. Ma – come si è accennato – è anche possibile stratificare rispetto a due o più variabili. Se, per esempio, volessimo aggiungere il genere come seconda variabile di stratificazione, dovremmo poi campionare un certo numero di maschi che abbiano conseguito la maturità scientifica, di femmine che abbiano il titolo di diploma tecnico per il turismo, di maschi con maturità classica e via dicendo.

La tecnica di campionamento a stadi è una tecnica che ci viene in soccorso in assenza della lista di campionamento che, pur non offrendo una maggiore efficienza rispetto a quello casuale, ne semplifica alcuni problemi, riducendo anche i costi, soprattutto se la popolazione di

[10] La procedura proporzionale è quella canonica. Tuttavia, il ricercatore potrebbe avere degli obiettivi particolari che potrebbero spingerlo a optare per quote *non proporzionali* rispetto alla popolazione di riferimento, aumentando il numero di alcuni strati in ragione delle proprie necessità e diminuendo quelli con i valori proporzionali più alti. In genere questa seconda strategia si adotta per focalizzare l'attenzione in fase di analisi su alcuni elementi che, nella popolazione, risultano minoritari ma non per questo meno importanti.

riferimento è distribuita su un territorio particolarmente vasto e, di fatto, difficilmente reperibile. Esso consiste nel suddividere la popolazione su più livelli, ordinati gerarchicamente. Dapprima si estraggono unità dal livello più alto, per poi scendere ai livelli successivi e campionare casualmente soltanto all'ultimo livello. Per fare un esempio, immaginiamo di voler campionare i dirigenti delle ASL nazionali. Al primo stadio potremmo ripartire la popolazione per regioni; al secondo per provincie; al terzo si campionano casualmente le ASL di ciascuna provincia e, di conseguenza, il suo dirigente. Questa procedura ci permette di risparmiare sulla formazione della lista di campionamento. Va da sé che il numero di stadi dipenderà dalle dimensioni strutturali della popolazione di riferimento.

Il campione a grappoli è un'ulteriore semplificazione del campione a stadi e rappresenta, in alcune situazioni, un ottimo compromesso tra efficienza e contenimento dei costi. È un tipo di campione che può risultare utile soprattutto quando si vuole risparmiare sulla logistica organizzativa. Un esempio è quello per cui, in occasione di un'indagine condotta sugli studenti delle scuole superiori italiane, si ripartisce dapprima il territorio rispetto alle diverse regioni, poi alle provincie, infine alle scuole. In un campione a stadi, il campionamento avviene casualmente rispetto agli studenti delle scuole selezionate; in quello a grappoli si prende, invece, un'intera classe. La differenza tra i due, dunque, sta nell'ultimo passaggio, che – nel caso del campionamento a grappoli – permette di economizzare sulla selezione dei casi. C'è però da pagare un prezzo per questa semplificazione: i casi che saranno estratti, come nell'esempio riportato relativo alle classi scolastiche, tenderanno a essere più omogenei tra loro e, quindi, a ridurre la variabilità del campione. Tuttavia, a fronte di questo inconveniente, va messo in luce un vantaggio sostantivo: quello di poter studiare, nel caso in cui – per esempio – il grappolo da estrarre sia un'intera famiglia (intesa come nucleo convivente), «le interrelazioni fra i membri della stessa» (Corbetta, 2014, p. 333), sicché «l'analisi può passare dal livello individuale a quello familiare e viceversa» (*ibidem*).

