Rapporto n. 1 2005
dmsia unibg.it

S. Biffignandi, D. Toninelli

Campionamento per le Indagini via Web

Serie Didattica

Dipartimento
di Matematica, Statistica,
Informatica e Applicazioni
"Lorenzo Mascheroni"

RSITÀ DEGLI STUDI DI BERGAMO

Indice

Abstract

Le indagini campionarie (*sample surveys*) via Internet e via web offrono numerose potenzialità interessanti. In particolare, è opinione diffusa che questo strumento di rilevazione consenta di raggiungere gli intervistati a costi relativamente bassi. Non si entrerà, in questa pubblicazione, nei dettagli sull'analisi dei costi delle indagini via web. Merita tuttavia di essere ricordato che, pur dovendosi confrontare con investimenti sicuramente non irrilevanti per le fasi di impostazione del questionario, di reperimento e controllo qualitativo delle liste, ecc., i costi marginali per raggiungere un numero elevato di unità, anche molto disperse sul territorio, sono molto bassi, se non nulli.

Obiettivo di questa pubblicazione è mettere in evidenza le principali problematiche e i punti di approfondimento che il ricercatore deve affrontare quando decide di condurre una indagine campionaria via Internet. La sempre più veloce e capillare diffusione in tutto il mondo della rete Internet, infatti, da un lato ha offerto innumerevoli nuove opportunità e nuovi strumenti di indagine, dall'altro ha posto all'attenzione degli studiosi nuovi problemi da risolvere e nuove tematiche di enorme importanza da affrontare.

Nella prima parte di questa pubblicazione, vengono richiamate le principali definizioni riguardanti i concetti essenziali legati alle indagini campionarie.

Nella seconda parte si effettua un breve *excursus* storico sulle origini delle indagini campionarie per arrivare ad introdurre il campionamento nell'ambito delle indagini via web.

Nella parte successiva, dopo aver brevemente illustrato i vantaggi che l'utilizzo delle indagini campionarie permette di conseguire, si prosegue affrontando il problema della scelta tra le due differenti strategie di indagine: la strategia censuaria o campionaria.

Si continua, poi, affrontando i differenti approcci alle tecniche di campionamento e prendendo in considerazione le opportunità, le caratteristiche e le problematiche peculiari a ciascuna di esse.

Le diverse tecniche di campionamento sono successivamente esaminate con specifico riferimento alle indagini via web. Le metodologie applicabili alle ricerche condotte via Internet si rifanno, in linea generale, a quelle utilizzate nelle indagini di tipo tradizionale. Anche in questo caso, dunque, è opportuno distinguere in primo luogo le tecniche di campionamento in due categorie principali: probabilistiche da un lato (Par. 6.1), non probabilistiche dall'altro (Par. 6,2).

A fronte di una classificazione che riprende quella tradizionale, vi sono, però, alcune importanti considerazioni riguardanti, nello specifico, l'utilizzo del campionamento nell'ambito di indagini via web (Cap. 7).

In particolare si affronteranno, oltre ai problemi legati alla scelta tra un approccio probabilistico o non probabilistico, quelli legati alle liste di campionamento, alla rappresentatività del campione e alle differenti modalità di selezione dei rispondenti.

Il primo aspetto di rilievo riguarda la disponibilità delle liste di campionamento: non sempre è possibile recuperare liste appropriate, e, anche nel caso le si ottenga, non sempre si rivelano essere opportunamente costruite, aggiornate, o utilizzabili per la

ricerca in questione. In particolare, nel Par. 7.1 si analizzano le principali fonti utilizzate per costruire liste di campionamento, sottolineando le problematiche che ciascuna di esse comporta. Nel Par. 7.2 si mette invece in evidenza una possibile classificazione delle fonti a seconda della connotazione probabilistica o non-probabilistica del campione che da esse può essere selezionato.

Anche nel caso si abbiano a disposizione liste di campionamento opportunamente costruite, non si possono trascurare alcune tematiche fondamentali tipiche del campionamento per indagini via web. In primo luogo gli individui che hanno accesso alla rete sono caratterizzati da particolari e ben precise connotazioni socio-demografiche ed economiche. Quindi, la applicabilità ad una certa popolazione di una indagine via web e l'approntamento di una opportuna tecnica di campionamento non può prescindere dalla conoscenza approfondita sia di coloro che hanno la possibilità di accedere alla rete Internet sia di chi non la ha. In secondo luogo, anche restringendo il campo di analisi a determinati argomenti o ad una particolare popolazione, vanno considerati con attenzione i potenziali problemi legati alla presenza di rispondenti professionisti e alla questione dell'onestà delle risposte. Questi temi sono affrontati nel Cap. 9. Per un loro approfondimento è anche possibile consultare il Quaderno di Ricerca "Inferenza nelle Indagini via Web" [1].

Per uno studio più approfondito del campionamento per le indagini via web si propone, nella parte conclusiva, oltre ad una classificazione sistematica delle differenti tipologie di indagine (Cap. 8), anche una classificazione delle diverse tipologie di rispondenti che è possibile coinvolgere in esse (Par. 9.2).

Questa pubblicazione si rivolge a tutti coloro che sono interessati ad implementare, in ambiente web, indagini campionarie: verranno affrontati i problemi di maggiore rilievo e si cercherà di presentare una panoramica esauriente delle tecniche di approccio, in modo che il lettore sia in grado di scegliere quella che risulta essere più appropriata al caso preso in considerazione dalla ricerca.

Parole Chiave

Campione, tecniche di campionamento, campionamento probabilistico, campionamento non-probabilistico, liste di campionamento, indagini campionarie, indagini via web, tipologie di indagine, rispondenti di indagini via web.

[1] Biffignandi, Toninelli (2006); si veda Bibliografia.

1 Introduzione

La raccolta di dati tramite indagini può utilizzare due diversi approcci: l'uno attinge informazioni da tutte le unità statistiche (in tal caso si parla di rilevazione censuaria o totale), l'altro considera solo una parte delle unità statistiche (si effettua cioè una rilevazione parziale - detta anche campionaria - delle unità).

Nel primo caso l'informazione raccolta descrive in modo completo la popolazione oggetto di indagine; eventuali errori o imprecisioni dipendono dalla non esatta copertura della popolazione obiettivo, da dati mancanti, e così via.

Nel secondo caso, invece, l'informazione osservata è parziale e il dato relativo alla popolazione non è noto. Tuttavia a seconda della tecnica statistica utilizzata per l'individuazione delle unità statistiche del campione, è possibile stimare il dato riferito all'intera popolazione. Le tecniche di inferenza statistica consentono, infatti, da un lato di pervenire alle stime e dall'altro di stabilirne il livello qualitativo.[2]

Nell'impostazione di ogni indagine sul campo la scelta della strategia di rilevazione (totale o campionaria) è cruciale ed è connessa ad una serie di fattori quali: i costi di rilevazione, la tempestività con cui è necessario raggiungere i risultati, la generalizzazione delle conclusioni.

L'utilizzo di indagini campionarie consente, infatti, di ridurre i costi di realizzazione di una indagine (Chisnall, 1990). In tal modo è possibile destinare maggiori risorse per attività quali il controllo della qualità del processo di raccolta e trattamento dei dati. Le indagini campionarie, inoltre, richiedono tempi di realizzazione più brevi, consentono, quindi, di ottenere risultati più tempestivi e permettono di trattare argomenti molto ampi e approfondire maggiormente i temi oggetto di ricerca. Infine, utilizzare un campione che presenta caratteristiche definite in modo preciso e dettagliato è un approccio più appropriato che operare su una intera popolazione di cui si sa molto poco. In alcuni casi, poi, è indispensabile servirsi di rilevazioni campionarie: ad esempio quando non è possibile contattare tutte le unità statistiche (quando cioè non si ha un loro elenco o non tutte sono disponibili a partecipare all'indagine), o quando l'osservazione del carattere oggetto di studio comporta la distruzione delle unità osservate.

[2] Per approfondimenti al riguardo si veda il Quaderno "Inferenza in Indagini via Web" Biffignandi, Toninelli (2006); si veda Bibliografia.

2 Alcune definizioni

Nelle indagini statistiche lo scopo è studiare un *fenomeno* che ha la tendenza a variare. Il fenomeno oggetto di studio si manifesta sulle *unità statistiche*, le quali vengono anche chiamate *supporti* del fenomeno, nel senso che questo si manifesta in un determinato modo (tra quelli possibili) su ciascuna di esse. Con *modalità* si definisce ciascuno dei possibili modi con cui il fenomeno può manifestarsi su una unità statistica.

Le unità statistiche possono essere sia individui, che oggetti, istituzioni, società, sistemi, e così via.

Nel loro insieme le unità statistiche oggetto di studio compongono la *popolazione di riferimento* (o *universo*, o *popolazione obiettivo*), ovvero l'insieme di soggetti relativamente a cui si vogliono raccogliere informazioni e a cui le conclusioni dell'indagine devono riferirsi od essere estese.

La popolazione (quindi, l'insieme di unità statistiche) può essere *finita* o *infinita*. Nel primo caso è composta da un numero finito di unità, anche se molto grande; nel secondo è costituito da un numero infinito di unità (ad esempio tutti i pezzi prodotti da un macchinario, potenzialmente infiniti se il macchinario non si rompe mai).

2.1 Indagini campionarie e campione

Per valutare come un fenomeno si presenta all'interno di una popolazione di riferimento, si osserva come esso si manifesta sulle singole unità statistiche.

L'osservazione o la raccolta dei dati può essere estesa all'intera popolazione universo; in questi casi si parla di *rilevazione censuaria* o *indagine censuaria*. Ciò di solito avviene quando si vogliono analizzare gruppi di unità molto piccoli e difficili da individuare o contattare, oppure quando si vuole condurre l'analisi ad un elevato livello di dettaglio (come avviene nei censimenti della popolazione).

In molti casi, però, non è necessario, opportuno o possibile raccogliere dati da tutte le unità che compongono la popolazione di riferimento. Non si procede, dunque, ad una rilevazione censuaria, ma si raccolgono informazioni solo su una parte delle unità statistiche oggetto di studio, cioè su un loro particolare sottogruppo, definito *campione*.

Si parla, in questo secondo caso, di rilevazione di tipo campionario o *parziale* (o anche *rilevazione campionaria*). Se la rilevazione è di tipo campionario, anche l'indagine che si sta conducendo viene definita *indagine campionaria*.

Nelle indagini campionarie, dunque, si rilevano informazioni in modo standardizzato osservando come il fenomeno che si vuole studiare si manifesta sulle unità oggetto di ricerca appartenenti ad un *campione rappresentativo*.

Il termine "rappresentativo" indica che le unità che compongono il campione vanno scelte in modo che questo rappresenti in scala ridotta l'universo di riferimento, almeno relativamente agli aspetti del fenomeno oggetto di studio o ad altre particolari caratteristiche. Più le caratteristiche del campione si avvicinano a quelle dell'universo di riferimento, più i risultati dell'indagine campionaria potranno essere considerati delle stime affidabili dell'andamento del fenomeno all'interno dell'intera popolazione.

La rilevazione parziale deve consentire di "risalire con sufficiente approssimazione alle caratteristiche complessive del fenomeno, comprese quelle relative ai casi non considerati nel campione" (Marbach, 1992). La scelta del piano di campionamento è, per questi fini, determinante.

2.2 Campionamento

Con il termine *campionamento* si indica la procedura in base alla quale, da un insieme comprendente le unità oggetto di studio (popolazione di riferimento), si estrae un numero ridotto di casi. Questi devono essere selezionati in modo tale che sia poi possibile generalizzare i risultati ottenuti dal campione all'intera popolazione, cioè in modo tale che, complessivamente, siano rappresentativi dell'universo (campione rappresentativo).

Punto di partenza per il campionamento della popolazione è la *lista di campionamento* (*sampling frame*), ovvero l'elenco contenente gli individui che risultano essere idonei ad essere coinvolti nell'indagine e che sono raggiungibili o possono essere contattati. In primo luogo è necessario accertarsi che la lista sia appropriata per la popolazione obiettivo che si vuole studiare sottoponendola ad un attento vaglio. Ciò permette di limitare il più possibile l'influenza di errori sistematici o originati da una eventuale distorsione del campione che porterebbe all'ottenimento di risultati distorti. È utile, a questo scopo, fissare dei *criteri di inclusione e di esclusione*, che permettano di stabilire senza ambiguità se una unità statistica può o meno essere inserita nella lista di campionamento.

Vi sono differenti tecniche di campionamento utilizzate per ottenere campioni rappresentativi della popolazione universo. Nel seguito (Cap. 6) verranno segnalate le principali, classificate in base al fatto che si tratti di tecniche probabilistiche (Par. 6.1) o meno (Par. 6.2).

Si definisce *numerosità campionaria* il numero di casi scelti, cioè il numero di unità statistiche che entrano a far parte del campione. Per piccole popolazioni, un campione può rappresentare anche il 10-30% della popolazione obiettivo. Per popolazioni molto grandi (in genere si considerano "popolazioni grandi" quelle composte da più di 150.000 individui) la popolazione campionata può aggirarsi attorno all'1% circa di quella totale.

Vi sono semplici formule che permettono di calcolare la numerosità campionaria in relazione al piano di campionamento scelto (per una trattazione più approfondita di questo argomento si rimanda a De Carlo e Robusto, 1996).

Per stabilire quale sarà la dimensione del campione, nella maggior parte dei casi è importante conoscere alcune caratteristiche della popolazione, informazioni che saranno molto utili anche nella fase di analisi dei dati. Per questo motivo, ad esempio, spesso è molto utile rilevare, nell'ambito di una indagine, anche dati riguardanti la conformazione demografica della popolazione obiettivo: età, sesso, professione, e così via. In linea generale si adotta il seguente criterio: maggiore è la variabilità che la popolazione presenta riguardo a queste variabili, maggiore deve essere la numerosità campionaria.

Operando su un campione relativamente molto numeroso, sarà possibile catturare larga parte della variabilità della popolazione, e questo è importante anche in fase di analisi.

Infatti molti test statistici sono basati sul principio della variabilità e, utilizzando campioni che la catturano in larga parte, ci si assicura di potere sfruttare tutte le loro potenzialità.

3 Dal 1800 all'avvento di Internet

Il primo rudimentale approccio alle indagini campionarie avvenne nel 1786, quando Laplace stimò la popolazione della Francia utilizzando un campione. Laplace, inoltre, calcolò, per la prima volta, stime probabilistiche dell'errore: queste non erano espresse negli stessi termini dei moderni intervalli di confidenza, ma attraverso la dimensione del campione necessaria ad assicurare un particolare limite superiore all'errore di campionamento, fissato un certo livello di probabilità.

L'utilizzo delle prime indagini campionarie risale, però, all'ultimo ventennio del XVII secolo. Il primo tentativo di implementazione fu intrapreso nel 1880 da K. Marx, il quale inviò 25.000 copie di un questionario ai lettori della rivista *Rèvue socialiste*. Lo scopo era di indagare sulle condizioni di vita dei lettori attraverso una serie di domande aperte.

Anche M. Weber tra il 1880 e il 1910 fece più volte ricorso a questionari per studiare problemi sociali riguardanti tematiche quali "le condizioni di lavoro nella Prussia orientale" o "gli effetti del lavoro nella grande industria sulla struttura della personalità e degli stili di vita degli operai". Nel primo caso vennero inviati questionari postali ai proprietari agricoli e ai pastori protestanti; nel secondo i questionari vennero sottoposti sia ad osservatori privilegiati che ad un campione di operai.

Attorno ai primi anni del 1900, però, si continuava a ritenere che, per raggiungere conclusioni su un fenomeno attinente ad una popolazione finita, l'unico metodo di ricerca affidabile fosse osservare il fenomeno su tutte le unità componenti la popolazione stessa. Solo più tardi, comunque nei primi decenni del '900, si ebbe un importante salto di qualità nell'utilizzo delle indagini campionarie. La storia delle teorie riguardanti le indagini campionarie è, dunque, relativamente breve: basti considerare che la maggior parte dello sviluppo metodologico relativo ad esse è avvenuto negli ultimi 50 anni.

A. N. Kiaer, direttore del *Norwegian Bureau of Statistics*, fu uno dei precursori delle moderne tecniche di indagine. A partire dalle riunioni dell'Istituto Internazionale di Statistica (ISI) tenutesi a Berna nel 1895, egli propose le metodologie da lui utilizzate e i risultati ottenuti contestualmente a studi riguardanti alcuni caratteri della popolazione Norvegese. Per la prima volta si ebbe l'occasione di confrontarsi con l'utilizzo dei campioni per la raccolta di dati statistici. Kiaer, allo scopo di ottenere un campione rappresentativo, poneva come presupposti il coinvolgimento di un gruppo numeroso di unità distribuite su tutto il territorio considerato; non era necessario che la selezione di queste unità avvenisse secondo i principi di casualità, ma poteva svolgersi in accordo con uno schema logico basato sui risultati generali di indagini statistiche precedenti.

Del metodo di indagine adottato da Kiaer si parlò anche successivamente. Nell'incontro dell'ISI, tenutosi a St Petersburg nel 1899, Kiaer sostenne che la rappresentatività relativa al campione poteva essere ottenuta nel caso in cui esso rispecchiasse, relativamente ad alcune variabili, la composizione della popolazione universo. Uno dei modi di operare ritenuti opportuni per ottenere risultati corretti era basato sul confronto con i risultati di indagini censuarie (quali, ad esempio, i censimenti della popolazione): se, rispetto alle variabili su cui si erano raccolti i dati, il campione risultava essere simile

alla popolazione universo, allora probabilmente esso poteva essere rappresentativo anche riguardo ad altre variabili non espressamente considerate nella ricerca censuaria.

Nonostante queste premesse rivoluzionarie, non si poteva però ancora parlare di una teoria rigorosa che avrebbe consentito di estendere i risultati della raccolta di dati campionari all'intera popolazione; inoltre non poteva aver luogo, in tale modo di operare, una valutazione oggettiva della qualità dei dati e delle conclusioni raggiunte dalla ricerca.

Solo nel 1926 Bowley, rifacendosi all'approccio Bayesiano e al teorema centrale del limite di Edgeworth, fu, per primo, in grado di stabilire la precisione delle stime di grandi campioni selezionati con procedimento casuale da popolazioni finite di grandi dimensioni.

Le teorie di Bowley (basate sulla selezione casuale del campione) e quelle di Kiaer (che nelle procedure di selezione del campione utilizzava metodi tutt'altro che casuali), però, non erano, all'epoca, facilmente integrabili tra di loro.

Solo attorno al 1933, grazie all'importante contributo di J. Neyman, si ebbe un notevole sviluppo degli studi riguardanti le tecniche di campionamento: la pubblicazione *"On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection"* (1934) è considerata alla base dell'attuale impostazione delle teorie campionarie. Le procedure di estrazione considerate sono basate su principi di casualità, che consentono di assicurare la rappresentatività non di un singolo campione, ma del complesso dei campioni che possono, con tali procedure, venire selezionati. Fu, quindi, attorno agli anni '30 che si cominciò a parlare del concetto di "rappresentatività" e prese avvio la diffusione della convinzione che, per conoscere la distribuzione di un fenomeno (o di una serie di variabili) all'interno di una determinata popolazione e per ottenere risultati affidabili non era necessario studiare tutti gli individui che la compongono. Neyman aveva infatti sottolineato come un campione, scelto opportunamente, fosse in grado di produrre risultati altrettanto accurati di quelli di una rilevazione censuaria, o addirittura migliori, visto che le risorse risparmiate con la riduzione dell'ampiezza della rilevazione potevano essere impiegate per migliorarne la qualità.

Nel 1936 le indagini campionarie vennero utilizzate, con esiti alterni, in occasione delle elezioni presidenziali degli Stati Uniti. Per cercare di prevedere quale dei due candidati (F.D. Roosvelt o A. Landon) avrebbe vinto, vennero condotte due indagini. La prima, realizzata dal Literary Digest, prevedeva l'invio di 10 milioni di fac-simile delle schede elettorali a nominativi estratti dagli elenchi telefonici e dai registri automobilistici. Le risposte ottenute furono circa 2 milioni e portarono a prevedere la vittoria di Landon (59%) sul rivale Roosvelt (41%). Per una analoga indagine progettata dalla Gallup vennero realizzate migliaia di interviste rivolte ad elettori estratti casualmente dall'intera popolazione. In questo secondo caso i risultati si rivelarono opposti: Roosvelt era dato vincente con il 60% delle preferenze, a fronte del 40% ottenuto, nelle interviste, da Landon. I risultati finali delle elezioni diedero ragione a questa seconda indagine: Roosvelt vinse con il 61% dei voti sul rivale.

Cosa non era andato per l'indagine del Literary Digest? Che problemi avevano portato a risultati tanto lontani dalla realtà? Da casi come questo si comprende quanto sia importante studiare la qualità dei dati raccolti e dei risultati a cui si è pervenuti, tenendo conto di quali metodologie sono state adottate per la ricerca. Nel particolare episodio

delle elezioni presidenziali USA certamente è possibile che abbiano avuto un grosso influsso gli errori di copertura e l'auto-selezione del campione. Di questi e di altri fattori che possono essere all'origine della distorsione dei risultati campionari si parlerà diffusamente nel seguito e, in modo più approfondito, nel Quaderno di Ricerca "Inferenza nelle Indagini via Web" [3].

Le attività sviluppate nella seconda metà del '900, successivamente alle teorie elaborate da Neymann, hanno prevalentemente approfondito gli aspetti metodologici delle tecniche di campionamento e hanno riguardato, in particolar modo, le differenti tecniche di selezione del campione e le problematiche relative alla stima dei parametri della popolazione.

Tra la fine degli anni '80 e l'inizio del decennio successivo, prima che si verificasse una diffusione su larga scala di Internet, presero avvio gli studi sulla possibilità di effettuare indagini tramite posta elettronica. Il sistema, sfruttando le *e-mail*, permetteva la trasmissione dell'indagine a costi molto bassi (quando, addirittura, non nulli). Il limite delle *e-mail* fu, però, presto evidente, trattandosi di codice ASCII che permetteva unicamente di inviare, via Internet, del semplice testo.

Le indagini via web cominciarono a divenire di diffuso utilizzo a partire dalla metà degli anni '90: velocemente questo tipo di indagine soppiantò le *e-mail* come veicolo per realizzare indagini via Internet. Mentre le prime *e-mail* erano basate sul codice ASCII, il web offrì la possibilità di indagini multimediali contenenti allegati audio e video, dando la possibilità di ottenere un coinvolgimento maggiormente interattivo del rispondente, che rendeva le fasi di risposta più piacevoli ed accattivanti. Inoltre si superò l'ostacolo della necessità di conoscenza dell'indirizzo di posta elettronica dei rispondenti.

Negli ultimi anni, la sempre più capillare diffusione della rete Internet tra la popolazione di tutto il mondo ha avuto una brusca accelerazione. Questo da un lato ha offerto innumerevoli nuove opportunità e nuovi strumenti di cui servirsi per realizzare le indagini campionarie, dall'altro ha posto all'attenzione degli studiosi nuovi problemi da risolvere e nuove tematiche di enorme importanza su cui dibattere.

Nonostante tutti i difetti e i problemi che, allo stato attuale, possono presentare, le indagini campionarie via web sono considerate uno strumento versatile e molto utile che, nel futuro, potrà essere usato ancora più diffusamente. Internet consente, infatti, di raggiungere una ulteriore ottimizzazione delle indagini, soprattutto dal punto di vista dei tempi (notevolmente compressi) e dei costi (possono subire un netto abbattimento) di realizzazione.

L'utilizzo di metodologie di raccolta dati che si basano sul supporto di Internet sta già avendo ripercussioni radicali sulle modalità di svolgimento delle ricerche su campo.

---

[3] Biffignandi, Toninelli (2006); si veda Bibliografia.

4 I vantaggi delle Indagini Campionarie

Si è già accennato, in modo sintetico, ai vantaggi che comporta l'utilizzo di indagini campionarie, in questo paragrafo si presenteranno in modo più approfondito alcuni dei motivi principali per cui si ricorre all'utilizzo di indagini campionarie.

In primo luogo un'indagine campionaria è indispensabile quando non è possibile effettuare una rilevazione totale: non si possono contattare tutte le unità statistiche della popolazione universo, non si ha un elenco in base a cui poterle rintracciare o non tutte sono disponibili a partecipare all'indagine. Si pensi, ad esempio ai turisti stranieri arrivati in Italia nell'arco di un anno: se si volesse fare una indagine su di essi sarebbe impossibile riuscire a rintracciarli ed intervistarli tutti.

Si ricorre alla rilevazione parziale anche quando l'osservazione del carattere oggetto di studio comporta la distruzione delle unità statistiche osservate, come succede molto di frequente nelle rilevazioni finalizzate al controllo statistico della qualità. Si pensi ai test effettuati per stabilire la durata media delle lampadine prodotte da un impianto. Per testare la durata, la lampadina viene resa inutilizzabile: è chiaro che non è possibile bruciare tutte le lampadine prodotte. Oppure, analogamente, si pensi ai test per verificare la resistenza alla trazione di una serie di componenti meccaniche. Anche in questo caso per rilevare il fenomeno oggetto di studio rende inutilizzabile l'unità oggetto della rilevazione.

L'utilizzo di indagini campionarie consente inoltre, come già si accennava in precedenza, di ridurre i costi di realizzazione di una ricerca: "il costo di una indagine effettuata su un campione di individui è inferiore a quello di una rilevazione complessiva (censimento) anche se il costo per unità elementare intervistata può risultare superiore in conseguenza della necessità di impiegare intervistatori esperti, delle spese amministrative di progettazione del campione e di impostazione generale del lavoro" (Chisnall, 1990). Spesso, quindi, sono necessità esclusivamente legate a costi di realizzazione che spingono a servirsi di un campione: potrebbero, cioè, non essere disponibili fondi necessari per una rilevazione censuaria. Ad esempio, se si vuole conoscere il livello di soddisfazione dei consumatori di una diffusa marca di detersivo, una cosa è intervistare tutti i consumatori (operazione che richiederebbe costi molto elevati), un'altra è sottoporre un questionario ad un campione rappresentativo di essi, molto ridotto in termini numerici. I costi (stampa del questionario, invio dei questionari, inserimento ed elaborazione dati, ...) saranno, nel secondo caso, molto minori. La riduzione dei costi relativi alla realizzazione dell'indagine consentirà, in questo modo, di utilizzare maggiori risorse per altre attività fondamentali, quali il controllo della qualità del processo di raccolta ed elaborazione dei dati.

