UNIVERSITA' CATTOLICA DEL SACRO CUORE
MILANO

Dottorato di ricerca in Sociologia e Metodologia della ricerca sociale
ciclo XXIII (S.S.D: SPS/07)

L'uso delle reti sociali per la costruzione di campioni probabilistici: possibilità e limiti per lo studio di popolazioni senza lista di campionamento

Coordinatore: Ch.mo Prof. Laura Bovone

Tesi di Dottorato di :
Alberto Vitalini
Matricola: 3610661

Anno Accademico 2009/2010

INDICE

PREMESSA

L'idea di questa tesi è nata abbastanza casualmente durante un vagabondaggio nella rete. Mentre cercavo articoli che affrontassero in modo scientifico il legame fra creatività e uso di sostanze stupefacenti nel mondo dell'arte mi sono imbattuto in un articolo dal titolo "Finding the beat: Using respondent-driven sampling to study jazz musicians" che descriveva l'uso di una strategia di campionamento "innovativa" (parole dell'autore) per identificare e intervistare musicisti jazz in quattro città americane. La strategia era descritta in questi termini: "Sulla base delle informazioni iniziali fornite da parte del coordinatore e dallo staff cittadino, sarà selezionata una mezza dozzina di musicisti in ogni città come "semi" a partire dai quali iniziare le catene di reclutamento. Questi musicisti saranno informati riguardo alle finalità dello studio, intervistati faccia a faccia usando un questionario a 114 domande, e verrà data loro, successivamente, l'opportunità di reclutare al massimo quattro musicisti ognuno. I reclutati saranno poi intervistati a loro volta e verrà data loro l'opportunità di reclutare altri colleghi. Questo processo continuerà finché l'obiettivo di 300 musicisti sarà raggiunto in ogni città. I musicisti saranno pagati per le loro interviste e per ognuno delle persone reclutate dopo che quest'ultime avranno completato l'intervista." (Heckathorn - Jeffri, 2001, p. 317)

Nulla di nuovo, già visto si dirà. Il ricercatore propone una strategia di selezione dei casi che va sotto il nome di campionamento a valanga e l'utilizzo di incentivi monetari per assicurarsi una maggiore collaborazione da parte degli

intervistati. Dov'è l'aspetto innovativo? Nel fatto che l'autore affermava di aver sviluppato una teoria statistica che rendeva per la prima volta possibile ottenere sia stime non distorte dei parametri della popolazione sia misure della precisione di queste stime a partire da dati raccolti con un campionamento a valanga.

Non c'è bisogno di sottolineare le potenzialità di una simile "scoperta". Il mio interesse si era acceso. Ho iniziato così, non senza una certa dose di scetticismo, a documentarmi sul metodo proposto. Secondo il vecchio adagio che l'appetito viene mangiando, il mio interesse si è, poi, via via allargato allo studio generale dei disegni di campionamento che sfruttano i legami sociali e alle possibilità del loro utilizzo per la costruzione di campioni di tipo probabilistico.

Il presente lavoro è il frutto di questo studio e può essere considerato il resoconto di un viaggio in un territorio poco esplorato della metodologia delle scienze sociali. Il linguaggio che si è cercato di utilizzare nel testo è stato il più possibile discorsivo, riducendo al minimo l'uso della formalizzazione matematica, per mettere in risalto i fondamenti logici delle tecniche campionarie trattate e le loro possibilità di utilizzo nelle scienze sociali. I dettagli delle procedure statistico matematiche e di calcolo sono stati, invece, rimandati nelle note e nelle appendici.

INTRODUZIONE

Gli individui in società sono legati gli uni con gli altri da relazioni ed interazioni che ne influenzano le percezioni, gli atteggiamenti e i comportamenti. Lo studio dei legami fra i membri di un gruppo si è rivelato così importante che si è sviluppato un approccio teorico e metodologico che va sotto il nome di analisi delle reti e che cerca di identificare, misurare, testare ipotesi riguardo le forme strutturali e ai contenuti sostanziali delle relazioni fra gli attori (Mattioli, 1995; Piselli, 1995; Chiesi, 1996, 1999; Gribaudi, 1996; Mutti, 1996; Knoke - Yang, 2008).

I legami che uniscono le persone possono però essere considerati, in un'accezione maggiormente pragmatica, anche come mezzi per trovare casi "interessanti" da inserire nella ricerca. Considerando questa seconda accezione essi sono applicati in situazioni molto diverse in riferimento sia agli obiettivi della ricerca che alle tecniche usate per la raccolta dei dati e delle informazioni. Alcuni esempi chiariranno il punto.

Nell'ambito delle tecniche non standard i legami sociali sono utilizzati per guadagnare l'accesso al campo studiato in ricerche che fanno uso dell'osservazione partecipante (Whyte, 1955; Patrick, 1973; Sclavi, 1994). Fa parte della storia sociologica il racconto di Whyte (1955) che fu accettato dal gruppo di giovani di Cornerville, solo dopo che Doc, un giovane che godeva di credibilità e prestigio nel quartiere, lo aveva portato in giro e presentato come

amico. Il nome dello stesso Doc era stato a sua volta suggerito a Whyte da un'assistente sociale del quartiere.

Oltre che per guadagnare l'accesso al campo, i legami sociali sono spesso utilizzati per individuare soggetti da coinvolgere nella realizzazione di interviste biografiche (Bichi, 2007): ad ogni persona intervistata viene chiesto di suggerire il nome di altre persone da intervistare. Per fare un esempio, con questo sistema in una ricerca sulla criminalità (Barbagli, 1995) si riuscirono ad intervistare oltre sessanta autori di reati individuati a partire dalle conoscenze personali di due intervistatori.

Nell'ambito delle tecniche standard, i legami sociali sono utilizzati per costruire campioni di persone, appartenenti a popolazioni di cui non si possiede la lista di campionamento (es. immigrati clandestini), alle quali somministrare un questionario. La procedura è simile a quella utilizzata per le interviste biografiche: ad ogni persona intervistata viene chiesto di segnalare il nome di altre persone da intervistare che, oltre ad entrare a far parte del campione ed essere intervistate, diventeranno informatori per l'individuazione di ulteriori contatti e così via... La differenza rispetto alle interviste biografiche è nella numerosità campionaria: l'intervista con questionario comporta numerosità campionarie solitamente elevate. In genere i ricercatori che utilizzano procedure di questo tipo si limitano a selezionare un congruo numero di persone, alle quali somministrare il questionario, solo al fine di esplorare il fenomeno studiato senza porsi come obiettivo la generalizzazione dei risultati, spesso per preparare il campo a studi successivi (Caselli, 2005).

Poniamoci la seguente domanda: nel caso i ricercatori non volessero limitare gli obiettivi della ricerca ad un studio esplorativo e intendessero generalizzare i risultati alla popolazione studiata, sarebbero in grado di farlo, partendo da campioni selezionati sulla base di procedure che sfruttano i legami sociali? La riflessione che ha cercato di rispondere a questa domanda ha avuto un vigoroso sviluppo, sia per le applicazioni pratiche sia per gli aspetti metodologici, soprattutto a partire dalla seconda metà degli anni '80 (per una rassegna cfr. Atkinson e Flint, 2001), in concomitanza con l'emergere di una richiesta sempre più consistente di tipi di campionamento "alternativi" a quelli,

tradizionalmente usati nella ricerca sociale, basati sull'estrazione casuale a partire da liste di campionamento.

Questa crescente richiesta è attribuibile, in ultima istanza, alle difficoltà che le tecniche di campionamento tradizionali, basate su liste di nominativi, incontrano nel risolvere problemi di reperimento casuale posti dalle grandi trasformazioni che hanno attraversato le società occidentali (ad es. mutamenti delle forme di convivenza familiare, elevata mobilità lavorativa e territoriale e processi migratori) (Bichi, 2007); dall'espansione della ricerca sociale in settori molto delicati della vita individuale (ad es. la ricerca sociale nella sanità) (Lanzetti, 2004; Lanzetti *et al.*, 2008) e dalla crescente preoccupazione circa le problematiche della privacy e della riservatezza.

Il presente contributo si inserisce in questo filone di riflessione: più precisamente l'obiettivo di questo lavoro consiste nel chiarire le possibilità e i limiti di utilizzo, nelle indagini campionarie, dei legami sociali per la costruzione di campioni probabilistici per lo studio di popolazioni di cui non si possiede la lista di campionamento e che, di conseguenza, non sono facilmente campionabili utilizzando le strategie di campionamento "classiche" quali, ad esempio, il campionamento casuale semplice, stratificato e a stadi. Alcuni esempi daranno concretamente l'idea del tipo di popolazioni per le quali l'utilizzo delle riflessioni esposte in questo testo potrebbero rivelarsi utili: immigrati irregolari, persone con redditi molto alti o molto bassi, persone colpite da gravi malattie, scienziati disabili, famiglie che possiedono solo il telefonino; bambini dai sei ai dieci anni, tossicodipendenti, membri di sette religiose, sieropositivi, collezionisti di francobolli, membri di community on line, senza tetto, collezionisti di auto d'epoca, veterani dell'Iraq.

Una precisazione è necessaria. Nella riflessione metodologica delle scienze sociali a popolazioni di questo tipo sono spesso associati gli aggettivi "elusive", "nascoste" e altri sinonimi che servono a sintetizzare le difficoltà che caratterizzano il loro studio. Questi aggettivi non verranno utilizzati nel testo, dal momento che c'è mancanza di chiarezza circa le definizioni di "nascosto" e "elusivo". Questi termini sono utilizzati non in modo univoco e chiaro. Sono

stati definiti, a seconda degli autori, come popolazioni "elusive" o "nascoste" i più svariati tipi di popolazioni: ad es., giovani donne single e disoccupate, donne vittime di violenze domestiche, minoranze etniche, persone che hanno contratto il virus del HIV, prostitute, consumatori di sostanze psicotrope, anziani, disabili, senza tetto, persone con redditi sotto la soglia di povertà, evasori fiscali, membri di piccole comunità religiose, uomini di affari, residenti non ancora iscritti all'anagrafe, omosessuali (Kish, 1991; Atkinson - Flint, 2001; Brackertz, 2007). L'ampia casistica a cui sono associati i termini "elusivo" e "nascosto"e il loro uso non coerente minano la loro utilità nell'ambito di una riflessione metodologica.

La tesi è formata da due parti: la prima teorica e la seconda empirica.
Nella parte teorica si cercherà di chiarire le possibilità e i limiti di utilizzo, nelle indagini campionarie, dei legami sociali per la costruzione di campioni probabilistici di popolazioni di cui non si dispone degli elenchi dei membri.
Le diverse strategie di campionamento di questo tipo verranno ricondotte ad un'unica formalizzazione matematica: il campionamento in grafo. Questa formalizzazione, oltre ad aver un valore in sé perché fornisce un quadro concettuale parsimonioso per affrontare diverse problematiche sollevate dalla riflessione metodologica su questi temi, è funzionale alla comprensione di un particolare disegno di campionamento chiamato *respondent-driven sampling* (Heckathorn, 1997). Questo disegno si propone come uno degli approcci più innovativi e promettenti per lo studio delle popolazioni di cui non si dispone degli elenchi degli appartenenti. La prima parte terminerà considerando criticamente gli aspetti del *respondent-driven sampling* che meritano ulteriori approfondimenti e studi.
La seconda parte consiste in una valutazione empirica del r*espondent-driven sampling*. Questo tipo di campionamento si è rivelato relativamente semplice, economico e flessibile, ed è stato utilizzato con successo per studiare, in più di una dozzina di nazioni, diversi tipi di popolazioni di cui non si possiede la lista dei membri: in particolare prostitute, tossicodipendenti, omosessuali (per una rassegna Malekinejad *et al*., 2008).

Il favore con cui è stato accolto evidenzia la capacità del r*espondent-driven sampling* di rispondere ad un bisogno diffuso nella comunità scientifica, ma non deve far dimenticare che una sua completa accettazione richiede un'approfondita valutazione dell'accuratezza delle stime .

Il presente lavoro cercherà di valutare le performance del r*espondent-driven sampling* sia attraverso simulazioni sia svolgendo un'indagine su una popolazione reale, una community Internet, di cui è stata ricostruita la struttura delle relazioni che legano le persone fra loro. L'utilizzo combinato di simulazione e indagine-test consentirà di sviluppare una comprensione qualitativamente più profonda di questa forma di campionamento.

PARTE PRIMA

1. *IL CAMPIONAMENTO PROBABILISTICO NELLE SCIENZE SOCIALI: ALCUNE PRECISAZIONI*

Prima di iniziare a considerare forme di campionamento che sfruttano i legami sociali sono necessarie alcune precisazioni sugli aspetti generali del campionamento rispetto ai quali le riflessioni di un ricercatore in scienze sociali si possono rivelare utili e fondate (par. 1); sui principi logici che consentono di fare inferenza statistica a partire da campioni estratti casualmente (par. 2); sulla ragione per cui le tecniche classiche di campionamento probabilistico sono in difficoltà nel trattare popolazioni di cui non si possiede la lista di campionamento (par. 3); e sull'espressione "campionamento a valanga" che si applica comunemente, nella riflessione metodologica delle scienze sociali, alla strategia che sfrutta i legami fra le persone per individuare i soggetti da inserire nel campione (par. 4).

Queste precisazioni non sono oziose o frutto di un ragionamento bizantino, servono per inquadrare le riflessioni successive evitando possibili fraintendimenti e confusioni. Quando si affronta in ambito sociologico una riflessione sul campionamento non è raro imbattersi in diverse definizioni degli stessi concetti. Basti pensare ai diversi significati attribuiti ai termini "rappresentativo" e "casuale" utilizzati in frasi molto comuni del tipo «il mio campione è rappresentativo perché è stato estratto casualmente» (Marradi, 1997; Palumbo – Garbarino, 2004).

1. Il campionamento probabilistico: sintassi e semantica

E' bene, quando si parla di campionamento probabilistico nelle scienze sociali, tenere distinti due piani: quello della formalizzazione matematica, che si potrebbe definire grammaticale (o sintattico), e quello della sua applicazione nella realtà, che si potrebbe definire semantico.

Il piano grammaticale è quello del modello matematico cioè della rappresentazione formale, espressa in linguaggio matematico, di un fenomeno (Israel, 2002). In generale, una descrizione "completa" di un fenomeno sarebbe un interminabile (quanto impossibile) discorso aderente a tutte le pieghe dei fatti, nessuna esclusa. Per descrivere un fenomeno è necessario fare delle scelte, selezionarne degli aspetti, in una parola semplificare. Anche il modello matematico risente di questo limite: esso, infatti, è una rappresentazione semplificata di un fenomeno. Il modello e il fenomeno studiato sono, di conseguenza, in un rapporto di analogia, non di identità: questo va ben tenuto presente quando dal piano sintattico si passa a quello semantico.

Il piano semantico è quello dell'applicazione al fenomeno studiato del modello matematico e delle deduzioni che si possono trarre da esso. Affinché quest'operazione sia sensata, il modello deve essere in grado di descrivere adeguatamente il sistema.

