Tipi di campione

Termini

Popolazione: intendiamo per popolazione un insieme ***N*** (*ampiezza della popolazione*) di unità statistiche/unità d'analisi che costituiscono l'oggetto del nostro studio (anche --> **universo**).
Commento: il termine "popolazione" evoca un insieme di esseri umani, ma in statistica ha un significato molto più ampio e si riferisce ad un qualsiasi insieme di oggetti (uomini, abitazioni, aziende, territori, manufatti, organismi viventi, oggetti inanimati, eventi, ecc.). Di più, si tratta di un insieme che ha una caratteristica in comune (costante).

Campione: l'insieme degli ***n*** (*ampiezza del campione*) casi (unità campionarie) selezionati tra gli ***N*** che compongono la popolazione, allo scopo di rappresentarla ai fini del nostro studio.

Campionamento: procedimento attraverso il quale si estrae, da un insieme di unità (*popolazione*) costituenti l'oggetto di studio, un numero ridotto di casi (*campione*) scelti con criteri tali da consentire la generalizzazione all'intera popolazione dei risultati ottenuti studiando il campione. In altri termini è la procedura che seguiamo per scegliere le ***n*** unità campionarie dal complesso delle ***N*** unità della popolazione.

Parametro: qualsiasi statistica calcolata relativamente ad una o più caratteristiche di tutte le unità d'analisi appartenenti alla popolazione (universo).

Stima: qualsiasi statistica calcolata relativamente ad una o più caratteristiche di tutte le unità d'analisi appartenenti al campione estratto. (Si dice anche --> **stimatore**). Un parametro è il valore esatto di una statistica calcolato su tutti i casi interessati, una stima è un valore approssimato calcolato su un campione estratto da quell'insieme di casi, che tenta appunto di stimare il parametro.

Errore di campionamento: Se chiamiamo ***V*** un dato parametro di una popolazione e ***v*** la stima di esso effettuata per mezzo di un campione, sarà $\boldsymbol{e = V - v}$, dove **e** è *l'errore di campionamento*.

Livello di fiducia: la probabilità che si ha di essere nel vero quando si afferma che ***V*** (valore del *parametr*o) è compreso nell'intervallo $\boldsymbol{v \pm e}$, dove ***e*** è un valore scelto in base a tale probabilità.

Intervallo di fiducia: l'intervallo $\boldsymbol{v \pm e}$ che con un dato *livello di fiducia* contiene ***V*** (parametro della popolazione).

Errori di selezione: errori di copertura (la lista della popolazione è imprecisa, mancano casi, alcuni sono ripetuti); errori di campionamento (vedi); errori di non-risposta (autoselezione dei rispondenti).

Campione probabilistico: un campione si dice probabilistico quando ogni unità è estratta con una probabilità nota a priori.

Campione non probabilistico: un campione si dice non probabilistico quando il metodo utilizzato per la selezione delle unità campionarie non consente di conoscere a priori la probabilità di estrazione di ogni unità.

Lista di campionamento: lista delle unità della popolazione da cui viene estratto il campione

Ponderazione: la procedura con la quale si modifica tramite operazioni matematiche la composizione di un campione già estratto onde renderla più prossima alla distribuzione della popolazione.

Campione autoponderante: è autoponderante il campione costituito da unità selezionate con uguale probabilità.

Efficienza di un campione: un tipo di campionamento è più efficiente di un altro se si ottiene la stessa precisione nella stima con un numero inferiore di casi, oppure maggior precisione con lo stesso numero di casi. Il tipo preso come paragone è il campionamento casuale semplice (vedi).

A. Tipi di campioni probabilistici

1. **Campione causale semplice: a. con ripetizione**
   b. senza ripetizione
2. **Campione sistematico**
3. **Campione stratificato: a. proporzionale**
   b. non proporzionale
   c. ottimale
4. **Campione a stadi (unità primarie e secondarie)**
   a. a grappoli
   b. per aree

A1. Campione casuale semplice

Come si fa
La selezione di un campione di questo tipo si effettua molto semplicemente. Immaginiamo di avere la lista degli studenti iscritti alla facoltà di Scienze della formazione (ipoteticamente, N=3000 – con 'N' si indica l'ampiezza della popolazione). Supponiamo di volere un campione di n=300 soggetti (con 'n' si indica l'ampiezza campionaria). In teoria si tratta di assegnare a ciascuno studente ciascuno dei numeri da 1 a 3000, inserire in un'urna i numeri da 1 a 3000 ed estrarne 300. I possessori dei 300 numeri estratti entreranno a far parte del campione.

Se la numerazione dei soggetti è un'operazione sempre necessaria, il ricorso all'urna non lo è. Esistono programmi di computer che consentono la generazione di numeri casuali (o, meglio, pseudocasuali) in tutto e per tutto analoghi a quelli che si estrarrebbero da un'urna. In alternativa, si ricorre ad apposite tavole dei numeri casuali, prodotte da programmi come quelli citati, riportate in genere nei manuali di metodologia (anche nel vostro – per comodità una pagina è allegata anche a questa lezione). Quando i computer erano meno diffusi venivano pubblicati volumi di numeri casuali, ora ovviamente la pubblicazione di una pagina ha piuttosto lo scopo di mostrare l'uso manuale di numeri casuali. In sostanza si tratta di decidere la riga e la colonna da cui partire e la direzione di scorrimento, ovviamente prima di guardare la pagina. Si procederà poi a rilevare le cifre in gruppi dell'ampiezza richiesta dal campione: se il campione è di 900 casi, le cifre andranno rilevate a tre a tre, se fosse di 9000, a quattro a quattro.