Le indagini censuarie richiedono tempi di realizzazione molto più lunghi di quelli necessari per indagini campionarie; questo è un aspetto di grande rilievo, se si considera che quasi sempre i risultati di una ricerca devono essere forniti in tempi brevi per essere di qualche utilità. Ad esempio, se si devono realizzare delle interviste volte a rilevare la soddisfazione in merito ad alcuni provvedimenti presi dallo Stato riguardo ad alcune tematiche sociali, non è opportuno intervistare l'intera popolazione dello Stato, ma è preferibile selezionare un campione rappresentativo di individui. In questo modo non

solo le risorse richieste, ma anche i tempi necessari per ottenere dei risultati significativi sono molto minori. Effettuare una indagine censuaria, invece, potrebbe portare a risultati tardivi che, una volta disponibili, potrebbero non essere più significativi, utili o sufficientemente aggiornati.

In alcuni contesti, poi, non è opportuno (o addirittura è controproducente) realizzare indagini censuarie. Si pensi a test di mercato effettuati per il lancio di un nuovo prodotto. Per questioni di segretezza nei confronti della concorrenza o per il fatto che alcune variabili di marketing non sono ancora nettamente definite non è opportuno operare con una ricerca estesa all'intera popolazione.

Altro vantaggio che si consegue dall'utilizzare rilevazioni campionarie è il fatto che, in virtù della riduzione di tempi e costi, si possono, allo stesso tempo, trattare argomenti molto ampi e, nell'ambito della ricerca, si può di andare più in profondità, anche allo scopo di individuare informazioni di interesse per gli scopi dell'indagine.

Infine, il fatto di poter fare indagini su un campione che presenta caratteristiche definite in modo preciso e dettagliato è non di rado un approccio più appropriato che operare su una intera popolazione di cui si sa molto poco. Come si trova riscontro in Marbach (1992, citazione di Giusti), non è detto, poi, che la rappresentatività di rilevazioni censuarie sia superiore in ogni caso a quella di indagini campionarie: "L'errore di osservazione nelle rilevazioni complete è altrettanto importante o forse più importante che nelle rilevazioni parziali". Il fatto che il campionamento sia affetto da un errore dovuto alla natura parziale della rilevazione "non può e non deve fare concludere che i risultati censuari siano di qualità superiore. [...] Anzi, nelle indagini per campione molte fonti di errore possono essere meglio tenute sotto controllo".

Degli errori campionari e non campionari ci occuperemo, in particolare, nel Quaderno di Ricerca "Inferenza nelle Indagini via Web" [4].

[4] Biffignandi, Toninelli (2006); si veda Bibliografia.

5 Le Indagini Campionarie via Web

Con la diffusione sempre più capillare di Internet in tutto il mondo, le innovative opportunità offerte dalla rete hanno portato ad un utilizzo parzialmente differenziato delle tecniche di indagine tradizionali. Nonostante nell'ambiente web l'approccio alle indagini campionarie sia, in larga misura, simile a quello tradizionale (in particolar modo considerando le tecniche di campionamento, come si vedrà più avanti), l'implementazione di indagini che si avvalgono del supporto della rete propone all'attenzione degli studiosi alcuni problemi che sono specificamente caratteristici di esso.

In questo capitolo verranno affrontati, in particolare, i vantaggi e gli svantaggi legati alla realizzazione di indagini campionarie via web.

5.1 Vantaggi delle indagini via web

Le indagini campionarie via web consentono, rispetto alle modalità di indagine tradizionali, di conseguire una serie di vantaggi non trascurabili, vantaggi ancora più di rilievo se si pensa al costo relativamente molto basso di questo tipo di approccio.

I tre principali vantaggi che è possibile ottenere riguardano la copertura geografica, la possibilità di utilizzare campioni più numerosi e quella di conseguire una copertura maggiore della popolazione.

Copertura Geografica

Con un campionamento via web non si hanno limiti di spazio: la copertura territoriale è (almeno potenzialmente) molto più ampia di quella ottenibile, ad esempio, con le interviste faccia a faccia o telefoniche. Il questionario, inviato via *e-mail* o pubblicato sulle pagine di un sito Internet, può raggiungere qualunque parte del mondo senza che vi siano elevati costi di diffusione dello stesso. Inoltre la supervisione, il coordinamneto e l'eventuale invio del questionario sono operazioni che possono essere fatte da un unico luogo fisico, senza la necessità di aprire sedi-satellite per "avvicinarsi" alla popolazione obiettivo.

Ciò non accade, ad esempio, nell'ambito delle interviste faccia-a-faccia, dove è invece necessario che l'intervistatore contatti direttamente i rispondenti e quindi l'ambito territoriale di interesse deve essere per forza di cose circoscritto; inoltre, per le interviste, si ha bisogno di intervistatori ben istruiti e sul loro operato devono essere effettuati dei controlli: la dispersione territoriale del campione non consente di effettuarli agevolmente.

Anche le interviste telefoniche consentono di ottenere una copertura geografica piuttosto ampia e possono essere condotte da un'unica sede centrale, ma in questo caso i costi per le chiamate a lunga distanza possono rivelarsi proibitivi da sostenere. Inoltre, nonostante il grado di diffusione raggiunto dalle linee di telefonia fissa, non è certo che la popolazione obiettivo abbia un accesso alla linea telefonica (ma questo è un problema comune anche per le indagini via web); possono verificarsi evidenti problemi, ad

esempio, nel caso in cui larga parte dei potenziali rispondenti risieda all'estero o abbia esclusivamente un apparecchio telefonico mobile.

Anche nell'utilizzare indagini postali il campione è raggiungibile, da una sede centrale, pressoché ovunque, ma bisogna aspettarsi tempi di risposta molto più lunghi rispetto a quelli richiesti dalla indagini via web; grazie ad Internet è infatti possibile raccogliere e registrare, in tempo reale, le risposte ad un questionario compilato in luoghi molto remoti. Inoltre nelle indagini postali i costi di spedizione costituiscono una voce di rilievo, mentre con le indagini via web è possibile abbattere notevolmente i costi marginali di contatto del campione: voci di costo di rilievo quali costi di spedizione, di contatto telefonico o dovuti agli spostamenti dell'intervistatore sono completamente azzerati.

Numerosità dei campioni

Il costo unitario dell'invio di un questionario via *e-mail* è pressoché nullo: una stessa *e-mail* può essere inviata, approssimativamente allo stesso costo, a 10 come a 5.000 indirizzi. Queste caratteristiche agevolano lo studio di gruppi molto più numerosi e anche di più gruppi differenti, riducendo notevolmente i problemi legati al *budget*.

Maggiore copertura della popolazione campione

Identificata una popolazione obiettivo, con le indagini via web è possibile ottenere una copertura potenzialmente maggiore, se si trascurano i problemi di uniforme diffusione della rete in tutta la popolazione (problemi che comunque si stanno progressivamente riducendo). Da un lato il fatto che si tratti di questionari auto-somministrati incoraggia i rispondenti restii a partecipare all'indagine per non mettere in pericolo la propria *privacy*. Dall'altro si supera il problema delle persone che sono riluttanti a rispondere a questionari telefonici perché hanno il timore che gli si voglia vendere qualcosa.

Inoltre le indagini via web sono risultate particolarmente adatte per ottenere dati da categorie di individui di solito piuttosto avverse a partecipare a ricerche, quali medici, politici, amministratori di aziende, e così via. Si tratta di persone difficili (o impossibili) da contattare per interviste faccia-a-faccia o telefoniche, dato che non hanno molto tempo libero. Il questionario, inoltre, non deve essere rispedito per posta e può essere compilato in qualsiasi momento della giornata, quando è più comodo per chi deve rispondere, senza che vi sia la necessità di fissare appuntamenti telefonici o per indagini *face-to-face*.

5.2 Svantaggi delle indagini via web

Le indagini via Web comportano anche alcuni problemi che ne limitano la applicabilità o che ne condizionano gli ambiti di utilizzo. Questi problemi vanno studiati con attenzione, per evitare di ottenere dati distorti o non utili ai fini della ricerca che si sta effettuando.

Le principali difficoltà sono legate alla costruzione del questionario, alla sua amministrazione e alle tecniche di campionamento. In questo paragrafo ci si occuperà di alcuni problemi connessi a questo ultimo aspetto: la disponibilità delle liste di campionamento, gli errori di copertura, l'approccio probabilistico, le distorsioni dovute alle mancate risposte, il tasso di risposta, il grado di alfabetizzazione, le lingue, l'identità del rispondente.

Disponibilità e rappresentatività delle liste

Esiste il rischio molto frequente di non avere liste di campionamento disponibili, oppure che quelle che si hanno siano incomplete o inaccurate; la conseguenza è che i risultati che si ottengono dall'indagine non potranno poi essere estesi al resto della popolazione. In questi casi può essere necessario ricorrere a modi alternativi per identificare e contattare il campione. Ad esempio, possono essere condotte delle indagini censuarie sulla popolazione obiettivo per stabilire qual è la popolazione da cui il campione deve essere estratto o quale sia la migliore tecnica di campionamento da utilizzare. È chiaro che, in questi casi, il dispendio di risorse risulta ingente.

Spesso, anche avendo a disposizione liste di campionamento, queste non consentono di ottenere campioni rappresentativi della popolazione obiettivo. Per tale ragione, è sempre indispensabile fare controlli accurati sulle liste stesse. Frequentemente esse comprendono solo individui con certe caratteristiche particolari e non sono in grado di rappresentare o coprire in modo significativo una popolazione più ampia. Il problema si aggrava ancor più quando non ci sono dati (o non ce ne sono a sufficienza) per contraddistinguere, all'interno delle liste di campionamento, determinate categorie di unità che sarebbero di elevato interesse per gli scopi della ricerca.

Una discussione più approfondita sulle problematiche riguardanti le liste di campionamento e sulle loro modalità di reperimento nell'ambito delle indagini via Web verrà affrontata nel Par. 7.1.

Errore di copertura (*Coverage error*)

Si definisce con questa espressione la differenza tra la lista di campionamento (*sampling frame*) e la popolazione obiettivo (*target population*). L'importanza dell'errore di copertura dipende dalla popolazione con cui si ha a che fare.

Se, ad esempio, si volesse svolgere una indagine campionaria sui clienti registrati ad un sito web utilizzando l'elenco di indirizzi *e-mail* da essi forniti al momento della registrazione, non si avrebbe alcun errore di copertura selezionando il campione della lista citata; infatti la lista di campionamento "copre" tutte le unità della popolazione obiettivo, e quindi consentirebbe di estrarre un campione rappresentativo. Invece, nel caso, ad esempio, si voglia fare un'indagine campionaria via *e-mail* riguardante tutti i cittadini italiani che votano per le elezioni politiche, l'errore di copertura potrebbe condurre a risultati distorti. Infatti, da una parte non esiste una lista completa di tutti gli indirizzi *e-mail* degli elettori, dall'altra non tutti i potenziali elettori hanno a disposizione un indirizzo di posta elettronica o hanno la possibilità di connettersi a Internet.

Non bisogna dimenticare, inoltre, che per redigere una lista di campionamento per indagini via web è necessario essere in grado di identificare tutte le unità che andrebbero incluse. Di solito questa situazione si verifica quando si ha a che fare con popolazioni numericamente non consistenti: ad esempio, gruppi di individui di cui si conoscono gli indirizzi *e-mail*, visitatori di determinati siti web, e così via. Nella maggioranza dei casi, però, non è possibile conoscere ogni membro della popolazione universo, soprattutto quando questa è molto numerosa; oppure non esiste una lista che possa essere utilizzata per il campionamento. Si pensi, ad esempio, al caso in cui si voglia svolgere un'indagine sugli adulti residenti negli Stati Uniti: non vi è alcuna lista che comprende tutti gli individui oggetto di studio.

Va anche tenuto conto, inoltre, che non tutti i potenziali rispondenti hanno la possibilità di connettersi alla rete Internet per partecipare all'indagine. La *NUA Ltd.*[5], ad esempio, nel Febbraio 2002 ha stimato che in tutto il mondo aveva accesso alla rete solo l'8,96% della popolazione (corrispondente a 544,2 milioni di individui), mentre in Nord America la possibilità di connettersi ad Internet riguardava circa il 58,5% della popolazione (164,4 milioni di individui). Negli ultimi anni la diffusione delle connessioni al web si è ulteriormente incrementata, ma, nonostante ciò, non è ancora possibile riuscire a contattare, ad esempio, popolazioni estese come, ad esempio, quella di uno stato come gli Stati Uniti.

Il problema degli errori di copertura verrà affrontato più diffusamente nel Quaderno di Ricerca "Inferenza nelle Indagini via Web" [6].

Approccio probabilistico nella selezione del campione

Si vedrà nel seguito (Cap. 6) che l'approccio per la fase di selezione del campione può essere di tipo probabilistico o non probabilistico. Il campionamento probabilistico consente di utilizzare gli strumenti dell'inferenza statistica per estendere i risultati ottenuti da un campione all'intera popolazione universo. Nel caso delle indagini via web, però, l'approccio probabilistico ai metodi di campionamento è piuttosto arduo da ottenere.

Distorsione dovuta alle mancate risposte

Anche nel caso in cui vengono contattati tutti i soggetti selezionati da una lista di campionamento opportunamente costruita, si avrebbe poi a che fare con la distorsione originata dalle mancate risposte. Infatti non tutti i soggetti contattati con successo potrebbero volere partecipare all'indagine o sarebbero in grado di farlo. L'importanza della distorsione da mancata risposta dipenderà da due fattori: da una parte l'incidenza delle mancate risposte, dall'altra come i rispondenti differiscono dai non rispondenti riguardo alle variabili di interesse. I parametri di interesse potrebbero così risultare confusi da variabili che determinano la partecipazione all'indagine.

Vi è da dire che il problema delle mancate risposte è in comune con le tradizionali tecniche di indagine, ma è ancor più accentuato per le indagini via web, dato che i tassi di risposta sono generalmente piuttosto bassi e che si hanno grosse difficoltà a selezionare un campione casuale di rispondenti.

Inoltre, per le indagini condotte via web, le mancate risposte potenzialmente sono più numerose a causa della a volte carente confidenza dei rispondenti nei confronti della tecnologia: essi, infatti, devono possedere una specifica cultura di base che comprenda la capacità di muoversi in rete, di usare un mouse o capire come compilare un questionario (come selezionare le possibilità di risposta da un menu, come indicare in modo corretto la propria risposta, e così via). Le possibili difficoltà legate al *background* culturale dei rispondenti possono essere aggravate anche da problemi tecnici come l'incompatibilità del questionario con determinati *browser* o la lentezza della connessione (Couper e altri, 2001).

---

[5] http://www.nua.ie/surveys/how_many_online/index.html

[6] Biffignandi, Toninelli (2006); si veda Bibliografia.

Un altro aspetto di rilievo legato a questo problema è che l'accesso a Internet (e la padronanza degli strumenti utilizzati) tende ad essere correlato con caratteristiche demografiche ben definite (reddito, età, residenza geografica, ...) e l'indagine può portare a risultati distorti se queste caratteristiche demografiche hanno effetto sulle variabili di interesse.

Per minimizzare i problemi dovuti alla distorsione originata dalle mancate risposte e per incrementare la rappresentatività del campione si possono ponderare i risultati con sistemi di pesi semplici o utilizzando il sistema di ponderazione *propensity score*[7]. I sistemi di ponderazione potrebbero essere una buona soluzione al problema, ma la difficoltà principale che si riscontra nelle indagini via web è il fatto che si hanno a disposizione pochissime informazioni sulla categoria dei non rispondenti. Spesso è quindi opportuno costruire un appropriato sistema di pesi grazie all'integrazione con altre tipologie di tecniche di indagine.

Basso tasso di risposta

Uno dei problemi più significativi legati alle indagini via web è il basso tasso di risposta. Nell'impostare e pianificare la strategia campionaria non si può non tenere conto di questo aspetto: in caso contrario si rischia che alcune fasce della popolazione obiettivo non siano presenti nel campione. Vi è da dire, però, che nelle indagini via web l'ampiezza del campione può essere incrementata senza che i costi di realizzazione dell'indagine ne risentano particolarmente. Inoltre, per incrementare il tasso di risposta, si possono utilizzare diversi accorgimenti come l'invio di *e-mail* di presentazione dell'indagine, gli incentivi, i contatti telefonici, personali o postali con i potenziali rispondenti, e così via.

Grado di alfabetizzazione

In comune con altre tecniche tradizionali (e in particolare con le indagini postali) per le indagini via web vi è un ulteriore rischio, quello di non riuscire ad intercettare le persone analfabete, o che, comunque, non sono in grado di rispondere al questionario. Il problema può rivelarsi di importanza non trascurabile, se il tasso di analfabetismo è particolarmente elevato (negli Stati Uniti, ad esempio, è attorno al 20%).

A parte chi non è in grado di leggere, ci sono determinate categorie di individui (anziani, persone con la vista debole, dislessici) che potrebbero considerare eccessivo lo sforzo di leggere un questionario e che, quindi, probabilmente rinunceranno a partecipare all'indagine o compileranno solo parzialmente i questionari.

I due problemi visti, combinati tra di loro, possono portare il ricercatore a dover rinunciare a dati provenienti da fette di popolazione anche piuttosto consistenti.

Una possibile soluzione a queste difficoltà può essere quella di integrare i dati dell'indagine via web con altri rilevati tramite tecniche alternative (ad esempio con interviste faccia a faccia su un campione mirato).

Lingue

---

[7] Questi argomenti sono approfonditi nel Quaderno di Ricerca "Inferenza nelle Indagini via Web" (Biffignandi, Toninelli, 2006); si veda Bibliografia.

Un altro problema che possono presentare, in generale, i questionari da auto-compilarsi è quello della lingua in cui sono scritti. Vi sono casi piuttosto particolari, ma in cui ci si può imbattere di frequente: ad esempio nella città di Los Angeles il 13% della popolazione non parla inglese e il 28% parla un'altra lingua come principale o unica lingua utilizzata in casa. In situazioni come queste si rende necessario tradurre il questionario in differenti versioni, correndo il rischio di snaturare lo scopo di alcune domande. Quindi, l'utilizzo di questionari auto-somministrati per popolazioni multi-lingue deve essere fatto con molta cautela e in molti casi va addirittura evitato.

Identità del rispondente

Nonostante tutto ciò che si possa fare per selezionare un campione opportuno della popolazione universo, le indagini via web sono soggette ad un altro serio problema: l'impossibilità di effettuare uno stretto controllo sul rispondente.

Da una parte, infatti, non è possibile controllare che esso risponda in modo appropriato, seriamente e onestamente alle domande del questionario; non è nemmeno possibile fornire chiarimenti nella fase di compilazione o prendere nota degli atteggiamenti e delle reazioni non verbali ai quesiti (tutte cose possibili ad esempio con intervite faccia a faccia).

Dall'altra, fatto ancora più importante, non si può controllare se per rispondere alle domande l'individuo coinvolto nell'indagine si consulti con altre persone: in questa situazione alcune risposte potrebbero non essere spontanee o non rispecchiare la realtà.

Infine, non si può controllare l'identità del rispondente: nessuno infatti assicura che chi risponde al questionario sia realmente la persona selezionata per fare parte del campione.

È importante che si tenga conto di tutti i problemi sopra esposti nella fase di pianificazione di un'indagine ed è essenziale che, nella discussione dei risultati finali, si faccia almeno un accenno alle difficoltà in cui ci si può essere imbattuti, al riguardo. Non è bene trascurare questi aspetti, in particolare, nell'ambito della scelta dell'opportuno piano di campionamento, fase di estrema delicatezza per ottenere dall'indagine risultati corretti e utili agli scopi conoscitivi che la ricerca si prefigge.

Nel successivo Cap. 6 si prenderanno in considerazione i principali metodi di campionamento.

6 Metodi di Campionamento

Il modo più opportuno di procedere per la selezione di un campione varia da contesto a contesto: è bene scegliere il metodo di campionamento che meglio si adatta al caso su cui si sta operando, metodo che deve essere definito in modo rigoroso.

In primo luogo è necessario avere ben chiaro quali sono gli obiettivi che l'indagine si propone: descrivere un fenomeno, comparare due fenomeni, prendere coscienza di una particolare situazione, e così via. Può essere utile, a questo scopo, porsi delle domande specifiche, riguardanti la ricerca, che aiutino a capire quali sono gli scopi conoscitivi che ci si propone.

In secondo luogo vanno specificati chiaramente i *criteri di eleggibilità*, ovvero le caratteristiche che un rispondente deve avere per poter essere incluso nell'indagine e nel campione, e i *criteri di esclusione*, che invece consentono di identificare gli individui non idonei ad essere coinvolti. I criteri di eleggibilità e di esclusione vanno applicati, prima di procedere con la tecnica di campionamento scelta, alla popolazione universo. Questa procedura consente di arrivare ad avere una lista di unità (denominata *sampling frame*) comprendente i soggetti "eleggibili". Chiaramente le unità vanno identificate in modo che siano utili agli scopi conoscitivi che ci si propone di raggiungere. È necessario, quindi, selezionare quelle in grado di fornire le informazioni che realmente servono, evitando un inutile dispendio di mezzi, tempo e risorse per analisi meno focalizzate.

"Sia la definizione della popolazione, sia la scelta delle unità (entità geografiche, gruppi sociali o istituzioni, famiglie, unità di abitazione, persone, fatti, e così via) costituiscono i cardini di ogni procedura di estrazione" (De Carlo, 1983).

Per fare un esempio, se il governo Turco vuole conoscere le abitudini dei fumatori negli individui di sesso maschile, un primo passo sarà quello di applicare, alla popolazione universo (cioè la popolazione della Turchia), criteri di eleggibilità ed esclusione: si selezioneranno, così, gli uomini e i fumatori (criteri di eleggibilità), mentre verranno escluse le donne e coloro che non fumano (criteri di esclusione).

Infine la scelta di un opportuno metodo di campionamento non può essere fatta senza considerare la popolazione obiettivo che verrà coinvolta.

Va detto però che, nonostante tutto quanto si possa fare, nessun campione può essere ritenuto perfetto: in genere è sempre presente un certo grado di distorsione od errore[8].

Alcuni accorgimenti preliminari (o semplici regole operative) come quelli sopra indicati consentono, però, di ridurre al minimo le cause della distorsione.

Scopo di questo capitolo è concentrarsi sulla fase successiva all'identificazione della *sampling frame*, cioè sulla scelta dell'opportuno metodo di campionamento.

---

[8] L'argomento verrà trattato diffusamente nel Quaderno di Ricerca "Inferenza in Indagini via Web" (Biffignandi, Toninelli, 2006); si veda Bibliografia.

Una prima classificazione delle tecniche di campionamento può essere fatta distinguendo tra campionamento probabilistico e campionamento non probabilistico. Del primo si parlerà subito, il secondo verrà affrontato nel Par. 6.2.

Le differenti tecniche di campionamento (probabilistico o non probabilistico) che verranno illustrate possono anche essere utilizzate in maniera combinata tra di loro (Malhorta, 1999).

6.1 Campionamento Probabilistico

Le tecniche di campionamento probabilistiche si pongono come fine quello di ottenere un campione rappresentativo della popolazione obiettivo allo scopo di raggiungere conclusioni o previsioni che siano valide per l'intera popolazione universo. La soggettività nella scelta del campione viene eliminata: infatti vi è una selezione casuale (*random selection*) delle unità che faranno parte di esso.

Ogni membro della popolazione di riferimento ha probabilità nota e diversa da zero di essere incluso nel campione. Al contrario, nel campionamento non probabilistico (di cui si discuterà nel Par. 6.2), invece, questa probabilità di inclusione non è conosciuta.

Il fatto di conoscere le probabilità di inclusione "consente di poter fornire stime delle caratteristiche della popolazione da cui il campione è stato selezionato nel rispetto del massimo di efficienza compatibile con il limite di spesa prefissato" (Tassinari, 1993) e consente di "valutare il grado di precisione delle stime campionarie" (De Luca, 1990).

Presupposto indispensabile per poter procedere alla selezione del campione, è disporre di una lista di campionamento accuratamente preparata, completa di tutte le unità statistiche che compongono l'universo.

Le principali tecniche di campionamento probabilistico sono:

- Campionamento Casuale Semplice (*Simple Random Sampling*)
- Campionamento Casuale Stratificato (*Stratified Sampling*)
- Campionamento Sistematico
- Campionamento per Grappoli (*Cluster Sampling*) o per Aree
- Campionamento a due o più stadi
- Campionamento a due fasi (Doppio)
- Campionamento con ripetizioni (*Replicated Sampling*)

Va tenuto presente che le differenti tecniche che saranno illustrate possono essere utilizzate in maniera combinata tra di loro. Dalla combinazione delle tecniche probabilistiche di campionamento si possono ricavare 32 metodologie differenti (Malhorta, 1999).

6.1.1 Campionamento Casuale Semplice

Il campionamento casuale semplice, allo scopo di ottenere un campione rappresentativo, prevede che le unità statistiche che faranno parte del campione siano selezionate in modo casuale, una alla volta e in modo indipendente una dall'altra. Ciò significa che, una volta estratta una di esse all'interno dell'insieme delle unità "eleggibili" (costituenti

la *sampling frame*), la stessa unità non potrà più essere estratta una seconda volta. Si parla, cioè, di estrazione senza re-inserimento e il piano viene definito *campionamento casuale semplice senza ripetizione* o *WOR* (*Without Replacement*).

Esiste, però, anche una seconda modalità di estrazione casuale, che prevede il reinserimento dell'unità statistica estratta nella lista di quelle tra cui verrà scelta la successiva. In questo secondo caso si parla di *campionamento casuale semplice con ripetizione* o *bernoulliano*, o *WR* (*With Replacement*). Con questo secondo sistema le probabilità di estrazione non variano da una estrazione alla successiva.

In ogni caso, è evidente che la probabilità di estrazione di ciascuna unità, sia nel campionamento con reinserimento che nel campionamento senza ripetizione, è nota a priori.

È anche evidente che in ogni estrazione la probabilità che ciascuna delle unità venga estratta è diversa da zero e uguale a quella di tutte le altre unità tra cui si deve scegliere.

Così, se ad esempio si vorranno estrarre 15 unità da una popolazione di 1.500 individui, alla prima estrazione ciascuna unità avrà probabilità (nota) di essere estratta pari a 1 su 100; quindi, diversa da zero e uguale a quella di ogni altra unità della popolazione di "eleggibili".

Per i motivi appena visti, il campionamento casuale semplice ha le caratteristiche dei metodi di campionamento probabilistici.

Operativamente l'estrazione può essere fatta solo se si ha a disposizione una lista completa delle unità della popolazione di riferimento. Le unità possono, ad esempio, essere numerate in senso progressivo. In questo modo, con l'ausilio di una tavola di numeri casuali[9] o di una lista casuale di numeri generata da un computer (molti *software* come, ad esempio, *Excel* consentono questa operazione) si possono scegliere le unità da estrarre: saranno quelle la cui posizione, nella lista, corrisponde ai numeri casuali scelti. Esistono, però, anche metodi di estrazione più rudimentali: ad esempio si può utilizzare un'urna contenente palline riportanti i numeri corrispondenti a tutte le unità eleggibili; si estrarrà, tra queste, una quantità di palline corrispondente al numero di individui che si vuole compongano il campione.