Riportando questa riflessione generale al tema specifico del campionamento, il dibattito metodologico nelle scienze sociali si muove essenzialmente all'interno del piano semantico: cioè analizza, discute, problematizza se, e quanto, il modello matematico del campionamento possa applicarsi nella reale pratica della ricerca; se il modello matematico (sintassi) è in grado di descrivere con l'approssimazione richiesta il sistema reale che si desidera studiare. Ad esempio, il modello matematico alla base del campionamento di tipo probabilistico richiede che tutte le persone estratte siano intervistate e che rispondano in modo accurato. Il dibattito metodologico nelle scienze sociali valuta se queste condizioni si riscontrano nella reale pratica delle indagini

campionarie e quali sono le conseguenze sull'accuratezza delle stime di una loro eventuale deviazione.

Il contributo di questa tesi si muove nell'ambito della semantica: fin dove, con quali limiti si possono utilizzare nella ricerca le riflessioni sviluppate nell'ambito della teoria dei campioni. Come già anticipato nella premessa, il linguaggio che si è cercato di utilizzare nel testo è stato il più possibile discorsivo, riducendo al minimo l'uso della formalizzazione matematica, per mettere in luce i fondamenti logici delle tecniche campionarie trattate (la sintassi). Sono stati considerati gli aspetti sintattico-formali strettamente necessari per capire le applicazioni pratiche a problemi di ricerca che sono "sociologicamente" rilevanti come ad esempio lo studio delle popolazioni devianti o delle minoranze etniche. I dettagli delle procedure statistico matematiche e di calcolo sono stati, invece, rimandati nelle appendici e nelle note a piè di pagina.

2. Il campionamento probabilistico e la probabilità di inclusione: un connubio indivisibile

Il campione è una parte selezionata di un tutto dalla cui analisi si traggono informazioni sull'insieme. Accettata questa definizione di campione siamo di fronte ad un tipico problema di inferenza induttiva: da una proposizione particolare (ad es., la maggioranza delle persone nel campione ritiene che non dovrebbero esserci discriminazioni sul lavoro nel caso di donne con figli) si vuole giungere ad asserzioni generali con un certo livello di fiducia (ad es., si è ragionevolmente sicuri che la maggioranza delle persone nella popolazione ritiene che non dovrebbero esserci discriminazioni sul lavoro nel caso di donne con figli). Per rispondere a questo tipo di problema una parte degli studiosi di scienze sociali, che aderisce al paradigma positivistico, fa ricorso agli strumenti e alle procedure matematiche dell'inferenza statistica, i quali sono stati messi a punto nell'ambito di una branca della matematica che va sotto il nome di teoria delle probabilità, o più specificatamente, teoria dei campioni. Per poter

applicare in un'indagine campionaria reale gli strumenti dell'inferenza statistica è necessario attribuire ad ogni individuo una probabilità nota, non nulla di venire a far parte del campione. La probabilità non deve essere uguale per tutte le unità; al limite ogni unità della popolazione può avere una probabilità diversa, purché sia conosciuta e diversa da zero.

Differenti probabilità possono risultare da alcune caratteristiche della procedura di campionamento (ad es., campionamento a stadi) oppure possono essere imposte deliberatamente dal ricercatore per ottenere migliori stime, includendo unità con particolari caratteristiche con una maggiore probabilità (ad es., campionamento stratificato in cui il ricercatore seleziona le donne con una probabilità doppia rispetto agli uomini).

Perché è così importante conoscere la probabilità di un'unità della popolazione di entrare a far parte del campione? L'importanza è dovuta al fatto che, nel modello matematico alla base del campionamento probabilistico, le probabilità sono indispensabili per il calcolo di stimatori non distorti[1] del parametro studiato e della sua varianza campionaria (Kish, 1992; Stuart, 1996) (cfr. appendice 1). In un linguaggio più discorsivo si può comprendere la necessità della conoscenza delle probabilità, se si riflette sul fatto che qualsiasi metodo di stima campionaria si fonda sul seguente principio: le unità comprese nel campione rispondono al questionario in rappresentanza delle rimanenti unità della popolazione che non sono entrate a far parte del campione. Tale principio si realizza praticamente attribuendo a ciascuna unità inclusa nel campione un peso che può essere visto come il numero di persone della popolazione "rappresentate" dalla persona che risponde al questionario. Se, ad esempio, ad un'unità campionaria di sesso femminile di età inferiore ai quarant'anni viene

[1] «Per evitare malintesi è bene ricordare che le affermazioni della teoria dei campioni acquistano significato solo se riferiti all'universo dei campioni e non riguardano, se non indirettamente, il singolo campione osservato o la stima da esso ricavata» (Herzel, 1991, p.628). In statistica con il termini di stimatore "non distorto" o "corretto" si intende che se noi potessimo estrarre da una popolazione data, ma del tutto arbitraria, con un dato piano di campionamento casuale un milione di campioni diversi di una data numerosità e calcolassimo un milione di percentuali (ad es. della variabile genere) lo stimatore percentuale è corretto se la media del milione di percentuali del genere (una per ogni campione) è uguale al valore della percentuale nella popolazione studiata. In altre parole il concetto di "corretto" o "non distorto" non è associato ad una singola percentuale (ad esempio quella calcolata nel cinquantesimo campione), il cui valore potrebbe essere anche molto differente dal valore della popolazione.

attributo un peso pari a 32 , questo indica che essa risponde ad una domanda per se stessa e per altre 31 donne con meno di quarant'anni che fanno parte della popolazione, ma che non sono state selezionate per partecipare all'indagine; è come se la rispondente avesse ricevuto la delega a rispondere per altre 31 donne. Il peso da attribuire ad ogni soggetto viene calcolato a partire dalla probabilità dell'unità di entrare a far parte del campione[2] (senza la probabilità non è possibile calcolare il peso e senza quest'ultimo non si possono calcolare stime non distorte).

3. Popolazioni senza lista di campionamento: il meccanismo si inceppa

Stabilita l'importanza di conoscere la probabilità di essere estratti, diventa importante capire come calcolarla. In tutte le forme di campionamento che potremmo definire "classiche", per poter calcolare la probabilità di un'unità di essere inclusa nel campione è necessario disporre della lista di tutte le unità che fanno parte della popolazione studiata[3]. La lista di campionamento dovrebbe essere, nei limiti del possibile, completa cioè contenere tutte le unità di analisi che compongono la popolazione oggetto di studio[4].

La situazione ideale sarebbe quella di avere una lista di campionamento in cui sono elencati tutti i membri della popolazione oggetto di studio. Se rappresentiamo come due quadrati di uguali dimensioni la lista di campionamento e la popolazione di riferimento, la situazione ideale è quella di

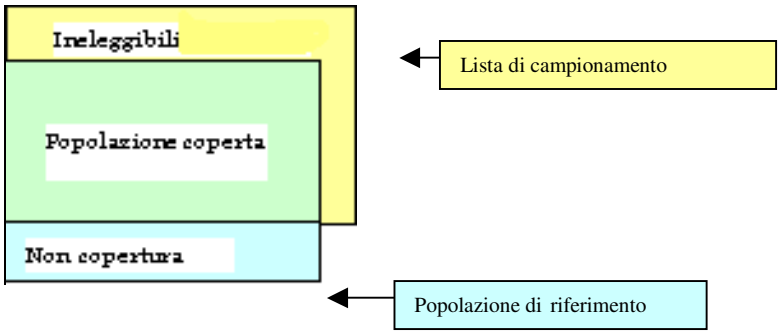
[2] Esso non è altro che il reciproco della probabilità di entrare a far parte del campione. Va sottolineato che nelle indagini effettive il peso da attribuire a ciascuna unità è ottenuto non solo in base ad una procedura di calcolo che corregge il peso determinato a partire dalla probabilità di entrare a far parte del campione, ma anche per attenuare l'effetto distorsivo sulle stime dovuto ad eventuali errori di mancata risposta e di non copertura.

[3] Nel campionamento a stadi è necessario avere la lista delle unità che possono essere selezionate ad ogni stadio.

[4] In termini teorici è possibile estrarre casualmente i membri di una popolazione anche «in assenza di una lista preventiva, purché esista un luogo dove tutta la popolazione sia localizzata ed il ricercatore possa passare in rassegna, nel corso della selezione, tutti i soggetti» (Corbetta, 1999, p.p.332-333) Se, ad esempio, tutte le persone devono passare da un punto preciso (ingresso di un museo, sportello di un ufficio pubblico) e io ne intervisto, ad es. una ogni cinque, il campione finale risulta essere di tipo casuale. La rarità di queste applicazioni conferma comunque l'importanza di una lista di campionamento preventiva.

perfetta sovrapposizione. In realtà nella maggioranza dei casi non esiste una perfetta sovrapposizione, la situazione nella pratica della ricerca è rappresentabile come in figura 1. In ogni caso, se in una ricerca reale si può ragionevolmente sostenere che la copertura è elevata e le differenza fra membri "non coperti" e "coperti" è limitata, si è legittimati ad utilizzare la lista di campionamento, altrimenti si cerca di costruirne una completa (magari unendo due o più liste parziali) oppure si ridefinisce la popolazione oggetto di studio per adattarla alle caratteristiche della lista di campionamento disponibile (Caselli, 2005).

Fig. 1 - Copertura della lista di campionamento della popolazione di riferimento



Fonte: Rielaborazione personale da Wayne E., www.idready.org/courses/2005/spring/survey_SamplingFrames.pdf

Oltre alla lista è necessario utilizzare una procedura di selezione che garantisca a tutte le unità la possibilità (anche molto piccola, ma diversa da zero) di essere estratte. La procedura migliore per garantire questa possibilità consiste nel selezionare casualmente un certo numero di unità dalla lista[5]. Per

[5] «La teoria dei campioni, nella sua accezione tradizionale, si fonda comunque esclusivamente sul campionamento casuale, nelle sue svariate forme. Più precisamente, questa teoria studia le proprietà dell'insieme costituito da tutti i campioni che possono essere estratti, ossia osservati, da una popolazione data, ma del tutto arbitraria, con un dato piano di campionamento casuale» (Herzel, 1991, p. 626). A patto di aver estratto un campione casuale (e solo a questa condizione), la teoria dei campioni (piano sintattico) permette di valutare, sulla base dei soli dati campionari, il grado di attendibilità delle stime. Praticamente permette di calcolare un intervallo nella forma:

stima +/- margine di errore campionario

in modo da avere "un'elevata sicurezza" che il valore del parametro della popolazione è da qualche parte tra i due valori, inferiore e superiore, dell'intervallo. Si sottolinea che questa affermazione è veritiera, in una reale indagine campionaria (piano semantico), solo nel caso siano di entità trascurabile altri tipi di errori, chiamati "non campionari" (ad es.: errori dovuti alla mancata partecipazione delle persone campionate, a formulazioni poco chiare delle domande, allo scarso impegno o cattiva volontà degli intervistatori e degli intervistati)..

chiarire questi due punti (lista di campionamento e procedura di selezione casuale) si può ricorrere ad un esempio molto utilizzato nei manuali di statistica: quello dell'estrazione da un contenitore. Si crea una lista degli elementi della popolazione oggetto di indagine (es. famiglie residenti in una provincia, persone maggiorenni residenti in un comune). Ad ogni elemento della popolazione si associa una pallina con un numero progressivo. Le palline vengono messe in un contenitore e mescolate più volte. Si estrae una pallina e la si mette da parte, si rimescolano le rimanenti palline e si estrae un'altra pallina…, così via fino a che si raggiunge il numero di elementi desiderato. I numeri sulle palline identificheranno gli elementi della lista che faranno parte del campione. Questa è una procedura di selezione casuale e la probabilità di ogni pallina di entrare a far parte del campione è uguale a n/N dove n è il numero di persone estratte nel campione e N è il numero di persone presenti nell'elenco.

In conclusione, sulla base della teoria dei campioni, per poter fare inferenza statistica (generalizzare le conclusioni tratte dal campione alla popolazione, quantificando la fiducia che abbiamo in questo trasferimento) la condizione necessaria è quella di conoscere la probabilità delle unità campionate di venire a far parte del campione. Per poter calcolare questa probabilità si deve avere a disposizione la lista delle unità della popolazione e utilizzare una qualche procedura di selezione casuale delle unità dalla lista[6].

Per molte popolazioni che sono oggetto di studio da parte dei ricercatori in scienze sociali la lista non esiste oppure non è disponibile. «Salvo i casi in cui i soggetti studiati facciano parte di associazioni formalmente costituite (per es. iscritti ad albi professionali, a partiti, associazioni, ecc.), o appartengano ad un'istituzione (studenti, ricoverati in ospedali, ecc.) o comunque siano a

[6] Relativamente alla possibilità di fare inferenza solo nel caso di campioni probabilistici è particolarmente illuminante la citazione seguente: «A meno di utilizzare metodi di campionamento casuali (o probabilistici, n.d.r.) non v'è assolutamente alcuna base per l'impiego di procedure inferenziali. […] essi rimuovono le distorsioni da selezione e ci mettono in grado di stimare correttamente. Se il lettore può alle volte avere la sensazione che lo statistico, con la sua insistenza sui metodi di campionamento casuale, stia semplicemente persuadendosi a fare un lavoro, dovreste ricredervi pensando che in, assenza di casualizzazione nel campionamento, l'intero apparato inferenziale dal campione alla popolazione viene a cadere, lasciando chi effettua il campionamento senza basi scientifiche per l'inferenza che desidera effettuare» (Stuart, 1996, pp. 34-35).

qualche titolo entrati in contatto con la pubblica amministrazione (iscritti alle liste di collocamento, proprietari di auto, ecc.), per la maggioranza degli studi di segmenti particolari della popolazione non siamo in possesso della lista della popolazione: si pensi a studi su operai, disoccupati, lavoranti a domicilio, artisti, immigrati, casalinghe, […], anziani che vivono soli, conviventi non sposati, ecc. » (Corbetta, 1999, p. 333).

Si è sottolineato come la presenza della lista di campionamento sia una condizione necessaria nelle tecniche di campionamento tradizionali, in quanto essa permette di adottare piani di estrazione casuale, i quali a loro volta forniscono la possibilità di calcolare le probabilità di inclusione delle unità campionate della popolazione. Nel caso di popolazioni di cui non si possiede o non si riesce a costruire un'accurata lista dei membri, l'intera argomentazione entra in crisi e l'assegnazione ai membri della popolazione delle probabilità di entrare a far parte del campione risulta un'operazione estremente difficile se non del tutto arbitraria.

4. Campionamento a valanga: una definizione che sta stretta

Passiamo ora ad un'espressione, "campionamento a valanga", che viene applicata comunemente nella riflessione metodologica delle scienze sociali alle diverse strategie che sfruttano i legami fra le persone per individuare i soggetti da inserire nel campione. In generale, il "campionamento a valanga" si può considerare una procedura in cui «occorre individuare, mediante tutte le informazioni e contatti di cui si dispone, almeno un soggetto appartenente alla popolazione che si intende studiare. Dopodiché, una volta somministrato a questo il questionario gli si chiede se può segnalare altri soggetti dotati delle caratteristiche che interessano il ricercatore. I soggetti segnalati diventeranno a loro volta, oltre che parte del campione, informatori per l'individuazione di ulteriori contatti e così via, secondo la metafora appunto della valanga che, pur con un'origine molto limitata, assume rapidamente proporzioni sempre più ampie» (Caselli, 2005, p. 154).