Quando si usano i numeri casuali, può accadere che il numero corrispondente ad una persona venga estratto più di una volta. Cosa si fa in questo caso? Se si include il soggetto più volte nel campione si effettua un campione **con reimmissione** (nel caso si usi un urna, si rimette la pallina nell'urna dopo averla estratta). Se si ignora la successiva occorrenza di uno stesso numero si effettua un campionamento **senza reimmissione** (nel caso dell'urna non si rimette la pallina nell'urna stessa). Poiché nelle scienze sociali non si usa intervistare due volte la stessa persona, si opera sempre con il campione senza reimmissione.
Potrebbe sembrare che vi sia un problema legato alle probabilità di inclusione. Nel caso di reimmissione la probabilità di inclusione è facile da calcolare: se si estraggono 300 soggetti da una popolazione di 3000, la probabilità di inclusione è di 1/10. Ogni soggetto ha una probabilità di 1/3000 ad ogni estrazione, le estrazioni sono 300, dunque la probabilità totale di inclusione è di 300*1/3000=1/10. Nel caso di campione senza reimmissione la probabilità alla prima estrazione è di 1/3000, ma alla seconda è di 1/2999 dal momento che la prima pallina è stata tolta dall'urna, ecc. Questa differente probabilità di selezione ad ogni successiva estrazione non modifica tuttavia la probabilità di inclusione di ogni soggetto nel campione che è costante ed è ancora pari a 1/10. Infatti, il caso selezionato alla seconda estrazione ha sì una probabilità di 1/2999 di uscire, ma ha contemporaneamente una probabilità di 2999/3000 di essere ancora nell'urna (avrebbe potuto infatti essere selezionato come primo). Quindi la sua probabilità di inclusione è di

1/2999*2999/3000 cioè di 1/3000 ed essendo 300 i casi da estrarre la probabilità totale è sempre di 300/3000. L'unica differenza tra i due campionamenti è che la formula dell'errore campionario nel secondo caso dovrebbe contenere il fattore di correzione per popolazioni finite di cui abbiamo parlato nella lezione precedente.

La lista

Si tratta della procedura più semplice di campionamento. L'unico problema su cui bisogna porre grande attenzione è quello della lista di campionamento (la lista della popolazione). E' chiaro che tale lista deve contenere tutti i membri della popolazione considerata e soltanto loro. Inoltre ogni componente della popolazione deve figurare una sola volta, altrimenti la probabilità di estrazione sarebbe variabile da caso a caso (vedi la prossima lezione per un esempio di campione a probabilità variabili).

Applicazione

Non è la procedura più usata per la difficoltà di reperire le liste di campionamento soprattutto nei casi in cui la popolazione è distribuita in un territorio esteso: ad esempio, l'anagrafe del comune di Bologna ha la lista di tutti i cittadini, ma nessuno ha la lista dei cittadini della provincia di Bologna. Per avere quest'ultima occorrerebbe assemblare le liste dei cittadini di tutti i comuni della provincia e ciò ha dei costi non indifferenti (oltre ad ovvi problemi di accessibilità delle liste legati alla privacy). In casi come questi si preferisce il successivo tipo di campionamento.

A2. Campione sistematico

Il campionamento sistematico si considera equivalente a quello casuale semplice e si può usare tutte le volte che si usa quest'ultimo. Differisce dal casuale semplice solo per la tecnica di estrazione

Come si fa

Si scorre la lista di campionamento e si seleziona un'unità ogni *k*, dove *k* è il 'passo di campionamento', o 'intervallo di campionamento'. Il valore di *k* è pari a *N/n*, dove *N* è l'ampiezza della popolazione e *n* l'ampiezza desiderata del campione. Nell'esempio degli studenti di Scienze della formazione *n* = 300 e *N* = 3000, perciò k = 3000/300 = 10. Si seleziona pertanto uno studente ogni 10, ad esempio il primo, l'undicesimo, il ventunesimo ecc. Non si deve partire dal primo, altrimenti tutti i campioni sistematici estratti da questa lista sarebbero uguali tra loro. Per ottenere un campione esattamente di 300 casi è sufficiente partire da uno qualunque degli studenti compresi tra il primo e il decimo. Di solito si estrae a sorte un numero compreso tra 1 e *k* (10) e si inizia dal soggetto corrispondente al numero estratto. E' evidente che l'estrazione di questo tipo di campionamento è indipendente dal supporto su cui si trova la lista (vanno bene anche schedari con una scheda per unità) e non è necessario numerare le unità.

Lista

Valgono le considerazioni già fatte per il campionamento casuale semplice. In questo campionamento c'è un problema aggiuntivo da considerare. La lista non deve contenere delle ricorrenze che abbiano lo stesso passo del campionamento. Ad esempio, se k = 10 e la lista comprende militari elencati per squadra, prima il sergente, poi invariabilmente 10 militari semplici, è chiaro che si selezionano soltanto i sergenti, oppure soltanto militari semplici, secondo il punto di partenza dell'estrazione. Di solito, l'elenco alfabetico esclude periodicità di questo tipo.

Applicazione

La particolarità di questo tipo di campionamento è che può essere usato senza una preventiva lista di campionamento. Viene usato ad esempio negli exit-polls, i sondaggi effettuati all'uscita dal seggio elettorale. Si intervista un elettore ogni k tra quelli che escono dal seggio tra l'apertura e la chiusura del seggio. Come è evidente, non è necessario procurarsi preventivamente la lista degli elettori o dei votanti. La stessa cosa si può fare per campionare i clienti di un supermercato (ovviamente nessuno dispone della lista che li elenca tutti).

A3a. Campione stratificato proporzionale

Supponiamo di avere una popolazione di 9 soggetti distribuita in questo modo secondo età e reddito (in migliaia di lire):

n. soggetto	1	2	3	4	5	6	7	8	9
età	30 anni	30 anni	30 anni	40 anni	40 anni	40 anni	50 anni	50 anni	50 anni
reddito	2000	2100	2200	3000	3100	3300	4000	4100	4200

Le medie e le deviazioni standard relative alla popolazione complessiva e ai tre gruppi di età sarebbero le seguenti:

	30 anni	40 anni	50 anni	Totale
Media	2100	3100	4100	3100
Dev.standard	81,65	81,65	81,65	820,57

Come si vede, la variabilità nella popolazione è assai superiore a quella nei singoli strati di età. Nella popolazione ci sono due tipi di variabilità: quella interna ai singoli strati di età e quella tra gli strati. In altri termini, i trentenni hanno stipendi diversi tra loro, così come i quarantenni e i cinquantenni (variabilità interna agli strati). D'altra parte i trentenni hanno stupendi molto differenti rispetto ai quarantenni e ai cinquantenni (variabilità esterna, tra gli strati). Se si guarda alla distribuzione dei valori nella popolazione, si vede anche a colpo d'occhio che la variabilità interna è assai inferiore a quella esterna: le differenze tra i trentenni sono molto inferiori alle differenze tra questi e i quarantenni/cinquantenni.