3. I campioni non probabilistici

Come si è detto, l'altra famiglia di tecniche di campionamento, quella delle tecniche non probabilistiche, non può puntare sulla casualità dell'estrazione, ma – ciò nondimeno – può risultare utilissima in molte occasioni. Innanzitutto, è una soluzione efficace quando non si dispone delle liste delle unità di studio (che, per capirci, si traduce nell'impossibilità – ad esempio – di avere accesso a tutti i numeri di telefono di una certa popolazione). Ma alcune delle tecniche di questa famiglia sono anche molto utili quando si cerca di ottimizzare alcune caratteristiche della popolazione di riferimento. È il caso del campione per quote che, nella procedura, è pressoché identico a quello stratificato, differendo soltanto nella fase finale: quella in cui i casi vengono estratti a sorte (nel campione stratificato), ovvero scelti ad uno ad uno (in quello per quote). I vantaggi sono evidenti: soggetti più facilmente raggiungibili, risparmio di tempo e costi. Quando si vuole campionare la popolazione di riferimento in ragione delle quote generate da due o più variabili criterio, si può decidere di ricorrere a un campione cosiddetto fattoriale[11], che può prendere due direzioni; in una di esse, si selezioneranno i casi *proporzionalmente* alla loro distribuzione nella popolazione di riferimento. L'altra soluzione prevede invece che i casi vengano presi *tutti nella stessa quota*, in modo da non creare cluster minoritari.

Supponiamo, per esempio, di voler condurre una ricerca sulla dieta mediatica dei giovani italiani di età compresa tra i 15 e i 35 anni, usando come ulteriore variabile criterio anche il sesso. La prima cosa da fare è ottenere i dati relativi alla popolazione italiana in generale, come in Tabella 5.3.

[11] Va comunque chiarito che «la logica del disegno fattoriale non discende da quella del campione, ma da quella dell'esperimento» (Corbetta 2014, p. 349). Vale a dire che l'uso canonico di questo tipo di campionamento si basa sulla necessità di avere a disposizione due gruppi perfettamente confrontabili rispetto alle proprietà di maggior interesse (per esempio età, titolo di studio, zona geografica di residenza). In questo modo è possibile – entro determinati margini – verificare il diverso impatto che una certa variabile può avere rispetto a due gruppi, uno detto "di trattamento", l'altro "di controllo". La variabile viene fatta entrare in azione soltanto rispetto al primo gruppo e, da lì, si passa ad analizzare le differenze riscontrate tra i due gruppi.

Tabella 5.3 Popolazione italiana in età compresa tra i 15 e i 35 anni, ripartita per sesso (fonte: ISTAT)

	Sesso		
Età	***Maschi***	***Femmine***	***Totale***
15 anni	294.580	277.657	572.237
16 anni	293.142	275.768	568.910
17 anni	297.515	278.257	575.772
18 anni	307.687	283.278	590.965
19 anni	308.983	280.274	589.257
20 anni	315.863	281.026	596.889
21 anni	313.089	282.209	595.298
22 anni	313.272	284.122	597.394
23 anni	311.589	285.082	596.671
24 anni	313.208	290.785	603.993
25 anni	317.679	298.606	616.285
26 anni	330.699	312.624	643.323
27 anni	328.767	312.183	640.950
28 anni	335.470	320.066	655.536
29 anni	332.568	322.363	654.931
30 anni	340.386	330.375	670.761
31 anni	333.167	325.459	658.626
32 anni	334.356	327.765	662.121
33 anni	343.850	339.063	682.913
34 anni	349.443	345.482	694.925
35 anni	355.461	351.315	706.776
	6.770.774	6.403.759	13.174.533

A questo punto calcoliamo la numerosità campionaria (con criteri che vedremo in seguito), che risulta essere pari a 1.067 individui, con un intervallo di confidenza del 95% e un margine di errore del 3% (torneremo in seguito su questi concetti: al momento al lettore si chiede un atto di fede). Il numero di casi per ciascuna combinazione (uomini 15anni, donne 15enni, uomini 16enni, donne 16enni, eccetera) dovrebbe essere selezionato con una semplice proporzione[12], ottenendo quanto si può vedere in Tabella 5.4.