E' importante sottolineare in questo paragrafo che il campionamento a valanga è ritenuto comunemente un tipo di campionamento che non permette di fare inferenza statistica (Corbetta, 1999; Palumbo - Garbarino, 2004; Bruschi, 2005; Caselli, 2005; Marradi, 2007; Babbie, 2010). Si avrà modo di approfondire in un secondo tempo la questione, dal momento che essa è il cuore di questa tesi. In ogni caso si ritiene utile una breve anticipazione. Nell'ormai classico articolo del 1961 di Goodman, dal titolo *Snowball Sampling*, si legge: «*The data obtained using a snowball sampling procedure can be utilized to make statistical inferences about various aspects of the relationships present in the population*» (Goodman, 1961, p. 148). Per quale motivo allora il campione a valanga, nella riflessione metodologica delle scienze sociali, è associato in maniera quasi esclusiva a campionamenti che non permettono di fare inferenza statistica?

Questa associazione è il riflesso di una pratica consolidata di utilizzo del campione a valanga per condurre ricerche di tipo qualitativo. L'uso del campionamento a valanga, associato a tecniche non standard e a partire da elementi iniziali scelti sulla base di criteri di convenienza, è talmente consolidato nella pratica della ricerca sociale, al punto che questo tipo di campionamento è diventato, *sic et simpliciter*, non probabilistico. Nel resto della tesi si cercherà di mostrare come questa conclusione sia eccessivamente semplicistica.

Il termine "campionamento a valanga", seppur molto utilizzato in letteratura, sarà sostituito, da qui in avanti, da espressioni neutre del tipo "campionamento tramite legami sociali" o "strategie che sfruttano i legami sociali". Il motivo di questa scelta consiste nel tentativo di fare *tabula rasa* di tutte le immagini, le idee, le procedure, richiamate nel lettore dal termine suggestivo, che rischierebbero di essere di intralcio nella comprensione di molti concetti presentati in seguito.

Una nota di cautela nei confronti delle tecniche di campionamento esposte nel resto della tesi va a questo punto fornita. Non esistono semplici soluzioni ai problemi di campionamento posti dallo studio di popolazioni di cui non si possiede la lista di campionamento e le forme di campionamento che sfruttano i

legami sociali non sono sempre necessariamente migliori dei disegni campionari standard. Come sarà ripetuto più volte le condizioni che devono essere soddisfatte affinché le strategie proposte permettano di raggiungere i risultati desiderati sono talvolta troppo stringenti (in alcuni casi, persino impossibili) da incontrare nella pratica della ricerca. Inoltre va anche sottolineato che le tecniche di campionamento considerate possono introdurre una complessità aggiuntiva nelle procedure di ricerca che rischiano di aumentare in modo elevato gli effetti degli errori di non campionamento sulle stime. Questo va detto per obiettività e per non coltivare illusioni apportatrici di discredito e di sfiducia in tecniche che, se ben comprese ed opportunamente utilizzate, possono rivelarsi utili per studiare le caratteristiche di alcune popolazioni delle quali altrimenti si rischierebbe di non poter dire nulla.

2. *IL CAMPIONAMENTO CHE SFRUTTA I LEGAMI SOCIALI: LA CORNICE DI RIFERIMENTO*

1. La terminologia: uno, nessuno, centomila

Una rassegna della letteratura sulle diverse tecniche di costruzione del campione che utilizzano i legami sociali (Thompson, 1997; 2002) rivela una realtà multiforme che rimanda ad una varietà di procedure di campionamento molto diverse fra loro a seconda della modalità di scelta delle persone che fanno parte del gruppo iniziale (semi); del numero di volte che il processo viene ripetuto (ondate) e delle modalità di selezione delle persone che entrano nel campione ad ogni ondata. Alcune brevi considerazioni serviranno a dare un'idea della varietà dei disegni. Immaginiamo un disegno campionario in cui gli individui che fanno parte del gruppo iniziale di persone (semi) siano scelti in modo casuale da una lista di campionamento. Ad ognuno di questi viene chiesto di indicare altre persone che conoscono con le caratteristiche richieste. Tutte le persone indicate entreranno nel campione; saranno a loro volta intervistate e si chiederà loro di indicare altri conoscenti con le caratteristiche richieste. Il processo sarà ripetuto finché non viene fornito alcun nuovo soggetto.

A questo disegno si possono applicare diverse varianti. Per quanto riguarda i semi essi potrebbero essere scelti non casualmente: ad esempio sulla base di una scelta ragionata. Per quanto riguarda le persone indicate in ogni ondata dai soggetti che entrano a far parte del campione, non è necessario che tutte quelle indicate entrino a loro volta nel campione e siano intervistate; si potrebbe

sceglierne un numero limitato (al limite solo una) sulla base di una selezione casuale oppure ragionata. Per quanto riguarda il numero di ondate il processo potrebbe essere interrotto solo dopo alcuni passaggi (al limite alla prima ondata).

I termini associati alle diverse procedure sono anch'essi molto diversi e rischiano di creare più confusione che ordine. Nella letteratura specialistica in lingua inglese si possono incontrare termini come: *snowball sampling* (Goodman, 1961; Kalton - Anderson, 1986; Snijders 1992; Frank - Snijders, 1994), *network sampling* (Birnbaum - Sirken, 1965; Granovetter, 1976; Sirken, 2006), *multiplicity sampling* (Birnbaum - Sirken, 1965), *random walk* (Klovdahl, 1989), *adaptative cluster sampling* (Thompson, 2002), *chains* (Erikson, 1979), *respondent-driven sampling* (Heckathorn, 1997), *targeted sampling* (Watters - Biernacki, 1989).

L' elevato numero di termini non sarebbe di per sé un problema se le definizioni fossero chiare, precise ed univoche. L'aspetto problematico è che spesso lo stesso termine viene utilizzato per definire cose totalmente diverse.

Si consideri, ad esempio, il termine *snowball sampling* che può riferirsi sia a procedure che obiettivi diversi. In Kalton and Anderson (1986) con *snowball* si intende una procedura che si articola nei seguenti passi: viene estratto un campione iniziale non probabilistico e si chiede ad ogni membro di nominare il massimo numero di conoscenti, il tutto ripetuto fino a quando nessun nuovo nominativo viene suggerito. Il fine è quello di costruire una lista dei membri di una popolazione affetta da malattie rare, a partire dalla quale estrarre, in un secondo momento, un campione di tipo probabilistico. In Goodman (1961) con *snowball* si intende una procedura che si articola, invece, in questo modo: viene estratto un campione casuale di individui (si possiede pertanto la lista) ai quali si chiede di identificare un numero fisso di altri individui, i quali a loro volta segnalano altre persone; il tutto ripetuto per un numero prefissato di ondate. Il fine è quello di fare inferenza riguardo vari aspetti delle relazioni presenti nella popolazione come il numero di "mutue relazioni" o "circoli sociali". In Frank - Snijders (1994) con *snowball* si intende una procedura che si articola in un modo ancora diverso: ad un campione ragionato di tossicodipendenti ottenuto a partire dalle liste dagli archivi della polizia e degli assistenti sociali di una città

olandese (l'estrazione è fatta a partire da un sottoinsieme della popolazione dei tossicodipendenti) viene chiesto di elencare i nomi di altri tossicodipendenti (ci si ferma alla prima ondata). Il fine è quello di stimare, a partire dai nomi segnalati e dalle loro ripetizioni, il numero in valore assoluto di tossicodipendenti nella città.

Il termine *snowball sampling* non è, comunque, l'unico il cui utilizzo può essere fonte di confusione, anche quello di *network sampling* non è usato in letteratura in modo univoco. Con esso si può indicare sia una forma di campionamento introdotta per lo studio di popolazioni rare (Birnbaum - Sirken, 1965) che una forma di campionamento che mira a raccogliere informazioni sul numero medio di persone conosciute da ogni persona e sulla "densità delle reti" (Granovetter, 1976).

A questo punto una modalità di presentazione potrebbe consistere in un lungo elenco di termini, in una riflessione sul loro utilizzo nei diversi autori. Questa soluzione risulta dispersiva e poco fruttuosa. Le diverse strategie di campionamento che sfruttano i legami sociali verranno, invece, ricondotte ad un'unica formalizzazione matematica: il campionamento in grafo. Questa formalizzazione è in grado di fornire un quadro concettuale parsimonioso e coerente per affrontare le diverse problematiche sollevate dalla riflessione metodologica sull'uso dei legami sociali per la costruzione di campioni casuali, in particolare nel caso, centrale in questa tesi, di popolazioni di cui non si possiede la lista di campionamento.

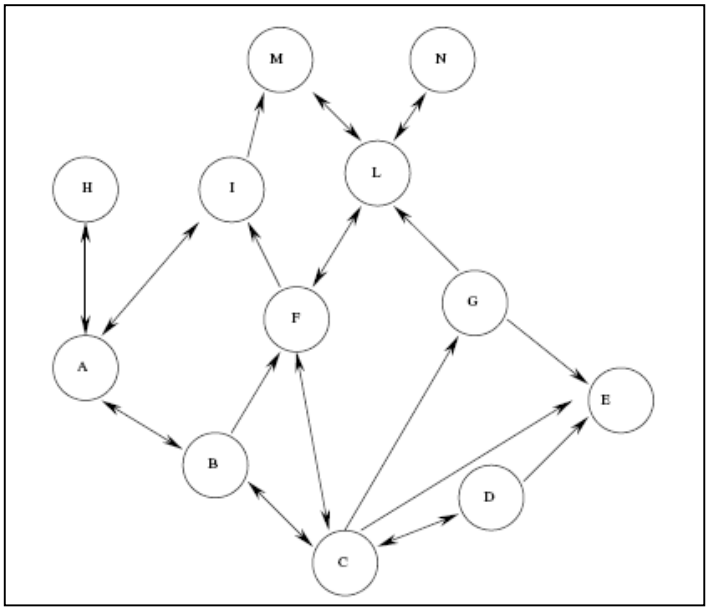
2. La formalizzazione: il campionamento in grafo

Nonostante l'estrema varietà, tutti i disegni di campionamento che sfruttano i legami fra persone possono essere ricondotti ad un'unica formalizzazione matematica: essi possono essere considerati come esempi di campionamento in grafo (; Wasserman - Faust, 1994; Thompson, 2002, 2006).

In matematica per grafo si intende una struttura costituita da: a) oggetti semplici, detti nodi; b) collegamenti tra i nodi, detti archi; c) da eventuali informazioni associate ai nodi e/o ai collegamenti (si veda appendice 2, per una definizione formale). In una rete sociale i nodi corrispondono alle persone e gli archi rappresentano relazioni, legami sociali del tipo più diverso (abitanti di uno stesso condominio, amici, compagni di scuola, membri dello stesso club, partecipanti ad uno stesso forum in internet).

Il grafo viene generalmente rappresentato, in modo visivamente molto immediato, con un disegno (fig.1) in cui ai nodi si fanno corrispondere cerchi (o altre figure geometriche) e agli archi segmenti che collegano i cerchi. Due nodi i e j possono essere uniti da una segmento con doppia freccia (< --- >) che rappresenta una relazione di tipo simmetrico (es. abitanti di uno stesso condominio) oppure da una freccia orientata da i verso j (--- >) oppure da j verso i (< ----), che rappresenta una relazione di tipo asimmetrico (es. essere figlio di).

Fig. 1 – Rappresentazione grafica di un grafo costituito da 12 nodi legati da relazioni sia simmetriche che asimmetriche



Nota: le frecce bidirezionali indicano relazioni di tipo simmetrico (se C nomina F, F nomina C), mentre le frecce direzionali indicano relazioni di tipo asimmetrico (se C nomina G, G quando intervistato non nomina C)

Le relazioni fra i nodi di un grafo possono essere rappresentate, oltre che con un disegno, anche con una matrice, detta di adiacenza binaria, costituita da N righe per N colonne in cui ogni elemento a_{ij} è uguale a 1 se c'è una relazione fra il nodo i e il nodo j e 0 altrimenti (tab. 1). Gli elementi sulla diagonale sono per convenzione fissati a 0. Per il nodo i, la somma dei valori sulla riga $a_{i.}$ fornisce il numero di nodi ai quali i è legato, detto "grado in uscita" (il numero di persone che conosce), mentre la somma dei valori per colonna $a_{.j}$ fornisce il numero di nodi che sono legati a i, detto "grado in entrata" (il numero di persone dalle quali è conosciuto).

Tab. 1 – Matrice di relazioni binarie che rappresenta le relazioni di figura 1

i \ *j*	***A***	***B***	***C***	***D***	***E***	***F***	***G***	***H***	***I***	***L***	***M***	***N***	$a_{i\cdot}$
A	0	1	0	0	0	0	0	1	1	0	0	0	**3**
B	1	0	1	0	0	1	0	0	0	0	0	0	**3**
C	0	1	0	1	1	1	1	0	0	0	0	0	**5**
D	0	0	1	0	1	0	0	0	0	0	0	0	**2**
E	0	0	0	0	0	0	0	0	0	0	0	0	**0**
F	0	0	1	0	0	0	0	0	1	1	0	0	**3**
G	0	0	0	0	1	0	0	0	0	1	0	0	**2**
H	1	0	0	0	0	0	0	0	0	0	0	0	**1**
I	1	0	0	0	0	0	0	0	0	0	1	0	**2**
L	0	0	0	0	0	1	0	0	0	0	1	1	**3**
M	0	0	0	0	0	0	0	0	0	1	0	0	**1**
N	0	0	0	0	0	0	0	0	0	1	0	0	**1**
$a_{\cdot j}$	**3**	**2**	**3**	**1**	**3**	**3**	**1**	**1**	**2**	**4**	**2**	**1**	

3. Le probabilità di inclusione e le popolazioni senza lista: una ridefinizione del problema in termini di campionamento in grafo

Dopo aver formalizzato la rete sociale di una popolazione attraverso il grafo si può formulare con maggiore precisione il problema centrale di questa tesi con la seguente domanda: «E' possibile selezionare, utilizzando gli archi (le relazioni), un sotto-insieme di nodi di una popolazione di cui non si abbia la lista, attraverso l'uso di procedure che permettano di calcolare le probabilità di ogni unità di entrare a far parte del campione e, di conseguenza, consentano di stimare, in modo non distorto e con un livello di precisione desiderato, alcune caratteristiche della popolazione stessa, quali ad esempio la distribuzione percentuale del titolo di studio o l'età media?».

Questa domanda merita alcuni commenti. Semplificando al massimo si può affermare che le procedure di costruzione di campioni che sfruttano i legami sociali si possono porre due obiettivi conoscitivi distinti:

1) «identificare, misurare e testare ipotesi riguardo la forma strutturale e ai contenuti sostantivi delle relazioni sociali tra attori» (Knoke – Yang, 2008, p. 4) . Questo obiettivo caratterizza l'approccio di rete (*network analysis)*[7];

2) identificare, misurare e testare ipotesi riguardo la distribuzione di variabili individuali degli attori; anche quando si studiano proprietà del collettivo esse derivano, nella maggioranza dei casi, da proprietà dei singoli componenti il gruppo.

La differenza fra le due prospettive è sintetizzata in Chiesi (1996, p. 60): «L'approccio classico della survey tratta l'individuo come un soggetto isolabile dal suo contesto sociale, di cui conserva alcune caratteristiche, le quali vengono concepite in modo statico, come degli attributi, che prendono la forma di variabili individuali e alcune di queste riguardano eventualmente la disposizione dell'individuo a intrattenere rapporti sociali. L'approccio di rete, invece, rovescia la prospettiva: lo scopo è di raccogliere dati su qualche tipo di relazione concreta tra individui e stimare le caratteristiche individuali sulla base della forma delle relazioni».