Cosa accade quando campioniamo da questa popolazione (n=3). Possiamo procedere in due modi:

1. Selezioniamo tre casi dai nove complessivi con campionamento casuale semplice o sistematico.
2. Selezioniamo un caso su tre entro ciascuno strato di età (un trentenne tra i trentenni, un quarantenne tra i quarantenni, ecc.) con un separato campionamento casuale semplice (o sistematico, qui non utilizzabile dato che estraiamo un solo caso).

Nel primo campione l'errore teorico di campionamento è pari a +/- 825 (in migliaia di lire). Il calcolo è effettuato con la formula già vista nella precedente lezione, assumendo un livello di fiducia del 95%.
Nel secondo caso l'errore teorico di campionamento in ciascun sottocampione per strato è pari a +/- 107 (in migliaia di lire). Poiché l'errore complessivo in questo tipo di campionamento è la media degli errori dei singoli sottocampioni, quello appena indicato è anche l'errore sul campione di 3 soggetti selezionato con campioni separati da ciascuno strato.
Come si vede a parità di numerosità campionaria (3) l'errore campionario è molto minore nel secondo caso. Questo secondo caso è appunto costruito con la tecnica del campione stratificato proporzionale che si dimostra più efficiente del campionamento casuale semplice effettuato sulla intera popolazione.
La maggiore efficienza dipende dal fatto che la numerosità ottimale di un campione è proporzionale alla variabilità della popolazione e spesso la variabilità nei singoli strati è di dimensioni notevolmente più ridotte rispetto a quella dell'intera popolazione.

Come si fa

Si tratta di sfruttare le informazioni disponibili sulla popolazione. Se disponiamo nella lista di campionamento delle informazioni circa una variabile correlata a quelle oggetto di studio (l'età correlata al reddito) possiamo suddividere la popolazione in strati secondo i valori di questa variabile. In altri termini, dividiamo la lista di campionamento in liste separate per ciascuno strato. Effettueremo campioni separati da ciascuna di queste liste.

Torniamo all'esempio degli studenti di Scienze della formazione. Nella lista compare implicitamente o esplicitamente il sesso dello studente. Supponiamo di avere 2000 femmine e 1000 maschi. Separiamo le due liste poi stabiliamo quante femmine dobbiamo estrarre dalla lista delle femmine e quanti maschi da quella dei maschi. Il campione complessivo deve essere formato da 300 studenti, cioè da 1/10 della popolazione. Estraiamo pertanto 1/10 delle femmine (200) e 1/10 dei maschi (100). In questo modo il nostro campione è proporzionale: nel senso che femmine e maschi vi compaiono nelle identiche proporzioni in cui compaiono nella popolazione (femmine 200:300 nel campione, 2000:3000 nella popolazione).

Applicazione

Trattandosi di un campione più efficiente di quello casuale semplice, vi si ricorre sempre quando si dispone delle informazioni necessarie nella lista di campionamento. Ricordiamo ancora una volta che per selezionare le quote di soggetti da ciascuno strato si può utilizzare la tecnica del campione casuale semplice oppure quella del campione sistematico.

A3b/c. Campione stratificato non proporzionale e ottimale

Se in una popolazione uno strato è di dimensioni ridotte seguire il criterio della proporzionalità potrebbe portare alla formazione di uno strato campionario di ampiezza tanto ridotta da rendere poco significative le percentuali calcolate in questo strato. Esempio: se in una popolazione di 10000 soggetti stratificata per confessione religiosa, gli ebrei sono 300 (3%) e il campione da estrarre ha dimensioni n=500, nel campione gli ebrei dovrebbero essere 15 (500*3/100). La distribuzione percentuale di qualunque variabile entro lo strato degli ebrei sarebbe calcolata su un totale di 15 e pertanto non sarebbe significativa (il totale minimo su cui percentualizzare è compreso tra 50 e 100). In questi casi si preferisce selezionare un numero uguale di soggetti per ogni strato: ad esempio, se le confessioni religiose sono 4 si selezioneranno 125 casi da ciascuno strato.

Così facendo, si risolve il problema della numerosità minima su cui percentualizzare, ma se rende il campione non rappresentativo. Nell'esempio, infatti, gli ebrei sarebbero sovrarappresentati, mentre gli altri strati sarebbero di conseguenza sottorappresentati.

Per ovviare a questo secondo problema, si procedere in sede di elaborazione alla ponderazione del campione (vedi più avanti).

Questa non è l'unica ragione per ricorrere ad un campionamento stratificato non proporzionale. Sappiamo dalle formule relative all'ampiezza ottimale di un campione che quest'ultima è direttamente proporzionale alla variabilità della popolazione da cui lo si estrae – tanto più omogenea la popolazione tanto più piccolo potrà essere il campione. Se 1) possiamo dividere la popolazione in strati, 2) di tali strati conosciamo la deviazione standard (la variabilità) e 3) gli strati hanno variabilità diverse, possiamo effettuare un campionamento in cui selezioniamo da ogni strato un numero di casi proporzionale alla variabilità dello strato: più casi dagli strati più eterogenei (meno casi rispetto a quelli che avremmo selezionati col criterio della proporzionalità delle quote). Questo tipo di campionamento è noto come campionamento stratificato *ottimale.* Questo tipo di campionamento è ancora più efficiente di quello stratificato proporzionale (meno casi per ottenere la stessa precisione).