[12] Il calcolo è semplicissimo: per esempio, per ottenere i 24 maschi 15enni, basterà dividere il numero di maschi 15enni presente nella popolazione (294.580) per il totale della popolazione (13.174.533) e moltiplicare il quoziente ottenuto per la popolazione desiderata (1.067)

Tabella 5.4 Campione di 1.067 casi di italiani in età compresa tra i 15 e i 35 anni, ripartita per sesso

	Sesso		
Età	***Maschi***	***Femmine***	***Totale***
15 anni	24	22	46
16 anni	24	22	46
17 anni	24	23	47
18 anni	25	23	48
19 anni	25	23	48
20 anni	26	23	48
21 anni	25	23	48
22 anni	25	23	48
23 anni	25	23	48
24 anni	25	24	49
25 anni	26	24	50
26 anni	27	25	52
27 anni	27	25	52
28 anni	27	26	53
29 anni	27	26	53
30 anni	28	27	54
31 anni	27	26	53
32 anni	27	27	54
33 anni	28	27	55
34 anni	28	28	56
35 anni	29	28	57
	548	519	1.067

Questa procedura vale sia per il campione per quote che per quello stratificato. La differenza è che in quello per quote la scelta delle unità da selezionare non avviene casualmente, bensì in modo mirato (se si tratta di individui, il criterio più frequente è quello della conoscenza diretta).
Se il campione non è – come in questo caso – proporzionale, si selezionerà lo stesso numero di casi per ciascuna cella, quantunque – nell'esempio riportato – le differenze tra i diversi valori di cella siano già irrisorie (si va da un minimo di 22 casi a un massimo di 29).

Esistono altre tecniche di campionamento che possono comunque tornare utili. Una di queste è il campionamento a valanga, o a palla di neve. Il campione, metaforicamente, si forma a partire da un nucleo iniziale (in genere, i contatti dell'equipe di ricerca) e, progressivamente, si gonfia attraverso il passaparola oppure, nelle forme che oggi ottimizzano le risorse del web, facendo partire mail verso destinatari che potrebbero potenzialmente rispondere ai requisiti del campione, ai quali viene chiesto, a loro

volta, di estendere il contatto ad altri. Se poi il campionamento è pensato in funzione di un questionario da autocompilare sul web, si possono ulteriormente sfruttare le potenzialità dei social, per esempio collocando annunci (anche dietro il pagamento di piccole cifre) a target appositamente definiti in base alle caratteristiche richieste dal ricercatore (per esempio, femmine sposate del sud Italia), ammesso e non concesso che il gestore della piattaforma social disponga di queste informazioni (Facebook, per esempio, le possiede).

Tanto appare semplice realizzare un campionamento a valanga, quanto evidenti possono risultare i suoi limiti. Il primo è quello dell'autoselezione dei casi: rispondono le persone che *decidono* di rispondere, che sono inclini a farlo, che vogliono aiutare la ricerca, che hanno tempo e ragioni simili. Questo primo limite ha un corollario: i casi che vengono selezionati soffrono spesso di un *bias* a monte, legato al fatto di essere connessi tra loro e, quindi, di avere una certa somiglianza: laureati che frequentano altri laureati; lettori del *Corriere della Sera* che frequentano persone con inclinazioni simili e via discorrendo. Tuttavia, questa tecnica di campionamento non va sottovalutata perché torna estremamente utile in condizioni di ricerca in cui si vogliano intercettare individui legati a fenomeni sommersi: la prostituzione, l'alcolismo, il gioco d'azzardo, il lavoro minorile, l'evasione fiscale, lo spaccio di sostanze stupefacenti, lo scambismo, il lavoro nero e tanti altri. In tutti questi casi non si dispone di alcuna lista di riferimento né, tantomeno, si conoscono le dimensioni del fenomeno (quante sono le persone coinvolte?), se non attraverso stime numeriche piuttosto aleatorie. Dunque, qualora si volessero intercettare soggetti che rispondano alle caratteristiche di cui necessita l'obiettivo della ricerca, non rimarrebbe che affidarsi al passaparola per cui gli individui che fungono da nuclei di partenza ne indicheranno altri che, a loro volta, faranno altrettanto. A latere va precisato che, come è facile immaginare, non tutte le tecniche di campionamento sono applicabili a prescindere dall'unità di analisi. Il campionamento a valanga è una di queste. Se si volesse condurre una ricerca sul tema della rappresentazione dell'adolescenza nell'ultimo mezzo secolo di cinematografia italiana, non potremmo certo aspettarci che un film ce ne suggerisca un altro...