Considerando la storia del campionamento a valanga come strumento che permette l'inferenza statistica, le diverse procedure di costruzione del campione sono state introdotte e formalizzate matematicamente (Coleman, 1958; Goodman, 1961) per identificare, misurare e testare ipotesi riguardo alla forma strutturale e ai contenuti sostantivi delle relazioni sociali tra attori[8].

Le diverse tecniche di costruzione del campione oggetto di studio in questa tesi, si discostano da questo approccio e mirano a studiare le caratteristiche delle persone o, più in generale, dei nodi della rete. La prospettiva con cui si considera il campionamento rimanda, pertanto, alla tradizione di ricerca centrata sulle variabili e sugli individui e non sulle caratteristiche della rete; in altre parole si vogliono usare i campioni costruiti sfruttando i legami sociali per fare stime che abitualmente si fanno con i campioni tradizionali.

[7] Per un approfondimento in lingua italiana circa l'approccio di rete si rimanda ai già citati Chiesi, 1999; Mutti, 1996, Gribaudi, 1996; Piselli, 1995; Mattioli, 1995.

[8] Anche il classico contributo di Granovetter (1976) si situa nell'ambito dell'approccio di rete.

Stabiliti, nel primo capitolo, alcuni punti fermi circa il campionamento probabilistico, ricondotte, in questo secondo capitolo, tutte le forme di campionamento che sfruttano i legami sociali a forme di campionamento in grafo possiamo finalmente cominciare a rispondere alla domanda posta all'inizio del paragrafo.

Semplificando, si possono individuare due strategie di risposta: la prima, che si può definire "approccio tradizionale", sfrutta una proprietà dei campioni che si sviluppano a partire dai primi soggetti campionati (semi) estratti casualmente dalla popolazione (cap. 3); la seconda sfrutta alcune proprietà di un processo di selezione dei nodi che può essere descritto formalmente come una catena markoviana (cap. 4).

3. Il campionamento che sfrutta i legami sociali: l'approccio tradizionale

Tutte le forme di campionamento che verranno presentate in questo capitolo si fondano sul seguente principio generale: se si estraggono i primi nodi in modo casuale e poi si amplia il campione sfruttando in modo appropriato i legami di questi nodi è possibile calcolare la probabilità di inclusione dei nodi, requisito che, come abbiamo sottolineato nel cap. 1, par. 2, caratterizza la possibilità di applicare gli strumenti dell'inferenza statistica ai campioni selezionati (Frank, 1977; Frank - Snijders, 1994; Thompson, 2002; Sirken, 2006).

Si potrebbe obiettare che la possibilità di estrarre un campione iniziale casuale implica la presenza di una lista di campionamento della popolazione oggetto di studio, che non possediamo. Questo è vero, però, e qui sta l'intuizione alla base delle strategie campionarie esaminate in questo capitolo, non è necessario che i primi soggetti selezionati vengano estratti dalla lista della popolazione di riferimento (che non possediamo). Si immagini che la popolazione di riferimento sia legata ad una seconda popolazione di cui possediamo la lista di campionamento ("lista ausiliaria"). Quest'ultima può essere utilizzata per selezionare un campione probabilistico iniziale e successivamente si può opportunamente estendere la selezione dei soggetti sulla base di appropriati legami sociali fra le due popolazioni. A questo punto della trattazione, le affermazioni appena fatte possono risultare vaghe e non comprensibili; un loro chiarimento richiede alcune precisazioni circa il modello matematico che formalizza questa strategia.

1. Relazioni asimmetriche versus relazioni simmetriche

Come sottolineato in precedenza, un campione che legittima l'utilizzo delle procedure inferenziali fornite dalla teoria delle probabilità richiede che ogni individuo abbia una probabilità nota e non nulla di entrare a far parte del campione. Le probabilità possono essere differenti per i diversi individui. In questo caso si devono bilanciare, nel procedimento di stima, con pesi adeguati.

Immaginiamo di seguire la seguente strategia di campionamento che si ripete per tre ondate. Per semplicità argomentativa ipotizziamo di possedere la lista di campionamento della popolazione oggetto di studio e che il numero di nodi campionati sia marcatamente inferiore a quello della popolazione. Nella prima ondata selezioniamo un insieme di nodi (i semi) casualmente. Dopodiché tutti i nodi legati ai semi da relazioni in uscita (freccia →) entreranno a far parte del campione. Nella seconda ondata entreranno nel campione tutti i nodi legati da relazioni in uscita (freccia →) ai nodi individuati nella prima ondata; ed infine, nella terza ed ultima ondata, entreranno nel campione tutti i nodi legati da relazioni in uscita (freccia →) ai nodi individuati nella seconda ondata. La procedura di generazione del campione a questo punto si interrompe.

Non è difficile immaginare che, se il numero di nodi campionati in un grafo è inferiore a quello della popolazione, nodi con un alto grado in entrata hanno probabilità maggiori di essere inclusi nel campione rispetto a nodi con gradi in entrata inferiori (in una popolazione reale questi ultimi hanno meno persone che potrebbero indicarli nelle varie ondate[9]). E' possibile calcolare le probabilità di inclusione dei nodi-soggetti individuati con la procedura presentata, ripetuta per tre ondate?

La risposta è, in linea teorica, affermativa a patto di conoscere il grafo della popolazione, cioè di conoscere la struttura dei legami (chi è legato a chi) e se

[9] Solo quando tutti i nodi della popolazione sono campionati, la sovrarappresentazione dei nodi con più elevati gradi di entrata cessa: infatti viene selezionata tutta la popolazione è tutti nodi hanno probabilità di entrare a far parte del campione uguale a 1. Comunque quest'ultima osservazione rischia di risultare, oltre che lapalissiana, di scarso valore pratico, dal momento che lo sforzo di selezionare tutti i membri di una popolazione non è realizzabile nella stragrande maggioranza delle situazioni.

questi sono di tipo asimmetrico o simmetrico. Tale conoscenza è una condizione proibitiva e difficilmente riscontrabile nella realtà. Scartiamo questa possibilità.

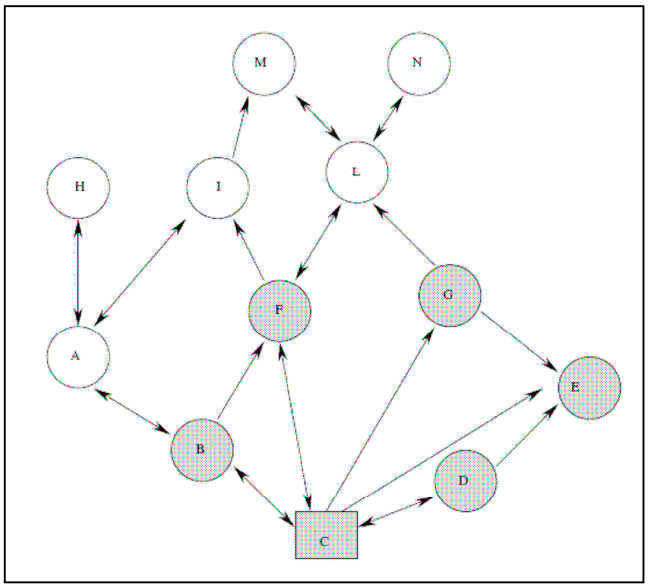
La domanda successiva che ci si può porre è la seguente: «E' possibile calcolare le probabilità di entrare a far parte del campione sulla base dei soli dati campionari, senza pertanto conoscere la struttura delle relazioni nella popolazione, ma solo la struttura della relazioni all'interno del campione? ».

La risposta a questa domanda non è immediata e richiede di approfondire la comprensione del tipo di legame che unisce i diversi nodi in un grafo: simmetrico o asimmetrico.

Semplifichiamo l'esempio precedente. Si immagini un campione che si sviluppa in due ondate: la prima, in cui si estrae casualmente un seme e la seconda in cui si selezionano tutti i nodi che sono legati al seme. Questa situazione estremamente semplificata permette di mantenere la trattazione ad un livello elevato di chiarezza. Le deduzioni, le considerazioni si possono comunque estendere anche a situazioni più generali. Per rendere più concreto il ragionamento applichiamo questa procedura di campionamento al grafo rappresentato in forma diagrammatica nel precedente capitolo. Estraiamo casualmente un seme (nella realtà potrebbe essere un studente) e seguiamo tutte le frecce in uscita che conducono agli altri nodi (nella realtà questo potrebbe corrispondere a chiedere allo studente di indicare il nome dei suoi amici di scuola); tutti i nodi individuati entrano a far parte del campione. Il processo è raffigurato in fig. 1 dove, inizialmente, è stato selezionato casualmente il nodo C, il quale a sua volta punta verso i nodi B, D, E, F, G.

Qual è la probabilità di essere incluso nel campione del nodo C?

Fig. 1 - Campione selezionato a partire da nodo C



Nota: i cerchi grigi indicano i nodi, verso cui punta il nodo C (rettangolo grigio) e che entrano nel campione (nella realtà segnalati ed intervistati) e i cerchi bianchi sono nodi di cui non conosciamo l'esistenza. Le frecce bidirezionali indicano relazioni di tipo simmetrico (se C nomina F, F nomina C) mentre le frecce direzionali indicano relazioni di tipo asimmetrico (se C nomina G, G quando intervistato non nomina C)

Il nodo C può entrare a far parte del campione se è estratto direttamente come seme oppure indirettamente se uno dei nodi B, D, F è stato scelto come seme iniziale. In termini di campionamento in grafo la probabilità di C di entrare a far parte del campione dipende dal suo grado in entrata. In termini più concreti la probabilità dell'unità C di venire a far parte del campione dipende dal numero di unità nella popolazione che, una volta estratte ed intervistate, faranno il suo nome. In linea del tutto teorica, in una ricerca, per conoscere questo valore si dovrebbe porre agli intervistati una domanda del tipo: «Quante altre persone farebbero il tuo nome se richieste? ». E' subito chiaro che solo in rarissimi casi ci potremmo aspettare che una domanda di questo tipo fornisca risposte valide ed affidabili.

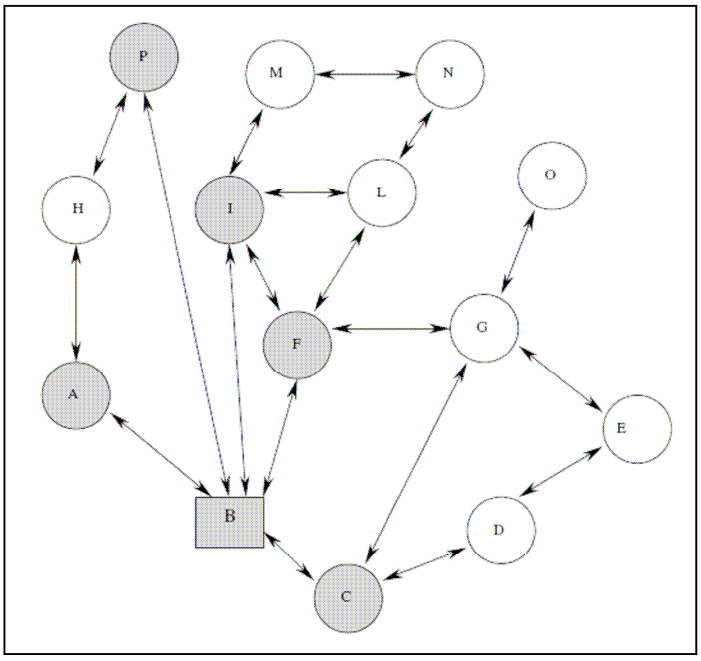
Cosa possiamo concludere allora a partire dall'esempio? Che è estremamente difficile, se non impossibile, calcolare le probabilità di inclusione a partire dai dati campionari per ogni unità nel campione. A cosa è dovuta questa difficoltà? Essa è una conseguenza delle asimmetrie di alcuni legami fra i nodi, così che, relativamente ad alcune unità nel campione, il ricercatore non sa quante altre

unità nella popolazione avrebbero potuto potenzialmente indirizzare il ricercatore verso di loro.

Esiste una soluzione a questo problema, un escamotage che ci permetta di aggirarlo? Un tentativo di soluzione consiste nel considerare, per la selezione dei soggetti nelle ondate successive alla prima, solo legami di tipo simmetrico.

Si osservi il grafo rappresentato in fig. 2: esso contiene solo legami simmetrici. In una ricerca questo risultato si può raggiungere ricostruendo una rete sociale a partire da domande che fanno emergere solo legami simmetrici (nella realtà, ad es., si potrebbe chiedere ad uno studente di nominare i suoi compagni di classe). In questo caso il nodo B può entrare a far parte del campione, direttamente, se è estratto come seme oppure, indirettamente, se uno dei nodi A, C, F, I, P è stato scelto come seme. La probabilità di B di entrare a far parte del campione dipende ancora dal suo grado in entrata, ma in un grafo contente solo relazioni simmetriche (definito "grafo non diretto") il grado in entrata è uguale a quello in uscita. In termini più concreti la probabilità del nodo B di venire a far parte del campione dipende dal numero di nodi nella popolazione che indicherà. Per conoscere questo valore, in una ricerca, si potrebbe porre all'intervistato una o più domande circa il numero di persone che conosce, frequenta e che hanno le caratteristiche volute dal ricercatore. Come avremo modo di approfondire ampiamente in seguito, questo tipo di domande, pur non essendo esente da problemi di validità ed affidabilità delle risposte, è molto utilizzato nella ricerca sulle reti sociali.

Fig. 2 - Campione selezionato a partire da nodo B



Nota: i cerchi grigi indicano i nodi, verso cui punta il nodo B (rettangolo grigio) e che entrano nel campione (nella realtà segnalati ed intervistati) e i cerchi bianchi sono nodi di cui non conosciamo l'esistenza. Le frecce biderizionali indicano una relazione d tipo simmetrico (se uno viene nominato, nominerà, una volta intervistato, la persona che lo ha nominato).

2. Dalla sintassi alla semantica: le condizioni di applicazione nella ricerca sociale

Per capire come le considerazioni un po' astratte del paragrafo precedente possano tradursi in qualcosa di concreto ed utilizzabile per un ricercatore in scienze sociali verranno descritti due esempi, estremamente semplificati, di possibili indagini campionarie. Successivamente questi esempi verranno analizzati in dettaglio per individuare le caratteristiche comuni. Queste ultime non sono altro che le condizioni necessarie per poter calcolare le probabilità di inclusione.