Vediamo con un esempio come si procede. Dobbiamo estrarre un campione di 500 casi. Innanzitutto si calcola la proporzione di casi che appartengono a ciascuno strato nella popolazione. Gli ebrei sono 300 su 10000, quindi la proporzione è di 0,03. Supponiamo che la deviazione standard della variabile oggetto di studio sia pari a 83,3 nello strato degli ebrei. Moltiplichiamo la proporzione dei casi appartenenti allo strato per la deviazione standard di questo stesso strato (83,3*0,03= 2,49), allo stesso modo procediamo per i restanti strati. Alla fine sommiamo tra loro tutti i prodotto ottenuti. Supponiamo di avere ottenuto come totale 10. Calcoliamo la proporzione di ciascun prodotto rispetto a quest'ultimo totale: per gli ebrei 2.49/10=0,249. Applichiamo questa proporzione all'ampiezza dei campione e per ogni strato otteniamo il numero dei soggetti da selezionare: per gli ebrei avremo 0,249*500 = 124,5, cioè dovremo intervistare 125 soggetti.

Ovviamente anche questo campionamento richiederà una ponderazione in sede di elaborazione per correggere il sovra o sottodimensionamento dei singoli strati.

A4. Campione a stadi (unità primarie e secondarie)

Si ricorre a questo tipo di campionamento per una o entrambe le seguenti ragioni: 1) quando manca la lista della popolazione e 2) quando la popolazione è distribuita su un territorio molto ampio e quindi l'indagine comporterebbe consistenti costi di trasferimento per gli intervistatori.

Come si fa
Se vogliamo condurre un sondaggio tra gli elettori ricorrendo ad un campione di 2000 soggetti, non possiamo certo effettuare un campionamento casuale semplice dalla lista degli italiani iscritti nelle liste elettorali, che comunque non esiste. Possiamo però procedere in questo modo: estrarre 25 province dalle 100 totali. Costruire la lista dei comuni di ciascuna delle venti province estratte. Da ciascuna di queste liste estrarre 5 comuni, in ciascun comune estrarre 4 seggi elettorali (i seggi come sapete sono numerati). Dalle liste elettorali di ciascuno dei 4 seggi, disponibili nell'ufficio elettorale di ciascun comune, estrarre 4 elettori. L'ampiezza del campione sarà quindi pari a 25*5*4*4 = 2000 come richiesto. Si possono usare 25 intervistatori, ognuno dei quali copre una provincia (cioè i 5 comuni estratti) si incarica di reperire le liste dei seggi e degli elettori dei seggi estratti e infine effettua 80 interviste. Con un campione casuale semplice avremmo avuto bisogno di più intervistatori o di far viaggiare molto gli intervistatori: non si sarebbe certo potuto assegnare a ciascuno 80 intervistati abitanti in soli 5 comuni di una stessa provincia.
Come si vede abbiamo risolto entrambi i problemi segnalati all'inizio: l'assenza della lista delle unità finali (gli elettori) e la dispersione sul territorio degli intervistati. Naturalmente tutto ha un prezzo: un campionamento come questo è meno efficiente dei tipi precedenti.
Perché è a stadi?
Non si tratta di un unico campionamento, ma di più campionamenti, per così dire, a cascata. C'è un primo stadio in cui le unità da selezionare sono le province: dalla lista di queste si estraggono 25 unità con campionamento casuale semplice, sistematico o stratificato. C'è un secondo stadio, quello dei comuni, anche qui può essere usata una delle tecniche di campionamento che abbiamo visto in precedenza. C'è un terzo stadio, quello dei seggi elettorali. C'è infine un quarto e ultimo stadio in cui si selezionano i soggetti da intervistare, sempre con una delle tecniche precedenti.
Notate che ad ogni stadio le singole unità complesse da estrarre (province, comuni, ecc.) in termini di eterogeneità della variabile oggetto d'indagine dovrebbero essere simili tra loro e mantenere il massimo di eterogeneità al loro interno. Nel linguaggio usato in precedenza, la variabilità esterna dovrebbe essere pressoché nulla, dovrebbe essere elevata la variabilità interna (diversamente dal campionamento stratificato). Se così non fosse, escludendo una provincia in cui gli elettori sono molto diversi da quelli delle altre, escluderemmo in via definitiva dal campione questo tipo di elettori. Lo stesso rischio possiamo correrlo ad ogni stadio successivo. Per questo il campione è meno efficiente. Spesso inoltre questo tipo di campione richiede procedure di ponderazione assai complesse, a meno che non si usi un particolare piano di campionamento di cui un esempio è dato nella prossima lezione.

A4a. Campionamento a grappoli

Supponiamo di voler costruire un campione di italiani. Potremmo partire dai comuni, estrarne un certo numero, reperire gli elenchi telefonici dei comuni estratti (esistono anche su supporto informatico,) e con campionamento sistematico selezionare un certo numero di abbonati (negli elenchi su supporto informatico è possibile distinguere le utenze private da quelle commerciali) . Ogni abbonato corrisponde ad una famiglia. Se si decide di intervistare tutti i membri della famiglia si sta usando un campione a grappoli, nel senso che si sono intervistati tutti i membri del grappolo famiglia. Se invece si intervista un solo membro per famiglia il campione è un 'normale' campione a tre stadi (comune-famiglia-individui). Per decidere quale membro della famiglia intervistare si usa di solito la tecnica del compleanno più prossimo alla data di rilevazione, nel senso che si intervista il membro della famiglia che ha compiuto gli anni per ultimo.

A4b. Campionamento per aree

Quando manca la lista della popolazione, si può procedere a stadi, selezionando unità geografiche sempre più piccole, fino ad arrivare ai soggetti da intervistare. Ad esempio, l'Istat ha diviso l'Italia in sezioni di censimento (ciascuna contiene circa 500 abitanti). Si può procedere così: si divide ogni provincia in tre aree: comune capoluogo, comuni entro 30 Km dal capoluogo, altri comuni. Abbiamo così circa 300 aree territoriali. Entro ciascuna possiamo campionare un certo numero di territori comunali. All'interno di ciascun territorio comunale estratto, posiamo campionare le sezioni censuarie dell'Istat. All'interno delle sezioni censuarie potremmo selezionare i numero civici, poi i piani delle abitazioni e così via.