Grazie alle sue caratteristiche (e in particolare al fatto che ogni unità ha la stessa probabilità di essere estratta), il campione ottenuto con il campionamento casuale semplice è considerato relativamente poco distorto. In particolare, come è logico desumere, più è ampia la numerosità campionaria, più il campione sarà rappresentativo della *target population*.

In particolare il funzionamento del campionamento casuale semplice si è dimostrato particolarmente efficiente quando si ha a che fare con popolazioni relativamente piccole, indipendenti e chiaramente definite. Per fare un esempio, non sarebbe opportuno fare un campionamento casuale semplice di tutta la popolazione degli Stati Uniti (che risulta essere troppo numerosa), ma con una città di circa 60.000 abitanti questa tecnica può rivelarsi particolarmente adatta: ci si può più agevolmente procurare una lista dei residenti e si può estrarre (da un'urna, tramite una tavola di numeri casuali o in altri modi), una prefissata proporzione della popolazione (ad esempio il 10 o il

[9] Si rimanda, al proposito, ad un qualunque manuale di Statistica.

20%), assicurandosi eventualmente di effettuare un prelievo ugualmente cospicuo da ogni lettera dell'alfabeto dell'elenco, affinché il campione sia il più rappresentativo possibile.

Vantaggi

Il campionamento casuale semplice è molto apprezzato perché consente, senza grosse difficoltà tecniche ed in modo semplice ed immediato, di ottenere un campione relativamente non distorto. Indipendentemente dai fini della ricerca, infatti, il campione, a causa della connotazione casuale, tendenzialmente assume le stesse caratteristiche dell'universo rispetto alle variabili che lo caratterizzano. La distorsione, inoltre, può essere diminuita, come visto, aumentando la numerosità campionaria.

Il campionamento casuale semplice, in secondo luogo, non richiede conoscenze specifiche dell'universo e, anzi, può essere utilizzato anche nel caso non si conosca nulla della popolazione in esame.

Inoltre il procedimento di estrazione casuale consente, attraverso l'utilizzo di particolari strumenti statistici, di stimare gli errori di campionamento e di comprendere se le differenze tra risultati campionari e dati relativi all'intera popolazione sono dovute al caso o ad errori di impostazione del piano di campionamento.

Svantaggi

Uno dei rischi che si corrono utilizzando il campionamento casuale semplice è quello di escludere un sottogruppo della popolazione che si differenzia dal resto delle unità statistiche per alcune caratteristiche particolari e che, invece, sarebbe di grande interesse avere nel campione. Analogamente, dato che l'estrazione è casuale, si rischia che alcuni particolari sottogruppi della popolazione universo non siano selezionati nelle proporzioni adeguate. In entrambe queste situazioni si può arrivare a risultati che non rispecchiano la realtà. Per ridurre tale rischo è possibile ricorrere al campionamento casuale stratificato (se ne parlerà nel seguente Par. 6.1.2).

Va aggiunto che la tecnica di selezione casuale non permette di utilizzare alcune informazioni che è possibile avere sulla popolazione oggetto di indagine, informazioni che potrebbero servire per migliorare il livello qualitativo dei risultati ottenibili con il piano di campionamento.

Non va dimenticato, infine, che l'utilizzo del campionamento casuale semplice si rende impraticabile nel caso si abbia a che fare con popolazioni particolarmente numerose: i costi di rilevazione dei dati potrebbero rivelarsi proibitivi da sostenere. Inoltre non sarebbe possibile risolvere alcuni problemi, come, ad esempio, sostituire soggetti che si rifiutano di cooperare con altri ritenuti ragionevolmente simili.

6.1.2 Campionamento Casuale Stratificato

La tecnica di campionamento casuale stratificato prevede che la popolazione di riferimento venga suddivisa in *strati* (o sottogruppi significativi) in base ad una o più caratteristiche di interesse o in qualche modo correlate agli scopi della ricerca.

Attraverso la stratificazione si riduce la varianza presente all'interno di ogni strato, rispetto a quella presente nella popolazione complessiva.

Per la costruzione dei gruppi si può ricorrere alla letteratura disponibile su casi analoghi o al parere di esperti. In altri casi, nella popolazione possono già essere presenti sottogruppi omogenei naturali che possono essere utilizzati per le estrazioni basate sulla stratificazione.

In ogni caso, al contrario del campionamento casuale semplice, sono necessarie informazioni sugli individui da cui verrà selezionato il campione. Qualora si conoscesse il comportamento di numerose variabili all'interno della popolazione, bisognerebbe, per la suddivisione in gruppi, almeno tenere conto di tutte quelle più significative.

La *cluster analysis* (analisi dei gruppi) può essere utile per costruire gruppi di unità quanto più possibili omogenei al loro interno.

In generale, più è alto il numero dei gruppi in base ai quali si stratificano le unità, maggiore è l'efficienza delle stime. Se si stratificasse in base alla variabile oggetto di studio, la situazione limite ideale sarebbe quella di avere gruppi costituiti da singole unità. In realtà, però, per la stratificazione ci si basa su variabili ausiliarie, solamente correlate con la variabile oggetto di studio. Quindi non è opportuno superare un certo numero di gruppi nelle procedure di stratificazione, perché i vantaggi conseguibili con un numero eccessivamente elevato di gruppi potrebbero rivelarsi eccessivamente modesti, mentre i costi dell'indagine potrebbero lievitare considerevolmente.

Nelle operazioni di stratificazione, oltre a gruppi composti da un'unica unità, possono essere individuati anche gruppi particolari per i quali tutte le unità comprese vengono comunque selezionate per fare parte del campione. In questo caso si parla di *unità auto-rappresentative.*

Una volta impostati gli strati, gli elementi della popolazione universo (eterogenea) vengono distribuiti all'interno di essi, in base alle variabili utilizzate per la classificazione: si ottengono, così, sub-popolazioni al loro interno omogenee. Da ciascun gruppo viene successivamente selezionato un campione casuale.

Aspetto di rilievo nel progettare un campionamento stratificato è la numerosità campionaria di ciascun gruppo. Il campionamento va strutturato in modo tale che tenga conto di come i gruppi individuati sono presenti all'interno della popolazione complessiva. Se gli strati sono tutti presenti presumibilmente con lo stesso peso nella popolazione obiettivo, si può utilizzare un metodo di ripartizione uniforme. Da ciascuno strato si estrae, cioè, un numero di unità pari al rapporto tra la numerosità campionaria (ad esempio, n) e il numero degli strati (ad esempio, k).

Questo sistema di allocazione della numerosità campionaria è però piuttosto grossolano. Per considerarne altri si rende necessario introdurre il concetto di frazione di campionamento (f_i), che, per ogni strato i (i = 1, 2, …, k), indica quante unità (n_i), tra quelle dello strato (N_i), vanno estratte.

$f_i = \frac{n_i}{N_i}$; i indica lo strato i-esimo.

Se la frazione di campionamento è uguale per ogni strato, si parla di campione stratificato proporzionato (*proporzionate stratified sample*) e la tecnica rappresenta uno sviluppo del campionamento casuale semplice. I differenti strati della popolazione sono

rappresentati nel campione proporzionalmente al loro peso numerico originario e la frazione di campionamento viene definita *uniforme* o *proporzionale*. Di conseguenza, il numero di unità statistiche del campione è distribuito tra gli strati in modo proporzionale al peso che ciascuno strato ha all'interno della popolazione. La frazione di campionamento uniforme in genere viene utilizzata quando non si hanno a disposizione informazioni attendibili sulla popolazione e quando si ritiene che la distribuzione all'interno dello strato presenti variabilità piuttosto ridotta.

Vi sono alcune situazioni in cui si preferisce utilizzare frazioni di campionamento differenti: si parla di *disproportionate stratified sample*. Ad esempio, nel caso in cui la media e la varianza tra i vari strati siano particolarmente differenti, si usa una frazione di campionamento *variabile* (o *non proporzionale*). Se in uno strato la variabilità è maggiore, si tende a considerare una proporzione di unità maggiore per quello strato. Questo avviene perché si parte dal presupposto che più è variabile la popolazione all'interno di uno strato, più è difficile rappresentarla con l'estrazione di un campione uniforme di ridotte dimensioni. Per popolazioni particolarmente eterogenee, quindi, questo secondo approccio che prevede frazioni di campionamento variabili permette di ottenere una precisione maggiore nei risultati.

Neyman, in alternativa ai metodi di ripartizione uniforme e proporzionale, ha introdotto il concetto di *ripartizione ottima* della numerosità campionaria tra gli strati. Secondo lo studioso il miglior sistema per distribuire un campione tra gli strati è farlo in modo proporzionale allo scarto quadratico medio (della variabile considerata) che caratterizza ciascuno strato. Le unità estratte da uno strato (n_i), secondo il teorema di Neyman, sono proporzionali al prodotto tra la dimensione dello strato (N_i) e il suo scarto quadratico medio (σ_i). La difficoltà maggiore è che non sempre queste informazioni sullo scarto quadratico medio dei vari gruppi sono disponibili, anche se a volte le si può desumere da ricerche precedenti.

Sia nel caso si utilizzi il metodo uniforme, proporzionale o di Neyman, una volta ricavati i risultati per ciascuno strato, per arrivare ad una misura del fenomeno riferibile all'intera popolazione i risultati devono essere ponderati proporzionalmente alla grandezza relativa dei singoli strati.

Se, invece, lo scopo è analizzare separatamente i differenti gruppi (ad esempio, per confrontarne le caratteristiche), non è necessario intervenire col ri-proporzionamento o con la ri-ponderazione (attraverso la frazione di campionamento) dei risultati dei singoli strati.

Vantaggi

Il campionamento casuale stratificato ha il vantaggio di mantenere la connotazione casuale delle estrazioni: all'interno dei gruppi la selezione delle unità è effettuata con lo stesso sistema visto per il campionamento casuale semplice. Quindi il campione risulta relativamente poco distorto (a causa della selezione casuale) e, in più, può essere corretto tenendo conto della distribuzione della popolazione per *variabili ausiliarie* quali genere, età, razza, ...o considerando altre variabili importanti ai fini dell'indagine.

In questo modo il campione rispecchia meglio la struttura della popolazione originaria e, grazie alla stratificazione, è possibile inserire in esso piccoli sottogruppi di interesse. Se, ad esempio, una popolazione comprende solo il 10%

di individui di sesso femminile, con l'estrazione casuale si rischia di scegliere un campione che sottostima la componente femminile (o che addirittura la esclude). Non è opportuno, quindi, ricorrere ad un'unica lista da cui estrarre un campione casuale, ma è opportuno operare una selezione basata sulla tecnica di stratificazione.

La stratificazione, per analoghe ragioni, è anche una opportuna scelta di operare quando la distribuzione della variabile oggetto di studio nella popolazione è fortemente asimmetrica, oppure nel caso vi siano dei valori estremi che potrebbero essere raggruppati in un numero esiguo di strati con varianza interna minima.

In sostanza, il campionamento stratificato consente di ottenere risultati più attendibili e precisi nelle ricerche, evitando la necessità di aumentare la numerosità campionaria, e risulta maggiormente efficiente del campionamento casuale semplice. Infatti a parità di numerosità campionaria, l'errore campionario del campionamento stratificato sarà inferiore a quello di un campione ottenuto tramite campionamento casuale semplice se la stratificazione è stata fatta in base ad uno o più caratteri correlati con gli argomenti oggetto di ricerca. Quindi da ciò consegue che la stratificazione consente anche di contenere in maggiore misura i costi dell'indagine.

In effetti spesso le variabili naturali su cui si basa la stratificazione sono facilmente identificabili e questo, oltre ad aiutare a contenere i costi, può essere considerato anche un vantaggio in termini gestionali ed organizzativi.

Utilizzando il campionamento stratificato è, infine, anche possibile fare delle analisi distinte e più approfondite relativamente ai differenti gruppi risultato della stratificazione e ottenere stime distinte per le diverse realtà.

Svantaggi

Il campionamento stratificato è più complicato e laborioso da realizzare rispetto al campionamento casuale semplice. È necessario, infatti, individuare un criterio appropriato per identificare gli strati, nonchè operare in modo che la composizione degli strati sia giustificata dai fini della ricerca. Inoltre la situazione si aggrava nei casi in cui non vi siano informazioni disponibili sulle variabili utilizzate per la stratificazione.

Inoltre è indispensabile calcolare la numerosità campionaria necessaria per ogni sottogruppo e questa operazione richiede un ulteriore impiego di tempo.

Va anche considerato che in molti casi la stratificazione non riesce a soddisfare finalità differenti: se è efficace per la stima di una grandezza, non è detto che lo sia anche per altre cui si può essere interessati.

Infine vi è un ulteriore evidente rischio: quello di arrivare all'utilizzo di un numero eccessivamente elevato di sottogruppi. In questa situazione l'indagine può divenire difficile da controllare, complessa, lunga e particolarmente dispendiosa.

6.1.3 Campionamento Sistematico

Il presupposto principale per applicare la tecnica di campionamento sistematico è la disponibilità di una lista ordinata e numerata delle unità selezionabili per la ricerca.

Una volta che ci si è procurati la lista, si suddivide il totale delle unità componenti la popolazione obiettivo (ad esempio, N) per il numero di unità da campionare (numerosità campionaria: n). Si individua, in tale modo, il *passo di campionamento* (p), ovvero la periodicità con cui, da un determinato punto della lista originaria, si selezionano le unità che faranno parte del campione.

$$p = \frac{N}{n} \qquad (\textit{passo di campionamento})$$

Il punto sulla lista da cui si cominciano a selezionare le unità (e quindi la prima unità da estrarre) viene scelto in modo casuale: ad esempio, tramite il lancio di un dado o tramite estrazione a sorte, ma anche in altri modi. Si procede, poi, seguendo un passo di campionamento p costante per la scelta delle successive n-1 unità. Ad esempio, se si seleziona a caso l'unità k della lista, le successive che verranno estratte saranno la $(k + p)$, la $(k + 2p)$, la $(k + 3p)$, e così via, fino a che si estrae un numero di unità pari alla numerosità campionaria (n). Nel caso in cui k non sia intero, in genere viene approssimato per eccesso.

Il campionamento sistematico può essere concettualmente ricondotto al campionamento a grappoli (si veda Par 6.1.4): ogni possibile campione, fissato un passo p, forma un grappolo tratto da una popolazione composta da k grappoli.

Vantaggi

Il campionamento sistematico è semplice da realizzare, veloce, economico e pratico.

È simile al campionamento casuale semplice quando la lista di campionamento non ha un ordine interno ricorrente inerente ad una variabile di interesse, quando copre interamente la lista, ma anche quando è possibile riordinare la lista o aggiustare in qualche modo gli intervalli di campionamento. Se la lista è organizzata casualmente, il grado di affidabilità delle tecniche di campionamento sistematico e di quella casuale semplice è uguale.

La selezione sistematica spesso viene utilizzata anche quando non si conosce a priori la numerosità campionaria (come succede ad esempio, con i visitatori che accedono ad un sito Web). In questo caso il passo di campionamento è, di solito, fissato a multipli di 5 o di 10 unità e il numero di partenza viene fissato casualmente; nel caso l'indagine si ripetesse nel tempo, sarebbe opportuno che il numero di partenza subisse una rotazione. In questa situazione si parla di *campionamento sistematico non casuale* (vedi Par. 6.2.6).

È possibile applicare la tecnica sistematica anche ad una popolazione stratificata, qualora si vogliano ottenere stime differenti per ogni strato. In questi casi la popolazione viene stratificata secondo un determinato criterio e poi viene estratto un campione sistematico da ogni strato. Se vi sono molti strati, è anche possibile estrarre un campione sistematico di essi e prelevare, poi, campioni sistematici all'interno degli strati selezionati.

La selezione sistematica è particolarmente vantaggiosa nel caso in cui la popolazione sia particolarmente numerosa e non convenga effettuare l'estrazione casuale semplice; oppure quando la lista delle unità da campionare viene costruita nella fase di attuazione di una indagine (ad esempio: gli utenti che utilizzano un determinato servizio via Web).

Svantaggi

Ovviamente, il campionamento sistematico può essere applicato solamente nel caso in cui sia disponibile una lista completa delle unità che devono essere campionate, e ciò, soprattutto nel caso delle indagini via web, non sempre è possibile.

Il rischio più grosso in cui si incorre in seguito all'utilizzo di una selezione di tipo sistematico è che non vi sia, per ciascuna unità, una uguale probabilità di essere inserita nel campione. Il campione, in questo modo, potrebbe risultare in definitiva non probabilistico. Ad esempio, se una lista prevede che i soggetti della popolazione eleggibile siano ordinati per nome e si estrae un campione con il metodo sistematico, la probabilità che un nome venga estratto non è uguale a quella degli altri nomi. Ci sono, anzi, alcuni nomi che saranno molto poco diffusi e rischieranno di non essere estratti.

Per risolvere il problema riguardante la presenza poco diffusa di alcune modalità di una variabile, si può ricorrere a differenti soluzioni: aggiustare gli intervalli di campionamento oppure ricorrere al campionamento a grappoli o a due stadi (verranno visti entrambi più avanti: rispettivamente Par. 6.1.4 e Par. 6.1.5).

Prima della selezione è importante controllare attentamente che gli elenchi utilizzati siano, oltre che ordinati e completi, anche ordinati casualmente rispetto alle variabili di analisi. Se le N unità della popolazione non sono in ordine casuale i risultati non sono significativi o possono verificarsi errori sistematici, in particolare se l'ordine delle unità è correlato a una delle caratteristiche che si vogliono studiare.

La selezione sistematica non è utilizzabile, invece, quando la periodicità di campionamento è una componente naturale della lista. Se, ad esempio, si vogliono estrarre sistematicamente i giorni della settimana di un anno per tenere sotto controllo la produzione di una azienda metalmeccanica e la periodicità di campionamento risulta pari a 7, si otterrà un campione che non è per nulla rappresentativo in quanto, scelto il primo giorno a caso (poniamo il Mercoledì), i successivi giorni che saranno inseriti nel campione saranno sempre gli stessi (tutti Mercoledì). Per risolvere quest'ultimo problema è possibile riordinare la lista in base ad altri criteri (o variabili) o modificare il passo di campionamento.

6.1.4 Campionamento a Grappoli e per Aree

Il campionamento a grappoli (*cluster sampling*) è piuttosto conveniente e veloce da realizzare, in quanto è basato su unità *naturali* (i *grappoli*, appunto, detti anche *cluster*) composte da una serie più o meno numerosa di unità statistiche. Con l'aggettivo *naturali* si vuole indicare che i grappoli sono composti da unità già raggruppate in natura. I criteri di aggregazione possono basarsi su differenti caratteristiche. Ad esempio possono derivare dalla contiguità spaziale delle unità (individui facenti parte della stessa

circoscrizione elettorale, dello stesso comune, della stessa città, della stessa regione, ...) o dal fatto che esse si rivolgono a determinate strutture eroganti particolari servizi (scuole, università, ospedali, ...) e da altri criteri di vario tipo.

Chiaramente, perché la tecnica del campionamento per grappoli funzioni a dovere, devono essere definiti criteri ben precisi che consentono di assegnare senza dubbi e ambiguità le unità della popolazione ad un grappolo piuttosto che ad un altro e ogni unità può appartenere solo ad uno dei gruppi disponibili.

Nel campionamento a grappoli si selezionano casualmente alcuni gruppi e tutti i membri che ne fanno parte vengono inseriti nel campione. Questo modo di procedere viene anche definito *piano ad un solo stadio*. Ad esempio nell'ambito delle elezioni amministrative se si effettua un campionamento per grappoli si possono selezionare 5 delle 100 sezioni elettorali di un grosso centro abitato italiano e, poi, verranno intervistati tutti gli elettori che si recano presso i seggi elettorali di quelle sezioni.

Il campionamento a grappoli è dunque simile allo stratificato, ma si differenzia per il fatto che è basato su una coerenza che non è studiata a tavolino o derivata dalla distribuzione di una serie di variabili, ma che ricorre naturalmente. Inoltre il campionamento stratificato prevede una fase di estrazione casuale all'interno dei gruppi che si sono formati, mentre nel campionamento a grappoli sono selezionate tutte le unità facenti parte di un gruppo naturale una volta che questo è stato scelto. Va aggiunto, infine, che nel campionamento stratificato si cerca di costruire gli strati in modo che siano al loro interno il più possibile omogenei, mentre per il campionamento a grappoli i *cluster* dovrebbero essere il più possibile simili tra loro e dovrebbero avere, al loro interno, variabilità piuttosto elevata, in modo che ciascuno rispecchi, idealmente, quella dell'intera popolazione.

Per il modo in cui i grappoli vengono identificati, in genere è difficile che risultino composti dallo stesso numero di unità, anche se la tendenza è quella di creare *cluster* della stessa ampiezza.

Gli usuali campi di applicazione del campionamento a grappoli e per aree sono, in particolare, studi riguardanti grandi dimensioni territoriali. Ad esempio, nel caso si stia effettuando una indagine sui turisti che viaggiano dagli Stati Uniti verso l'Europa, anziché ricorrere alla laboriosa costruzione di liste (che risulteranno particolarmente lunghe) tra cui selezionare le unità che faranno parte del campione, si possono, più agevolmente, selezionare solo un certo numero di agenti di viaggio e intervistare tutti i viaggiatori che si sono rivolti ad essi. Analogamente, per una indagine sui pazienti di tutti gli ospedali di una città, piuttosto che ricorrere alla selezione da una lista molto lunga (e quindi laboriosa da costruire) di tutti i pazienti, è preferibile redarre o procurarsi una lista di tutti gli ospedali e, tra di essi, sceglierne alcuni casualmente. Verranno poi intervistati tutti i pazienti degli ospedali scelti.

Invece di scegliere tutte le unità dei grappoli selezionati, si può procedere ad una scelta casuale o sistematica di esse: si estrae cioè a sorte tra le unità dei grappoli o ci si basa su elenchi per procedere ad una selezione sistematica. In questo caso si parla di *piano di campionamento a due stadi*. Dai grappoli scelti (*unità primarie* o *di primo stadio*) vengono scelte (in modo casuale o sistematico) le unità finali su cui si svolgerà l'indagine (*unità secondarie* o *di secondo stadio*). L'integrazione del campionamento a grappoli all'interno del campionamento a più stadi rende le operazioni più laboriose, ma in genere la precisione dei risultati che si riesce a conseguire è maggiore.

Particolare tipologia del campionamento per grappoli è il *campionamento ad aree*, che viene utilizzato prevalentemente in indagini di stampo economico o sociale. I grappoli sono costituiti, in questo caso, da unità territoriali, in base alle quali viene suddivisa la popolazione oggetto di studio: per analisi di questo tipo in genere ci si serve dell'ausilio di carte topografiche. La dimensione delle aree è bene che rispetti la conformazione del territorio e la natura dell'universo su cui si effettua l'indagine.

Una volta identificate in modo preciso le aree, il primo passo è estrarre un gruppo di queste, ad esempio utilizzando un criterio di estrazione casuale che si basi sul criterio della equiprobabilità. Effettuata l'estrazione, è poi possibile procedere effettuando un ulteriore sub-campionamento basato su aree di dimensione minore oppure redigendo una lista delle unità elementari che fanno parte delle aree estratte e procedendo con differenti metodi di estrazione. Kish (1965) suggerì che non è indispensabile determinare in modo esatto la misura di grandezza delle aree di base, ma essa deve essere approssimativamente proporzionale al numero di elementi contenuti. La dimensione, comunque, può essere approssimativamente costante; oppure si può seguire il criterio di massima omogeneità interna, che garantisce maggiore attendibilità nelle procedure di stima (Hansen e Hurwitz, 1942). Per ottenere risultati più attendibili, poi, è essenziale impostare le divisioni tra aree in modo tale che popolazioni dalle caratteristiche particolare o aree particolarmente "a rischio" (sotto un certo punto di vista) risultino essere a se stanti.

Vantaggi

Il campionamento a grappoli può essere utilizzato in indagini piuttosto ampie e laboriose, in particolare quando risulta impraticabile compilare una lista esaustiva delle unità che compongono la popolazione obiettivo, perché si tratterebbe di un'operazione eccessivamente lunga, costosa o impossibile da realizzare. Analogamente la metodologia è utile quando non sono disponibili liste affidabili od aggiornate. La lista, eventualmente, risulta essere necessarie solo per i grappoli scelti.

L'efficienza derivante dall'utilizzo di questo tipo di campionamento permette comunque di compensare eventuali perdite in precisione dei risultati derivanti da una eventuale non soddisfacente composizione dei grappoli (De Cristofaro, 1979).

Altro vantaggio derivante dall'utilizzo del campionamento a grappoli riguarda i costi e i tempi di rilevazione. Entrambi sono piuttosto ridotti, data la minore dispersione territoriale delle unità selezionate, soprattutto nel caso si tratti di ricerche in cui è prevista la presenza di un intervistatore.

Per diminuire ulteriormente i costi di realizzazione dell'indagine e per avere un campione che presenti un grado di distorsione particolarmente basso si scelgono in genere *cluster* di ampiezza ridotta che siano eterogenei al loro interno e omogenei tra di loro.

Nonostante il fatto che il campionamento a grappoli/per aree sia particolarmente economico, i requisiti probabilistici sono rispettati e la precisione delle informazioni rilevate è elevata, relativamente ai bassi costi.

Per quanto riguarda, nello specifico, il campionamento ad aree, in genere non è nemmeno necessario disporre di *sampling frame*; si evita, in questo modo, di

incorrere nelle distorsioni originate dall'utilizzo di liste che possono essere incomplete o non aggiornate.

Anche nel caso di campionamento per aree, per ottenere stime più corrette sarebbe opportuno ridurre la dimensione delle aree aumentandone il numero. Agendo in questo modo, però, da un lato si diminuisce il numero di unità da estrarre da ognuna di esse e, dall'altro, si ottiene un aumento sensibile dei costi per il contatto delle unità campionate (che risulteranno essere maggiormente disperse). Per aumentare l'efficienza delle stime, al contrario, andrebbero selezionate aree di vaste dimensioni. La soluzione ideale a questo *trade-off* è selezionare una dimensione che concili i due aspetti, permettendo di ottenere stime più efficienti e di contenere anche i costi.

Svantaggi

In primo luogo tutti i *cluster* che fanno parte della lista dovrebbero avere caratteristiche perlomeno simili, e dovrebbero rispecchiare quelle della popolazione; altrimenti, selezionando solo un particolare tipo di grappolo, si rischia che alcuni casi di interesse vengano esclusi dall'indagine. Per questo motivo i grappoli in genere dovrebbero essere impostati in modo tale da avere, al proprio interno, la stessa variabilità della popolazione universo. Meno omogenei saranno i grappoli, più alto dovrà essere il numero di *cluster* da estrarre. De Cristofaro (1979) sostiene che, in effetti, il campionamento a grappoli è tanto meno efficiente del campionamento casuale semplice (a parità di dimensioni) quanto più bassa è la variabilità interna ai grappoli e quanto più differenti sono essi tra di loro.