Chiudiamo questa rassegna con quella che è certamente la meno nobile tra tutte le tecniche di campionamento: il campionamento accidentale (o "a casaccio"). Con esso si intervista chi si trova e/o chi offre la propria disponibilità (oppure, se le unità di analisi sono altre, le prime che

troviamo: articoli di riviste prese a caso, imprese campionate senza criterio, venditori di auto usate selezionati in base alla prossimità col luogo di lavoro). Siamo al grado zero della rappresentatività statistica e della casualità, ma in compenso – persino optando per una scelta simile – potremmo trarre delle indicazioni preliminari per altre possibili ricerche. Basterebbe pensare che esistono anche casi estremi – come i cosiddetti studi di caso o le indagini pilota – che servono proprio a tastare il terreno in assenza di risorse e strumenti più idonei.

4. La ponderazione del campione

Quanto più il campione costruito è lontano rispetto alla distribuzione con cui la variabile criterio si presenta nella popolazione di riferimento, tanto più si rischia di ottenere informazioni che ne riproducono in maniera distorta le caratteristiche. Nonostante la consapevolezza di problemi come questo, tecniche di campionamento come quello a valanga sono diventate sempre più diffuse da quando l'accessibilità alle web survey ha assunto i connotati della nuova panacea metodologica, poiché consentono di raccogliere molte risposte con il minimo sforzo, con evidenti ricadute sulla qualità dei dati. In questi casi, i ricercatori più avveduti apportano dei correttivi ex post, quando ciò è fattibile, ponderando il campione, ossia ristabilendo le proporzioni tra ciò che si è raccolto e la quota corrispondente nella popolazione. La ponderazione, che – se usata con la dovuta accortezza – può tornare utile in quasi tutte le tecniche di campionamento viste in precedenza, si basa semplicemente su una proporzione. Se, ad esempio, abbiamo una quota di maschi che nel campione è del 64% mentre nella popolazione di riferimento è del 48%, ogni intervistato di sesso maschile varrà 48/64, ossia 0,75. Analogamente, ogni femmina varrà 1,44. Insomma, è come se si aggiustassero le proporzioni sui dati complessivi, operazione che può essere compiuta con una certa agilità anche nel caso in cui si volessero combinare le stesse quote di più variabili note nella popolazione: una certa quota di donne in età compresa tra i 18 e i 24 anni, residenti nel Centro Italia e con titolo di studio licenza media superiore, e così via elencando.

Supponiamo, per esempio, di voler ponderare il campione rispetto al territorio di residenza (Nord, Centro, Sud e isole) e al sesso. Per prima

cosa prendiamo i dati Istat rispetto alla popolazione che ci interessa; quindi, calcoliamo la proporzione (anche in percentuale) di ciascun tipo (maschi del Nord, femmine del centro, e via elencando per tutte e sei le combinazioni) rispetto al totale della popolazione che fa da riferimento. Facciamo poi la stessa cosa con il nostro campione. A quel punto otteniamo quanto possiamo vedere nella Tabella 5.5.

Tabella 5.5 Esempio di ponderazione su due variabili criterio a partire da dati Istat

	Popolazione		**Campione**		**Ponderazione**	
	Maschi	*Femmine*	*Maschi*	*Femmine*	*Maschi*	*Femmine*
Nord	25%	26%	14%	29%	1,77	0,92
Centro	11%	12%	20%	24%	0,54	0,49
Sud e isole	13%	13%	13%	12%	0,96	1,08

Nel passaggio all'analisi dei dati, le risposte di ogni maschio del Nord (un tipo fortemente sottorappresentato nel campione) verranno moltiplicate per un fattore di ponderazione di 1,77; ogni donna del centro (fortemente sovrarappresentata nel campione) verrà invece dimezzata (il fattore di ponderazione, in questo caso, è di 0,49) e così via. È evidente che la ponderazione deve essere utilizzata con raziocinio, e non come soluzione per campioni costruiti senza alcun senso. In ogni caso, deontologia vuole che, quando si ricorre a questo artificio statistico, il ricercatore dichiari di averne fatto uso.