B. Tipi di campioni non probabilistici

1. Campionamento per quote
2. Disegno fattoriale
3. Campionamento a scelta ragionata
4. Campionamento bilanciato
5. Campionamento a valanga
6. Campionamento telefonico
7. Campionamento di convenienza

In alcuni casi la reperibilità di liste delle unità da campionare è troppo dispendiosa in termini di denaro e/o di tempo (senza le liste non si può in genere usare un campione probabilistico) e non sempre è possibile ricorrere a quei tipi di campionamento probabilistico che possono essere adottati anche in assenza delle suddette liste, come il campionamento sistematico e quello a stadi.
In tali casi si ricorre a tecniche non probabilistiche di formazione del campione.
Nelle indagini campionarie accade spesso che una quota rilevante del campione risulta irreperibile o rifiuta di partecipare all'indagine. Nel caso di questionari autosomministrati tale quota arriva anche al 40% del campione. Nelle indagini telefoniche è usuale una caduta d'intervista pari al 20%. Questo fenomeno dipende naturalmente anche dalla composizione del campione e dalla natura dell'indagine. Di solito, si ovvia a questo problema predisponendo un campione di riserva dal quale si attingono casi che sostituiranno quelli che non possono essere intervistati. Tuttavia, le cadute iniziali hanno una conseguenza assai indesiderabile: la perdita di rappresentatività statistica del campione originale. Le sostituzioni non neutralizzano questo effetto. Sapendo che le cose stanno così, molti ricercatori preferiscono scegliere un piano di campionamento meno dispendioso, che non necessita di liste e di complesse operazioni di selezione. Adottano cioè piani di campionamento non probabilistici che, pur usando alcune regole circa la scelta delle persone da intervistare, consentono di scegliere gli intervistati stessi 'a caso' (non casualmente come nel campionamento probabilistico.
Naturalmente questa non è una ragione per abbandonare il campionamento probabilistico sempre e comunque. La selezione dei casi attraverso procedure oggettive garantisce comunque che gli intervistatori non scelgano gli intervistati nella cerchia dei loro conoscenti, o, comunque, tra le persone più facilmente reperibili, con gravi ed evidenti effetti distorsivi. Questo accade invece spesso nei campioni non probabilistici, nonostante le regole più o meno rigide introdotte dai ricercatori circa la scelta dei soggetti da intervistare.
Insomma, la scelta tra non probabilistico e probabilistico va fatta caso per caso, valutando sia i costi delle due soluzioni, sia i vantaggi e gli svantaggi in termini di rappresentatività del campione in quel particolare caso.

B1. Campionamento per quote

Il campionamento per quote è del tutto analogo al campionamento stratificato, proporzionale o non proporzionale. La differenza consiste solo nel fatto che, una volta divisa la popolazione idealmente oggetto d'indagine in strati, le persone da intervistare in ogni strato vengono selezionate 'a caso' (le prime che si incontrano).

Come si fa

Della popolazione non occorre avere la lista. E' necessario tuttavia conoscere la distribuzione della popolazione negli strati individuati da una variabile connessa con quella oggetto di studio (come nel campionamento stratificato), è necessario cioè conoscere le 'quote' di popolazione che appartengono ai diversi strati. Si fissa l'ampiezza del campionamento, quindi si calcolano le quote di soggetti da intervistare in ogni strato. Ad ogni intervistatore verranno assegnate quote di soggetti da intervistare, lasciandoli liberi di contattare chi credono, nel rispetto di queste quote. Gli strati possono risultare dall'uso di molteplici variabili. Questo tipo di campionamento è molto usato nelle indagini di mercato. Le variabili di stratificazione più usate sono: il sesso, l'età, il titolo di studio e la dimensione del comune di residenza, considerate contemporaneamente. Agli intervistatori si chiede di intervistare, ad esempio, 3 soggetti di sesso femminile, tra i 30 e 50 anni, con laurea e residenti in comuni di 10000-50000 abitanti, ecc. Il campione complessivo naturalmente riprodurrà nel campione la distribuzione della popolazione nei differenti sottogruppi.

B2. Disegno fattoriale

Quando si usa

Se si lavora con piccoli campioni e si è interessati a studiare specifiche relazioni tra un numero ridotto di variabili piuttosto che alla rappresentatività del campione. A rigore, può essere sia probabilistico, sia non probabilistico, secondo che per la selezione dei casi si adotti il campionamento stratificato o quello per quote. Dati gli scopi – studio di relazioni invece che interesse alla rappresentatività – di solito si adotta uno schema non probabilistico, cioè il campionamento per quote.

Come si fa

Supponiamo di voler studiare la relazione tra pratica religiosa e le due variabili indipendenti 'istruzione' e 'genere'. Se si usa un campione casuale semplice o un campione stratificato (per istruzione e genere) proporzionale, di solito si ottiene un campione in cui non c'è alcuna certezza che i maschi e le femmine abbiano la stessa istruzione media. Con simili campioni il problema è il seguente. Supponiamo di avere accertato che i soggetti più istruiti sono meno praticanti dei meno istruiti. Se ora studiamo la relazione tra genere e pratica religiosa, potremmo trovare che i maschi sono meno praticanti delle femmine. Non sapremmo però se questa è una relazione genuina, oppure se è prodotta semplicemente dal fatto che i maschi sono anche più istruiti delle femmine (i più istruiti praticano di meno). In altri termini la relazione tra genere e pratica potrebbe essere *spuria*.

Per ovviare a questo problema si adotta un disegno di campionamento non proporzionale rispetto alla popolazione. Se consideriamo genere e istruzione, abbiamo quattro possibili gruppi: maschi molto istruiti, maschi poco istruiti, femmine molto istruite, femmine poco istruite. Se il campione deve essere di 200 casi, è sufficiente selezionare 50 casi per ognuno dei sottogruppi. In questo modo quando metteremo a confronto maschi e femmine sapremo che metà dei maschi è molto istruita e metà poco e che la stessa cosa si può dire delle femmine. Se troveremo, nonostante questo campionamento, una relazione tra maschi e femmine, sapremo che si tratta di una relazione genuina e non spuria (dovuta cioè alla diversa composizione dei due gruppi secondo l'istruzione).