In secondo luogo ci possono essere, in alcuni casi, problemi legati alla scelta dei *cluster*: "alcuni di essi possono condurre a risultati più precisi di altri. Ad esempio, dovendo formare un campione di comuni, si può procedere ad un loro raggruppamento secondo la Provincia o la Regione di appartenenza" (De Carlo e Robusto, 1996).

Infine, per il campionamento ad aree, è opportuno che le unità siano piuttosto piccole, allo scopo di contenere i costi di rilevazione; ma ciò non sempre avviene. Questa tecnica di campionamento, inoltre, è inadatta nel caso si abbia a che fare con elementi rari, cioè unità particolari che costituiscono all'incirca l'1-2% della popolazione indagata.

6.1.5 Campionamento a Due o più Stadi

Il *campionamento a due stadi* è considerato una estensione del campionamento a grappoli: in un primo tempo si selezionano i grappoli, poi, attraverso la tecnica di campionamento casuale semplice o con altre tecniche, si effettua una ulteriore selezione delle unità dai grappoli selezionati.

Gli stadi possono anche essere più di due (in questo caso si parla di *campionamento a più stadi*), e i raggruppamenti e il campionamento casuale possono essere fatti ad ogni stadio successivo. In un campionamento a tre stadi, ad esempio, si avranno grappoli di primo ordine (anche detti *unità primarie*) all'interno dei quali vengono selezionati i grappoli di unità elementari (grappoli di secondo ordine, o *unità secondarie*); da questi ultimi, al terzo stadio, vengono estratte le *unità elementari*.

La procedura di campionamento in generale è, quindi, di tipo gerarchico: l'insieme comprendente le unità finali è incluso in un gruppo di unità superiore che, a sua volta, risulta inglobato in un altro gruppo di livello ulteriormente elevato.

Per procedere con il campionamento a più stadi è necessario (De Carlo e Robusto, 1993):

- Decidere il numero di stadi che saranno utilizzati. Per fare cioè vanno considerate la disponibilità e il costo delle liste di campionamento e la disponibilità di informazioni da utilizzare per la stratificazione.
- Individuare le variabili da utilizzare per la stratificazione delle unità di primo stadio.
- Decidere quante unità verranno selezionate ad ogni stadio.
- Scegliere le modalità di selezione delle unità statistiche che verranno utilizzate ad ogni stadio.

Al campionamento a più stadi si ricorre in prevalenza nel caso in cui si abbia a che fare con una popolazione distribuita su un territorio ampio e quando i tempi a disposizione per svolgere l'indagine sono notevolmente ristretti.

L'obiettivo primario che ci si prefigge di raggiungere è quello di massimizzare il numero dei gruppi, in modo da far diminuire la dimensione del campione all'interno di ciascuno di essi.

Vantaggi

Il campionamento a due stadi consente di ottenere gli stessi vantaggi del campionamento a grappoli, nel caso si abbia a che fare con indagini piuttosto ampie e laboriose o con ricerche per cui compilare la lista delle unità eleggibili risulta un'operazione laboriosa o, al limite, impossibile da realizzare. In effetti è sufficiente che siano disponibili le liste delle sub-popolazioni contenute nelle unità selezionate al livello superiore e le liste contenenti le unità elementari tra cui saranno selezionate quelle da inserire nel campione.

I costi di implementazione risultano inferiori a quelli derivanti dall'utilizzo del campionamento stratificato; ciò consente di utilizzare campioni che all'ultimo stadio presentino numerosità campionaria più elevata. Inoltre la tecnica garantisce un elevato controllo sull'universo considerato e la massima rappresentatività del campione.

Non va dimenticata la elevata flessibilità che si può raggiungere: si possono utilizzare, nei differenti stadi, diverse tecniche di selezione. Questa particolarità fa sì che la tecnica possa essere adattabile a ricerche svolte in situazioni molto differenti.

Svantaggi

Se i costi richiesti dal campionamento a più stadi sono inferiori rispetto a quelli necessari per il campionamento stratificato, i risultati ottenuti sono però meno precisi. In realtà, la minore precisione dei risultati può essere in parte compensata dalla possibilità di utilizzo, a costi minori, di numerosità campionarie più elevate.

Rispetto alla variabilità, il campionamento a più stadi è meno efficiente di quello casuale, ma questo aspetto può essere migliorato con l'utilizzo di informazioni ausiliarie e valutando gli errori di campionamento.

6.1.6 Campionamento a Due Fasi (Doppio)

Il *campionamento a due fasi* (*two phases sampling*), introdotto nel 1938 da Neyman, non va confuso con il campionamento a due stadi. Mentre in quest'ultimo si considerano, nelle fasi successive, unità differenti, nel campionamento a due fasi le unità considerate sono sempre le stesse, ma a *sub-campioni* di alcune di esse vengono chieste maggiori informazioni, allo scopo di approfondire alcuni punti della ricerca. Queste informazioni possono essere utilizzate, tra le altre cose, per aumentare la validità delle stime (se usate in concomitanza con gli altri dati rilevati), possono essere oggetto di indagine indipendente o possono essere utilizzate al fine di valutare l'attendibilità dei risultati dell'indagine.

La procedura di campionamento utilizzata nella prima e nella seconda fase può essere fissata in base agli obiettivi della ricerca, nel modo che sembra più opportuno al ricercatore. Ad esempio si può decidere di selezionare il campione con il procedimento casuale semplice e il sub-campione col metodo della stratificazione.

In genere questo tipo di campionamento viene utilizzato per contenere i costi di una indagine e, allo stesso tempo, per avere la possibilità di ottenere dati molto particolareggiati o che approfondiscano determinati aspetti di un fenomeno.

Nella prima fase si possono raccogliere informazioni su un campione relativamente ampio; successivamente si possono utilizzare questi dati per la stratificazione. Sulla base della stratificazione si procede, poi, alla selezione del sub-campione; i parametri del sub-campione rappresentano la stima ponderata di quelli corrispondenti riferiti alla popolazione di riferimento.

Molto di frequente il campionamento a due fasi è utilizzato nell'ambito di indagini su popolazioni rare: nella prima fase si seleziona il più alto numero possibile di elementi facenti parte di tali popolazioni. In genere, dopo la prima fase, le unità vengono inizialmente allocate in due strati: uno positivo (che contiene gli elementi rari della popolazione) ed uno negativo (con le unità non-rare). Il passaggio successivo è l'estrazione del campione dallo strato positivo per arrivare ai risultati finali.

Il *campionamento multifasico* è la naturale estensione della tecnica a due fasi: da un campione originario vengono rilevate alcune informazioni di base che poi sono approfondite ed integrate con altre ottenute da una serie di sub-campioni ricavati dal campione iniziale di ampiezza differente.

6.1.7 Campionamento con Ripetizioni

Il campionamento con ripetizioni (*replicated sampling* o *interpenetrating sampling*) è stato utilizzato per la prima volta nel 1946 da Mahalanobis ed è ritenuto particolarmente adatto a situazioni in cui è necessario ottenere velocemente i risultati dell'indagine (senza che ciò vada a discapito della qualità dei dati) o nelle fasi preliminari di una indagine in cui è possibile valutare le prestazioni dei rilevatori o individuare le parti critiche di un questionario; ma si ricorre frequentemente a questa tecnica anche per indagini sull'evoluzione di determinati fenomeni nel tempo.

Il campione ottenuto da una popolazione obiettivo viene suddiviso in sub-campioni che siano rappresentativi della popolazione stessa, indipendenti tra di loro e costituiti dallo stesso numero di individui. In genere l'ampiezza dei sub-campioni (detti anche *reticoli di elementi compenetranti*) viene determinata tramite il rapporto tra l'ampiezza campionaria e il numero di rilevatori che verranno utilizzati nella fase di raccolta dati.

Operando in questo modo, i parametri della popolazione possono venire stimati utilizzando, come punto di partenza, le statistiche di ciascun sub-campione.

La procedura, inoltre, agevola il calcolo dell'errore standard, che permette di operare il controllo della qualità dei dati.

Vantaggi

L'utilizzo di una serie di differenti sub-campioni permette di avere una visione dettagliata e attendibile del fenomeno ottimizzando l'uso delle risorse disponibili: il numero di rilevatori e le dimensioni dell'indagine possono risultare notevolmente ridotti. Infatti l'analisi è limitata a sottoinsiemi del campione esaminato e non estesa a tutte le unità facenti parte di questo.

La tecnica, inoltre, può essere applicata a differenti contesti e piani di campionamento: campioni stratificati, a uno o più stadi o fasi, per aree, e così via.

Altro vantaggio che comporta il campionamento con ripetizioni è che facilita il calcolo degli errori standard per i sub-campioni; in questo modo si ottengono risultati maggiormente attendibili.

Svantaggi

Se il numero di unità comprese nei sub-campioni indipendenti è particolarmente basso, non è possibile effettuare una analisi dettagliata degli errori dovuti al rilevatore.

Dato, inoltre, che ogni sub-campione deve essere rappresentativo, nel catodi indagini *face-to-face* ciò comporta per i rilevatori spese cospicue per gli spostamenti infatti dovranno percorrere tragitti piuttosto lunghi per raggiungere le unità da intervistare (Fabbris, 1989).

6.2 Campionamento non probabilistico

Non sono rari i casi in cui, utilizzando un sistema di selezione casuale delle unità, si arriva a delle conclusioni che non sono per nulla applicabili alla popolazione obiettivo. In queste situazioni le tecniche di campionamento non probabilistico possono servire per selezionare in modo opportuno delle unità che siano rappresentative dell'universo.

Nel campionamento non probabilistico la scelta delle unità statistiche che verranno incluse nell'indagine non è casuale, ma è effettuata considerando le caratteristiche della popolazione obiettivo e gli scopi dell'indagine. Per questo motivo, a differenza del campionamento probabilistico, alcuni membri della popolazione potranno essere estratti (o scelti), mentre altri no. Ciò significa che le singole unità non hanno uguale probabilità di essere selezionate per alcune di esse: la probabilità di inclusione nel campione potrebbe essere pari a zero.

Il campionamento non probabilistico è preferibile alle tecniche di campionamento probabilistiche principalmente nei seguenti tre casi:

1. Quando si sta svolgendo un'indagine su gruppi difficili da identificare, cioè non si ha a disposizione una lista completa degli elementi che compongono l'universo. Ciò succede, ad esempio, quando si vuole indagare su un fenomeno sommerso (come il lavoro in nero) o quando è difficile che si ottenga la collaborazione dei rispondenti (su temi particolarmente delicati, come l'uso di sostanze stupefacenti); nella stessa situazione ci si trova anche nel caso in cui si trattino temi delicati o particolarmente difficili o nel caso in cui le unità della popolazione siano diffidenti (come in indagini riguardanti immigrati clandestini, i quali potrebbero temere di venire espulsi). Le unità di difficile rilevazione possono essere sostituite con altre considerate equivalenti, grazie al fatto che non è necessario rispettare i criteri di casualità delle tecniche probabilistiche. Questo aspetto può contribuire, in molti casi, a ridurre i costi della ricerca.
2. Quando si deve fare una indagine su gruppi specifici di individui: motivi etici o riluttanza ad affrontare tutti i potenziali rispondenti portano a preferire queste tecniche a quelle probabilistiche.
3. In indagini pilota: quando si vuole effettuare la raccolta di dati sperimentali su gruppi di studio e quando i dati raccolti non sono destinati ad essere diffusi, ma all'attività di pianificazione di una indagine o di una ricerca. Il campionamento non probabilistico tra le altre cose consente di ridurre notevolmente le spese di rilevazione e i tempi di ricerca.

Una limitazione delle tecniche di campionamento non probabilistico è il fatto che in generale non consentono (come succede con i modelli probabilistici) l'individuazione di un intervallo, intorno al valore medio campionario, entro il quale è possibile collocare, con ragionevole fiducia, il valore medio incognito della popolazione che si vuole conoscere (De Luca, 1990). Inoltre vi è una particolare esposizione ad errori di tipo sistematico.

Le principali tecniche di campionamento non probabilistico sono:

- Campionamento di convenienza (*Convenience sampling*)
- Campionamento a valanga (*Snowball sampling*)
- Campionamento per quote
- *Focus Group*
- Campionamento ragionato (*Judgement* o *purposive sampling*)
- Campionamento Sistematico non Casuale
- Campionamento auto-selettivo (*Self-selection sampling*)
- Campionamento di elementi rappresentativi
- Campionamento basato su aree barometro
- Campionamento per plausibilità (*Plausibility sampling*)

Nel seguito si passerà in rassegna singolarmente ciascuna tecnica.

6.2.1 Campionamento di convenienza (*Convenience sampling*)

Il *campionamento di convenienza* (dalla terminologia inglese *convenience sampling*), denominato anche *accidentale*, è basato sulla selezione di un gruppo di individui che si sono dichiarati pronti e disponibili a fare parte del campione (*metodo della convenienza*). Si tratta della tecnica di selezione forse più rudimentale, in quanto si selezionano "le prime unità che capitano" (Bailey, 1985): si intervistano, cioè, tutti gli individui che si dichiarano disponibili a partecipare all'indagine o che hanno in qualche modo dimostrato interesse per la ricerca; è possibile, quindi, definire questi individui "volontari".

La selezione del campione avviene, quindi, senza seguire alcun criterio, se non l'ordine con cui le unità statistiche vengono contattate o si propongono spontaneamente, fino ad arrivare al numero richiesto dalla numerosità campionaria.

A differenza delle popolazioni per cui si utilizza il campionamento a valanga (che verrà trattato tra breve), per il campionamento di convenienza si fa riferimento a gruppi numerosi e facilmente accessibili (come i rispondenti a questionari proposti da una rivista, gli studenti di un corso di laurea, i pazienti di un medico, ...).

Vantaggi

La tecnica di campionamento prevede costi di implementazione piuttosto bassi ed un impiego di tempo ridotto.

Svantaggi

Potrebbe darsi che il gruppo di volontari differisca da chi invece non accetta di partecipare all'indagine, se si considerano particolari variabili che sono oggetto di interesse per la ricerca. Il campione rischia, quindi, di non essere rappresentativo di tutte le sfumature della popolazione obiettivo e può portare a risultati distorti.

Il rischio di distorsione, quindi, è molto forte. Per questo motivo l'utilizzo del campionamento accidentale è in genere limitato alle indagini esplorative. Considerati questi notevoli rischi, i risultati ottenuti da un campione di tale tipo vanno maneggiati con estrema cautela. Essi, inoltre, sovente sono applicabili solo a gruppi di persone simili a quelle che hanno accettato di partecipare all'indagine: questa similitudine, ovviamente, viene definita in base ad alcune variabili che possono essere in qualche modo legate agli scopi della ricerca.

6.2.2 Campionamento a valanga (*Snowball sampling*)

Il *campionamento a valanga* viene definito, in inglese, anche come *network sampling*, *chain sampling* o *reputational sampling*. I nomi che gli sono attribuiti ne rispecchiano la modalità di funzionamento: i membri di un gruppo, selezionati in quanto corrispondenti ai requisiti richiesti per partecipare all'indagine, indicano altri individui della popolazione adatti a partecipare alla ricerca (in quanto rispondenti agli stessi requisiti). Il campione, in questo modo, inizia con un piccolo gruppo di rispondenti e poi "cresce a valanga", aumenta costantemente (*snowball*) dato che ciascun rispondente della popolazione di partenza indica altri individui (*snowball* significa anche "palla di neve"). Costoro, a loro volta, diventano "informatori" e segnalano altre unità idonee ad essere coinvolte nella ricerca.

L'utilizzo del campionamento a valanga è limitato prevalentemente a popolazioni "rare", molto piccole o particolarmente disperse sul territorio e si basa sul presupposto che chi appartiene ad una popolazione così connotata (dove spesso gli individui sono a stretto contatto tra di loro) molto probabilmente conosce altre persone appartenenti alla stessa categoria.

Vantaggi

Questa tecnica può essere proficuamente utilizzata quando non sono disponibili liste in base a cui effettuare il campionamento o risulta difficile costruirle o procurarsele. È quanto succede, ad esempio, nel caso di particolari unità statistiche che non vogliono "mettersi in evidenza" (come gli immigrati clandestini) o unità che non hanno un recapito noto o, più in generale, sono molto difficili da rintracciare o contattare.

In genere la segnalazione da parte di alcune unità è anche segnale che queste ultime poi avranno la volontà di collaborare.

Svantaggi

Il sistema delle segnalazioni su cui si basa il campionamento a valanga potrebbe portare ad avere un campione distorto.

Inoltre il ricercatore ha un controllo minimo (se non, addirittura, inapplicabile) sulle unità statistiche che vengono citate e, quindi, inserite nella ricerca: spesso, cioè, non ha la possibilità (se non *ex-post*) di verificare direttamente che rispondano ai requisiti per partecipare all'indagine.

Altro possibile difetto della metodologia consiste nel fatto che le persone più conosciute vengono spesso citate, mentre c'è il rischio di escludere dalla ricerca individui più marginali o universalmente meno noti.

6.2.3 Campionamento per quote

Il campionamento per quote viene principalmente utilizzato nell'ambito delle indagini di mercato e nei sondaggi di opinione. La popolazione universo che deve essere studiata viene suddivisa in un certo numero di sottogruppi, omogenei al loro interno, in base ad una serie di variabili.

Ad esempio, una popolazione di individui potrebbe essere suddivisa in differenti gruppi in base a sesso ed età, se queste variabili sono significativamente collegate agli scopi della ricerca. Utilizzando una semplice classificazione di questo tipo, si possono, a titolo di esempio, ottenere quattro sottogruppi: maschi/anziani, maschi/giovani, femmine/anziane, femmine/giovani. Impostate le categorie si ricava (da fonti esistenti o da una eventuale ricerca su campo *ad-hoc*) la distribuzione quantitativa della popolazione obiettivo tra i vari gruppi.

Nella fase successiva, il campione va selezionato in modo che rifletta, proporzionalmente, la distribuzione della popolazione oggetto di studio. La numerosità campionaria viene così suddivisa tra i gruppi individuati in modo proporzionale a quanto accade nell'universo: si perviene, così, al numero di unità che vanno osservate per ciascuna delle classi costruite.

Naturalmente si può ricorrere al metodo "per quote" solo nel caso in cui si abbiano informazioni o dati utili per effettuare la suddivisione delle unità nei differenti gruppi. Inoltre, affinché le proporzioni siano realizzate in modo oculato e accurato, i record contenenti i dati relativi alle unità statistiche devono essere costantemente aggiornati, altrimenti si corre il rischio di ottenere risultati distorti, soprattutto nel caso si tratti di ricerche che si prolungano nel tempo o che vengono realizzate a cadenza periodica.

Il campionamento per quote può essere utilizzato congiuntamente ad altre tecniche, ad esempio con il campionamento stratificato: quest'ultimo è utilizzato per selezionare le unità di primo stadio, mentre il campionamento per quote è usato per reperire gli intervistati (unità del secondo stadio).

Vantaggi

La tecnica di campionamento per quote è efficace se le proporzioni sono individuate accuratamente e considerando variabili legate agli scopi dell'indagine. Nel caso in cui le unità risultino essere già assegnate ai differenti gruppi, non è necessaria la lista delle unità elementari: è possibile scegliere indifferentemente un certo numero di soggetti o sostituire parte di essi (che si sono rivelati, ad esempio, difficilmente o non raggiungibili) con altri equivalenti, senza correre il rischio di ottenere risultati particolarmente distorti.

Nella maggior parte dei casi, inoltre, il campionamento per quote consente di risparmiare su tempi e costi dell'indagine, dato che non è necessario costruire una lista per l'estrazione casuale o sistematica.

Svantaggi

In alcuni casi non si hanno i dati necessari per classificare le unità nei differenti gruppi. Anche qualora siano disponibili, non sempre si ha sufficiente accuratezza nella costruzione delle ripartizioni: infatti le proporzioni dei gruppi possono variare in modo significativo da contesto a contesto, oppure si può registrare una mobilità piuttosto elevata tra i gruppi che rischia di inficiare o distorcere la qualità dei dati ottenuti dal campione.

I rischi di distorsione possono essere dovuti al fatto che le unità reperibili per la ricerca o l'intervista sono categorie ben determinate e non rappresentano un campione casuale della popolazione.

Infatti se, ad esempio, ci si riferisce all'attività di un intervistatore, questo tenderà ad intervistare esclusivamente persone molto disponibili, a trascurare quelle difficilmente raggiungibili o che si rifiutano di partecipare e a fare la selezione anche in base al comportamento del rispondente; oppure, cosa ancor più grave, potrebbe essere propenso ad intervistare solo propri conoscenti e/o parenti. Questo può portare ad una distorsione dei risultati anche particolarmente considerevole. Per tali motivi Marbach (1992) ha introdotto alcuni principi che rendono meno arbitraria la scelta delle unità elementari oggetto di indagine.

L'intervistatore, inoltre, potrebbe non insistere con chi si rifiuta di rispondere; ciò potrebbe condurre ad una sottostima della variabilità del fenomeno studiato (infatti, le persone contattate rischiano di essere molto simili tra di loro). Va anche considerato, a questo proposito, che non sono previsti (o possibili)

controlli su chi non collabora, e questa mancanza di informazioni può essere fondamentale per arrivare al conseguimento dei risultati dell'indagine.

Per tutte queste ragioni, Hansen, Hurwitz e Madow (1953) ritengono il campionamento per quote non molto utile per effettuare indagini approfondite o che richiedono risultati che abbiano un alto tasso di precisione. Anche Deming (1960) ritiene il campionamento per quote anacronistico ed inefficiente.

Altri studiosi (Moser e Stuart, 1953) invece considerano un mezzo per arrivare a risultati validi; il suo valore è conseguenza, però, della qualità delle informazioni possedute sulle unità statistiche e delle esigenze (anche in termini di grado di precisione) che si propone la ricerca.

6.2.4 *Focus Group*

Il *focus group* è composto da un numero di persone (che varia, in genere, tra le 10 e le 20) scelte in modo tale da rappresentare una particolare popolazione. Ad esempio: giovani, potenziali consumatori di un certo prodotto, soggetti membri di una particolare professione, individui appartenenti ad una determinata classe socio-economica o demografica, e così via.

La tecnica basata sui *focus group* è principalmente usata nell'ambito delle ricerche di mercato, in particolare nel caso si vogliano identificare e valutare le esigenze di certe categorie di una specifica popolazione e quando si vuole prevedere quali saranno le dinamiche future dei consumi degli individui che ne fanno parte.

Il gruppo di persone selezionate si ritrova in un medesimo luogo e, sotto la supervisione di un intervistatore (o di uno dei ricercatori che coordinano il progetto di ricerca), discute liberamente sul tema proposto (o sui temi proposti). Gli argomenti che possono essere trattati e approfonditi con i *focus group* riguardano la salute, le condizioni sociali, il gradimento di un certo prodotto o servizio, e così via. In genere sono oggetto di studio il consumatore (o il paziente/utente) e le sue aspettative, le sue esigenze, i suoi problemi, le sue preferenze, ...

Spesso i *focus group* vengono utilizzati nella fase di test delle indagini. Sono un modo molto valido per valutare su un ristretto, ma teoricamente rappresentativo, gruppo di persone l'efficacia, la comprensibilità e altre caratteristiche di un questionario che poi, in fasi successive della ricerca, verrà somministrato ad un popolazione più ampia.

È ovvio che uno degli aspetti fondamentali per ottenere buoni risultati da questa tecnica è che il gruppo selezionato rispecchi il gruppo più numeroso che compone la popolazione obiettivo; per questo sono necessari studi approfonditi sia del gruppo di indagine che della *target population*.

Vantaggi

Le indagini basate sui *focus group* consentono di ottenere un quadro accurato e dettagliato di bisogni, aspettative, esigenze, ma anche di problemi e motivi di insoddisfazione di un gruppo specifico di individui (ad esempio i consumatori di un certo prodotto o gli utenti di un determinato servizio) in tempi relativamente brevi e mantenendo costi di ricerca piuttosto contenuti.

Svantaggi

Se il gruppo selezionato si contraddistingue in un certo modo (tutti i componenti hanno un livello di istruzione piuttosto elevato, o un reddito molto superiore alla media, o sono tutti appartenenti allo stesso sesso oppure alla stessa fascia d'età o di reddito, ...) i risultati ottenuti dall'indagine potrebbero, in seguito, non essere estendibili ad una popolazione più ampia. Per questo motivo, la scelta degli individui che compongono il *focus group* deve essere fatta in modo estremamente accurato e oculatamente, in modo da prevenire la non rappresentatività del campione; altrimenti si corre il rischio di pervenire a risultati distorti.

6.2.5 Campionamento ragionato

Utilizzando il *campionamento ragionato* (in inglese *judgement sampling* o *porposive sampling*), il ricercatore che ha il compito di impostare la ricerca si rivolge, col questionario, a persone che si crede appartengano alla popolazione "media" o a un gruppo considerato obiettivo dell'indagine (Bradley, 1997). Il giudizio del ricercatore si sostituisce, in questa tecnica di ricerca, al caso. Infatti, per la scelta degli individui componenti il gruppo ci si può basare su casi precedenti, sulla letteratura disponibile, su esperienze personali, su informazioni disponibili riguardo la popolazione, e così via.

L'uso del campionamento ragionato è prevalentemente fatto nel contesto di indagini esplorative.

Vantaggi

Il campionamento ragionato può essere utilizzato anche quando non si ha alcuna informazione sul fenomeno preso in considerazione, ma si può presumere che il gruppo di individui scelto possa essere in qualche modo rappresentativo della popolazione obiettivo.

Inoltre (Vianelli, 1986) si preferisce il campionamento ragionato per indagini molto ampie, ma che comprendono poche unità territoriali (ad esempio per i test di prova di un prodotto).

Svantaggi

Non si ha la certezza che il campione scelto sia rappresentativo: il metodo di selezione, pur basandosi sull'opinione di esperti o su casi precedenti, non è detto che rispecchi la popolazione universo di riferimento.

Inoltre la tecnica non prevede l'applicazione di procedimenti probabilistici (che, ad esempio, permettono di calcolare il grado di affidabilità dei risultati).

Se, infine, si può arrivare a risultati sufficientemente garantiti per quanto concerne aspetti conosciuti della popolazione e caratteri fortemente associati a questi, non è detto che la stessa affidabilità nei risultati caratterizzi aspetti meno noti (o che non si comportano concordemente a quelli noti).

6.2.6 Campionamento Sistematico non Casuale

Il *campionamento sistematico non casuale* è, nelle modalità di funzionamento, molto simile al campionamento sistematico casuale: può essere considerato una sua variante, soprattutto in determinati ambiti operativi. Ciò che cambia è che non si dispone di una lista di unità da cui potere estrarre il campione poichè le unità si rendono disponibili via

via che trascorre il tempo. In questa situazione non è, quindi, nemmeno possibile assicurare l'equiprobabilità di estrazione delle unità.