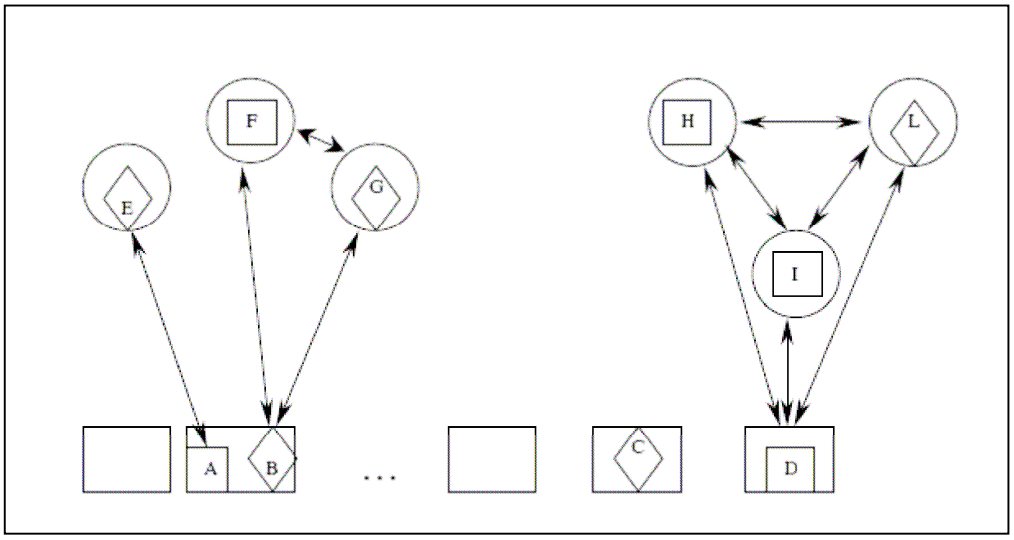
2.1. Primo esempio: un'indagine sulla popolazione affetta da diabete

In un'indagine che si propone di studiare la popolazione di persone affette da diabete in una città, un campione casuale semplice di un centinaio di

famiglie è selezionato a partire dalla lista anagrafica e agli adulti che risiedono nelle famiglie selezionate viene chiesto di rispondere se essi stessi soffrono di diabete oppure se qualcuno dei loro fratelli e sorelle che vive nella stessa città (nel caso esistano) soffre della stessa malattia. Le famiglie sono le unità di selezione, le persone adulte sono le unità di rilevazione. Il grafico di fig. 3 rappresenta le relazioni riscontrate nel campione.

- In novantasette famiglie selezionate (indicate da rettangoli bianchi in fig. 3) nessun componente ha la malattia e non possiede fratelli o sorelle in città. Nel caso di tre famiglie la situazioni è più interessante: o un componente soffre di diabete oppure uno o più dei suoi fratelli o sorelle che vive in città in un'altra famiglia soffre di diabete (rombi in fig. 3).
- Nella famiglia composta da A e B, A non soffre di diabete ma ha un fratello, E, che vive in città in un'altra famiglia e che soffre di diabete; B, invece, ha la malattia; inoltre dei suoi due fratelli (F, G) che vivono in diverse famiglie in città, G soffre di diabete, mentre F no.
- C vive in famiglia da solo, soffre di diabete ed è figlio unico.
- D vive da solo, ha tre fratelli H, I, L che vivono nella stessa città, ciascuno in famiglie diverse. Fra i fratelli uno solo soffre di diabete ed è L.

Fig.3 - Rappresentazione grafica del primo esempio



Nota: i rettangoli indicano famiglie selezionate nel campione iniziale. I cerchi indicano fratelli /sorelle che vivono in altre famiglie nella stessa città. Il rombo indica persone affette dalla malattia, mentre il quadrato persone non affette dalla malattia. La doppia freccia indica un legame simmetrico (essere fratello o sorella).

2.2. Secondo esempio: un'indagine sui bambini e la TV

Una ricerca si propone di stimare il numero di ore che i bambini dai sei agli unici anni di una città passano davanti alla TV. La lista di campionamento che possiamo utilizzare è la lista elettorale. Quest'ultima però contiene solo l'elenco delle persone maggiorenni residenti in città e non quello dei ragazzi dai sei agli undici anni. In ogni caso si estrae un campione di un migliaio di nominativi, ai quali si chiede la composizione della loro famiglia (si intende rilevare tutte le persone che coabitano abitualmente con l'intervistato) e la loro rispettiva età. Si individueranno così un certo numero di famiglie, di cui fanno parte alcuni bambini dai sei agli undici anni. In questo caso si cercherà di contattare all'interno della famiglia una persona che sia in grado di rispondere circa il numero di ore che i bambini passano davanti alla TV. Alla fine dell'indagine sono stati individuati circa duecento bambini dai sei agli undici anni, relativamente a ciascuno dei quali è riportato il numero di ore passato davanti alla Tv.

2.3 Le condizioni per il calcolo delle probabilità di inclusione

Cerchiamo di individuare le caratteristiche comuni che contraddistinguono questi due esempi. Esse possono essere riassunte in cinque punti.

1. Non si possiede la lista della popolazione oggetto di studio; è disponibile, comunque, una lista di campionamento (lista ausiliaria), a partire dalla quale è possibile estrarre in modo casuale un campione iniziale di casi (semi); nel primo esempio era la lista anagrafica, nel secondo esempio la lista elettorale.
2. Utilizzando opportune domande è possibile ricostruire "una" rete sociale (grafo). L'articolo indeterminativo è evidenziato per rimarcare che, per una stessa popolazione, a differenti domande corrispondono differenti reti sociali (grafi). Ad ogni individuo campionato inizialmente (seme) viene

chiesto di indicare, nel primo esempio, se ha uno o più fratelli/sorelle che vivono in città in una famiglia diversa dalla sua; e, nel secondo esempio, gli altri membri della sua famiglia.

3. Le relazioni individuate dalle domande sono di tipo simmetrico. La relazione è definita, nel primo esempio, come "essere fratello/sorella di", mentre, nel secondo esempio, come "coabitare con persone cui si è legati da legami di parentela".
4. Sulla base delle domande poste e' possibile dividere la popolazione oggetto di studio in un numero mutuamente esclusivo di reti sociali, chiamate "grappoli". Nel primo esempio gli individui sono raggruppabili in grappoli di fratelli che vivono in famiglie diverse (chi non ha fratelli costituisce anch'esso un grappolo, è fratello di se stesso), nel secondo esempio la popolazione oggetto di studio si può dividere in grappoli di persone coabitanti (famiglie). Ogni persona della popolazione oggetto di studio può appartenere solo ad un grappolo sia esso un gruppo di fratelli che vivono in famiglie diverse oppure un gruppo di persone coabitanti legate da legami di parentela. Considerando il primo esempio si avranno novantotto grappoli di un'unità (di cui uno, C, affetto da diabete); un grappolo di due unità (A, E); un grappolo di tre unità (B, F, G) ed, infine un grappolo di quattro unità (D, I, H, L).
5. Tutte le persone che fanno parte di un grappolo sono individuate in un numero di ondate finito: in entrambe gli esempi alla prima ondata (il numero di ondate potrebbe essere superiore).

Queste cinque condizioni permettono il calcolo delle probabilità di inclusione per ogni soggetto selezionato. Quest'ultima, tenendo in considerazione i ragionamenti fatti in precedenza (par. 1 in questo capitolo), è proporzionale al numero di persone che formano il grappolo di cui il soggetto selezionato fa parte. Se una persona ammalata di diabete appartiene ad un grappolo costituito da quattro unità (es. L in fig. 3) avrà una probabilità di entrare nel campione quattro volte maggiore di chi appartiene ad un grappolo di un solo elemento (es. C in fig. 3). Così, nel secondo esempio, un bambino di sei anni che vive in

una famiglia costituita dai suoi genitori, dal nonno paterno e una prozia materna avrà una probabilità di essere incluso nel campione quattro volte superiore a quella di un bambino che vive con un solo genitore.

Più precisamente la probabilità di inclusione di un'unità della popolazione oggetto di studio è data dalla probabilità che il grappolo sia intersecato dal campione iniziale. Per un intervistato inserito in un grappolo di una persona, la probabilità di intersezione è semplicemente la probabilità di essere estratto nel campione iniziale. Per grappoli più ampi la probabilità di intersezione è data dalla probabilità che una o più delle unità del grappolo siano incluse nel campione iniziale.

Da un punto di vista statistico (o sintattico come scritto nel primo capitolo), stime campionarie delle variabili di interesse possono essere a questo punto calcolate, utilizzando come pesi i valori delle probabilità di inclusione negli usuali stimatori di medie e percentuali (per un approfondimento formale cfr. appendice 3).

3. Il *multiplicity sampling* : l'applicazione nella ricerca

Esiste una strategia campionaria consolidata in letteratura che capitalizza sulle considerazioni appena enunciate: il *multiplicity sampling* . Questo tipo di campionamento è stato introdotto da Birnbaum - Sirken (1965) e si caratterizza per il fatto che i rispondenti campionati non rispondono solo per se stessi, ma vengono utilizzati quali informatori proxy per raggiungere (o per ottenere informazioni su) altre persone legate loro da una relazione chiaramente definita (es. fratelli), che non vivono nella stessa abitazione.

Una precisazione terminologica alla quale già si è accennato in precedenza è a questo punto doverosa per evitare fraintendimenti. Questa strategia di campionamento è spesso chiamata anche *network sampling*; essa non va confusa con quella studiata e considerata sotto lo stesso nome da Granovetter (1976) il quale si proponeva, a partire da un campione di individui estratti casualmente da una lista di campionamento, di stimare il numero medio di persone

conosciute da ogni persona campionata e di calcolare intervalli di confidenza attorno a queste stime.

Il *multiplicity sampling* è spesso utilizzato per lo studio di sottopopolazioni "rare", poco numerose. Una definizione univoca di cosa sia una sottopopolazione rara non esiste in letteratura; come regola generale si possono comunque considerare come "rare" sottopopolazioni che raggiungono il 10% o meno della popolazione. L'individuazione di procedure di campionamento probabilistico di queste popolazioni è un compito tutt'altro che semplice (per una sintetica rassegna si veda Kalton, 2009). La procedura più diffusa consiste nello "screening" di grandi campioni estrattati da liste di popolazione generali per individuare gli appartenenti alla popolazione di riferimento. Ad esempio, si estrae un campione di numeri telefonici, si contattano gli individui e si chiede se il rispondente o le persone che abitano con lui appartengono alla popolazione oggetto di studio. Se la popolazione di riferimento comprende una percentuale ragionevolmente elevata della lista di campionamento e i tassi di risposta sono sufficientemente elevati la strategia di screening risulta efficiente e i costi contenuti[10]. I costi comunque crescono rapidamente al diminuire della percentuale di membri della popolazione di riferimento nella lista di campionamento. Ad esempio, per selezionare un campione di 800 persone con un tasso di risposte del 60%, si deve sondare un campione iniziale di oltre 3.300 persone, nel caso la popolazione di riferimento ammonti al 40% della lista di campionamento. Se la percentuale si abbassa al 6%, il campione iniziale che deve essere sottoposto a screening supera le 20.000 unità.

Il *multiplicity sampling* in questi casi può rivelarsi molto utile: infatti è stato spesso utilizzato per espandere l'approccio standard di screening utilizzando le persone campionate come proxy per ottenere informazioni relativamente a persone che sono legate loro da una specifica e chiara relazione. Nella ricerca sono spesso utilizzati legami di parentela come essere genitore, fratello o figlio (Sudman *et al.* 1988; Sirken 2006).

[10] Se la percentuale della popolazione di riferimento è pari a p della lista di campionamento, la numerosità della popolazione target è uguale a n e il tasso di risposte è pari a r, il numero di unità da sondare n_screening per raggiungere la numerosità n voluta è uguale a: n_screening = n / (p x r) [Cervantes - Kalton, 2008, p. 115]

In uno studio pilota sui veterani del Vietnam, Rothbart *et al.* (1982) utilizzarono il *multiplicity sampling*. Inizialmente venne estratto un campione di nominativi telefonici che furono contattati. Una domanda standard di screening veniva posta nelle primissime fasi della chiamata per identificare se chi rispondeva era un veterano del Vietnam o meno. Gli intervistatori inoltre chiedevano se figli, fratelli o nipoti del rispondente erano veterani del Vietnam, se vivevano con il rispondente o in altre famiglie. I potenziali soggetti definiti come veterani del Vietnam venivano contattati direttamente (o telefonicamente o di persona) in un secondo momento per verificare la loro appartenenza alla popolazione di riferimento e, in caso affermativo, intervistati. Alla fine dell'indagine pilota furono identificati direttamente circa 500 veterani del Vietnam su 8700 interviste di screening fatte a partire dai numeri telefonici campionati (circa il 6% dei rispondenti al telefono) e indirettamente 476 veterani a partire dalle informazioni sui figli, fratelli e nipoti.

Simile è l'applicazione del *multiplicity sampling* fatta da Sudman e Freeman (1988). All'interno di un'indagine telefonica sull'accesso alle risorse sanitarie fu necessario sovracampionare un gruppo di persone con malattie croniche o molto gravi. Per raggiungere lo scopo, durante l'iniziale intervista telefonica al rispondente, si chiedevano informazioni per identificare lo stato di salute anche di eventuali coniugi, genitori, fratelli, nipoti. L'uso del *multiplicity sampling* permise di accrescere il numero di ammalati cronici o con malattie invalidanti, compresi nel campione, di circa il 30%.

Il *multiplicity sampling* non serve solo a potenziare l'approccio di screening. Alcune forme di relazioni sociali permettono di raggiungere individui che non sono nella lista di campionamento originale .

Per esempio, Brick (1990) ha descritto un'indagine pilota per verificare la possibilità di utilizzare il *multiplicity sampling* all'interno di un'indagine nazionale telefonica, la *National Household Education Survey* (NHES), per accrescere il campione di soggetti di 14-21 anni che vivevano in abitazioni senza telefono e che avevano interrotto gli studi (*drop-outs*). In un sottocampione di famiglie, a tutte le donne dai 28 ai 65 anni venivano chieste informazioni sui loro figli dai 14 ai 21 anni, che non vivevano in famiglia. Alcuni dei figli

vivevano in abitazioni con il telefono e alcuni in abitazioni senza telefono. Mentre i primi avevano una probabilità doppia di entrare a far parte del campione rispetto ai ragazzi che vivevano con la madre in un'abitazione con il telefono, i secondi non avrebbero avuto alcuna possibilità di entrare nel campione considerando come lista di campionamento l'elenco dei numeri telefonici, a meno di utilizzare opportunamente i legami sociali dei rispondenti (in questo caso essere figlio di). Il disegno campionario riuscì a ridurre la non copertura dei giovani dai 14 ai 21 anni che vivevano in abitazioni senza telefono e che avevano interrotto la scuola (*drop-outs*) dal 30%[11] al 19%.

Sempre considerando la possibilità di applicare il *multiplicity sampling* nell'ambito delle indagini telefoniche si può annoverare tra i più recenti ed articolati lavori quello di Tortora *et al.* (2008). Gli autori hanno cercato di testare il *multiplicity sampling* per costruire campioni rappresentativi di persone che non sono abbonate a linee telefoniche fisse e utilizzano solo il cellulare. Questa indagine sarà ripresa in modo analitico successivamente nella sezione sulle considerazioni di ordine metodologico perché è, per molti aspetti, emblematica delle sfide poste dalla trasformazioni tecnologiche e sociali alle forme di campionamento tradizionali e delle possibilità, ma soprattutto dei limiti, delle strategie di campionamento basate su legami sociali per rispondere ad esse.