5. La numerosità campionaria

Va da sé che quanto più si ricorre a tecniche di campionamento che non riescono a riprodurre, neppure a spanne, le caratteristiche della popolazione, tanto meglio l'aumento del numero di casi selezionati riuscirà a tamponare le carenze stesse del campione[13]. Il che pone inevitabilmente una domanda: quanto deve essere grande un campione? Dall'esempio tratto dall'episodio del 1936 sulle elezioni presidenziali a stelle e strisce, abbiamo visto che le dimensioni non sono garanzia di precisione e che la qualità del campionamento – anche quando esso viene costruito tramite tecniche non probabilistiche – conta più della dimensione. Ciò nondimeno, la statistica fissa alcuni parametri per fornire una risposta che, se si vuole seguire filologicamente le strada dei campioni probabilistici, è pienamente sensata. Se la variabile criterio per la definizione della numerosità campionaria è di tipo cardinale e se ne conosce il valore nella popolazione (l'esempio più canonico è l'età), la formula per calcolare la numerosità campionaria è questa, che – come si può notare – include la deviazione standard (σ):

$$n = \frac{\frac{z^2 \times \sigma \times (1-\sigma)}{e^2}}{1+\frac{z^2 \times \sigma \times (1-\sigma)}{e^2 N}}$$

Tuttavia, quando invece non si dispone della deviazione standard della popolazione e la variabile criterio è di tipo categoriale (o non ne vien adottata alcuna), la variabile criterio è di tipo cardinale, il prodotto $p \times q$ viene sostituito dal quadrato della deviazione standard[14] e la formula diventa la seguente:

[13] Ciò è possibile per due ragioni: la prima riguarda il fatto che all'aumentare della numerosità campionaria aumenta inevitabilmente la variabilità; la seconda è che su un campione di più ampie dimensioni è possibile intervenire più efficacemente ex post mediante la ponderazione.

[14] La deviazione standard è lo scarto medio di una certa distribuzione rispetto alla sua media. Quanto più è alta la deviazione standard, tanto più sarà eterogeneo il campione (e viceversa). Questo ci porta a un paradosso per cui se si volesse campionare rispetto, poniamo, a una popolazione di giovani aventi tutti la stessa età, usando proprio l'età come variabile criterio, basterebbe prendere un solo elemento. Il paradosso spiega bene in quale senso campionare rispetto a una certa variabile non implica che anche tutte le altre siano, come si dice, "rappresentative" e che il processo di inferenza deve fare i conti con questo balordo assunto di fondo.

$$n = \frac{z^2\,Npq}{e^2(N-1)+z^2pq}$$

Vediamo nel dettaglio questi parametri, partendo proprio da quest'ultimo, perché esso non sempre è calcolabile. Quando ciò accade, la formula – come si è detto – diventa proprio quella riportata, nella quale *pq* è il prodotto tra la quota percentuale di una variabile categoriale dicotomica e il suo complemento a 1. Se, per esempio, si conosce la percentuale di maschi (supponiamo, 48%) e di femmine (52%) di una determinata popolazione, questo prodotto è pari a 0,52 per 0,48. Se anche questo parametro è ignoto, dobbiamo metterci nella condizione più sfavorevole possibile, ossia supporre che la variabile criterio sia ripartita al 50%[15]. Analogamente, se abbiamo scelto come variabile criterio una variabile categoriale divisa in più di due modalità, il prodotto *p* x *q* sarà uguale alla quota proporzionale di una delle modalità per il complemento a 1 delle altre due.