Al campionamento sistematico non casuale si può ricorrere anche nel caso in cui non possano essere determinate le quote campionarie, a causa della mancanza di dati sulla popolazione oggetto di indagine. Ad esempio, considerando i clienti di un distributore di benzina, si può decidere di selezionarne, per un'intervista, uno ogni 10 che vogliono fare rifornimento. Questa operazione può essere eseguita, ad esempio, tenendo conto della valutazione empirica della curva di affluenza nel corso della giornata (Marbach, 1992).

Come si può constatare, in tali situazioni la numerosità della popolazione e la numerosità campionaria non sono note a priori; quindi il passo di campionamento deve essere deciso arbitrariamente dal ricercatore. Nella scelta bisogna, però, tenere conto della numerosità approssimativa che si vuole ottenere (o che si può ottenere) in relazione alle caratteristiche della popolazione di riferimento.

6.2.7 Campionamento auto-selettivo

Nel *campionamento auto-selettivo* (*self-selection sampling*) i rispondenti si offrono volontariamente per partecipare alla ricerca (Bradley, 1997). Questo è quanto succede nella maggior parte delle indagini che vengono realizzate via Web, come si vedrà meglio nel seguito.

Sono evidenti i rischi di una procedura di questo tipo: si possono ottenere risultati anche notevolmente distorti, in particolare per il fatto che gli individui che accetteranno di partecipare alla ricerca sono caratterizzati da connotazioni ben specifiche e potrebbero essere rappresentativi solamente di una piccola parte della popolazione obiettivo o, al limite, potrebbero addirittura non coincidere nemmeno parzialmente con essa.

6.2.8 Campionamento di elementi rappresentativi

L'utilizzo del campionamento di elementi rappresentativi avviene in particolare quando le scadenze di realizzazione dell'indagine sono piuttosto ristrette e non si ha il tempo di selezionare o utilizzare un campione numericamente ampio.

Una volta identificata la popolazione obiettivo, il *campionamento di elementi rappresentativi* procede con una ricerca riguardante alcune sue caratteristiche, eventualmente prendendo anche in considerazione i risultati di indagini precedenti. In una seconda fase viene costituito un campione di individui che, riguardo le variabili su cui si è indagato nella prima fase, si avvicinano ai valori medi rilevati sull'intera popolazione. Gli individui selezionati sono considerati *elementi rappresentativi* dell'intera popolazione obiettivo.

È necessario porre molta attenzione nella scelta dei caratteri determinanti per considerare un individuo come rappresentativo della popolazione. In primo luogo deve trattarsi di variabili in qualche modo correlate con gli obiettivi della ricerca; in secondo luogo bisogna accertarsi che i dati statistici su cui ci si basa siano corretti ed aggiornati.

Uno dei principali campi di applicazione per il campionamento di elementi rappresentativi è quello geografico: un'area campione viene considerata (per le caratteristiche che presenta) rappresentativa di un'area più vasta, oppure di un'insieme di aree differenti ma simili a quella considerata.

6.2.9 Campionamento basato su aree barometro

Le *aree barometro* sono alcune piccole aree (*microaree*) che, riguardo a determinati aspetti, si presentano con caratteristiche simili a quelle di zone più estese (*macroaree*).

Il *campionamento basato su aree barometro* si basa essenzialmente sulla capacità delle microaree di essere rappresentative delle macroaree.

I due principali campi di applicazione di questa tecnica di campionamento non probabilistico sono quello politico e quello delle indagini di mercato.

Nel primo contesto, un'area di estensione ridotta (ad esempio, un comune) può presentare, se opportunamente scelta, orientamenti elettorali simili a quelli dell'intera Nazione (macroarea): per avere un pronostico dei risultati delle elezioni, si potrà quindi effettuare un'indagine approfondita sulla volontà di voto degli abitanti della microarea, con notevoli risparmi di risorse e tempo.

Per quanto riguarda il secondo aspetto (indagini di mercato), si possono identificare microaree rappresentative, a livello di comportamenti di consumo, di zone più vaste. Così, per valutare il gradimento nei confronti di un determinato prodotto, sarà possibile testarlo sull'area barometro, apportando poi i cambiamenti più appropriati al caso, prima che venga distribuito su scala nazionale.

6.2.10 Campionamento per plausibilità (*Plausibility sampling*)

Il *campionamento per plausibilità* (o *plausibility sampling*) si basa sulla scelta di un campione selezionato per il fatto che "sembra plausibile che la sua popolazione sia rappresentativa di una popolazione più ampia. Non si ha, però, alcuna evidenza di questo fatto" (Talmage, 1988), come può avvenire invece nel caso della scelta di aree barometro (comportamenti simili attestati, ad esempio, da dati passati) o nel caso dell'utilizzo di elementi rappresentativi (che, cioè, presentano caratteri vicini alla media della popolazione obiettivo).

Nella seguente Tabella (Tab. 1) sono riportati i principali tipi di tecniche di campionamento, classificate come probabilistiche e non probabilistiche.

Tipo di campionamento	
Probabilistico	**Non probabilistico**
C. casuale semplice	C. di convenienza
C. sistematico	C. a valanga
C. casuale stratificato	C. per quote
C. a grappoli / a aree	*Focus Group*
C. a due o più stadi	C. ragionato
C. a due fasi (doppio)	C. sistematico non casuale
	C. auto-selettivo
	C. di elementi rappresentativi
	C. di aree barometro
	C. per plausibilità

Tab. 1. *Tecniche di campionamento probabilistico e non-probabilistico.*

7 Campionamento per Indagini via Web

Le principali tecniche di campionamento probabilistico e non-probabilistico possono essere applicate alle indagini via Web.

Così, ad esempio, per ottenere un campione utilizzando la tecnica del campionamento casuale semplice, basta utilizzare un elenco di indirizzi di posta elettronica opportunamente ordinati. Tramite una tabella di numeri casuali si seleziona un insieme di indirizzi *e-mail*, tra quelli disponibili. Oppure (utilizzando la tecnica di campionamento sistematico) si può fare in modo che un computer invii per posta elettronica un invito a partecipare all'indagine ad alcuni indirizzi selezionati in modo sistematico dalla lista, cioè utilizzando un determinato passo di campionamento. Nel caso si conoscano altre caratteristiche socio-demografiche od economiche delle unità statistiche che compongono la popolazione obiettivo, si può ricorrere ad una loro stratificazione. Ma è anche possibile, ad esempio in una indagine rivolta ai dipendenti degli ospedali di una città, estrarre dei grappoli (cioè solo alcuni ospedali) e poi intervistare tutte le unità che a essi fanno riferimento (i dipendenti di quegli ospedali) utilizzando le liste di indirizzi di posta elettronica dei dipendenti fornite dalle stesse strutture ospedaliere.

La popolazione obiettivo e le liste di campionamento, come pure le modalità di selezione dei rispondenti hanno, in molti casi, caratteristiche strettamente legate all'ambiente di Internet. In particolare, le differenze tra popolazione obiettivo e campione e le difficoltà in cui ci si imbatte nell'identificare le liste di campionamento sono all'origine di problemi di copertura. Inoltre, l'auto-selezione è un importante problema, specifico delle indagini condotte via Internet, e una tematica di rilievo relativamente al campionamento. Per questi motivi in molti casi è difficile ottenere un campione probabilistico e si rende, così, necessario ricorrere a campioni non probabilistici.

Da queste premesse si intuisce l'importanza di classificare, in via preliminare, le differenti fonti che consentono di selezionare i rispondenti che in seguito verranno coinvolti in indagini via Web. La comprensione delle caratteristiche che contraddistinguono i differenti tipi di liste di campionamento e del modo in cui queste sono costruite è infatti una fase fondamentale che deve precedere una attenta riflessione sui problemi, legati alle tecniche di campionamento, specifici delle indagini condotte in ambiente Internet.

7.1 Le liste di campionamento

La *lista di campionamento* (*sampling frame* o *sampling list*) è la lista che contiene le unità che soddisfano i criteri adottati per identificare la popolazione obiettivo. La lista di campionamento è considerata, nell'analisi empirica, una approssimazione della popolazione obiettivo. Quindi una buona lista deve coprire nel modo più completo possibile la popolazione obiettivo (Malhotra, 1999).

Per estensione, poi, con *lista di campionamento* si intende anche la lista di soggetti tra i quali si può estrarre il campione che verrà studiato.

Uno dei principali problemi per le indagini via Web, peraltro comune a molti altri tipi di indagine, è quello della reperibilità di liste di campionamento adatte al caso che si vuole studiare. Infatti solo nel caso in cui la lista di campionamento sia una buona approssimazione della popolazione obiettivo e, in definitiva, solo se la popolazione obiettivo (e quindi il *sampling frame*) può essere facilmente definita, identificata e contattata ha senso utilizzare un approccio probabilistico e, quindi, applicare criteri di inferenza statistica. Vi sono casi in cui, invece, la popolazione obiettivo non è ben definita o contattabile in modo immediato; in tale situazione sarà opportuno utilizzare metodi di selezione non probabilistici.

In questo capitolo si studieranno in modo dettagliato le possibili fonti (e le relative caratteristiche) cui si può ricorrere nel campo delle indagini via Web per costruire liste di campionamento opportune al caso che si vuole trattare; contestualmente si cercherà di effettuare una loro classificazione.

7.1.1 Modalità di selezione dei rispondenti

Nelle indagini via Web vi sono essenzialmente due tipi di approccio frequentemente utilizzati per arrivare alla costruzione di *frame*. In genere, utilizzando entrambi, si otterranno liste parziali, composte da individui auto-selezionati o che, comunque, non sono una buona approssimazione della popolazione obiettivo. La maggior parte delle liste potrà, infatti, essere considerata un campione non probabilistico della popolazione obiettivo e comunque da esse non sarà possibile estrarre campioni probabilistici.

I due tipi di approccio indicati sono l'"esterno" e l'"interno":

- **Approccio "interno"** (***internal***): i partecipanti all'indagine vengono "reclutati" in rete. Si tratta di visitatori di siti Web o di persone iscritte (per differenti ragioni) a *mailing-list*. Nella classificazione proposta da Alvarez ed altri (2003), l'approccio interno viene a sua volta suddiviso in due differenti metodologie, considerando il tipo di contatto utilizzato.
  - La prima comprende modalità di contatto avvenute tramite *Web advertisements*: il reclutamento avviene prevalentemente tramite annunci pubblicati su siti Web o rivolti a *newsgroup* e finalizzati ad incoraggiare la partecipazione all'indagine. Si tratta, evidentemente, di modalità di selezione che portano alla creazione di campioni non probabilistici.
  - Il secondo metodo è invece basato su *subscription* o *co-registration procedures*: a individui che si sono registrati per usufruire di altri servizi o hanno fatto una sottoscrizione, viene richiesto di fornire un recapito di posta elettronica. Successivamente queste unità verranno coinvolte nello svolgimento di indagini *on-line*.
- **Approccio "esterno"** (***external***): comprende altri metodi di selezione dei rispondenti. Il contatto viene stabilito utilizzando liste provenienti da *panel*, elenchi cartacei o di altro tipo ed avviene tramite forme di comunicazione differenti da quelle previste per l'approccio interno (ad esempio: lettere di invito, affissioni, contatto telefonico...). I potenziali rispondenti vengono invitati ad entrare in rete per partecipare all'indagine. Internet, in questo secondo caso, è considerato un tramite per la raccolta dei dati e non un modo per coinvolgere i potenziali rispondenti nell'indagine.

Sono anche previste forme di contatto considerabili un misto dei due approcci indicati.

7.1.2 Il contatto con i rispondenti

Considerando sia l'approccio interno che quello esterno, è in corso un acceso dibattito riguardante la fase di contatto con i rispondenti, molto delicata perché spesso da essa dipende il successo del tentativo di convincere l'individuo a collaborare all'indagine. Il concetto attorno a cui ruota il dibattito, introdotto da Godin (1999), è *permission* (permesso). In base al termine *permission* l'orientamento attuale è quello di contattare, per svolgere una ricerca, solo le persone che hanno dato il permesso o che si aspettano di essere contattate.

Il "permesso" su cui si basa il contatto dei rispondenti è risultato della legislazione vigente, delle norme che regolano la attività degli *ISP* (*Internet Service Provider*, i fornitori dei servizi Internet), della resistenza a partecipare alle indagini che vengono proposte (purtroppo sempre più diffusa e sempre meno facilmente penetrabile) e delle attuali e logiche buone regole di comportamento cui dovrebbe sottostare la fase di contatto del rispondente.

I principali tipi di contatto sono di due tipi: *opt-out* e *opt-in*.

- ***Opt-out***: si tratta della modalità di contatto utilizzata prevalentemente negli Stati Uniti. Considerando, a titolo esemplificativo, un utente che si sta registrando ad un sito web attraverso un *form*, l'approccio *opt-out* prevede che nel *form* stesso venga posizionato un *box*. Selezionando questo *box* il rispondente manifesta la sua non disponibilità ad essere contattato in futuro.
- ***Opt-in***: è l'approccio prevalentemente utilizzato nei paesi della Unione Europea; si basa su una logica lievemente differente da quella *opt-out*. Considerando lo stesso caso (visto sopra) di registrazione ad un sito web attraverso un *form*, il soggetto deve attivamente esprimere il proprio consenso e la propria disponibilità ad essere contattato in occasioni future. All'interno del *form* è inserito un *box* selezionabile che riporterà una indicazione del tipo: "vi autorizzo a contattarmi in futuro per eventuali indagini…".

Sia nel caso della modalità *opt-in* che per la *opt-out* deve comunque essere predisposta una procedura semplice, immediata e attendibile per permettere alla persona che si è registrata (e ha dato il consenso ad essere ricontattata in futuro) di revocare il permesso.

È acceso il dibattito anche sul fatto di inserire, nel *form* di registrazione, un *box* già pre-selezionato (o addirittura inserito fuori dall'area della pagina visualizzata, in modo che il consenso venga espresso in modo non consapevole). Da un lato questi stratagemmi possono aumentare il tasso di adesioni "non volontarie" o "passive", dall'altra però può provocare l'irritazione di chi cade nella "trappola" e può compromettere, come conseguenza, la volontà futura dei rispondenti (coinvolti subdolamente) a farsi ri-contattare o coinvolgere in altre indagini.

Esiste un terzo sistema di contatto dei potenziali rispondenti: il sistema "doppio *opt-in*".

- ***Double opt-in***: chi si sta registrando ad un sito web o vuole usufruire di un servizio ad esso collegato deve esprimere la propria disponibilità ad essere contattato (o ricontattato) seguendo un iter in due fasi. Al primo passo deve esprimere la propria disponibilità ad essere contattato in futuro e fornisce un

indirizzo *e-mail* valido. In un momento successivo viene richiesta, via *e-mail*, una conferma ulteriore della volontà di essere ri-contattato in ulteriori occasioni. Il vantaggio dell'approccio *double opt-in*, rispetto ai due precedenti, è che consente di ottenere una lista di indirizzi di posta elettronica più "raffinata", cioè composta da persone realmente interessate ad ipotesi di contatto future. La disponibilità a collaborare sarà maggiore di quella di altri che potrebbero avere selezionato per sbaglio od inconsapevolmente il *box* di consenso, oppure potrebbero avere fornito indirizzi e-mail non validi.

Le modalità di contatto basate sul consenso del potenziale rispondente sono, molto di frequente, alla base della costituzione di liste di campionamento utilizzate in molte indagini di tipo differente.

I modi (ed in contesti) in cui si può cercare di convincere e coinvolgere in indagini via Web i potenziali rispondenti sono numerosi. Nel paragrafo successivo si cercherà di effettuare una loro classificazione prendendo in considerazione i principali.

7.1.3 Come costruire le liste di campionamento

Le fonti che possono essere utilizzate per costruire le *sampling frame* sono molteplici: l'ambiente Web offre una svariata gamma di strumenti. Ovviamente i *frame* ricavati dalle differenti metodologie hanno vantaggi, svantaggi e caratteristiche differenti importanti in quanto in stretta relazione con le fasi di campionamento e di stima [10].

Nel seguito verranno presentati i principali criteri attraverso cui è possibile costruire liste di campionamento, considerandone le caratteristiche principali.

- **Annunci:** sono pubblicati sulle pagine di siti Web (*invitation tags*), inviati attraverso *e-mail* oppure diffusi attraverso altre forme di comunicazione (volantinaggio, pubblicazione sui giornali, affissioni e così via); gli annunci presentano l'indagine, incoraggiano la partecipazione e indirizzano il rispondente direttamente sulla pagina Web che ospita il questionario; oppure forniscono un punto di contatto attraverso il quale è possibile avere maggiori informazioni sulle modalità di svolgimento dell'indagine. Con gli indirizzi *e-mail* forniti è possibile costruire una lista utilizzabile anche in occasioni successive. Attraverso gli *invitation tags* (e con metodi analoghi) si ottiene un campione totalmente auto-selezionato che rischia, quindi, di produrre risultati distorti.
- **Inviti (*banner*, lettere):** l'invito a partecipare all'indagine avviene attraverso *banner* pubblicitari pubblicati sulle pagine di un sito Internet. I *banner* possono essere visibili in qualsiasi momento e a qualsiasi visitatore, oppure possono essere proposti solo a cadenza periodica: il *banner*, ad esempio, può essere visualizzato in base ad un prefissato intervallo temporale o ad un certo numero di visitatori che fanno il loro ingresso in una determinata pagina. L'efficacia dei *banner* (in alcuni casi associati a incentivi che incoraggiano la partecipazione) è maggiore di quella degli *invitation tags* e il loro utilizzo avviene, in genere, nell'ambito di indagini specifiche. Infatti possono essere posti all'interno di siti

[10] Questo aspetto viene trattato più diffusamente nel Quaderno di Ricerca "Inferenza nelle Indagini via Web" (Biffignandi, Toninelli, 2006); si veda Bibliografia.

(o pagine Web) che hanno un contenuto ben definito e possono raggiungere, quindi, gruppi di persone piuttosto ben selezionati e circoscritti. L'utilizzo dei *banner* non porta, però, a risultati molto confortanti in termini di tassi di adesione, in genere attorno allo 0,5%. Il costo, relativamente al tasso di partecipazione all'indagine, è piuttosto elevato. Sono evidenti, inoltre, i problemi legati alla auto-selettività del campione, anche se con il *banner* è possibile non specificare a priori l'argomento dell'indagine e quindi cercare di arginare questo fenomeno.

- **Link ipertestuali:** i *link* ipertestuali che rimandano direttamente al questionario dell'indagine sono posti sulle pagine Web, in motori di ricerca o anche in altri contesti. I *link* possono essere messi in evidenza con eventuali accorgimenti grafici.

- **Indagini Pop-up (*Pop-up surveys*):** in modo casuale viene selezionato un campione di visitatori di un sito Web. Il questionario appare direttamente sullo schermo al potenziale rispondente la navigazione del quale viene intercettata *on-line*. Il navigatore fa richiesta di una pagina specifica e, se viene selezionato dal meccanismo, il sistema provvede ad aprire, davanti alla pagina richiesta, una nuova finestra contenente l'invito all'indagine. L'accesso all'indagine può essere proposto tramite un invito (un'introduzione riportante scopi, caratteristiche e enti promotori dell'indagine) che rimanda al questionario vero, oppure può essere diretto (in questo caso il questionario appare direttamente al visitatore selezionato). Il metodo di selezione *Pop-up* del campione è usato piuttosto di frequente e in genere dà risultati migliori di sistemi basati sulla presenza di un *banner* fisso (Malhotra 1999). La presenza di *cookie* registrati dal personal computer del rispondente, inoltre, consente di evitare di ricontattare due volte la stessa persona. Vi sono però alcuni problemi che possono aumentare il rischio di distorsione: alcuni individui cancellano molto spesso i *cookie*, e quindi rischiano di essere invitati più volte a partecipare ad una indagine; inoltre rischiano di essere contattate ripetutamente anche persone con a disposizione due o più macchine da cui hanno accesso alla rete. Infine va considerato anche il problema opposto: vi sono individui che condividono lo stesso computer, e ciò comporta che alcuni di essi potenzialmente potrebbero non essere mai contattati; infatti, una volta che uno degli utilizzatori del computer ha partecipato all'indagine, grazie alla presenza dei *cookie* allo stesso terminale non verrà più richiesta la collaborazione. L'approccio *Pop-up* viene prevalentemente utilizzato in indagini sul profilo dei visitatori di siti web (caratteristiche, esigenze, esperienze, ...) o per esperimenti controllati finalizzati a testare alternative esperienze *on-line*. L'indagine, per avere successo, deve essere breve, presentata con un aspetto grafico attraente, corredata di un invito di qualità e deve riguardare un argomento il più possibile interessante. In alcuni contesti si utilizzano incentivi, ma Comley (2000) ha fatto notare che con questo tipo di incoraggiamenti alla collaborazione si rischia di attirare risponditori fraudolenti. Costoro, per poter essere di nuovo contattati e beneficiare di nuovo degli incentivi, cancellano frequentemente i *cookie* dai propri personal computer. Il tasso di collaborazione alle indagini proposte con il sistema *Pop-up* varia tra il 2 e il 15% (secondo

quanto evidenziato dalla *Fletcher Research* [11]); la Virtual Surveys [12] parla, per le sue ricerche, di un tasso di adesione di circa il 25%.

- ***Hijack* o *Browsing intercept Software*:** si tratta di un approccio simile a quello *Pop-up*. Il modo di contattare i potenziali rispondenti è, però, molto più intrusivo: specifici *software* intercettano la lettura di documenti e la navigazione su Internet. Un campione di visitatori, scelti in base a determinati criteri (simili a quelli visti per le indagini *Pop-up*), viene intercettato in modo attivo e ri-direzionato, invece che alla pagina richiesta, ad una pagina contenente il questionario da compilare o che propone l'indagine (Pfleiderer & Gente 1998). Buona norma vuole che in questa pagina sia inserito un *decline button*, che consente di *by*-passarla per raggiungere direttamente la pagina richiesta. L'approccio *hijack* non è molto raccomandato, in quanto è ritenuto dalla maggior parte dei ricercatori eccessivamente intrusivo e rischia di aumentare la futura resistenza dei rispondenti (anche in occasione di altre indagini). I risultati che consente di ottenere sono, però, piuttosto incoraggianti: una ricerca condotta nel Regno Unito dalla BBC riguardante i propri ascoltatori ha ottenuto un tasso di risposta dell'80%. Il successo dell'indagine, in realtà, è stato spiegato con l'ottimo rapporto che lega l'emittente pubblica inglese con i propri telespettatori.
- ***Opinion center.*** Alcuni degli *opinion center* più popolari sono quelli di AOL (*America On Line* [13]) e DMS (*Digital Marketing Services* [14]), ma in rete ce ne sono molti altri (si veda il sito: www.opinionplace.com). I navigatori (attraverso, ad esempio, l'uso di *banner*) vengono spinti ad accedere all'*opinion place*, un particolare luogo virtuale dove ciascun visitatore può esprimere la propria disponibilità a partecipare ad una indagine rispondendo ad un breve questionario. Una volta manifestato il consenso a fare l'intervista (ma senza conoscere l'argomento, che viene scelto in automatico dal sistema), i rispondenti decidono liberamente quando rispondere al questionario proposto (in genere si tratta di più di una intervista ogni settimana).
- ***Panel.*** I rispondenti vengono reclutati per partecipare a più indagini e/o a indagini longitudinali. Il contatto avviene tramite telefono (ma anche via *e-mail*, per lettera o in altri modi): viene chiesto il consenso all'inserimento in un gruppo a cui, in futuro, verrà richiesto (in genere via *e-mail*) di partecipare ad un certo numero di indagini. In base ad alcune caratteristiche di base e di stratificazione delle unità inserite nel *panel* (e richieste all'atto dell'inserimento) è possibile tracciare un profilo delle diverse unità, utile per individuare quelle da selezionare nelle specifiche indagini. L'approccio tramite *panel*, permettendo di contattare individui ben definiti o particolari sub-popolazioni, consente di operare con obiettivi di indagine anche molto specifici. Un problema significativo (Alvarez e VanBeselaere, 2003) è dovuto alla reperibilità dei partecipanti: gli indirizzi di posta elettronica vengono cambiati piuttosto di

---

[11] http://easily.co.uk/

[12] http://www.virtualsurveys.com/

[13] http://corp.aol.com/index.shtml

[14] http://www.dmsdallas.com/company_overview.html

frequente e problemi tecnici possono influire sulla possibilità di contattare i potenziali rispondenti. Le indagini basate su *panel* sono anche particolarmente adatte per valutare come variano nel tempo la soddisfazione e le esigenze di determinati gruppi di consumatori. Sono tutt'ora in corso studi per valutare come varia il comportamento del *panel* per indagini di lungo corso. Non sembra che vi sia alcuna particolare influenza della lunghezza dell'indagine sul comportamento del *panel* (Alvarez e VanBeselaere, 2003), anche se, con il prolungarsi di essa, il tasso di risposta tendenzialmente cala. Le *panel company* si occupano del reclutamento dei rispondenti e del contatto con essi, amministrano e progettano le interviste e fanno in modo che i campioni rimangano stabili nel tempo, sostituendo eventuali fuoriuscite dal *panel* con individui dal profilo analogo [15]. In genere si possono identificare due tipi di *panel*: piuttosto grandi, costituiti da *database* di indirizzi di coloro che hanno dato il consenso (corredati da poche informazioni essenziali) e più piccoli e meno focalizzati, composti da pochi membri di cui però si dispone di profili molto accurati e dettagliati. In genere, con un campione basato sul *database panel*, si hanno tassi di risposta del 30-40%.

- ***Phone invitation***. Una particolare modalità di reclutamento è nota come *phone invitation*, utilizzata, ad esempio, dalla *B2B Research* [16] per reclutare i panelisti e per ottenere campioni di consumatori raggruppati geograficamente. Un intervistatore contatta le possibili unità statistiche telefonicamente e verifica se corrispondono alla *target population*. Al soggetto viene poi spiegato il processo da seguire per la registrazione *on-line* e viene fornito un indirizzo URL per la registrazione nella *panel list*. Il tasso di risposta per questo tipo di indagini è piuttosto elevato, ma non sempre è giustificato dagli alti costi che il contatto telefonico richiede.
- **Liste di indirizzi E-mail**: gli indirizzi sono contenuti in grossi *database* costruiti per fini e con modalità di diversa natura. In sintesi si possono elencare i seguenti tipi di liste:
  a. Liste di persone o imprese che hanno dato il consenso ad essere ricontattate, in genere incidentalmente alla richiesta della fruizione di determinati servizi. Un esempio è il caso della *SSI* (*Survey Sampling Inc.* [17]). A differenza del *panel*, spesso un soggetto non sa di essere stato incluso in una lista di *e-mail* di questo tipo. Vi sono però dei problemi: in primo luogo, alcuni individui possono essere inseriti in più liste che poi vengono utilizzate per la stessa indagine (si rischia così di contattare due volte la stessa persona). Inoltre, per le liste di qualità, si ottengono in genere tassi di partecipazione tra il 2 e l'8%, quindi particolarmente bassi. Per avere un certo numero di interviste complete è necessario, dunque, contattare un numero cospicuo di indirizzi.