I legami di parentela non sono l'unico tipo di legame che è stato sfruttato e studiato. Nell'ambito di disegni campionari che ricadono nella categoria generale del *multiplicity sampling* una varietà di legami è stata esaminata per lo studio di popolazioni rare (Sirken, 1970, 2006; Bergsten - Pirson, 1982; Kalton - Anderson, 1986; Sudman *et al.*, 1988; Czaja - Blair, 1990; Thompson, 2002). Si considerano per la loro rilevanza due di essi: il legame di amicizia e quello di vicinato. Il ricorso al legame di amicizia risulta particolarmente allettante in sede di ricerca perché garantisce una rapida espansione delle reti e un'elevata copertura delle popolazioni rare. Esiste, comunque, il rovescio della medaglia: il ricorso ai legami di amicizia presenta problemi per il calcolo delle probabilità di inclusione, condizione *sine qua non* (cap. 1, par. 2) per lo sviluppo di forme di

[11] Questo valore era una stima calcolata a partire dai dati di un'altra indagine: *Current Population Survey*.

campionamento probabilistico. Spesso il legame di amicizia non è reciproco: la persona A può nominare come suo amico stretto la persona B, ma quest'ultima quando intervistata può non nominare A. La non reciprocità della relazione crea problemi nel calcolo delle stime perché, come abbiamo visto nel par. 1 di questo capitolo, alcuni individui entrano nel campione con probabilità sconosciute.

Sempre nell'ambito di ricerche volte a studiare caratteristiche rare, il legame di vicinato è stato utilizzato nella *Health Survey of England* (Erens *et al.*, 2001) e nella *British Crime Survey* (Bolling *et al.*, 2008) per sovracampionare minoranze etniche. La tecnica di *multiplicity sampling* che usa come legame quello di vicinato prende il nome di "*focused enumeration*" (Kalton, 2009).

La tecnica si articola nelle seguenti fasi. Si estrae un campione probabilistico di indirizzi. La lista di campionamento è costituita, pertanto, non da famiglie, ma da un elenco di indirizzi[12]. Si intervistano coloro che abitano all'indirizzo campionato e ci si accerta della loro appartenenza o meno alla popolazione oggetto di indagine. Inoltre si chiede all'intervistato di fornire informazioni sull'appartenenza alla popolazione studiata dei vicini definiti come coloro che abitano all'indirizzo direttamente precedente o successivo.

Nel caso all'indirizzo campionato non si riesca a trovare nessuno o che la persona si rifiuti di rispondere o affermi che non è in grado di dare informazioni sui vicini, l'intervistatore cerca di ottenere le informazioni dai vicini all'indirizzo campionato. L'obiettivo è ottenere le informazioni riguardo ai grappoli di tre famiglie: quelle che vivono all'indirizzo campionato e ai due indirizzi legati ad esso. Nel caso venga identificata una famiglia che potrebbe appartenere alla popolazione oggetto di studio, l'intervistatore cerca di intervistarla.

Per il suo valore di curiosità si può anche citare, nell'ambito di utilizzo del legame di vicinato per espandere l'approccio di screening, il lavoro di Bergsten e Pierson (1982), i quali chiesero ai rispondenti di identificare i loro vicini che

[12] In GB per le indagini campionarie è disponibile una lista di campionamento chiamata Postcode Address File che lista tutti gli indirizzi nazionali ai quali può essere inviata una lettera; essa ha una copertura del 100% degli indirizzi ed è regolarmente aggiornata (Gilbert, 2008, p. 174)

avevano un orto[13]. I ricercatori cercavano di studiare l'utilizzo, da parte dei possessori di un orto, di fango di liquami (*sewage sludge*) per la concimazione.

3.1 Una variazione sul tema: l'adaptative cluster sampling

Nell'ambito di disegni campionari che ricadono nella categoria generale del *multiplicity sampling* va segnalato anche un altro tipo di campionamento proposto in letteratura: l'*adaptative cluster sampling*, nel quale la selezione delle unità da includere nel campione dipende dai valori di una variabile di interesse osservata nei casi già selezionati (Thompson, 1997, 2002).

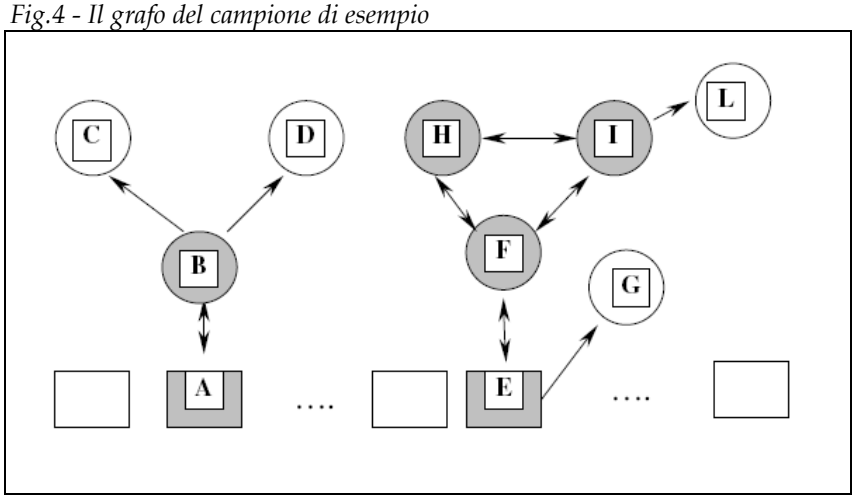
Per comprendere le caratteristiche dell'*adaptative cluster sampling* può essere utile un esempio. Si consideri un'indagine campionaria interessata a studiare la dipendenza da tabacco nei ragazzi che frequentano le scuole medie (12-14 anni) di una città. Si potrebbe inizialmente estrarre un campione casuale semplice di cinquecento ragazzi a partire dalla lista dei diecimila studenti delle scuole medie della città e, successivamente, intervistare gli estratti chiedendo loro se e quante sigarette fumano mediamente in una settimana. Questo disegno di campionamento non sarebbe però ottimale dal momento che il comportamento studiato è poco diffuso: verrebbero estratti pochi ragazzi fumatori. Per ottenere un campione più ampio di ragazzi fumatori è pero possibile sfruttare i legami sociali dei ragazzi che fumano, adottando un disegno *adaptative sampling*. Così, se e solo se, un ragazzo risponde di aver fumato delle sigarette nella settimana precedente l'intervista, gli viene chiesto il nome dei compagni di classe con i

[13] Più precisamente «Respondents living in single family dwellings were asked to identify vegetable gardens on the property next door to the right and to the left, but only for next door neighbors living in single family dwellings. Respondents living in structures containing 2 or 3 housing units were asked about the other dwelling(s) in the same structure. Respondents living in structures containing 4 or more units were not asked about their neighbors at all. It was thought that people living in large multifamily structures would be unlikely to have vegetable gardens, therefore there would have been little to gain from such querying. Even if this were not the case, the wide variety of housing configurations possible in large multifamily structures made defining a neighbor close enough to assure acquaintance very difficult. Respondents were asked only if their neighbors had vegetable gardens on their home property; no inquiries were made about gardens maintained at other locations. Respondents reporting a neighbor's vegetable garden were asked for the name, telephone number, and address of the referral household (Bergsten - Pierson, 1982, p.146) ».

quali ha discusso questioni personali nell'ultimo mese (non è necessario che indichi i compagni che fumano). I compagni nominati vengono a far parte del campione.

In fig. 4 è mostrato un possibile risultato della procedura. Il primo ragazzo (A) che ha affermato di fumare nomina un altro ragazzo (B) che, una volta intervistato, si rivela anch'esso un fumatore. Questa nuovo ragazzo a sua volta fornisce, oltre al nome di A, i nomi di due compagni di classe i quali dichiarano di non fumare (C e D). A questi ultimi, in base alla logica del *adaptative sampling*, non verrà richiesto di fornire i nominativi di altri compagni di classe. Il secondo ragazzo che fuma nel campione iniziale (E) nomina due nuovi compagni di scuola: uno dei quali , una volta intervistato, dichiara di non fumare (G) e l'altra sì (F). In base alla logica del *adaptative sampling*, non verrà richiesto a G di fornire i nominativi di altri compagni di classe. F riporta, oltre il nominativo di E, due nuovi nomi (H e I), entrambi fumatori. H una volta intervistato fornisce il nominativo di chi lo ha nominato e quello di I, mentre quest'ultimo segnala oltre il nominativo di F e di H, quello di L. Quest'ultimo si rivela un non fumatore e il processo si interrompe.

Fig.4 - Il grafo del campione di esempio



Nota. I rettangoli indicano persone selezionate nel campione iniziale, i cerchi indicano persone che sono nominate. Il colore bianco indica persone che non fumano e quello grigio indica persone che fumano. Le doppie frecce indicano un legame simmetrico mentre la freccia un legame asimmetrico.

E' ancora possibile calcolare in questo tipo di campionamento le probabilità di inclusione? Sì, infatti valgono le stesse condizioni che si verificano per il *multiplicity sampling*, in particolare:

- sulla base della relazione individuata dal legame ("compagni di classe con cui si è discusso questioni personali nell'ultima settimana") e' possibile dividere la popolazione oggetto di studio in un numero mutuamente esclusivo di grappoli di persone collegate fra loro da legami simmetrici;
- tutte le persone che fanno parte di un grappolo sono individuate in un numero di ondate finito.

Nell'esempio si avranno così 498 grappoli di un solo elemento (ragazzi che dichiarano di non fumare), 1 grappolo di due elementi (A e B) e 1 di quattro elementi (E, F, H, I). Nei grappoli non compaiono C, D, L, G dal momento che, in base alle regole di costruzione dell'*adaptative sampling*, queste unità è come se fossero legate alle altre da un legame asimmetrico e, pertanto, non sono considerate come elementi dei rispettivi grappoli.

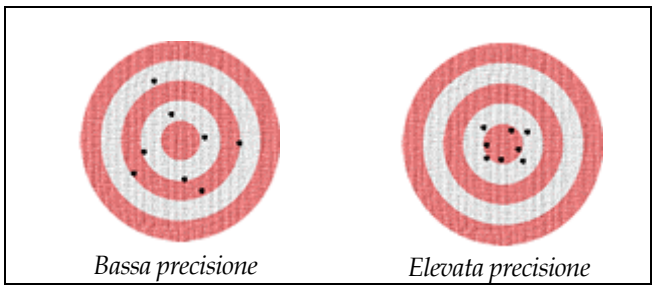
4. Il multiplicity sampling: considerazioni metodologiche

Dalla descrizione fatta fino ad ora potrebbe sembrare che il *multiplicity sampling* sia una soluzione semplice che non presenta problemi di applicazione maggiori rispetto a tecniche di campionamento tradizionali. Purtroppo le cose non stanno propriamente in questo modo. Come vedremo nei paragrafi successivi, il *multiplicity sampling* presenta tutte le problematiche associate alle forme di campionamento casuale tradizionali, alle quali ne aggiunge di specifiche legate alla sua natura.

4.1 Aspetti sintattici: l'effetto del disegno

Si è cercato di chiarire come esista, a determinate condizioni, la possibilità di ottenere campioni che si espandono sfruttando i legami sociali e che permettono di calcolare stimatori non distorti delle variabili di interesse. Tuttavia, la non distorsione di una stimatore non assicura che una particolare stima sia uguale al valore vero della popolazione. Proprio perché basiamo le nostre stime su campioni e non sull'intera popolazione, qualche volta le stime saranno troppo elevate, altre troppo basse; solo in media saranno uguali al valore reale della popolazione. Stando così le cose il fatto che uno stimatore sia non distorto non significa che sia "utile" in pratica. Per capire questo concetto si pensi alla seguente analogia dove i diversi tipi di campionamento sono dei tiratori che sparano ad un bersaglio con un fucile. Ogni colpo sparato corrisponde ad un campione estratto. Ogni tiratore dispone di otto colpi. Il centro del bersaglio è il parametro della popolazione (il valore reale che cerchiamo di stimare). Con ogni colpo-campione si colpirà il bersaglio più o meno lontano dal centro (fig. 5). I tiri tendono a distribuirsi in media attorno al centro di ciascuno dei bersaglio. E', pero, evidente una differenza. Il tiratore a sinistra colpisce un po' ovunque il bersaglio (fig. 5, a sinistra). I suoi colpi non sono precisi. Il tiratore a destra colpisce vicino al centro. I suoi colpi sono più precisi (fig. 5, a destra).

Fig.5 - Due rappresentazioni della precisione delle stime in un bersaglio



Un modo comune di valutare la precisione delle stime di un tipo di campionamento, in questo caso il *multiplicity sampling*, consiste nel confrontarlo con la precisione delle stime, che si ottiene con un campione casuale semplice di pari ampiezza campionaria. Usando l'analogia precedente è come se si

confrontasse il tiratore, *multiplicity sampling,* con il tiratore, campione casuale semplice. La perdita di efficienza risultante dall'uso del *multiplicity sampling* si può misurare con il fattore di disegno (deff) Kish (1965). Questa misura è definita come rapporto tra la varianza di uno stimatore ottenuto dal tipo di campionamento che si vuole valutare e quella ottenuta con un campione casuale semplice della stessa ampiezza.

$$\text{Deff} = \frac{\text{varianza campionaria dello stimatore (}\textit{multiplicity sampling}\text{)}}{\text{varianza campionaria dello stimatore (campione casuale semplice)}}$$

Generalmente, ma non sempre, l'effetto del disegno per un campionamento *multiplicity* è maggiore di 1 indicando che questo tipo di campionamento fornisce stime mediamente meno precise di quelle di un campione casuale semplice di pari numerosità (Thompson, 2002; Sirken, 2006)[14]. Questo risultato è coerente con la letteratura sui disegni campionari complessi, che generalmente accrescono la variabilità delle stime. Ci si potrebbe chiedere la ragione di questa minore efficienza. In termini intuitivi si può affermare che la maggiore inefficienza è legata al fatto che coloro che sono in un grappolo sono spesso molto simili fra loro in termini di caratteristiche studiate. Perché questa caratteristica comporta conseguenze negative? Per il motivo che la varianza delle stime campionarie è funzione della quantità di "informazione nuova" nel campione e non solo dell'ampiezza campionaria. Per illustrare questo punto si consideri l'esempio fatto per descrivere le caratteristiche dell'*adaptative cluster sampling*. I grappoli erano costituiti da "amici" appartenenti alla stessa classe e l'obiettivo dell'indagine consisteva nello studio della dipendenza al tabacco

[14] Il valore dell'effetto del disegno è importante perché permette di calcolare l'ampiezza campionaria effettiva che indica il numero di unità elementari che dovrebbero essere richieste in un campione casuale semplice per fornire la stessa precisione ottenuta con il campione prodotto dal *multiplicity sampling*. Essa è calcolata come rapporto fra la numerosità del campione multiplicity e l'effetto del disegno. Immaginiamo di aver estratto un campione di 1.000 casi utilizzando il *multiplicity sampling*. Ipotizzando che l'effetto del disegno, per la stima di una variabile, sia pari a 2 ci sarebbe bastato un campione di 500 casi in un campione casuale semplice per ottenere la stessa precisione, cioè lo stesso margine di errore della stima, di quello *multiplicity*.

nella popolazione di ragazzi fumatori di 12-14 anni. È probabile che ragazzi appartenenti alla stessa classe e che si scambiano confidenze abbiano comportamenti simili; di conseguenza campionando un membro supplementare dallo stesso grappolo si aggiunge meno "informazione nuova", circa i comportamenti oggetto di studio, di quanto avverrebbe con una selezione completamente indipendente.