Problemi
Si capisce ora perché lo si usa con piccoli campioni. Per dirimere la questione delle relazioni spurie si può procedere anche in un altro modo, cioè costruendo tabelle a tripla entrata (se l'esempio avesse considerato tre variabili indipendenti, anziché due, tabelle a quadrupla entrata). Ma la frammentazione sempre più spinta del campione in sottogruppi richiesta da questo procedimento richiede ampiezze elevate del campione, altrimenti ci troveremmo con sottogruppi poco numerosi in cui il calcolo percentuale non sarebbe significativo. Il disegno fattoriale, dice il vostro testo, elimina la necessità di ricorrere a tabelle a doppia entrata. Le relazioni infatti non possono essere spurie.
Vediamo ora attraverso alcuni esempi che ciò è vero se l'unico nostro interesse è evitare le relazioni spurie, ma che in generale non possiamo trascurare di costruire tabelle a tripla entrata, se non vogliamo rischiare di perdere importanti informazioni.

La tab. 1 mostra la relazione tra pratica religiosa, genere e istruzione. Le tab. 1.1 e la tab. 1.2 mostrano le relazioni bivariate tra pratica e istruzione e tra pratica e genere. Esaminando queste ultime vediamo che c'è una relazione tra istruzione e pratica e che c'è analogamente relazione tra genere e pratica
Ci possiamo chiedere se la relazione tra genere e pratica è genuina o spuria. Trattandosi di un campionamento stratificato proporzionale non ci siamo preoccupati uniformare maschi e femmine nei riguardi della variabile istruzione. Infatti, in tab. 1 si vede come i maschi istruiti siano 400 su 500 mentre le femmine nella stessa condizione siano solo 100 su 500. Stando così le cose è assai probabile che la relazione genere-pratica sia spuria. Cosa che è documentata dalla seconda parte di tab.1. I valori percentuali mostrano che se si tiene costante il livello d'istruzione non vi è alcuna differenza di pratica tra maschi e femmine. L'unico effetto visibile è quello dell'istruzione.

La tab. 2 mostra un campione di 1000 casi come il precedente, ma estratto colla tecnica del disegno fattoriale. Si può infatti notare che i quattro sottogruppi che si possono formare incrociando genere e istruzione hanno la stessa ampiezza, 250 soggetti in tutti i casi. Se osserviamo le tab. 2.1 e 2.2 osserviamo che c'è relazione sia tra istruzione e pratica, sia tra genere e pratica. In questo caso, tuttavia, siamo certi che la seconda relazione è genuina, data l'adozione del disegno fattoriale. La controprova la troviamo in tab. 2. I valori percentuali mostrano che a parità di istruzione permane una differenza di pratica tra maschi e femmine, quindi c'è relazione tra genere e pratica.

Consideriamo però le tab. 3 , 3.1 e 3.2. Qui non c'è relazione tra genere e pratica (tab. 3.2) . Per quanto abbiamo detto fino ad ora, siamo tentati di considerarlo un risultato genuino e a non spingerci oltre. Se tuttavia non mettiamo da parte la nostra diffidenza e controlliamo la tabella a tripla entrata (tab. 3) vediamo che c'è una relazione tra genere e pratica, magari un po' inusuale, ma c'è. Va in una direzione se l'istruzione è elevata, nell'altra quando l'istruzione è bassa. Questo particolare aspetto non si sarebbe potuto cogliere se non costruendo l'incrocio a tre variabili.

B3-3. Campionamento a scelta ragionata/bilanciato/a valanga/telefonico/di convenienza

Questi campionamenti, sono utilizzati dai ricercatori con minore frequenza rispetto ai precedenti. Le procedure di selezione che li caratterizzano sono di facile comprensione, salvo forse per i campioni a scelta ragionata e bilanciato che, peraltro sono estremamente rari in letteratura. Per tutte queste ragioni rinvio semplicemente al testo.

1) Pratica religiosa per genere e istruzione in un campione stratificato proporzionale

	Istr. superiore		Istr. inferiore		Istr. superiore		Istr. inferiore	
	maschi	femmine	maschi	femmine	maschi	femmine	maschi	femmine
Praticanti	100	25	75	300	25,0	25,0	75,0	75,0
Non praticanti	300	75	25	100	75,0	75,0	25,0	25,0
	400	100	100	400	100,0	100,0	100,0	100,0

1.1) pratica e istruzione

	Assoluti		Percentuali	
	Superiore	Inferiore	Superiore	Inferiore
Praticanti	125	375	25,0	75,0
Non praticanti	375	125	75,0	25,0
	500	500	100,0	100,0

1.2) pratica e genere

	Assoluti		Percentuali	
	maschi	femmine	maschi	femmine
Praticanti	175	325	35,0	65,0
Non praticanti	325	175	65,0	35,0
	500	500	100,0	100,0

2) Pratica religiosa per genere e istruzione in un campione per quote a disegno fattoriale

	Istr. superiore		Istr. inferiore		Istr. superiore		Istr. inferiore	
	maschi	femmine	maschi	femmine	maschi	femmine	maschi	femmine
Praticanti	50	100	150	200	20,0	40,0	60,0	80,0
Non praticanti	200	150	100	50	80,0	60,0	40,0	20,0
	250	250	250	250	100,0	100,0	100,0	100,0

2.1) pratica e istruzione

	Assoluti		Percentuali	
	Superiore	Inferiore	Superiore	Inferiore
Praticanti	150	350	30,0	70,0
Non praticanti	350	150	70,0	30,0
	500	500	100,0	100,0

2.2) pratica e genere

	Assoluti		Percentuali	
	maschi	femmine	maschi	femmine
Praticanti	200	300	40,0	60,0
Non praticanti	300	200	60,0	40,0
	500	500	100,0	100,0