---

[15] Un esempio di compagnie specializzate in indagini panel è la *Knowledge Networks* (http://knowledgenetwords.com).

[16] http://www.b2binternational.com/

[17] http://www.surveysampling.com/

b. Liste di unità registrate in archivi amministrativi quali, ad esempio, il *Business Register* o il *Social Security Database*. Il *database* amministrativo è costruito per raccogliere alcune informazioni di base a scopi amministrativi, ciononostante sono richiesti anche altri dati come l'indirizzo *e-mail* dei soggetti inseriti nel registro. Dato che in queste liste l'indirizzo di posta elettronica non è una variabile essenziale ai fini amministrativi, le informazioni possono essere spesso mancanti o non aggiornate. In molti paesi, inoltre, le leggi sulla *privacy* non consentono agli enti che gestiscono gli archivi amministrativi di diffondere le informazioni in essi contenute.

c. Liste di unità il cui indirizzo *e-mail* è stato raccolto attraverso le pagine di siti Internet. In questa categoria possono essere inseriti gli *Harvested addresses*, ovvero indirizzi registrati in modo automatico tramite appositi *software*, oppure manualmente. Il sistema di raccolta automatico di indirizzi può essere reso meno efficace dall'azione di *software* appositamente studiati. Uno di questi, attualmente di uso piuttosto diffuso, è *Spam bait creator*, che crea in modo automatico pagine web contenenti indirizzi fasulli. La raccolta di questi indirizzi falsi oltre a ridurre drasticamente la qualità del *database* può rendere inutilizzabili le liste di invio. Vi sono, poi, altri tipi di *software* finalizzati alla protezione degli indirizzi *e-mail* pubblicati sulle pagine di un sito web: questi possono essere utilizzati dai visitatori del sito, ma non possono essere copiati dai sistemi di raccolta automatici. Una ulteriore protezione agli indirizzi pubblicati in rete può essere l'utilizzo di *password* che consentono l'accesso esclusivo alle pagine del sito solo a determinati utenti. In questi casi l'impossibilità di raccolta degli indirizzi è piuttosto grave in quanto, trattandosi di recapiti di posta elettronica protetti, si presuppone che siano indirizzi funzionanti che potrebbero arricchire notevolmente il *database* che si vuole costruire. In alcuni Stati sono anche mantenute, per essere utilizzate in seguito, delle liste pubbliche nazionali di indirizzi *e-mail* (Batagelj, 1999). Alcuni siti web, infine, raccolgono indirizzi tramite *form* di registrazione utilizzati per accedere a determinate pagine e/o per usufruire di determinati servizi. Nel *form* di registrazione possono essere richieste anche numerose altre informazioni aggiuntive ad un livello di approfondimento più o meno elevato. Questi dati, che vengono registrati nel *database* insieme agli indirizzi, consentono non solo di fare analisi più approfondite (forniscono profili molto dettagliati dei visitatori di un sito), ma sono anche una valida base utilizzabile per la selezione di campioni. In realtà, però, anche questi dati vanno considerati con molta cautela: una ricerca della *Graphics Visualization Unit* ha infatti indicato che all'incirca la metà di coloro che si registrano sui *form* di un sito Internet dichiarano informazioni false.[18]

d. Liste di clienti registrati tramite carte fedeltà, sottoscrizioni, abbonamenti. In questi casi di solito all'atto della registrazione vengono richieste alcune informazioni, tra cui l'indirizzo *e-mail* e altri dati che permettono di elaborare i profili degli individui registrati. Oltre alle liste di indirizzi *e-mail* costruite per scopi commerciali, vi sono liste di indirizzi private (ad esempio:

[18] http://www.cc.gatech.edu/gvu/user_surveys/survey-1998-10/

i clienti di uno studio dentistico o i fornitori di un ipermercato). Queste liste sono, però, più problematiche da ottenere per un ricercatore esterno: deve essere chiesto il permesso per ogni loro utilizzo in campi che non sono strettamente attinenti al motivo per cui sono state costruite, richiedono di poter identificare i minorenni, devono prevedere una metodologia efficace e veloce di cancellazione, necessitano molta cura per una trascrizione corretta degli indirizzi e rendono indispensabili frequenti operazioni di controllo e aggiornamento. Inoltre l'utilizzo di tali liste è limitato all'osservazione di popolazioni finite e circoscritte. Lo stesso succede per gli archivi in cui numerose grandi aziende raggruppano gli indirizzi di posta elettronica dei propri dipendenti. In questo ultimo caso si è in grado però di ottenere campioni di individui con caratteristiche ben precise, in quanto di solito vengono registrate, negli stessi archivi, anche numerose informazioni anagrafiche o di altro tipo che possano rivelarsi molto utili ai fini della selezione del campione.

e. Liste di membri di gruppi Internet. I gruppi di discussione sono caratterizzati dal fatto che chi ne fa parte è accomunato agli altri membri da uno spiccato interesse per un determinato argomento: si tratta di comunità virtuali in cui i membri discutono di un argomento specifico. In merito sono stati fatti approfonditi studi da Witmer e altri (1999). Oltre ai gruppi di discussione, si possono citare i *newsgroup*, *Usenet* (cioè una rete che fornisce servizi ai gruppi di dibattito su Internet), le liste di discussione (*discussion list*), i gruppi di interesse (*Interest Groups*), le liste di discussione (*Discussion Lists*); si ricorre anche all'utilizzo di *software* come *Listserv* (elaboratore centrale di elenchi postali: si tratta di un servizio che consente di formare sul computer collegato in rete un elenco di indirizzi di posta elettronica) o *ListBot*. Una ricerca di Bradley (1999) ha dimostrato che utilizzando un campione selezionato tramite i gruppi Internet si ottiene un tasso di risposta maggiore e un tempo di risposta molto più veloce, rispetto a quelli riscontrati per indagini postali.[19] L'utilizzo dei gruppi Internet comporta una serie di problemi. In primo luogo uno stesso individuo può essere iscritto a più gruppi, e quindi può essere coinvolto nelle stesse indagini più volte. In secondo luogo gli stessi individui appartenenti ad una lista possono risultare iscritti ad essa due o più volte (con indirizzi di posta elettronica differenti). Infine vi sono alcuni casi di individui che, pur essendo inseriti in un gruppo di discussione, non avrebbero intenzione di esserlo, ma non riescono a farsi cancellare (o non ne hanno la possibilità). Vi sono, poi, altri rischi collegati all'utilizzo dei gruppi Internet: alcuni rispondenti possono inviare le loro risposte alla lista principale rischiando di influenzare gli altri e contribuendo ad accrescerne l'irritazione per il numero di *e-mail* ricevute); ciò può comportare una diminuzione del tasso di risposta. Il fatto, poi, che alcuni

[19] Il tasso di risposta (*response rate*) è definito da Talmage e altri (1988) come percentuale del numero di questionari adeguatamente compilati ottenuti nell'indagine sul numero degli individui che hanno i requisiti necessari per parteciparvi.

La velocità di risposta (*response speed*) è definita da Tse (1994) come l'intervallo di tempo intercorrente tra il momento della spedizione (o della ri-spedizione) di un questionario da compilare e il momento del suo rientro.

questionari vengano rispediti per posta, dimostra che alcuni rispondenti sono più esperti di altri nell'utilizzo del computer. [20]

f. Talvolta un metodo utilizzato per ampliare liste di indirizzi *e-mail* o liste di contatti si basa sul criterio adottato per il metodo di campionamento *snowballing*: le persone che già fanno parte della lista ne segnalano altre che possono essere interessate ad esservi inserite. Questo metodo trova ambito di applicazione principalmente nel caso si abbia a che fare con popolazioni piuttosto limitate numericamente, i cui membri sono difficili da contattare ma spesso si tengono in contatto tra di loro. È evidente il rischio della procedura di raccolta *snowballing*, simile alla auto-selezione del campione: sono gli appartenenti al campione stesso che giudicano se un altro soggetto è idoneo o meno a partecipare all'indagine che si sta svolgendo. L'influenza del ricercatore è ridotta ai minimi termini: al più può dare indicazioni su quali siano i soggetti che sarebbe ideale coinvolgere nella ricerca.

Da quanto detto, risulta evidente che, nei limiti del possibile, è opportuno in ogni caso passare al vaglio l'elenco di indirizzi che si ha a disposizione per cercare di eliminare le eventuali duplicazioni e per effettuare una attenta e oculata selezione degli elementi che si vogliono inserire nel campione.

7.1.4 Problemi aperti

Sia nel caso si utilizzi un sistema casuale o sistematico di selezione del campione (indagini *Pop-up*), sia si utilizzi un apposito *software* che intercetta la navigazione dei potenziali rispondenti (*Browsing Internet Software*), bisogna tenere presente che sussistono ancora molti problemi relativi allo studio delle caratteristiche dei visitatori di un sito web e alle informazioni disponibili su di essi. Si possono ottenere, in primo luogo, dati che cercano di misurare l'aspetto prettamente quantitativo delle presenze: si tratta, però, di informazioni eccessivamente rudimentali che non consentono uno studio approfondito del problema.[21] Inoltre i dati riguardanti i visitatori di un sito web tendono a sovrastimare o sottostimare il traffico in modo significativo. Questo grazie anche alla presenza di siti specchio, alla possibilità per il navigatore di "nascondere" la propria presenza, o all'utilizzo di particolari *frames* o congegni meccanici che falsano le statistiche sul numero di accessi ad un sito. Alcuni di questi problemi sono oggetto di dibattito in Smith (1998) e Foan & Read (1999).

---

[20] Uno studio sperimentale di Bradley (1999) ha dimostrato che l'utilizzo di gruppi di discussione è una buona strategia da utilizzare per realizzare indagini via web. Bradley ha rilevato come il tasso di risposta vari, a seconda della lista considerata, tra il 25 e il 54%, comunque superiore al livello generico ottenibile con indagini postali, pari a circa il 20% (Bourque e altri,1995). Anche il tempo di risposta rilevato varia tra le 34 e le 51 ore, a seconda della lista, ma è comunque inferiore ai tempi di risposta ottenuti per i questionari rientrati via posta (262 ore, pari a circa 10 giorni). Lo studio di Bradley si è concentrato su gruppi di discussione sulla grafologia. Questi gruppi sono risultati essere composti da una tipologia di soggetti in parte differente da quella composta da chi usualmente utilizza Internet: l'età media, ad esempio, è risultata attorno ai 52 anni, rispetto ai 39 rilevati da una ricerca di Forrester (Kumar e altri, 1999) sugli utilizzatori della rete. Questo fa dedurre che il profilo dei rispondenti, se come liste di campionamento vengono usate quelle contenenti gli iscritti ai gruppi Internet, varia notevolmente in relazione all'argomento cui il gruppo è collegato.

[21] Alcuni criteri di classificazione dei navigatori verranno visti nel Cap. 9.

Come si può notare, nella maggioranza dei casi visti il reperimento di liste di individui cui proporre l'indagine è quasi sempre basata sull'ottenimento di indirizzi di posta elettronica validi di cui ci si può servire per ricontattare i potenziali rispondenti. L'utilizzo degli indirizzi *e-mail* richiede, però, estrema cautela. Una parte degli utenti, infatti, cambia molto spesso il proprio *provider* e, contestualmente, decide di cambiare anche il recapito di posta elettronica, o di aprirne uno nuovo. Ciò può portare a conseguenze significative su una selezione del campione basata su indirizzari *e-mail*. Da una parte si possono avere in lista numerosi recapiti non più utilizzati, dall'altra, come già accennato, si rischia che più indirizzi facciano riferimento ad uno stesso soggetto. Il campione potrebbe quindi risultare, in definitiva, scarsamente rappresentativo; oppure potrebbe portare a sovrastimare la presenza di determinate categorie di soggetti. Sarà, ad esempio, più probabile selezionare un rispondente particolarmente sensibile ai prezzi praticati dai *provider* (e ai prezzi, in generale), più predisposto ad aprire nuovi *account* di posta elettronica o a cambiare di frequente il gestore della propria connessione piuttosto che, ad esempio, soggetti che vivono in condizioni economicamente più agiate e quindi sono meno propensi al cambiamento a causa di ragioni economiche.

A parte il discorso sulla mobilità e sulla volatilità degli indirizzi *e-mail*, rimane sempre aperto, poi, il problema di indirizzi non più usati (l'utente non ricorda più la *password*, ha cambiato *account*, si è rivolto a *provider* che offrono servizi migliori, ...), non usati per lungo tempo, controllati a scadenze particolarmente dilazionate, pieni (la casella di posta elettronica non viene svuotata regolarmente e non può ricevere ulteriori messaggi).

Tutti questi aspetti sottolineano l'importanza di una "manutenzione" regolare e di un controllo scrupoloso e assiduo degli indirizzari *e-mail* utilizzati per il contatto del campione.

7.2 Campionamento probabilistico e non-probabilistico

I differenti sistemi di costruzione delle *sampling frame* visti nel capitolo precedente possono essere all'origine di liste utilizzate sia per il campionamento probabilistico che per il campionamento non probabilistico.

A seconda del modo in cui sono reclutati i rispondenti delle indagini via Web, e quindi in base anche al tipo di *sampling frame* utilizzato, è possibile (o comunque conveniente) effettuare la scelta tra campionamento probabilistico piuttosto che non probabilistico, applicando le tecniche introdotte a livello più generale nel Cap. 6.

Una classificazione dei metodi di campionamento impostata da Couper (2000) mette in evidenza la differenza tra situazioni in cui è possibile utilizzare metodi probabilistici e situazioni in cui, invece, è opportuno utilizzare tecniche di tipo non probabilistico per il reclutamento del campione. Ci si può ricondurre ai principi del campionamento probabilistico nelle seguenti quattro situazioni: quando si ha a che fare con indagini basate sull'intercettazione della navigazione, su liste di indirizzi *e-mail*, su *panel* reclutati in precedenza o su metodi misti. Nel seguito si vedrà ciascuno di questi casi nel dettaglio.

1. **Indagini basate sull'intercettazione dei visitatori di particolari siti Web**: questa categoria comprende i campioni selezionati in indagini basate su sistemi *Pop-up* o *Hijack*. La "virtuale" lista di campionamento comprende tutti i visitatori di un sito o di una pagina Web, tra i quali, casualmente, viene selezionato il campione cui verrà

proposto di partecipare all'indagine. Il meccanismo di selezione del campione può variare da caso a caso, come visto in precedenza.

2. **Liste di *e-mail***: sono liste di vario tipo (come descritto al paragrafo precedente). La qualità dell'indagine statistica che è possibile condurre dipende in larga misura dalla copertura della lista. La situazione ideale, per effettuare un'indagine di tipo probabilistico, è quella in cui si ha una popolazione in cui la possibilità di connettersi a Internet è diffusa universalmente e in cui si può utilizzare una lista di indirizzi completa e senza duplicazioni. Nelle popolazioni demografiche e di imprese il grado di copertura della lista è un problema notevole: è molto difficile identificare liste che forniscano un *sampling frame* che rappresenti una buona approssimazione della popolazione. Il problema è particolarmente critico e deve essere seriamente affrontato di volta in volta. Meno rilevante è, invece, la questione della copertura nel caso di popolazioni piccole e chiuse (ad esempio, se si considerano gli studenti universitari di una determinata facoltà o gli impiegati di una grande azienda). In queste situazioni si ottengono campioni piuttosto ben mirati in quanto di solito vengono registrate, negli stessi archivi, anche parecchie altre informazioni aggiuntive sugli individui registrati.
3. ***Panel* pre-reclutati**: gli individui vengono contattati per essere inseriti in *panel*; la qualità dell'indagine dipende, quindi, dai criteri di selezione dei rispondenti e dal *frame* utilizzato. A titolo esemplificativo, il campione probabilistico (ad esempio selezionato telefonicamente con la tecnica del *RDD*) se adeguato alla popolazione obiettivo consente di selezionare direttamente le unità che saranno oggetto di studio.[22] Un problema significativo, come già visto, è dovuto alla reperibilità dei partecipanti. È inoltre improbabile che il *panel* mantenga nel tempo una buona qualità: va quindi modificato e aggiornato costantemente, col passare del tempo, e ciò comporta ingenti costi di mantenimento.
4. **Indagini basate su metodi misti**: prevedono l'utilizzo integrato di tecniche differenti tra quelle sopra elencate: l'indagine via web è solo uno dei modi di partecipare all'indagine, alternativo alla indagine telefonica o ad altre tecniche.

I quattro metodi di selezione visti sopra hanno in comune il fatto di mettere a disposizione una lista di indirizzi o un sistema di contatto con i potenziali rispondenti (come l'accesso ad un sito, le visite ad una determinata pagina) che consente una loro selezione casuale. Ad esempio, nel caso di liste di *e-mail* si potranno estrarre a caso numeri corrispondenti a ciascun indirizzo, nel caso di visitatori di un sito o di una pagina web si potrà effettuare una selezione del campione che si basi su un particolare meccanismo, nel caso di *panel*, tra gli individui che ne fanno parte o tra un loro particolare sottogruppo si potranno scegliere casualmente le persone da coinvolgere nell'indagine.

Va però sottolineato che nelle indagini via web, la selezione di un campione casuale tramite tecniche di campionamento probabilistiche è piuttosto difficoltosa e rischia di condurre a campioni non rappresentativi. In genere è possibile scegliere tra due differenti soluzioni. Per superare il problema della non rappresentatività è opportuno

---

[22] Knowledge Networks (http://knowledgenetwords.com) ha superato il problema della disponibilità di un accesso ad Internet offrendo ad un campione casuale di rispondenti l'accesso alla rete in cambio della partecipazione a indagini future.

restringere la popolazione obiettivo a coloro che utilizzano Internet e, una volta redatta una lista di campionamento, scegliere un metodo di scelta casuale delle unità che faranno parte del campione. Un'altra soluzione possibile è utilizzare un sistema di contatto dei rispondenti alternativo (ad esempio telefonico) che consenta di raggiungere un più diffuso insieme di individui; costoro vengono successivamente inseriti in un *panel* o in un gruppo che potrà essere coinvolto in eventuali indagini.

Nelle indagini telefoniche il sistema di *Random Digit Dialing* (*RDD*) facilita il compito al ricercatore: si presume che ogni famiglia possieda un telefono e, quindi, che la probabilità che un individuo venga selezionato sia uguale a quella degli altri e diversa da zero; in questo modo si ottiene un campione considerabile probabilistico. Ad esempio, negli Stati Uniti il 95% delle famiglie ha un apparecchio telefonico, quindi la copertura della rete telefonica può essere considerata totale.

La stessa situazione non si riscontra, però, per il campionamento per indagini via web. Considerando sempre il caso degli Stati Uniti, infatti, la percentuale di famiglie con connessione a Internet, una delle più elevate al mondo, è comunque inferiore al 50% (*U.S. Department of Commerce*, 2000). Quindi riuscire a selezionare un campione casuale rappresentativo anche con una tecnica di selezione casuale degli indirizzi *e-mail* analoga alla *RDD* risulta molto difficoltoso; tra l'altro, generare indirizzi di posta elettronica casuali non è facile come generare numeri di telefono. Da un lato, infatti, molti degli indirizzi generati potrebbero risultare inesistenti, non validi o non utilizzati, dall'altro le combinazioni che possono portare a indirizzi validi possono essere molto più articolate e complicate di quelle che portano alla combinazione di 7 numeri (come ad esempio succede per i numeri telefonici degli Stati Uniti). Per di più, anche nel caso in cui fosse possibile generare indirizzi di posta elettronica casuali, fare *spamming* (cioè inviare *e-mail* non richieste a lunghi elenchi di indirizzi) è vietato e molti server di posta elettronica provvedono a cancellare automaticamente le *e-mail* identificate come *spam*.

Dunque, risulta attualmente molto difficoltoso reclutare campioni casuali da una popolazione vasta e di cui, comunque, non esiste una lista completa di contatti; il rischio di ottenere campioni non rappresentativi è alto. Finché la diffusione di Internet non sarà universale (almeno nell'ambito della popolazione obiettivo considerata dalla ricerca) e finché non verranno stilati elenchi di indirizzi *e-mail* completi, risulterà piuttosto difficile ottenere campioni probabilistici rappresentativi usando esclusivamente sistemi di contatto via web.

Per ovviare a questo problema, alcuni ricercatori hanno sviluppato *indagini multi-modali*: i rispondenti vengono, ad esempio, contattati telefonicamente e in questo modo vengono persuasi a partecipare ad indagini via web. Per risolvere il potenziale problema di campioni non rappresentativi dovuti alla mancanza di attrezzature in alcuni casi è stato messo a disposizione del gruppo di rispondenti selezionato un accesso ad Internet chiedendo come contropartita la partecipazione a future indagini via Web.

Altra alternativa presa in considerazione è integrare i dati delle indagini via web con quelli ricavati da indagini telefoniche. Nonostante i costi di realizzazione aumentino notevolmente, tali rimedi hanno comunque permesso ai ricercatori di ottenere risultati

significativi, anche grazie alla possibilità di utilizzare appropriati sistemi di ponderazione [23].

Nonostante vi sia la possibilità di ricorrere a questi accorgimenti, nelle indagini via web viene utilizzo prevalentemente un approccio non probabilistico. Questo avviene per il fatto che è quasi sempre impossibile sia identificare i membri della popolazione obiettivo che contattare un campione probabilistico della popolazione.

Da ciò consegue che i risultati campionari non possano essere estesi, tramite l'utilizzo di tecniche statistiche, ad una popolazione più ampia: la fase di inferenza risulta quindi piuttosto difficoltosa.

Una classificazione delle metodologie di campionamento non probabilistiche (che comprende diversi metodi di coinvolgimento dei potenziali rispondenti tra quelli visti al paragrafo precedente) è stata elaborata da Couper (2000). La classificazione prevede tre distinte tipologie: indagini per intrattenimento, campione auto-selezionato, *panel* di volontari. Nel seguito si discuterà brevemente di ciascuna di esse.

1. **Indagini web finalizzate all'intrattenimento (*Entertainment surveys*)**: non hanno finalità scientifiche vere e proprie; in genere sono proposte ai visitatori di un sito che scelgono autonomamente se parteciparvi o meno. La finalità di questo tipo di indagini o sondaggi (molto diffusi in rete) è di solito quella di rendere più accattivanti, attraenti ed interattivi i contenuti di un sito web. Si possono citare esempi di siti web che chiedono di dare un voto agli ultimi film usciti, di segnalare la propria canzone o il proprio cantante preferito, il personaggio pubblico più amato, e così via.
2. **Campione auto-selezionato senza restrizioni**: si usano sistemi di invito pubblicati sulle pagine di un sito web (annunci, *invitation tags*, *banner*, *link* ipertestuali) per coinvolgere i rispondenti. Possono entrare a far parte del campione solamente i visitatori di uno specifico sito che esprimono in modo attivo la volontà di partecipare all'indagine.
3. ***Panel* di volontari**: come avviene per gli *opinion center*, si tratta di gruppi di individui che si propongono volontariamente per essere coinvolti in indagini di vario tipo[24]. I potenziali panelisti vengono invitati ad accedere ad una predefinita pagina web dove vengono loro richieste alcune informazioni fondamentali che li riguardano (tra cui l'indirizzo *e-mail*). Tutte le informazioni pervenute vengono immagazzinate in un *database* che consentirà, poi, l'estrazione di un opportuno campione sul quale effettuare la ricerca. Pur essendo, questo, un sistema di selezione di tipo non-probabilistico, è più probabile che conduca alla selezione di un campione rappresentativo rispetto ad altri metodi di selezione non probabilistici.

---

[23] Un approfondimento di questo argomento è disponibile sul Quaderno di Ricerca "Inferenza nelle Indagini via Web" (Biffignandi, Toninelli, 2006); si veda Bibliografia.

[24] Due esempi molto conosciuti di *panel* di volontari sono quelli della *Harris Interactive* (http://vr.harrispollonline.com/register/main.asp) e della *Greenfield Online* (http://greenfieldonline.com).

8 Tipologie di Indagini via Web

Una volta che è stata decisa la modalità di reclutamento del campione e una volta che è stata pianificata la metodologia di campionamento ritenuta più opportuna, è importante riflettere su come impostare l'indagine dal punto di vista della libertà di accesso consentita ai potenziali rispondenti.

Queste problematiche sono direttamente collegate ad un altro aspetto di rilievo, ovvero alle modalità di somministrazione del questionario. Si ritiene però utile accennare ai differenti approcci possibili, seppure in maniera sintetica, sia per una maggiore completezza della trattazione dell'argomento, sia per il fatto che, in base al tipo di metodo che si deciderà di usare, sarà opportuno procedere nella scelta tra uno dei metodi di reclutamento del campione indicati nel Cap. 6 e nel Cap. 7.

La caratteristica principale di tutte le indagine via web è che il questionario viene auto-compilato dal rispondente. Per la quasi totalità dei casi si tratta, quindi, di indagini *CASI* (*Computer Assisted Self-Completion Interviews*), cioè indagini in cui il soggetto risponde, direttamente su un terminale, alle domande che gli vengono poste nel questionario. Un vantaggio di questa metodologia di raccolta dati è che le informazioni raccolte risultano immediatamente inserite in un apposito *database*, pronte per essere "ripulite" ed elaborate. In altri casi, molto meno frequenti, i questionari possono essere stampati dal rispondente che provvede poi a compilarli e ad inviarli per posta, tramite fax o via posta elettronica.

Nonostante possano riguardare argomenti disparati e possano presentarsi in varie forme e formati, è possibile classificare quasi tutte le indagini via web, a seconda del formato in cui sono presentate, in due categorie generali: le interattive e le passive.

- **Indagini interattive (*interactive surveys*)**. Le indagini interattive sono somministrate pagina per pagina: ogni schermata propone un quesito cui il partecipante all'indagine deve rispondere. I vantaggi di questa modalità di somministrazione sono diversi: in primo luogo, anche nel caso in cui il rispondente non arrivi a completare tutto il questionario, le risposte già date non vanno perse; in secondo luogo è possibile impostare una serie di salti che consentono al rispondente di rispondere solo alle domande che gli competono. In una indagine sulle abitudini dei fumatori se, ad esempio, il soggetto dichiara di non essere un fumatore, il sistema automaticamente *by*-passerà tutte le domande rivolte espressamente ai fumatori. Questa impostazione contribuisce ad abbreviare il questionario ed i tempi di partecipazione all'indagine. L'impostazione interattiva, d'altro canto, comporta alcuni problemi. Il primo di questi è che in genere non è consentito al rispondente tornare su risposte già date o correggerle. Inoltre non si può avere sotto controllo la reale estensione del questionario e non ci si rende conto del punto di esso a cui ci si trova. Per ovviare a questo inconveniente è opportuno predisporre una barra di scorrimento che metta in evidenza lo stato di avanzamento. Altro inconveniente delle indagini interattive è che in genere richiedono il supporto di *software* specifici che possono trovare difficoltà di

funzionamento su terminali vecchi o poco potenti o su *browser* di vecchia generazione.