Diverse ricerche sono state fatte per studiare l'effetto del disegno del *multiplicity sampling* (Thompson - Seber, 1996; Thompson, 2002; Sirken, 2006). I risultati non sono comunque univoci, dal momento che l'effetto del disegno dipende da diversi fattori, in particolare dal tipo di variabile stimata. Si possono però presentare alcuni punti fermi ai quali sono giunti gli studi appena citati. Va sottolineato nuovamente, per evitare fraintendimenti, che una strategia di campionamento si considera più efficiente rispetto ad un'altra, da un punto di vista statistico, quando produce stime con una minore varianza campionaria. Il concetto pertanto non rimanda a valutazioni più ampie che possono tenere in considerazione altri aspetti come, ad es., il fatto che l'implementazione di un particolare disegno campionario può introdurre una complessità aggiuntiva nelle procedure di ricerca, che rischia di aumentare in modo elevato gli effetti degli errori di non campionamento sulle stime.

In sintesi, l'efficienza, in termini statistici, del *multiplicity sampling* rispetto ad un campione casuale semplice tende a accrescersi se:

1. per i caratteri studiati, il valore della frazione della varianza entro i grappoli sulla varianza totale è elevato; in altre parole all'interno dei grappoli il comportamento studiato deve avere un'elevata variabilità. Se, ad esempio, si volesse stimare il numero di sigarette fumate in una settimana dai ragazzi, questo comportamento dovrebbe essere molto variabile tra gli individui entro un grappolo. Questa condizione, seppur riscontrabile in particolari situazioni di ricerca, non sembra, comunque, riflettere la situazione generale poiché la formazione dei grappoli è basata su legami sociali ed è noto che le persone tendono ad associarsi e creare relazioni con altre persone dai comportamenti simili (McPherson *et al.*, 2001).

2. il numero di persone all'interno di ogni grappolo che appartiene alla popolazione oggetto di studio non è molto elevato: è meglio avere tanti piccoli grappoli costituiti da poche unità che pochi grappoli costituti da tante unità[15].

4.2 Aspetti semantici: gli errori non campionari

Il *multiplicity sampling* può essere utilizzato nelle indagini in due modi:

- per raccogliere informazioni "proxy" circa le caratteristiche di alcuni appartenenti alla popolazione target. In questo caso non è previsto un contatto diretto da parte dello staff di ricerca con le persone segnalate dagli intervistati;
- oppure per individuare alcuni appartenenti alla popolazione target, ehe verranno contattati direttamente da parte dello staff di ricerca in un secondo momento.

Le due modalità di utilizzazione sollevano problematiche diverse, pertanto saranno considerate separatamente nei sottoparagrafi 4.2.1. e 4.2.2. In ogni caso affinché il *multiplicity sampling* possa funzionare due condizioni generali sono necessarie in entrambe le situazioni:

1. individuare una popolazione di cui si possiede una lista di campionamento ragionevolmente completa, i cui legami con la popolazione target siano in numero elevato;
2. scegliere un legame di tipo simmetrico che permetta ai ricercatori di espandere il campione a partire dai semi e, contemporaneamente, di costruire grappoli di persone mutuamente esclusivi e non troppo ampi. Nel paragrafo precedente è già stato evidenziato che all'aumentare dell'ampiezza dei grappoli gli stimatori perdono in efficienza. Accanto a questa osservazione di tipo sintattico si deve

[15] Per un approfondimento formale di questi due punti si veda (Thompson, 2002, pp. 297-298).Si può notare come le considerazioni che si applicano al *multiplicity sampling* sono molto simili a quelle che si applicano al tradizionale campionamento a stadi e a grappoli (Vitalini, 2010a, 2010b).

aggiungerne un'altra di tipo semantico: legami che generano ampie reti sociali richiedono ai rispondenti un compito cognitivo di identificazione ed elencazione delle persone maggiormente gravoso, con tutto ciò che ne deriva in termini di accuratezza delle risposte.

4.2.1. I rispondenti come proxy

Particolarmente delicato è il problema delle risposte proxy nelle indagini sociali. Spesso si chiede all'intervistato di rispondere "al posto di" un'altra persona. La pratica delle risposte proxy non è associata solo a campionamenti che sfruttano i legami sociali come il *multiplicity sampling*; anche indagini basate su forme di campionamento tradizionale fanno ricorso ad essa (Istat, 2006). La differenza è che nel *multiplicity sampling* l'uso del proxy è connaturato alla strategia di campionamento, mentre in un campionamento tradizionale esso è accidentale. In quest'ultimo il ricorso al proxy è sconsigliato e da utilizzare solo nel caso che la persona che deve rispondere personalmente sia assente o impossibilitata per motivi oggettivi al momento dell'intervista.

Alla base di questa pratica c'è l'assunto implicito che il rispondente sia in grado di rispondere accuratamente circa comportamenti e atteggiamenti di un'altra persona. Assunto alquanto scabroso. In letteratura esistono numerosi studi che evidenziano come le risposte proxy generino distorsioni di tipo sistematico nei dati nei campionamenti di tipo tradizionale (cfr. Schwartz - Wellens, 1997). Questi risultati, aggiunti al fatto che nel *multiplicity sampling* l'uso del proxy è fondamentale, sembrano gettare una luce negativa sulla qualità dei dati che si possono raccogliere con il *multiplicity sampling* rispetto alle forme di campionamento tradizionali. Questa considerazione va valutata *cum grano salis*. La teoria e la pratica non sempre vanno d'accordo nelle ricerche basate su forme di campionamento tradizionale. Intervistare una persona assente (e in molti casi probabilmente non disponibile a rispondere) comporta il ritorno più di una volta sul luogo dell'intervista, una maggiore fatica e maggiori costi: il ricorso al proxy è una scorciatoia molto allettante per l'intervistatore in

questa situazione. La vera differenza fra il *multiplicity sampling* e le forme di campionamento tradizionale risiedono nel fatto che, nel primo, è esplicitato e registrato il ricorso al proxy e pertanto, è possibile tenerlo in considerazione nella valutazione della qualità dei dati; nel caso di indagini campionarie tradizionali il ricorso al proxy spesso non è segnalato da nessuna parte, lasciando chi legge e riflette sulle stime senza strumenti per valutare l'effettivo impatto di questa pratica in termini di distorsione.

4.2.2. Le persone segnalate sono intervistate direttamente

Diverse dalle precedenti sono le problematiche sollevate nel caso il *multiplicity sampling* sia utilizzato per individuare membri di una popolazione senza lista di campionamento, che saranno intervistati direttamente in un secondo tempo. In questo caso gli "informatori" devono essere disponibili[16] a fornire informazioni che permettano successivamente di contattare le persone segnalate. Questo punto solleva problemi etici per il ricercatore e il rispondente: in particolare quando si studiano popolazioni soggette a diverse forme di stigma.

Esistono vari studi che valutano, per diversi tipi di popolazioni, il grado di disponibilità degli informatori a rivelare informazioni per rintracciare le persone che fanno parte della propria rete sociale. Una risposta univoca anche in questo caso non può essere fornita dal momento che, spesso, la disponibilità dipende dal tipo di popolazione, dai temi trattati e dai legami che vengono utilizzati per espandere il campione (Sirken, 1970, 2006; Kalton - Anderson, 1986; Sudman *et al.*, 1988; Czaja - Blair, 1990).

Per comprendere le possibilità teoriche e i limiti pratici del *multiplicity sampling* è utile considerare, in dettaglio, una sua applicazione per lo studio di popolazioni di persone/famiglie non abbonate alla linea telefonica fissa e che utilizzano solo il cellulare (Tortora *et al.*, 2008).

[16] Si suppone che siano in grado di fornire le informazioni.

Prima però è necessaria una breve premessa. Le tradizionali indagini telefoniche utilizzano come lista di campionamento l'elenco degli abbonati alla telefonia fissa. Nonostante alcuni problemi la copertura della popolazione italiana, garantita da questa lista, è stata considerata, fino a pochi anni fa, ragionevolmente accettabile (Istat, 2001, 2006). La disponibilità della lista, unita al fatto che un'intervista telefonica comporta minori costi di una faccia a faccia, ha stimolato una larga diffusione delle indagini telefoniche. Negli ultimi anni la situazione sta cambiando: la non copertura della lista di campionamento sta diventando un problema sempre più grave a causa dell'aumento di persone/famiglie che non hanno abbonamenti alla linea fissa e utilizzano solo il cellulare. Attualmente in molte nazioni, fra cui l'Italia, non esiste un elenco aggiornato che contenga la lista della popolazione che utilizza il cellulare e non è abbonata ad una linea telefonica fissa (Tortora *et al.*, 2008). Di qui la necessità di sviluppare nuove strategie di campionamento per includere questa fetta della popolazione, pena l'abbandono dell'utilizzo delle indagini telefoniche o una drastica limitazione dell'ambito di generalizzazione dei risultati.

Fatta questa breve premessa, è possibile comprendere, più chiaramente, le finalità dell'indagine di Tortora *et al.* (2008). Gli autori si posero come obiettivo quello di testare la possibilità di applicare il *multiplicity sampling* a partire da un campione casuale estratto dalla lista degli abbonati alla telefonia fissa. L'idea era quella di chiedere alle persone campionate di identificare i genitori, fratelli e figli (grappoli di tre generazioni) che non possedevano il telefono fisso ma utilizzavano solo il cellulare[17]. In questo modo si sarebbe potuto calcolare la probabilità di ogni persona con solo il cellulare in funzione del numero di membri appartenenti al proprio grappolo tri-generazionale e in possesso di un telefono fisso[18].

[17] «The burdensome nature of the informant's task, cost of measurement, and weight variation concerns constrained the size of the network to immediate relatives: living parents, siblings, and children of 18 years of age and older who live in a mobile-only houseold». (Tortora *et al.*, 2008, p. 139)

[18] Il *multiplicity sampling* non permette, in questo caso, di conoscere le probabilità di estrazione di persone/famiglie che utilizzano solo il cellulare e che appartengono ad un grappolo di persone che non possiedono il telefono fisso.

Come campione iniziale fu utilizzato il Gallup Panel costituito da circa 14.000 famiglie estratte con *random digit dialing* (RDD) e residenti negli stati Uniti d'America; a 7.000 famiglie fu inviato un questionario postale e alle restanti 7.000 fu inoltrato un invito di partecipazione alla Web survey. Tutti i membri della famiglia appartenenti al panel dovevano rispondere al questionario (circa 20.000 individui). Presero parte all'indagine 3.042 rispondenti ai quali venne chiesto di riportare il numero di cellulare dei loro genitori, fratelli e figli adulti che non vivevano nella stessa abitazione e non avevano una linea telefonica fissa. I risultati non furono incoraggianti: solo il 10% dei rispondenti fornì almeno un numero di cellulare dei parenti; il 78% si rifiutò espressamente di fornire il numero di cellulare di almeno un parente; il 7% rispose di non conoscere il numero di cellulare di almeno un parente, ed infine il 5% non rispose nulla.

Oltre alla distorsione che si può presumere generata dall'alto tasso di non partecipazione del campione iniziale e da questo schema di non risposte, l'indagine riscontrò un'ulteriore problema quando furono contattate le persone che avrebbero dovuto avere solo il cellulare e non il telefono fisso e il cui numero di cellulare era stato indicato dagli intervistati: ci fu un elevato numero di "falsi positivi". A dispetto del tentativo di ottenere solo i numeri di telefono di famiglie senza la linea fissa, i numeri forniti dai rispondenti contenevano molte famiglie che possedevano sia il telefono fisso che il cellulare.

4.3 Considerazioni sull'applicabilità del multiplicity sampling

Una valutazione complessiva dell'applicabilità in contesti reali di ricerca del *multiplicity sampling* è tutt'altro che semplice. L'elemento principale a favore di questa forma di campionamento consiste nel fatto che consente la costruzione di un modello matematico ben definito. D'altro canto, non si può nascondere che le condizioni che devono essere soddisfatte, affinché il *multiplicity sampling* permetta di raggiungere i risultati desiderati, sono talvolta troppo stringenti da incontrare nella pratica della ricerca. Come sintetizza Kalton (2009, p. 136) il «*network sampling has not been widely used in practice for surveys of rare population*

members. […] *There is the risk that the sampled informant may not accurately report the rare population status of other members of the linkage, either deliberately or through lack of knowledge. Nonresponse for the main survey data collection is another concern. In addition, ethical issues can arise when sampled persons are asked about the rare population membership of those in their linkage when that membership is a sensitive matter. The benefits of network sampling are partially offset by the increased sampling errors arising from the variable weights that the method entails, and by the costs of locating the linked rare population members*».

5. Il campionamento che sfrutta i legami sociali e la scelta dei semi non casuale: un matrimonio fragile

Come si è detto all'inizio di questo capitolo le forme di campionamento fin qui presentate si fondano sul fatto che il campione dei semi sia estratto in modo casuale sfruttando successivamente i legami di questi semi con i membri della popolazione target. Questa strategia presuppone la disponibilità di una lista di campionamento ausiliaria per l'estrazione dei semi.

Nel caso non si possa utilizzare nessun tipo di lista di campionamento ausiliaria è possibile estrarre casualmente i semi? La riposta è negativa: non è possibile. L'unica cosa che si può fare è cercare di approssimare la condizione dell'estrazione casuale dei semi (Snijders, 1992).

Per comprendere sia le possibilità che i limiti che si incontrano seguendo questa soluzione si può considerare, come emblematico, il lavoro di Frank - Snijders (1994). Gli autori cercarono di stimare il numero dei consumatori di eroina in una città olandese. Ad un campione ragionato di tossicodipendenti, estratto a partire dagli archivi della polizia e degli assistenti sociali della città (liste affette da un elevato errore di non copertura), venne chiesto di elencare i nomi di altri tossicodipendenti (ci si fermò alla prima ondata). I due studiosi proposero e discussero diversi stimatori che avrebbero permesso, a partire dai nomi segnalati e dalle loro ripetizioni, di calcolare il numero in valore assoluto dei tossicodipendenti in città.

Gli stimatori considerati risulterebbero «non distorti a patto che il campione iniziale fosse frutto di una procedura di estrazione bernoulliana» (Frank - Snijders, 1994, p. 63), la quale garantisce, cioè, che ogni individuo della popolazione oggetto di studio abbia la stessa, anche se sconosciuta, probabilità di essere incluso nel campione e che le probabilità di inclusione siano indipendenti fra gli individui. Per approssimare ragionevolmente l'assunto, i due studiosi suggerirono di estrarre i membri del campione iniziale a partire da fonti di contatto con la popolazione oggetto di studio il più possibile diverse ed indipendenti fra loro, facendo attenzione di selezionare all'interno di ogni fonte pochi membri anch'essi non in relazione fra loro (Frank - Snijders, 1994, p. 66). Purtroppo però il grado di approssimazione dell'assunto può esser valutato solo attraverso giudizi soggettivi del ricercatore e la sua violazione introduce una distorsione la cui entità non è conosciuta e non è stimabile a partire dai dati del campione finale. In generale la violazione dell'assunto dovrebbe tradursi in una sovrarappresentazione degli individui più centrali (con più grandi gradi di entrata e di uscita) a scapito dei membri più periferici della popolazione e, di conseguenza, in una sottostima, di entità sconosciuta, della popolazione studiata.