3) Pratica religiosa per genere e istruzione in un secondo campione per quote a disegno fattoriale

	Istr. superiore		Istr. inferiore		Istr. superiore		Istr. inferiore	
	maschi	femmine	maschi	femmine	maschi	femmine	maschi	femmine
Praticanti	50	100	200	150	20,0	40,0	80,0	60,0
Non praticanti	200	150	50	100	80,0	60,0	20,0	40,0
	250	250	250	250	100,0	100,0	100,0	100,0

3.1) pratica e istruzione

	Assoluti		Percentuali	
	Superiore	Inferiore	Superiore	Inferiore
Praticanti	150	350	30,0	70,0
Non praticanti	350	150	70,0	30,0
	500	500	100,0	100,0

3.2) pratica e genere

	Assoluti		Percentuali	
	maschi	femmine	maschi	femmine
Praticanti	250	250	50,0	50,0
Non praticanti	250	250	50,0	50,0
	500	500	100,0	100,0

C. Ponderazione dei campioni non proporzionali

Esempio: popolazione di 2000 soggetti, di cui 1000 protestanti, 800 cattolici, 200 ebrei. Formare un campione di 300 soggetti con metodo stratificato

Gli ebrei sono 1/10 (200/2000) della popolazione. Se si rispettassero le proporzioni della popolazione dovremmo avere nel campione 300*1/10= 30 ebrei. Il numero è esiguo, tale da rendere poco affidabili le percentuali calcolate in questo sottogruppo. Per risolvere il problema della numerosità minima di uno strato, si preferisce campionare in parti uguali dai diversi strati: 100 protestanti, 100 cattolici e 100 ebrei.

Nella tab. a che segue vengono presentati ipotetici risultati ottenuti da un campione stratificato non proporzionale costruito con strati di pari ampiezza. I risultati si riferiscono alla relazione tra confessione religiosa e titolo di studio.

La ponderazione può essere effettuata semplicemente applicando al campione dei pesi pari all'inverso delle probabilità di inclusione dei differenti soggetti. Il calcolo di questi pesi è esposto di seguito alla tab. a.

I valori assoluti di tab. b sono ottenuti moltiplicando i corrispondenti valori di tab. a per gli appropriati pesi calcolati. Il vantaggio di questa procedura è la semplicità. Lo svantaggio consiste nel fatto che il totale dei soggetti nella tabella coincide con l'ampiezza della popolazione (2000) invece che con l'ampiezza del campione originario (300). (In realtà i pesi non vengono applicati ai risultati sintetici delle tabelle, ma alla matrice originaria dei dati, ogni protestante viene considerato come fossero 10 soggetti, e così via, Il risultato tuttavia è identico.)

Per ovviare al problema del totale assoluto della tabella, si ricorre a pesi corrispondenti al prodotto tra l'inverso della probabilità di inclusione e la probabilità di estrazione che si sarebbe usata in caso di campionamento stratificato proporzionale. Il calcolo di questi pesi è esposto di seguito alla tab. b.

I valori assoluti di tab. c sono ottenuti moltiplicando i corrispondenti valori di tab. a per gli appropriati pesi così determinati. Come si vede la tab. c contiene 300 casi, cioè l'esatta ampiezza campionaria.

Vediamo cosa accade se percentualizziamo le tabelle. Le colonne relative ai singoli strati (protestanti, cattolici, ebrei) sono identiche nelle tre tabelle. Cambia invece la colonna dei totali tra la tab. a da un lato e le tab. b e c dall'altro (i due metodi di ponderazione portano agli stessi risultati percentuali).

In conclusione, la ponderazione è necessaria - in caso di campionamento non proporzionale - soltanto se si devono utilizzare dati relativi al campione complessivo. Non è necessaria se ci si limita a confrontare tra loro i differenti strati. Notare che nelle tabelle ponderate le percentuali relative agli ebrei sono calcolate su un totale di 30. In questo caso le percentuali sono significative, perché in realtà i casi intervistati sono stati 100, ridotti a 30 per via matematica.

Si può ricorrere alla ponderazione anche nel caso che si operi una stratificazione a posteriori del campione, o quando si debba correggere l'effetto di mancate risposte distribuite in modo diseguale tra differenti sottogruppi del campione. In questi casi le procedure di ponderazione sono statisticamente poco fondate. Si rimanda comunque al testo per altri dettagli.

a) non proporzionale

	Valori assoluti				Percentuali			
Titolo	protestanti	cattolici	ebrei	totale	protestanti	cattolici	ebrei	totale
basso	20	23	19	62	20,0	23,0	19,0	20,7
medio	40	38	21	99	40,0	38,0	21,0	33,0
alto	40	39	60	139	40,0	39,0	60,0	46,3
totale	100	100	100	300	100,0	100,0	100,0	100,0

Calcolo pesi come inverso della probabilità di inclusione

	protestanti	cattolici	ebrei
1. Numero nella popolazione	1000	800	200
2. Numero nel campione	100	100	100
Rapporto 2/1	100/1000	100/800	100/200
Probabilità inclusione	1/10	1/8	1/2
Peso=Inverso p. inclusione	10	8	2

b) ponderazione usando un peso inverso alla probabilità di inclusione

	Valori assoluti				Percentuali			
Titolo	protestanti	cattolici	ebrei	totale	protestanti	cattolici	ebrei	totale
basso	200	184	38	422	20,0	23,0	19,0	21,1
medio	400	304	42	746	40,0	38,0	21,0	37,3
alto	400	312	120	832	40,0	39,0	60,0	41,6
totale	1000	800	200	2000	100,0	100,0	100,0	100,0