- **Indagini passive (*passive surveys*)**: il questionario viene presentato al rispondente tutto di seguito (su un'unica pagina) e di solito è prevista la presenza di una barra di scorrimento che permette di tornare alle risposte date in precedenza ed, eventualmente, correggerle. I dati compilati vengono trasmessi solo una volta che il questionario è completato e il rispondente seleziona il tasto predisposto per l'invio. Sia le procedure di compilazione che quelle di invio sono piuttosto semplici: si evita, così, di incorrere in grossi problemi tecnici o legati all'incompatibilità dei *software* disponibili per il rispondente.

Nel caso di indagini interattive il posizionamento del questionario, in genere, avviene su una pagina web, mentre per indagini passive questo può essere anche contenuto in un messaggio di posta elettronica, oppure può essere strutturato come combinazione delle due cose.

Le considerazioni che porteranno alla scelta della tipologia di indagine che si vuole impostare vanno viste in stretta connessione con le liste di campionamento disponibili e il tipo di popolazione obiettivo che andrà studiata. Ad esempio, nel caso in cui si abbia a che fare con un campione poco avvezzo all'utilizzo di Internet e con a disposizione attrezzature che possono non essere adeguate, si preferirà utilizzare un'indagine di tipo passivo. Invece, volendo effettuare una indagine sui dipendenti di una grande azienda collegati alla rete Intranet e sapendo di avere infrastrutture *software* e *hardware* di un certo livello che accomunano tutti i potenziali rispondenti, non si avranno limitazioni di ostacolo alle possibilità offerte da indagini di tipo interattivo.

8.1 Modalità di accesso all'Indagine

Una volta scelta la tipologia di questionario da proporre, in base a considerazioni sul tipo di campione, sulla popolazione obiettivo, sui tempi, sui costi e sui supporti tecnici disponibili, è anche necessario decidere che modalità di accesso prevedere per il questionario pubblicato *on-line*. Gli approcci che si vedranno tra poco possono essere utilizzati sia per indagini via web che per ricerche svolte su *network* di computer quali le reti *Intranet*, *Extranet* e così via; la scelta del modo di condurre l'indagine, inoltre, va effettuata con riferimento ad un ben specifico metodo di campionamento.

È possibile classificare i questionari posti su pagina web in tre differenti tipologie, a seconda del tipo di accessibilità consentita ai navigatori:

- ***Open web***: non c'è alcun controllo all'accesso dei visitatori, che vi sono indirizzati tramite *banner*, avvisi, inviti a partecipare pubblicati sulle pagine di un sito o inviati via *e-mail*, e così via. Il tipo di indagine è, quindi, spesso associato a metodologie di campionamento probabilistico (secondo la classificazione di Couper vista al Par. 7.2) in cui si intercettano casualmente i partecipanti all'indagine. Viste le modalità di selezione, il campione ottenuto potrebbe essere anche significativamente distorto rispetto alla popolazione obiettivo. Inoltre, la scarsità dei dati su coloro che rispondono e la tendenza a fornire dati falsi sulla propria identità impediscono quasi ogni tentativo di ri-

ponderare i risultati e di capire a che tipologia specifica di popolazione essi debbano essere riferiti.

- ***Closed web***: rispetto al caso precedente, le tipologie di invito a partecipare alla ricerca possono essere le stesse, ma l'accesso alla pagina contenente il questionario è consentito solo previo inserimento di una *password*. In questo secondo approccio, il campione può essere selezionato in modo diretto: si può chiedere ai potenziali rispondenti, in cambio della *password* per accedere al questionario, di fornire alcune informazioni che potranno poi rivelarsi utili ai fini della fase di campionamento e l'elaborazione dei dati. Vi è il rischio, però, che parecchi dei dati immagazzinati siano falsi.
- ***Hidden web***: il questionario appare solo a certi visitatori che navigano su un particolare sito Internet. I meccanismi di selezione possono essere casuali o sistematici. Il questionario, ad esempio, può essere proposto solo ai visitatori di una specifica data (o ora), a quelli che dimostrano il proprio interesse per una determinata pagina di un sito web (e quindi che sono interessati ad uno specifico argomento), a cadenza periodica (ogni *n* visitatori o ogni *n* minuti) e così via. Questo tipo di indagine può essere associata sia a metodi di selezione casuale (selezione sistematica o casuale dei rispondenti) che a metodi di selezione più mirata del campione (il questionario è proposto a chi chiede di visitare una certa pagina). Il numero di informazioni aggiuntive che possono essere richieste agli individui selezionati è, in genere, piuttosto basso; ciò non favorisce la possibilità di uno studio accurato delle caratteristiche del campione.

Si è detto che i questionari, oltre che su pagine web, possono essere diffusi anche via posta elettronica. In questo secondo caso, si possono definire tre ulteriori tipologie di indagine:

- ***Simple e-mail***: si invia ai rispondenti un semplice messaggio di posta elettronica contenente le domande a cui rispondere. Rispetto ai questionari pubblicati su siti web, la metodologia di selezione del campione è diretta: si possono scegliere a priori gli individui da contattare e, se si hanno informazioni su di essi, si possono scegliere coloro che meglio si adattano alla ricerca che si sta svolgendo o, comunque, quelli che consentono di ottenere un campione rappresentativo. È chiaro però che, per realizzare una indagine di questo tipo, come requisito minimo serve una lista di indirizzi *e-mail* accuratamente compilata.
- ***E-mail attachment***: ai potenziali rispondenti viene inviata una *e-mail* "di accompagnamento" che serve a presentare l'indagine e a sollecitare la partecipazione ad essa. In allegato alla *e-mail* c'è il questionario che deve essere compilato (in formato elettronico o facendone una stampa) e poi essere inviato al ricercatore. La restituzione può avvenire per posta o fax (se il questionario è stato stampato) oppure per posta elettronica (come allegato ad una *e-mail* di risposta). Anche in questo caso la selezione del campione è diretta ed attiva: il ricercatore può scegliere, in base alle informazioni che ha (se ne ha di disponibili) quali individui contattare.

- ***E-mail URL embedded***: agli indirizzi selezionati (direttamente, come succede nei due casi precedenti) viene inviata una *e-mail* contenente un invito alla partecipazione e un *link* o un indirizzo *URL* su cui cliccare per accedere direttamente alla pagina web su cui è pubblicato il questionario. Selezionando il *link*, viene attivato il *browser* e il rispondente compila la pagina web contenente il questionario. Questo, in tale situazione, è definito *web-based*: può essere disponibile ad ogni accesso, oppure il rispondente, per accedervi, può dover specificare una password (questo per evitare eventuali intrusioni di rispondenti estranei al campione selezionato).

Per approfondire vantaggi e svantaggi di ciascuno di questi approcci si può fare riferimento alle pubblicazioni di Witt e altri (1998) e Frost (1998).

Altra classificazione, più generale, che utilizza come criteri le modalità di accesso all'indagine è quella elaborata da Watt (1997), che prevede tre distinte categorie di campioni e che può essere associata alle modalità di accesso ai questionari prese in considerazione sopra:

- **Senza restrizioni (*Unrestricted*)**: l'indagine è aperta a tutti (corrisponde alla categoria *Open Web* vista sopra). Ha un problema: utilizza campioni di rispondenti che soffrono di scarsa rappresentatività.
- **Setacciato (*Screened*)**: tra tutti coloro che si propongono (o che si dichiarano disponibili) vengono selezionati solo individui con determinati requisiti o appartenenti a particolari categorie. In questo caso il campione può risultare più rappresentativo di quello selezionato senza restrizioni.
- **Reclutato (*Recruited*)**: si tratta di un campione costruito più "attivamente", cioè andando alla ricerca di quegli individui che solamente sono di interesse per portare a termine l'indagine. L'adesione alla ricerca non è volontaria, ma è proposta, stimolata ed incoraggiata da chi deve attuare l'indagine. Il campione "reclutato" è simile a quello che si ricava in indagini condotte con metodi *panel* (ovvero i metodi utilizzati con maggiore successo nelle indagini degli ultimi anni).

9 Rispondenti nelle Indagini via Web

Analogamente a quanto avviene nell'ambito delle indagini di mercato, dove è importante riuscire a costruire profili dei consumatori, anche nelle indagini via web è possibile considerare differenti criteri di classificazione degli utilizzatori di Internet. Questo tipo di studi è indubbiamente molto importante, al fine della predisposizione di opportune tecniche di campionamento, in quanto consente di delineare meglio le caratteristiche dei potenziali rispondenti cui è più opportuno rivolgersi.

9.1 Rischi e proposte di soluzioni

McDonald (1999) ha pensato ad un criterio di classificazione dei navigatori che riprende il modello classico utilizzato, nelle indagini di mercato, per spiegare come i consumatori reagiscono alla diffusione delle innovazioni. Si può così distinguere tra diversi tipi di utilizzatori di Internet: *innovatori*, *opinion leader*, *precoci*, coloro che fanno parte della *iniziale* o della *tarda maggioranza* e *ritardatari*.

Per dare un'idea di come le diverse categorie siano presenti in modo differente in un ipotetico campione casuale di navigatori, basta sottolineare che nel contesto delle indagini via web la categoria "ritardatari" rischia di essere sottostimata, in quanto proporzionalmente è poco rappresentata. La stessa tipologia di rispondenti, d'altra parte, è sovrastimata nelle indagini di mercato basate su interviste personali. Tenuto conto di questa differenza sostanziale, nella realizzazione di una indagine che comprenda questo tipo di rispondenti il problema può essere risolto utilizzando in modo combinato sia indagini via web che indagini basate su interviste *face-to-face*.

Oltre al rischio di sottostimare la presenza di alcune tipologie di rispondenti, si può anche incorrere nel rischio opposto, ovvero sovrastimare alcuni gruppi particolari. Per questo motivo Farmer (1988) ha introdotto la tecnica denominata *sifting* (ovvero "cernita"). Nel reclutare il campione per una indagine via web, in un primo tempo si accettano in modo indifferenziato le adesioni di partecipazione di tutti coloro che si propongono spontaneamente o volontariamente. In una seconda fase vengono scartati gli individui che possono essere definiti "ineleggibili" in base ad una serie di criteri prefissati. Il metodo di *sifting* proposto da Farmer può rivelarsi molto utile per aggiustare la composizione complessiva del campione finale e può consentire di superare alcuni problemi legati alla auto-selezione dei rispondenti. Infatti la scelta finale dei componenti del campione avviene in modo indipendente dal fatto che sia stato il rispondente a proporsi per l'indagine. Il *sifting*, però, non consente di contattare chi ha scelto di non partecipare all'indagine: questi individui vengono completamente persi, nonostante uno studio del loro gruppo possa essere determinante ai fini della ricerca.

Utilizzando la tecnica di *sifting* si corre un altro rischio, collegato alle tecniche di campionamento, cioè il pericolo di non ottenere un campione realmente probabilistico (in cui, cioè, gli individui siano scelti casualmente). Tenendo conto di questo problema alcuni studiosi hanno proposto l'utilizzo, applicata agli indirizzi *e-mail*, di una tecnica simile al *RDD* (*Random Digit Dialling*), la quale in genere viene applicata alle indagini telefoniche per la selezione casuale di numeri di telefono da contattare. A questo scopo

in via preliminare è necessario predisporre una piattaforma per la raccolta di indirizzi *e-mail*, prima del lavoro sul campo. In una fase successiva, tramite un sistema di selezione casuale di indirizzi *e-mail* si possono individuare quelli che vengono utilizzati abitualmente. Se, poi, agli indirizzi possono essere associati dei valori numerici, è possibile utilizzare un sistema simile all'*RDD* per la scelta casuale di indirizzi da contattare per proporre la partecipazione all'indagine.

Nonostante si possano adottare numerosi accorgimenti, spesso non si riesce a sopperire ai difetti di affidabilità delle liste di campionamento e si rischia, così, di incorrere in errori di copertura, nella sottostima di alcune particolari categorie di rispondenti o nell'utilizzo di campioni distorti (del problema ci si occuperà diffusamente nel Quaderno di Ricerca "Inferenza nelle Indagini via Web" [25]).

In alcuni casi (e in genere per particolari tipologie di individui) è opportuno, per questi motivi, ricorrere alla tecnica *Saturation Surveying*, introdotta da Turner (1989). Questa metodologia prevede che si tenti di osservare tutti i possibili *target* identificabili nella popolazione obiettivo. D'altra parte i costi delle indagine via web, a differenza di altri tipi di indagini più tradizionali, non aumentano considerevolmente incrementando in modo cospicuo il numero di persone da contattare e questa strada può essere agevolmente percorsa per ottenere questi risultati.

9.2 I "Navigatori" di Internet

Il "popolo dei navigatori" è connotato da caratteristiche socio-demografiche ed economiche piuttosto particolari.

A sottolineare questa distinzione, McDonald (1999) sostiene che in genere un campione di persone che utilizzano Internet è rappresentativo solo di utilizzatori di Internet. Andando più nei dettagli, Coomber (1997) ha sottolineato che la distorsione che si ha nelle indagini condotte via web è dovuta alle caratteristiche particolari dei navigatori, che si differenziano in modo netto rispetto ad una popolazione più generica (ad esempio quella di uno Stato).

Chi utilizza il web ha, in genere, un particolare *background* socio-demografico: appartiene prevalentemente al genere maschile, è di razza bianca, relativamente ricco e ben educato, ha un livello di istruzione superiore e può fruire di risorse particolari rispetto ad una popolazione più generica (Nielsen & CommercNet, 1995; Kehoe & Pitkow, 1996). Risulta pressoché impossibile, invece, contattare fasce di popolazione ben definite comprendenti, ad esempio, persone anziane, con reddito basso, con livello basso di istruzione, e così via.

Sempre Coomber ha però sottolineato che, grazie al diffondersi sempre più capillare (territorialmente e in tutte le fasce della popolazione) delle nuove tecnologie e grazie alle nuove opportunità offerte da connessioni a Internet sempre più veloci, questa distorsione riguardante il campionamento per indagini via web è destinata ad assottigliarsi progressivamente; in un prossimo futuro, quindi, i problemi saranno molto simili a quelli riscontrati per il campionamento in indagini tradizionali.

---

[25] Biffignandi, Toninelli (2006); si veda Bibliografia.

È ovvio che in qualsiasi ricerca che si avvalga del supporto di Internet è opportuno tenere conto delle peculiarità che connotano i navigatori. Altrimenti, il rischio che si corre è quello di ottenere risultati distorti, basati su una fascia specifica di popolazione che può differenziarsi in modo rilevante dalla popolazione obiettivo.

Questo limite delle indagini condotte via web, mentre esclude un loro possibile utilizzo in determinate circostanze, le rende però particolarmente adatte ad essere applicate ad altri casi specifici. La rappresentatività e la distorsione di una indagine via Internet è legata, in particolare, ai dati raccolti.

Andando maggiormente nei dettagli, il campionamento per indagini via web non incorre in grossi rischi di distorsione nel caso in cui:

- Parte della popolazione obiettivo include i navigatori (o gli utilizzatori di tecnologie avanzate di comunicazione) o si tratta di indagini riguardanti espressamente il web e chi ne fa uso.
- La popolazione obiettivo ha caratteristiche in comune con i navigatori (individui giovani, prevalentemente di sesso maschile, percettori di redditi piuttosto alti, con un grado di istruzione elevato, ...). Eventualmente i risultati, avendo a disposizione le opportune informazioni sulla composizione del campione, possono poi essere riponderati per arrivare a conclusioni più generali della ricerca.
- Vi è una certezza quasi assoluta che le risposte date dai navigatori non differiscono sostanzialmente da quelle che darebbero coloro che non hanno la possibilità di accedere ad Internet.

Per gli stessi motivi visti, è importante saper capire le differenze tra coloro che partecipano e non partecipano ad una indagine: solo così si comprenderà se va data maggiore o minore importanza al problema della distorsione del campione. Per la descrizione di coloro che non utilizzano frequentemente la rete, si rimanda a Coffey & Johnson (1998).

Nonostante il potenziale problema di distorsioni, va comunque sottolineato come in alcuni casi anche un campione "sbagliato" è in grado di fornire le risposte giuste ai quesiti che la ricerca si pone. In particolare, un campione sbagliato, utilizzato per sondare ad esempio una certa opinione, tende a fornire le risposte "giuste" se la distorsione nel campione è indipendente dalla visione che si ha dell'argomento oggetto di studio. Ad esempio, se si fa una indagine per conoscere la proporzione di persone mancine di una determinata popolazione, probabilmente una ricerca effettuata *on-line* (e quindi su un campione potenzialmente distorto) darà lo stesso risultato di un'altra condotta con le tradizionali metodologie di indagine *off-line*. Questo perché il fatto che una persona sia mancina o meno non è legato alla possibilità di avere a disposizione una connessione alla rete. Si può ragionevolmente ritenere che ciò succeda ogni volta che per oggetto di indagine si considerano variabili indipendenti dall'effetto-campione, cioè dalle distorsioni che una particolare composizione del campione può originare.

È bene, quindi, fare molta attenzione alla presenza di possibili correlazioni tra la caratteristica misurata dall'indagine e il fatto che questa venga misurata su un campione composto esclusivamente da abituali frequentatori della rete.

Vi sono casi, poi, in cui un campione non corretto fornisce risposte comunque non rispondenti alla realtà. Ciò accede se sono verificate le due seguenti condizioni:

- L'opinione riguardo ad un certo argomento differisce in un modo che rispecchia le differenti categorie e il loro peso relativo nell'accesso alla rete. Ciò accade, ad esempio, se l'opinione è distinta per età o classe sociale: i risultati saranno distorti. Può essere utile, in questi casi, controllare le varianze del campione con il sistema delle quote.
- L'opinione su un determinato argomento è differente tra persone che hanno e non hanno la possibilità di accedere alla rete Internet. Si consideri, ad esempio, una indagine che vuole valutare la reperibilità di *Compact Disc*: chi ha accesso alla rete avrà modo di procurarsi i *CD* molto più facilmente rispetto a chi si deve rivolgere ai tradizionali negozi: potrà, ad esempio, ordinarli su appositi siti web che si occupano di vendite on-line trovando in genere una maggiore gamma di scelta. In questa situazione con il metodo delle quote il problema non si allevia.

Le prime importanti considerazioni che vanno fatte, dunque, riguardano da un lato la composizione particolare del campione selezionabile con i metodi visti nei capitoli precedenti, dall'altra l'argomento oggetto di indagine. Valutando le connessioni che si possono avere tra i due aspetti, sarà possibile concludere se si ha a disposizione un campione in grado di dare risultati corretti (anche nel caso non sia rappresentativo della popolazione oggetto di indagine) o meno.

9.2.1 Livello della tecnologia

Un altro degli aspetti che vanno attentamente considerati per comprendere la tipologia di potenziali rispondenti con cui si ha a che fare nell'ambito del campionamento per indagini via web è stato considerato da una classificazione elaborata da Bradley (1997).

La classificazione tiene conto di un duplice punto di vista: da una parte le capacità possedute dall'utente, dall'altra le opportunità offerte dalla attrezzatura a sua disposizione. L'importanza di queste informazioni è evidente se si considera che sono rilevanti per approntare un opportuno piano di indagine: decisioni riguardanti il metodo di campionamento, i sistemi utilizzati per la diffusione dei questionari, e così via devono essere infatti prese in base allo scopo che ci si propone, a quelle che sono le proprie risorse, ma anche in relazione ai potenziali rispondenti che si vogliono coinvolgere. Il rischio che si corre è di non riuscire a raggiungere tutte le unità selezionate per far parte del campione.

Per quanto riguarda il primo aspetto (le capacità del rispondente), va considerato che coloro che navigano su Internet, non costituiscono, in definitiva, una popolazione omogenea. Ogni individuo ha un grado di conoscenza dell'utilizzo della rete differente da altri. Allo stesso modo l'abilità tecnica di un navigatore può variare notevolmente da caso a caso. Possono essere all'origine di particolari situazioni una serie di fattori differenti, sia culturali (l'iter educativo percorso, il *background* culturale), che economici (risorse a disposizione, tipo di connessione), psicologici (predisposizione ad intuire il funzionamento di *form*, questionari, *software* che fanno da interfaccia all'indagine), o di altro tipo.

Già quanto detto fa capire come le numerose statistiche a disposizione assumano un significato piuttosto ridotto e vadano prese con estrema cautela, visto che riguardano gruppi di navigatori caratterizzati da profili anche nettamente differenti. Queste statistiche in genere riguardano gli accessi alla rete (o ai singoli siti web), il numero di individui *on-line*, il numero di visitatori di un sito, il numero di indirizzi *e-mail* registrati, e così via. Offrono, quindi, un ritratto della situazione di connotazione prettamente "quantitativa" e piuttosto superficiale. Non consentono, invece, di avere dati che riguardino aspetti più rilevanti: ad esempio, non permettono di concludere quanti, tra coloro che hanno accesso alla rete, vi fanno ingresso esclusivamente per la lettura della posta elettronica, per navigare in rete o per entrambi gli scopi.

Se si considera, poi, il secondo aspetto di cui si parlava (il livello della attrezzatura a disposizione), si può notare come i potenziali intervistati avranno a disposizione attrezzature *hardware* o *software* di differente tipo e/o livello. Alcuni *Personal Computer*, ad esempio, possono essere usati esclusivamente per leggere la posta elettronica, oppure solamente per navigare in rete, oppure unicamente per accedere a reti *Intranet*. È rilevante, inoltre, anche il tipo di connessione a disposizione, in quanto condiziona pesantemente l'utilizzo di questionari particolarmente lunghi: chi ha una connessione lenta o relativamente costosa non sarà disposto a passare molto tempo in rete per rispondere alle domande di una indagine. Inoltre, a parità del livello delle attrezzature *hardware* e *software* a disposizione del rispondente, va considerato il tipo di utilizzo che ciascun individuo fa dei mezzi a sua disposizione, che può essere indicatore della sua abilità pratica.

Incrociando i due aspetti appena visti (relativi, quindi, alle capacità operative del navigatore e alle opportunità a lui offerte dall'equipaggiamento a disposizione) è possibile arrivare ad una classificazione di coloro che utilizzano la rete *Internet* o la posta elettronica in 13 differenti categorie, illustrate nella Tabella 2.

In ogni indagine ci si può imbattere in ciascun tipo di rispondente, tra i 13 indicati in Tabella 2. All'interno di popolazioni diverse, poi, le proporzioni con cui le differenti tipologie di individui sono presenti possono variare anche considerevolmente.

Una situazione che, potenzialmente, potrebbe minare il successo di una indagine via posta elettronica è rappresentata da una massiccia presenza di individui riconducibili alla tipologia 9. In particolare, la categoria 9 comprende gli individui che dispongono di un sistema in cui è istallato un *software* per la gestione della posta elettronica, ma non sanno usarlo: nel caso in cui gran parte della popolazione abbia queste caratteristiche, è necessario riflettere sull'opportunità di proporre un'indagine basata su questionari proposti unicamente via *e-mail*. Non è difficile, nella realtà, imbattersi in questa situazione: alcune aziende forniscono al personale recapiti e *software* per la posta elettronica che poi, di fatto, restano inutilizzati.

Anche la situazione evidenziata dal gruppo 7 rappresenta un estremo negativo per un gestore di questionari pubblicati su pagine web. Si tratta di persone che non hanno a disposizione un *browser* (come *Netscape*, *Internet Explorer*, ...) e, comunque, anche avendolo, non saprebbero come usarlo.

Diametralmente opposta la situazione evidenziata dal gruppo 1: non solo questi individui hanno la capacità di utilizzare *browser* per la navigazione e programmi di gestione di posta elettronica, ma sono anche dotati di un equipaggiamento che consente loro di sfruttare entrambe queste capacità. La persona contattata, dal canto suo, è in

grado sia di utilizzare la posta elettronica che di accedere e partecipare in modo appropriato a indagini proposte in rete.

Le situazioni intermedie che si possono individuare tra quelle più estreme comprendono un buon numero di altre categorie.

	Caratteristiche considerate			
	Capacità dell'utente		***Potenzialità equipaggiamento***	
Tipi di utente	*E-mail*	*Browser*	*E-mail*	*Browser*
1.	Sì	Sì	Sì	Sì
2.	Sì	Sì	Sì	
3.	Sì	Sì		Sì
4.	Sì		Sì	Sì
5.		Sì	Sì	Sì
6.	Sì	Sì		
7.	Sì		Sì	
8.	Sì			Sì
9.		Sì	Sì	
10.		Sì		Sì
11.			Sì	Sì
12.			Sì	
13.				Sì

Tab. 2. *Caratteristiche del navigatore e dell'equipaggiamento tecnico.*

Una prima analisi qualitativa della popolazione obiettivo che tenga conto degli aspetti appena evidenziati è molto utile in quanto consente di rendersi conto quanto, potenzialmente, i visitatori di un sito web o, comunque, i rispondenti selezionati per partecipare ad una indagine, possano essere in grado di parteciparvi.

Non sono solo, però, le capacità dei rispondenti e l'attrezzatura a disposizione a determinare se una indagine avrà successo o meno. A parità di capacità e avendo disponibili programmi e attrezzature necessarie, vanno considerate, infatti, anche altre limitazioni, legate al tipo di *software* o *hardware* a disposizione. Ad esempio, alcuni *software* per gestire la posta elettronica non sono in grado di attivare i *browser* (e questo fattore può essere un limite all'utilizzo di questionari inviati via *e-mail*), oppure non consentono la visualizzazione degli allegati (grosso *handicap* per le indagini in cui il questionario è allegato alle *e-mail* di invito). Vi sono, poi, alcuni ricevitori del segnale televisivo che, pur essendo in grado di accedere alla rete Internet, sono condizionati da limitazioni di rilievo. Altra situazione di cui non si può non tenere conto è l'utilizzo

piuttosto diffuso delle *web-based e-mail facilities* (come ad esempio *Yahoo Free Email Addresses*), che spesso non lasciano piena libertà di azione al rispondente.

In definitiva sono molteplici gli aspetti che andrebbero considerati nel costruire un campione per indagini via web: quanto frequentemente i soggetti coinvolti consultano il proprio recapito di posta elettronica, se l'uso del web è limitato alla lettura delle *e-mail*, con quale frequenza vengono visitati determinati siti web, che tipo di connessione si ha a disposizione, se più individui utilizzano la stessa attrezzatura, da quale punto di connessione avviene l'accesso alla rete (casa, posto di lavoro, *Internet point*, ...), se vi sono a disposizione molteplici indirizzi di posta elettronica, in che modo tali recapiti vengono utilizzati, se eventualmente sono condivisi con altri utenti, e così via dicendo. Tutti questi aspetti possono fare comprendere sia quanto velocemente perverranno le risposte all'indagine, sia se questa avrà alte probabilità di successo in termini di tassi di risposta e di grado di rappresentatività del campione. Si potrà, così, concludere in via preventiva se sarà possibile scegliere un campione giusto, se ci si deve aspettare risultati distorti, se si rischia di escludere una porzione rilevante dei rispondenti, e così via.