Vediamo anche gli altri due parametri, dando per scontato che non vi siano dubbi sulla numerosità della popolazione di riferimento (N): il livello di fiducia (chiamato anche intervallo di confidenza e indicato nella formula con *z*) e il margine di errore accettato (indicato con *e*). Il primo fornisce la probabilità con cui la stima di un certo valore rilevato nel campione ricade effettivamente nel valore corrispondente della popolazione. In altri termini, ci indica quanto possiamo fidarci delle stime che vogliamo ottenere attraverso il nostro campione. È evidente che quanto più pretendiamo di volerci fidare (passando, per esempio, da un livello di fiducia del 95% a uno del 99%), tanto più aumenterà, di conseguenza, la numerosità campionaria. Questo parametro si ottiene a partire da alcune tavole che indicano il valore corrispondente ai diversi intervalli di fiducia. Altrettanto accade con l'errore di campionamento. Quest'ultimo ci dice entro quali margini la stima che abbiamo operato su un certo parametro del campione sia precisa. Supponiamo che si voglia stimare l'intenzione di voto degli elettori rispetto al Partito Democratico. Dire che si accetta un margine di errore del 5% significherebbe, per esempio, che se la nostra

[15] Per capire la ragione per cui sia questa la situazione più sfavorevole – e dunque quella che ci garantisce il risultato più sicuro – possiamo ricorrere alla geometria. Immaginiamo di avere una figura di quattro lati perimetro pari a 4 metri. se i lati fossero tutti uguali (ossia tutti lunghi 1 metro), il prodotto sarebbe 1 mq. Se invece i lati fossero, per esempio di 0,2 metri (quindi 20 cm) e di 1,8 metri (dunque, 180 centimetri), il perimetro rimarrebbe ovviamente identico ma l'area si rimpiccolirebbe, passando a 0,36 mq.

rilevazione ci restituisse un valore del 19%, questo valore potrebbe di fatto oscillare tra il 14% e il 24% (ossia ±5%), mentre se fossimo più rigidi, tollerando un errore soltanto dell'1%, i due robbi della forchetta oscillerebbero tra il 18% e il 20%. Combinando l'errore di campionamento accettato e il livello di fiducia richiesto, potremmo produrre affermazioni di questo tipo: "esiste il 95% di probabilità che il Partito Democratico alla prossima tornata elettorale acquisisca tra il 14% e il 24% dei voti".

Da ultimo, qualche nota a proposito del campionamento nell'ambito delle tecniche digitali, oggi così in voga. Esse rappresentano un importante scantonamento rispetto alla teoria del campionamento, forse addirittura un cambio di paradigma. Attraverso i Big Data, infatti, la logica del campionamento viene sostanzialmente accantonata per concedere spazio a indagini che rilevano informazioni pressoché censuarie. A fronte di questo, che è certamente un vantaggio, va considerato il fatto che quando si opera su dati del genere non si sa esattamente quale popolazione si stia studiando, ossia chi li abbia prodotti. Anche volendo ricostruire i profili rispetto ad alcune variabili socioanagrafiche rilevanti (il genere, l'età), finiremmo per imbarcarci in «un'impresa tecnicamente ardua e dagli esiti incerti» (Airoldi 2017, p. 18). In questo caso, dunque, viene meno proprio la dimensione della rappresentatività, sulla quale – come abbiamo visto – si gioca gran parte del blasone della ricerca standard. Essa infatti «può essere considerata comune a tutte le prospettive di indagine, quelle standard e quelle non standard, in quanto unico criterio in grado di poter asserire qualcosa, anche se soltanto in via ipotetica o generale, su una classe di individui» (Di Franco, 2010, p. 70). Tuttavia, come abbiamo visto, la rappresentatività si presta anche ad ambizioni malriposte, così come induce a sovradimensionare il problema dell'inferenza statistica, lasciando sulle quinte quello della relazione tra variabili in termini esplicativi.