Calcolo pesi come prodotto tra inverso della probabilità di inclusione reale e probabilità proporzionale teorica

	protestanti	cattolici	ebrei	totale
1. Numero nella popolazione	1000	800	200	
2. Numero nel campione	100	100	100	
Rapporto 2/1	100/1000	100/800	100/200	
Probabilità inclusione	1/10	1/8	1/2	
3. Inverso p. inclusione	10	8	2	
4. Ampiezza campione				300
5. Ampiezza popolazione				2000
Rapporto 4/5				300/2000
6. Probabilità proporzionale teorica	0,15	0,15	0,15	0,15
Peso=Prodotto 3*6	1,5	1,2	0,3	

c) peso inverso al prodotto della probabilità effettiva di estrazione per la probabilita proporzionale teorica

	Valori assoluti				Percentuali			
Titolo	protestanti	cattolici	ebrei	totale	protestanti	cattolici	ebrei	totale
basso	30	27,6	5,7	63,3	20,0	23,0	19,0	21,1
medio	60	45,6	6,3	111,9	40,0	38,0	21,0	37,3
alto	60	46,8	18	124,8	40,0	39,0	60,0	41,6
totale	150	120	30	300	100,0	100,0	100,0	100,0

TAV. A1. *Numeri casuali*

10 09 73 25 33	76 52 01 35 86	34 67 35 48 76	80 95 90 91 17	39 29 27 49 45
37 54 20 48 05	64 89 47 42 96	24 80 52 40 37	20 63 61 04 02	00 82 29 16 65
08 42 26 89 53	19 64 50 93 03	23 20 90 25 60	15 95 33 47 64	35 08 03 36 06
99 01 90 25 29	09 37 67 07 15	38 31 13 11 65	88 67 67 43 97	04 43 62 76 59
12 80 79 99 70	80 15 73 61 47	64 03 23 66 53	98 95 11 68 77	12 17 17 68 33
66 06 57 47 17	34 07 27 68 50	36 69 73 61 70	65 81 33 98 85	11 19 92 91 70
31 06 01 08 05	45 57 18 24 06	35 30 34 26 14	86 79 90 74 39	23 40 30 97 32
85 26 97 76 02	02 05 16 56 92	68 66 57 48 18	73 05 38 52 47	18 62 38 85 79
63 57 33 21 35	05 32 54 70 48	90 55 35 75 48	28 46 82 87 09	83 49 12 56 24
73 79 64 57 53	03 52 96 47 78	35 80 83 42 82	60 93 52 03 44	35 27 38 84 35
98 52 01 77 67	14 90 56 86 07	22 10 94 05 58	60 97 09 34 33	50 50 07 39 98
11 80 50 54 31	39 80 82 77 32	50 72 56 82 48	29 40 52 42 01	52 77 56 78 51
83 45 29 96 34	06 28 89 80 83	13 74 67 00 78	18 47 54 06 10	68 71 17 78 17
88 68 54 02 00	86 50 75 84 01	36 76 66 79 51	90 36 47 64 93	29 60 91 10 62
99 59 46 73 48	87 51 76 49 69	91 82 60 89 28	93 78 56 13 68	23 47 83 41 13
65 48 11 76 74	17 46 85 09 50	58 04 77 69 74	73 03 95 71 86	40 21 81 65 44
80 12 43 56 35	17 72 70 80 15	45 31 82 23 74	21 11 57 82 53	14 38 55 37 63
74 35 09 98 17	77 40 27 72 14	43 23 60 02 10	45 52 16 42 37	96 28 60 26 55
69 91 62 68 03	66 25 22 91 48	36 93 68 72 03	76 62 11 39 90	94 40 05 64 18
09 89 32 05 05	14 22 56 85 14	46 42 75 67 88	96 29 77 88 22	54 38 21 45 98
91 49 91 45 23	68 47 92 76 86	46 16 28 35 54	94 75 08 99 23	37 08 92 00 48
80 33 69 45 98	26 94 03 68 58	70 29 73 41 35	53 14 03 33 40	42 05 08 23 41
44 10 48 19 49	85 15 74 79 54	32 97 92 65 75	57 60 04 08 81	22 22 20 64 13
12 55 07 37 42	11 10 00 20 40	12 86 07 46 97	96 64 48 94 39	28 70 72 58 15
63 60 64 93 29	16 50 53 44 84	40 21 95 25 63	43 65 17 70 82	07 20 73 17 90
61 19 69 04 46	26 45 74 77 74	51 92 43 37 29	65 39 45 95 93	42 58 26 05 27
15 47 44 52 66	95 27 07 99 53	59 36 78 38 48	82 39 61 01 18	33 21 15 94 66
94 55 72 85 73	67 89 75 43 87	54 62 24 44 31	91 19 04 25 92	92 92 74 59 73
42 48 11 62 13	97 34 40 87 21	16 86 84 87 67	03 07 11 20 59	25 70 14 66 70
23 52 37 83 17	73 20 88 98 37	68 93 59 14 16	26 25 22 96 63	05 52 28 25 62
04 49 35 24 94	75 24 63 38 24	45 86 25 10 25	61 96 27 93 35	65 33 71 24 72
00 54 99 76 54	64 05 18 81 59	96 11 96 38 96	54 69 28 23 91	23 28 72 95 29
35 96 31 53 07	26 89 80 93 54	33 35 13 54 62	77 97 45 00 24	90 10 33 93 33
59 80 80 83 91	45 42 72 68 42	83 60 94 97 00	13 02 12 48 92	78 56 52 01 06
46 05 88 52 36	01 39 09 22 86	77 28 14 40 77	93 91 08 36 47	70 61 74 29 41
32 17 90 05 97	87 37 92 52 41	05 56 70 70 07	86 74 31 71 57	85 39 41 18 38
69 23 46 14 06	20 11 74 52 04	15 95 66 00 00	18 74 39 24 23	97 11 89 63 38
19 56 54 14 30	01 75 87 53 79	40 41 92 15 85	66 67 43 68 06	84 96 28 52 07
45 15 51 49 38	19 47 60 72 46	43 66 79 45 43	59 04 79 00 33	20 82 66 95 41
94 86 43 19 94	36 16 81 08 51	34 88 88 15 53	01 54 03 54 56	05 01 45 11 76

Fonte: The Rand Corporation, *A Milion Random Digits*, Glencoe, Ill., The Free Press.