Oltre agli aspetti puramente tecnici, relativamente alla scelta di un determinato campione non va trascurato un altro importante fattore, ovvero il periodo temporale in cui viene proposta la partecipazione all'indagine. Si è notato che esiste una stretta relazione tra la scelta del periodo in cui l'indagine è proposta e la composizione del campione che si ottiene di conseguenza. È chiaro, infatti, che particolari fasce orarie (o periodi dell'anno) possono contribuire a connotare in modo decisivo la composizione del campione. In base agli orari (o ai giorni) scelti, alcune particolari categorie di rispondenti potranno essere facilmente reperibili, altre saranno più difficilmente contattabili o potranno partecipare con non poche difficoltà all'indagine. La conseguenza, anche in questo caso, sarà quella di vedere aumentare il rischio di ottenere risultati distorti, rischio molto probabile se l'oggetto di ricerca si connota diversamente per le specifiche categorie di rispondenti considerate.

9.2.2 Rispondenti professionisti

Altro problema in cui ci si imbatte frequentemente è quello dei cosiddetti *rispondenti professionisti*, cioè coloro che partecipano ad un grosso numero di indagini. Sono stati fatti numerosi tentativi di quantificare la presenza di rispondenti professionisti che partecipano ad una indagine e si è spesso cercato di studiarne le caratteristiche e di classificarli. L'importanza di un approfondimento di questi aspetti è legata al rischio di arrivare a risultati distorti e a conclusioni sbagliate.

Per dare un'idea della diffusione del fenomeno dei rispondenti professionisti, basti ricordare che nel Regno Unito una casalinga media partecipa a circa 11 interviste telefoniche all'anno. È stato inoltre stimato che negli Stati Uniti il 50% di tutte le risposte alle indagini e ai sondaggi provengono da solamente il 4% della popolazione. Si capisce ancora di più la portata del fenomeno se si pensa che vi sono alcuni individui che vanno alla ricerca di indagini cui partecipare e ci sono alcuni siti web[26] che

[26] Ad esempio www.surveys4money.com, www.doughstrett.com, www.rewardsites.com.

segnalano ai navigatori dove trovare indagini per cui si possono guadagnare dei soldi o vincere dei premi.

Se il problema dei rispondenti professionisti si presenta per le indagini tradizionali, esso è ancora più sentito per le indagini *on-line*, dove, come visto, già la rappresentatività del campione è un problema di non facile soluzione. Per cercare di incrementare i bassi tassi di partecipazione, spesso si introducono incentivi finalizzati a incoraggiare la collaborazione. Questa strategia, però, comporta rischi di un certo rilievo: vi saranno alcuni individui che risponderanno al questionario esclusivamente per ottenere l'incentivo. Per arginare il fenomeno alcuni istituti di ricerca (*AOL/Opinion place* e *Millward Brown IntelliQuest Tech Panel*, ad esempio) hanno precise regole sulla frequenza con cui coloro che fanno parte di un *panel* possono essere interpellati. In genere una frequenza di intervista ritenuta opportuna è pari ad una volta ogni due settimane e si cerca di fare in modo che chi fa parte di un *panel* non partecipi ad un numero superiore di indagini. Nonostante questo tentativo, per le *panel company* risulta però impossibile poter controllare se i propri rispondenti sono iscritti anche ad altri *panel*. D'altra parte uno studio della SSI ha concluso che quasi il 90% dei rispondenti vuole fare almeno una indagine alla settimana, il 60% anche più di una. A conferma della volontà di essere frequentemente intervistati vi è un altro studio di Wilke ed altri (2000) in cui si sottolinea che il 46% degli individui inseriti nel *panel BASES* della ACNielsen (persone che partecipano a più di 20 test per anno), vorrebbe essere intervistato più frequentemente.

Oltre a quanto detto, vi sono altri due rischi collegati ai rispondenti professionisti. In primo luogo può darsi che essi, partecipando a un numero molto elevato di indagini, diventino più sensibili a determinati argomenti e più accorti quando si imbattono in quesiti particolari (come le domande che vengono utilizzate nei questionari per il controllo di coerenza). In questo modo essi rischiano di non essere rappresentativi della popolazione reale. In secondo luogo è aperta la questione sull'onestà delle risposte, se l'unico scopo per cui sono date è per ottenere l'incentivo promesso.

Tenendo conto della frequenza di partecipazione alle indagini e del grado di onestà del rispondente, è stata elaborata una classificazione dei rispondenti ad indagini attuate via web (Fig. 1).

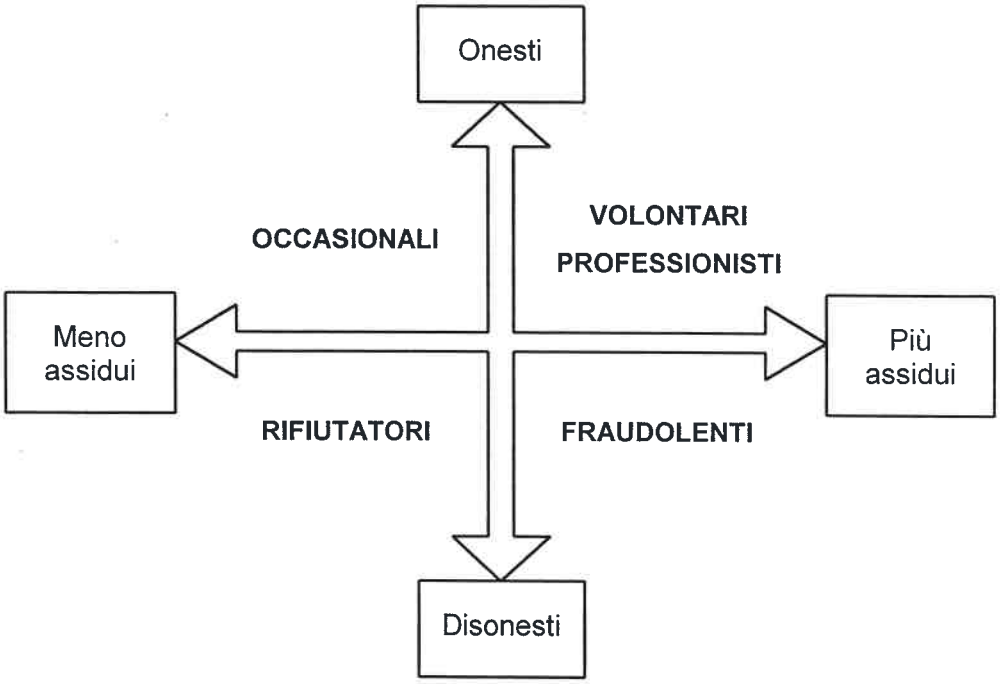

Fig. 1. *Classificazione dei rispondenti alle indagini via web.*

- Il problema più grosso che va affrontato è rappresentato dai rispondenti *fraudolenti*, cioè quegli individui che partecipano a numerose indagini e forniscono risposte false per aumentare il proprio guadagno e/o per malafede. Alcune ricerche, esaminando l'incoerenza delle rispose, hanno permesso di notare che il gruppo di rispondenti fraudolenti è, in genere, piuttosto ridotto.
- I *professionisti* partecipano molto di frequente alle indagini, motivati dagli incentivi, ma in genere sono sinceri nelle risposte. I *volontari*, invece, partecipano molto spesso ad esse, spinti dal desiderio di aiutare o perché le trovano divertenti. Uno studio della SSI (marzo 2000) ha permesso di constatare che il peso delle due categorie era pressochè equivalente all'interno del campione considerato: il 50% circa era rappresentato da rispondenti professionisti, il 50% da volontari.
- I *"rifiutatori"* sono individui che rifiutano di partecipare a qualsiasi tipo di indagine. Alcuni non accettano di fare indagini con nessun tipo di approccio, altri rifiutano solo quelle di un determinato genere: ad esempio possono non accettare di partecipare a ricerche *on-line* ma non ad altri tipologie (telefoniche o faccia a faccia). Da uno studio condotto dal *MRS Respondent Interviewer Interface* è stato notato che il numero di coloro che si rifiutano di partecipare alle indagini è in costante aumento. Altra considerazione aggiuntiva in merito ai "rifiutatori": in genere, se individui di questa tipologia vengono costretti a partecipare ad una indagine, spesso forniscono risposte false.

- Gli *occasionali* sono i rispondenti ideali: se gli si chiede al momento opportuno su un argomento gradito e nel giusto modo di partecipare ad una indagine, essi vi parteciperanno. Queste persone si trovano piuttosto di rado nei *panel* o nelle *mailing-list*, ma possono essere facilmente intercettate con modalità di contatto del tipo *Pop-up*.

Per comprendere con che tipo specifico di rispondente si ha a che fare, si possono inserire nel questionario domande strategiche quali: "a quante indagini ha partecipato negli ultimi sei mesi?" o "qual è il motivo principale per cui partecipa a questa indagine?". Questo tipo di quesiti può essere utile a capire quanti rispondenti di ciascuna della quattro tipologie sono presenti nel campione. I rischi, però, non sono completamente debellati: i risponditori fraudolenti più smaliziati potrebbero subodorare il meccanismo e la finalità delle domande-spia e riuscire ad eluderle vanificando la loro funzione conoscitiva.

10 Summary and Conclusions

Per indagini via Internet/Web è possibile utilizzare le medesime tecniche di campionamento utilizzate per indagini tradizionali. A fronte delle potenzialità (in termini di tempestività, possibilità di ampia estensione territoriale e costi comparativamente bassi) offerte da questo tipo di indagini vi sono però da affrontare alcune tematiche di rilievo connesse con il campionamento.

Date le particolari modalità secondo cui è ottenuto il contatto nelle indagini via Web, nell'approntare la tecnica di campionamento. Uno di problemi maggiori è legato alla disponibilità delle liste di campionamento. Nella maggior parte dei casi non si hanno liste di campionamento a disposizione. Quando invece se ne hanno, spesso queste non sono adatte al caso oggetto di studio, sono incomplete, inaccurate o non consentono di ottenere campioni rappresentativi. La conseguenza è che i risultati che si ottengono da una indagine basata su liste non appropriate non potranno poi essere estesi alla popolazione obiettivo.

Altro problema di rilievo legato alle liste di campionamento (quando se ne hanno a disposizione) è rappresentato dagli errori di copertura cioè dalle differenze tra liste di campionamento utilizzate e popolazione obiettivo. D'altra parte la *NUA Ltd.* [27] ha stimato (Febbraio 2002) che in tutto il mondo ha accesso alla rete solo l'8,96% della popolazione (544,2 milioni di individui), contro il 58,8% registrato nel Nord America (164,4 milioni di individui). Negli ultimi anni la diffusione delle connessioni al Web si è velocemente incrementata, ma visti i dati, si capisce come non sia ancora possibile riuscire a contattare, ad esempio, una popolazione generica come quella di uno Stato.

Altro punto critico per le indagini via Web è la difficoltà di utilizzo di tecniche di campionamento di tipo probabilistico, vista la difficoltà di reperimento delle liste e la loro potenziale non-rappresentatività. Per risolvere in parte le difficoltà legate a questo aspetto, in genere è possibile scegliere tra due *modus operandi*: restringere la popolazione obiettivo a coloro che utilizzano Internet o utilizzare un sistema alternativo di contatto dei rispondenti (ad esempio telefonico) che consenta di raggiungere un più diffuso insieme di individui da coinvolgere.

In realtà, nel campionamento per indagini via Web si utilizza prevalentemente un approccio non probabilistico. Spesso, infatti, risulta impossibile sia identificare i membri della popolazione obiettivo che utilizzare un sistema di selezione probabilistico. Anche servendosi di un sistema di generazione casuale degli indirizzi *e-mail* molti di quelli generati potrebbero essere inesistenti, non validi o non utilizzati. Inoltre fare *spamming* (cioè inviare *e-mail* non richieste a grandi quantità di indirizzi) è vietato e molti server di posta elettronica provvedono a cancellare automaticamente *e-mail* identificate come *spam*. Tutti questi fattori possono aumentare esponenzialmente il rischio di ottenere un campione non rappresentativo della popolazione e in alcuni casi anche accentuatamente differente da quello selezionato con il piano di campionamento. Per ovviare a questo inconveniente l'utilizzo del *mixed mode approach* diviene sempre più ricorrente e fondamentale (Couper, 2000). I rispondenti possono, ad esempio, essere

[27] http://www.nua.ie/surveys/how_many_online/index.html

contattati via telefono col proposito di convincerli a partecipare ad indagini via Web; oppure è possibile integrare i dati raccolti mediante indagini Internet con altri raccolti invece tramite tecniche di indagine più tradizionali.

Un problema notevole per impostare strategie di campionamento è la scarsità e l'estrema approssimazione delle informazioni a disposizione sui non rispondenti.

Anche nel caso in cui si riuscisse a selezionare, avendo a disposizione una opportuna lista di campionamento, un campione appropriato, vi è, inoltre, da tenere presente un ulteriore aspetto: la questione dell'identità del rispondente, problema peraltro ricorrente anche in altri tipi di indagini. Le indagini via Web non sono in grado di garantire che chi risponde al questionario proposto sia realmente la persona selezionata per fare parte del campione, né sono in grado di garantire che, per la compilazione del questionario, il rispondente non si consulti con altri individui.

Nel campionamento per le indagini via Web, poi, non si può trascurare un ulteriore problema di rilievo, ovvero la presenza dei *rispondenti professionisti*, cioè coloro che partecipano ad un grosso numero di indagini. Due rischi collegati ai rispondenti professionisti sono che essi, partecipando a un numero molto elevato di indagini, diventino più sensibili a determinati argomenti e quindi rischino di non essere rappresentativi della popolazione reale che si vuole studiare; inoltre è aperta la questione sull'onestà delle risposte, se l'unico scopo per cui sono date è ottenere l'incentivo promesso.

Per arginare il fenomeno, alcuni istituti di ricerca hanno precise regole sulla frequenza con cui coloro che fanno parte di un *panel* devono essere interpellati. Nonostante ciò, per queste compagnie *panel* risulta però impossibile poter controllare se i propri rispondenti sono iscritti anche ad altri *panel*.

Si capisce, da quanto detto, come sia importante, al fine di approntare una opportuna tecnica di campionamento, approfondire in prima istanza lo studio delle caratteristiche di coloro che navigano in rete, per andare oltre le essenziali e rudimentali informazioni che spesso si possono avere riguardo, ad esempio, il numero di visitatori di un sito Web. D'altra parte, la presenza di siti specchio, la possibilità per il navigatore di occultare la propria presenza, particolari *frames* o congegni meccanici possono falsare i dati statistici sul numero di visite (Smith, 1998 e Foan & Read, 1999).

Nel costruire un campione, inoltre, non bisogna dimenticare che il "popolo dei navigatori" è connotato da caratteristiche socio-demografiche ed economiche piuttosto particolari. Per questo McDonald (1999) ha sottolineato che in genere un campione di persone che utilizzano Internet è rappresentativo solo di utilizzatori di Internet. Anche Coomber (1997) ha evidenziato che la distorsione che si ha nelle indagini condotte via Web è dovuta alle caratteristiche particolari dei "navigatori", che si differenziano in modo netto rispetto ad una popolazione generica: chi utilizza il Web in genere ha un particolare *background* socio-demografico, appartiene prevalentemente al genere maschile, è di razza bianca, relativamente ricco e ben educato, ha un livello di istruzione superiore e può fruire di risorse particolari (Nielsen & CommercNet, 1995; Kehoe & Pitkow, 1996).

Risulta pressoché impossibile, invece, contattare fasce di popolazione ben definite comprendenti, ad esempio, persone anziane, con reddito basso, con livello di istruzione basso, e così via. Per un approfondimento sullo studio di coloro che non utilizzano

frequentemente la rete, si rimanda a Coffey & Johnson (1998). È anche vero (Coomber) che, grazie al diffondersi sempre più capillare delle nuove tecnologie e grazie alle nuove opportunità offerte da connessioni a Internet sempre più veloci, questa distorsione tipica del campionamento per indagini via Web è destinata, in un prossimo futuro, ad assottigliarsi.

L'efficacia e l'appropriatezza di un piano di campionamento per indagini via Web sono quindi, subordinate ad una conoscenza approfondita di coloro che navigano in rete e di coloro che non hanno la possibilità di accedervi, ma sono anche strettamente connesse ai dati che si vogliono raccogliere: se la popolazione oggetto di ricerca è costituita dagli utilizzatori di tecnologie avanzate di comunicazione o (ancor meglio) dagli utilizzatori di Internet, campionamento e indagini via web sono una scelta di approccio alquanto appropriata. Altrettanto si può dire se parte della popolazione obiettivo include i navigatori o se si tratta di indagini riguardanti espressamente Internet e chi ne fa uso, se la popolazione obiettivo ha caratteristiche in comune con i navigatori o se vi è una certezza quasi assoluta che le risposte date dai navigatori non differiranno sostanzialmente da quelle che darebbero coloro che non hanno la possibilità di accedere ad Internet.

È quindi essenziale, per garantire che si ottengano informazioni esatte, verificare che i dati raccolti riguardo un certo argomento non differiscano in un modo che rispecchia le differenze nelle modalità di accesso alla rete e che non siano differenti tra persone che hanno accesso o meno a Internet.

Per le tecniche di campionamento per indagini via Web, che pure sono l'implementazione delle analoghe tecniche tradizionali, si riscontrano, quindi, numerosi problemi di applicazione. Le metodologie tradizionali assumono, in questo contesto, connotazioni particolari, comportano una conoscenza approfondita della popolazione dei potenziali rispondenti e un loro appropriato utilizzo è limitano solo ad alcuni ambiti applicativi, mentre se ne sconsiglia la applicazione ad altri.

Testi di approfondimento

Per ulteriori approfondimenti sulle tradizionali tecniche di campionamento:

Babbie, E. (1990). *Survey Research Methods*. Belmont, California, USA: Wadsworth Publishing Company.

Barnett, V. (1991). *Sample Survey – Principles & Methods*. London, England: Edward Arnold.

Czaja, R., Blair, J. (1995). *Designing Surveys – A Guide to Decisions and Procedures*. Thousand Oaks, California, USA: Pine Forge Press.

Deming, W.E. (1990). *Sample Design in Business Research*. New York, USA: John Wiley & Sons Inc.

ISTAT (1989). *Manuale di Tecniche di Indagine – Vol. 6: Il sistema di Controllo della Qualità dei Dati*. Roma, Italia: ISTAT – Istituto Nazionale di Statistica.

Kish, L. (1987). *Statistical Design for Research*. New York, USA: John Wiley & Sons, Inc.

Maisel, R. (1996). *How Sampling Works*. Thousand Oaks, California, USA: Pine Forge Press.

Moser, C.A., Kalton, G. (1979). *Survey Methods in Social Investigation*. Aldershot, England: Gower Publishing Company Limited.

Parpinel, F., Provasi, C. (1999). *Probabilità e Statistica per le Scienze Economiche*. Torino, Italy: G. Giappichelli Editore.

Sarndal, C.E., Swensson, B., Wretman, J. (1992). *Model Assisted Survey Sampling*. New York, USA: Springer-Verlag.

Utts, J.M. (1996). *Seeing through Statistics*. Belmont, California, USA: Wadsworth Publishing Company.

Zikmund, W.G. (1997). *Exploratoring Marketing Research – Sixth Edition*. Orlando, Florida, USA: The Dryden Press.

Per ulteriori approfondimenti riguardo il campionamento per indagini via web:

WebSM (Web Survey Methodology). http://www.websm.org/.

Bibliografia

Alvarez, R. M., Sherman, R. P., VanBeselaere, C. E. (2003). *Subject Acquisition for Web-Based Surveys*. From http://survey.caltech.edu/encyclopedia.pdf

Biffignandi, S., Toninelli , D. (pubblic. prevista: 2006). *Inferenza in Indagini via Web*. Quaderno di Ricerca – Dipartimento di Matematica, Statistica, Informatica ed Applicazioni – Università degli Studi di Bergamo.

Bourque, L. B., Fielder, E. P. (1995). *How to Conduct Self-Administered and Mail Surveys - The Survey Kit, Vol. 3*. Thousand Oaks, California, U.S.: SAGE Publications, Inc. .

Bradley, N. (1997). *Sampling for Internet Surveys. An examination of respondent selection for Internet research*. University of Westminster. Retrieved 10.07.2005, from http://users.wmin.ac.uk/~bradlen/papers/sam06.html

Chisnall, P. M. (1990). *Le ricerche di marketing*. McGraw-Hill Libri Italia: Milano.

Cochran, W. (1977). *Sampling Techniques* (3rd edition). John Wiley and Sons: New York.

Comley, P. (2000). *Pop-up Surveys. What works doesn't work and what will work in the future*. Proceedings of the ESOMAR Worldwide Internet Conference Net Effects 3, Dublin, 2000.

COMMERCENET and NIELSEN MEDIA RESEARCH (1996). *Internet Hyper-Growth Continues*, Fuelled by New, Mainstream Users. Press Release, http://www.commerce.net/work/pilot/nielsen_96/press.html .

Coomber, R. (1997). *Using the Internet for Survey Research*. Sociological Reserach Online, vol. 2 no. 2: http://www.socresonline.org.uk/socresonline/2/2/2.html.

Couper, M. P. (2000). *Web Surveys: A Review of Issues and Approaches*. Public Opinion Quarterly, 64, pp. 464-494.

Couper, M. P., Traugott, M. W., Lamias, M. J. (2001). *Web Surveys Design and Administration*. Public Opinion Quarterly, 65, pp. 230-253.

Corbetta, P. (1999). *Metodologia e tecniche della ricerca sociale*. Il Mulino: Milano.

De Carlo, N.A. (1983). *La scelta del campione*. Liviana: Padova.

De Carlo, N.A., Robusto, E. (1996). *Teoria e Tecniche di Campionamento nelle Scienze Sociali*. Milano, Italy: LED – Edizioni Universitarie di Lettere Economia Diritto.

De Cristofaro, R. (1979). *Rilevazioni campionarie. Breve introduzione metodologica*. Clueb: Bologna.

De Luca, A. (1990). *Metodi statistici per le ricerche di mercato*. Utet Libreria: Torino.

Deming, W.E. (1960). *Sample Design in Business Research*. Wiley: New York.

Dillman, D. A. (2000). *Mail and Internet Surveys: The Tailored Design Method* (2nd edition). John Wiley and Sons: New York.

ESOMAR (1998). *Conducting Marketing and Opinion Research Using the Internet.* NL: Esomar Guideline.

ESOMAR (1998). *Internet Conference.* Paris (January 1998): papers 117-133, 165,182, 168-182, 219-232, .

ESOMAR (1999). *Internet Conference.* London (February 1999): papers 119-129.

Fabbris, L. (1989). *L'indagine campionaria. Metodi, disegni e tecniche di campionamento.* La Nuova Italia Scientifica: Roma.

Fink, A. (1995). *How to Sample in Surveys – The Survey Kit, Vol. 6.* SAGE Pubblication, Inc.: Thousand Oaks, California, U.S. .

Foan, R., Read M. (1999). *Website Auditing.* London: MRG Meeting (10-2-1999).

Giusti, F. (1989). *Statistica applicata.* Cacucci: Bari.

Godin, S. (1999). *Permission Marketing.* Simon & Schuster.

Hansen, M. H., Hurwitz, W. N., Madow, W. G. (1953). *Sample Survey Methods and Theory - vol. 1: Methods and applications; vol. 2: Theory.* Whiley: New York.

Hansen, M. H., Hurwitz, W. N., (1942). *Relative Efficiencies of Various Sampling Units in Population Enquiries.* Journal of the American Statistical Association, 37, pp. 89-94.

Jeavons, A. (1999). *Research on the Internet.* ESOMAR Handbook of Market and Opinion Research (pp. 1085-1101).

Kehoe, C.M., Pitkow, J.E. (1996). *Suveying the Territory: GVU's Fire WWW User Survey.* Worldwide Web Journal, vol. 1, issue 3 www.w3.org/pub/WWW/Journal/3/s3.kehoe.html .

Kish, L. (1965). *Survey Sampling.* Wiley: New York.

Knowledge Networks (2002). *Decoding the Consumer Genome.* http://www.knowledgenetworks.com/press/presskit.html.

Malhotra, N. K. (1999). *Marketing Research. An Applied Orientation.* (International Edition, 3rd ed.) London: Prectice Hall.

Marbach, G. (1992). *Le ricerche di mercato.* UTET: Torino.

McDonald, M., Wilson H. (1999). *E-Marketing: Improving Marketing Effectiveness in a Digital World.* London: Financial Times Prentice Hall.

Mugo, F.W., *Sampling in Research.* From http://www.socialresearchmethods.net/tutorial/Mugo/tutorial.htm

NIELSEN MEDIA RESEARCH and COMMERCENET (1995). *The CommerceNet Nielsen Internet Demographics Survey.* www.nielsenmedia.com .

Pfleiderer, R., Gente J. (1998). *A New Dimension of Internet Research.*

Poynter, R. (2001). *A Guide to Best Practice in Online Quantitative Research. ASC International Conference.* May, 2001. from http://www.asc.org.uk/Events/May01/Resource/Poynter2.pps

Smith, P.R. (1998). *Marketing Communications*. London: Kogan Page.

Talmage, P. A. (1988). *Dictionary of Market Research*. London: MRS/ISBA.

Tse, A., altri (1994). *A comparison of the effectiveness of mail and facsimile as survey media on response rate, speed and quality*. Journal of the Market Research Society, n. 36, pp. 349-355.

Turner, W.J. (1989). *Small Business data collection by area censusing: a filed test of "Saturation Surveying" methodology*. Journal of Market Reserach Society (Vol. 3, No.2, April 1989).

U.S. Department of Commerce (2000). *Faling Through the Net: Toward Digital Inclusion*. http://www.ntia.doc/ntiahome/fttn00/falling.htm.

WebSM (Web Survey Methodology). http://www.websm.org/.

Wilke, J., Lundy, S., Mustard D. (2000). *Burning Out Internet Respondents. Avoiding the Mistakes of the Past*. Proceedings of the ESOMAR Worldwide Internet Conference Net Effects 3, Dublin, 2000.

Witmer, D. F., Colman, R. W., Katzman S. L. (1999). *From Paper-and-Pencil to Screen-and-Keyboard: Toward a Methodology for Survey Research on the Internet*. Reproduced on pp. 145-162 in Jones, S. (1999). Doing Internet Research. Critical Issues and Methods for Examining the Net. California, U.S.: SAGE Publications, Inc. .

Il paper è stato realizzato nell'ambito del progetto di ricerca

WebSM *(Web Survey Methodology – 6th EU Framework Program)*

www.websm.org