Sul giudizio di "ragionevole approssimazione" si appuntano le critiche dei metodologici più rigorosi: affermare che il campione di semi sia di tipo casuale non è dovuto al fatto che il ricercatore ha utilizzato una procedura di selezione che garantisce la casualità dell'estrazione (come in un campionamento casuale semplice) ma è frutto, nel migliore dei casi, di un campionamento di tipo ragionato con tutti i limiti che ne seguono.

Si ipotizzi di rinunciare completamente a procedure di estrazione dei semi che diano risultati simili ad un campionamento casuale. Ci si potrà allora chiedere: «Partendo da un numero di semi scelti in modo non casuale (ad es. sulla base di una scelta ragionata), si può giungere a conoscere le probabilità di estrazione dei soggetti nel campione?». Il prossimo capitolo sarà dedicato a esaminare se esiste una risposta affermativa a questa domanda.

4. Il campionamento che sfrutta i legami sociali: una nuova prospettiva

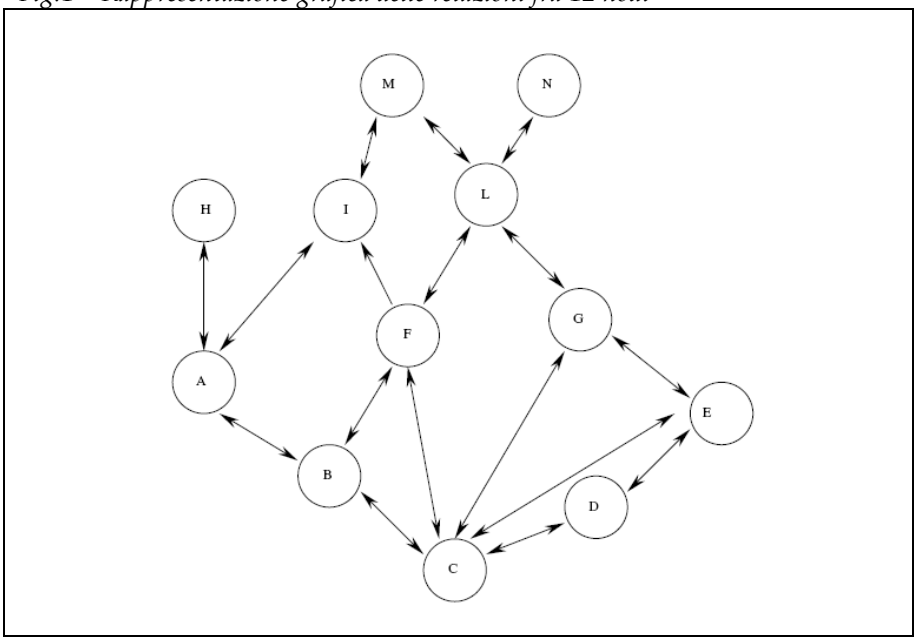
1. Le catene markoviane

Una possibile risposta affermativa alla domanda posta alla fine del capitolo precedente si è avuta considerando la possibilità di formalizzare matematicamente il processo di reclutamento basato sui legami sociali come una catena markoviana (Heckathorn, 1997, 2002; Salganik - Heckathorn, 2004; Volz - Heckathorn, 2008). In termini molto generali, quest'ultima è un processo probabilistico che descrive fenomeni che possono assumere un numero finito di stati e in cui la transizione da uno stato ad un altro non è deterministica, ma probabilistica; dipende solo dallo stato immediatamente precedente ed è indipendente da tutti gli altri stati passati.

Per quale motivo la formalizzazione matematica del processo di reclutamento basato sui legami sociali come una catena markoviana potrebbe risultare feconda ed utile per rispondere alla domanda circa la possibilità di stimare le probabilità di estrazione dei soggetti nel campione finale, partendo da un numero di semi scelti in modo non casuale? Perché le proprietà matematiche delle catene markoviane sono ampiamente studiate e permettono di sviluppare una teoria statistica (una sintassi) per il calcolo di stimatori corretti di percentuali e medie.

Per capire questo punto è necessario procedere per gradi. Innanzitutto si consideri il grafo in fig. 1.

Fig.1 – Rappresentazione grafica delle relazioni fra 12 nodi



Le relazioni sono tutte simmetriche e non esistono nodi isolati, cioè senza almeno un legame. Si immagini di partire da un nodo qualsiasi come seme (ad es. A), e di scegliere a caso uno dei nodi ai quali è legato (B, H, I), ad es. B. Questa è la prima iterazione. Partendo dal nodo B che entra a far parte del campione si sceglie casualmente uno dei nodi ai quali è legato (A, F, C). Si noti che anche il nodo A, che era già stato campionato, può essere scelto nuovamente in questa seconda iterazione del processo. Si ipotizzi che venga estratto il nodo C. Partendo da quest'ultimo si estrae casualmente (terza iterazione del processo) uno dei nodi ai quali è legato (B, F, G, D, E), ad es. B. A questo punto (quarta iterazione) si estrae uno dei nodi legati a B (A, F, C), ad es. F. Il processo si può ripetere per un numero infinito di volte. Questo tipo di processo è un processo markoviano: infatti il numero di stati è finito (i 12 nodi) e la transizione da uno stato all'altro (il passaggio da un nodo all'altro) è di tipo probabilistico (si sceglie un nodo a caso fra quelli legati al nodo selezionato nell'iterazione precedente). Infine la probabilità di un nodo di essere estratto ad una data iterazione dipende solo dallo stato immediatamente precedente ed è indipendente da tutti gli altri stati passati (la probabilità di essere estratto di un nodo ad una data iterazione dipende solo dal grado in uscita del nodo precedente).

Date le caratteristiche appena sottolineate si può dimostrare (Salganik - Heckathorn, 2004; Grinstead - Snell, 2006) che la frequenza di volte che un nodo è estratto nel "lungo termine" (avremo modo di discutere successivamente cosa questa espressione significhi) si stabilizza, ed è proporzionale al valore del grado in uscita del nodo, qualsiasi sia il nodo da cui si parte (seme). Ad es., qualsiasi sia il nodo di partenza, la frequenza di estrazione converge, per il nodo L, al valore di 4/34 (dove 4 è il numero di legami in uscita di L e 34 è il numero di legami in totale nel grafo); per il nodo I al valore di 3/34 e così via per tutti gli altri nodi.

Riformulando le considerazioni appena svolte in termini più formali, data la matrice di adiacenza binaria **A** (N righe per N) colonne in cui ogni elemento a_{ij} assume il valore 1, se c'è una relazione fra il nodo i e il nodo j e 0 se non c'è relazione, poniamo che Z_k identifichi l'unità (nodo) del grafo che è selezionata nella K-esima ondata. Se i è il nodo nella k-esima ondata, allora per l'ondata successiva (k + 1) uno dei nodi legato ad i è selezionato in modo casuale. L'insieme dei nodi selezionati $\{Z_0,Z_1,Z_2, \ldots\}$ risulta, di conseguenza, il prodotto di una catena markoviana la cui matrice di transizione, **Q,** è costituita dagli elementi $q_{ij} = P(Z_{k+1} = j/Z_k = i) = a_{ij}/a_{i.}$ Si può dimostrare (Salganik - Heckathorn, 2004) che in un grafo "connesso" (non esistono nodi isolati: ogni nodo del grafo è raggiungibile a partire da qualsiasi altro nodo) e "non diretto" (tutte le relazioni fra i nodi sono simmetriche), la catena markoviana è irriducibile e le probabilità di selezione, nel lungo periodo, dei singoli nodi sono $p_i \propto a_{i.}$ dove $\propto$ indica una relazione di proporzionalità diretta.

Tab. 1 – Matrice di transizione Q del grafo di fig.1 ($q_{ij} = a_{ij}/a_{i.}$)

	A	**B**	C	D	E	F	G	**H**	**I**	L	M	N	$a_{i.}$
A	0	**1/3**	0	0	0	0	0	**1/3**	**1/3**	0	0	0	*3*
B	**1/3**	0	**1/3**	0	0	**1/3**	0	0	0	0	0	0	*3*
C	0	**1/5**	0	**1/5**	**1/5**	**1/5**	**1/5**	0	0	0	0	0	*5*
D	0	0	**1/2**	0	**1/2**	0	0	0	0	0	0	0	*2*
E	0	0	**1/3**	**1/3**	0	0	**1/3**	0	0	0	0	0	*3*
F	0	**1/4**	**1/4**	0	0	0	0	0	**1/4**	**1/4**	0	0	*4*
G	0	0	**1/3**	0	**1/3**	0	0	0	0	**1/3**	0	0	*3*
H	**1/1**	0	0	0	0	0	0	0	0	0	0	0	*1*
I	**1/3**	0	0	0	0	**1/3**	0	0	0	0	**1/3**	0	*3*
L	0	0	0	0	0	**1/4**	**1/4**	0	0	0	**1/4**	**1/4**	*4*
M	0	0	0	0	0	0	0	0	**1/2**	**1/2**	0	0	*2*
N	0	0	0	0	0	0	0	0	0	**1/1**	0	0	*1*

A questo punto si può cercare di mettere in luce come le considerazioni appena svolte possano servire per lo studio di popolazioni di cui non si possiede la lista di campionamento e i cui membri siano legati da relazioni sociali.

Si consideri un'ipotetica popolazione di cui non si dispone degli elenchi dei membri. Questi ultimi devono essere in qualche modo in contatto fra loro, ad es. attraverso la partecipazione comune a particolari attività, e devono riconoscersi come membri della popolazione. Si supponga, inoltre, che le relazioni fra i membri della popolazione siano di tipo simmetrico e che ogni persona appartenente alla popolazione sia raggiungibile a partire da qualsiasi altra attraverso una catena di persone[19].

Si chieda ai rispondenti iniziali di identificare diversi contatti, uno dei quali è poi selezionato a caso per essere il successivo rispondente. Nel caso ideale il rispondente scrive la lista nominativa dei conoscenti che sono membri della popolazione studiata e poi ne seleziona, con un meccanismo casuale, uno affinché partecipi all'indagine.

[19] La condizione che ogni persona possa essere raggiunta a partire da ogni altra persona può apparire non realistica, specialmente per grandi popolazioni nelle quali alcune unità possono essere isolate dal resto della rete. In realtà, si è trovato che la maggior parte delle reti possiede una componente, definita gigante, che copre la maggioranza della popolazione. Da un punto di vista di inferenza statistica la generalizzazione dei risultati si dovrà limitare alla componente gigante ma se questa è molto grande la distorsione è ridotta. Problemi analoghi si riscontrano anche in campionamenti convenzionali che partono da liste che raramente sono esaustive della popolazione di riferimento per problemi di aggiornamento, duplicazione o omissione.

Se adottiamo una strategia di campionamento con reinserimento, il risultato sarà che l'insieme delle persone selezionate che vengono a far parte del campione sono il prodotto di un processo markoviano come quello che abbiamo descritto nell'esempio all'inizio del capitolo. Stando così le cose, nel lungo termine, indipendentemente dai semi iniziali scelti, la frequenza di selezione di ogni persona dovrebbe stabilizzarsi ed essere proporzionale al numero di persone nella popolazione studiata con le quali è in relazione (in termini formali al suo grado in uscita) e a partire dalle quali ha reclutato. Ad es., una persona con venti legami dovrebbe avere il doppio di probabilità di essere reclutata, in ogni iterazione del processo, di una con 10 legami.

Prima di passare oltre si ritiene utile riassumere le condizioni per poter applicare nella ricerca gli strumenti messi a disposizione dalla riflessione sulle catene markoviane. Esse possono essere sintetizzate nei seguenti cinque punti:

1. i rispondenti mantengono relazioni reciproche con gli individui che conoscono come membri della popolazioni oggetto di studio;
2. i rispondenti sono inseriti in una rete sociale dove ogni persona è raggiungibile a partire da qualsiasi altra attraverso qualsivoglia percorso;
3. le persone possono partecipare più volte all'indagine;
4. il reclutamento dei soggetti che fanno parte del campione è il frutto di una selezione casuale fra i componenti della propria rete di conoscenze;
5. è possibile conoscere il numero di persone alle quali il reclutatore è legato e a partire dalle quali ha reclutato.

2. Il *respondent-driven sampling*: dalla teoria alla pratica

Sul piano sintattico le argomentazioni risultano precise ed eleganti. Il problema che si pone per un ricercatore, che volesse utilizzarle per lo studio di una popolazione di cui non si dispone degli elenchi dei membri, si traduce in una domanda: «E' possibile organizzare una strategia di campionamento che si adatti alle richieste della teoria matematica? ».

Il tentativo più sistematico, e di maggior successo, di utilizzo di procedure di costruzione dei campioni a partire dai legami sociali, che rimandano alla formalizzazione delle catene markoviane fa capo ad Heckathorn (1997, 2002, 2007; Salganik - Heckathorn, 2004; Volz - Heckathorn, 2008), il quale ha coniato l'espressione *Respondent-Driven Sampling* (RDS in seguito).

In cosa consiste il disegno RDS? La procedura del RDS inizia individuando un numero ristretto di semi (in genere dai sei ai venti) appartenenti alla popolazione oggetto di studio (es. consumatori di droghe). Questa prima selezione avviene solitamente sulla base di un campionamento per obiettivi (Bichi, 2002). Ad ogni seme viene chiesto di reclutare, in modo completamente libero, da due a quattro conoscenti che fanno parte della popolazione studiata. Coloro che sono reclutati dai semi concordano, spesso telefonicamente, con gli intervistori dello staff data e luogo dell'intervista.

Gli intervistatori dello staff di ricerca, nelle prime fasi dell'intervista, verificano che le persone abbiano tutti i requisiti per poter partecipare all'indagine (si vedrà successivamente in dettaglio quali sono e come verificarli). Se tutti i requisiti sono incontrati i reclutati vengono intervistati.

Durante l'intervista, oltre alle domande sostanziali, si chiede all'intervistato di indicare il numero di persone della propria rete sociale, che fanno parte della popolazione studiata. Ad esempio, nelle indagini su popolazioni che fanno uso di droga si utilizza spesso la seguente domanda: «Quante persone che fanno uso di droga e che hai frequentato almeno una volta nell'ultima settimana conosci per nome?». La domanda sull'ampiezza della propria rete sociale nella popolazione target è una domanda chiave, dal momento che la risposta è

utilizzata come stima del numero di persone alle quali il reclutatore è legato e a partire dalle quali ha reclutato (in termini formali del grado in uscita).

Una volta finita l'intervista, viene chiesto all'intervistato di reclutare altre persone (da due a quattro). Si sottolinea che la differenza fra i semi e gli intervistati delle ondate successive è che i primi sono reclutati dal ricercatore mentre i secondi dai rispondenti.

Spesso sono previsti degli incentivi in forma monetaria per aumentare nelle persone coinvolte la motivazione a partecipare e a reclutare.

A certe condizioni (vedremo in seguito se facilmente o difficilmente riscontrabili nella realtà), è possibile formalizzare matematicamente il processo di reclutamento RDS come una catena markoviana e, di conseguenza, nel "lungo termine", indipendentemente dai semi iniziali scelti, la frequenza di selezione di ogni persona dovrebbe stabilizzarsi ed essere proporzionale al numero di persone nella popolazione studiata con le quali è in relazione e a partire dalle quali ha reclutato. Gli ideatori del RDS avanzano l'ipotesi che la probabilità di selezione per un nodo dovrebbe stabilizzarsi dopo un numero non eccessivo di passaggi (Heckathorn, 1997, 2002; Volz - Heckathorn, 2008). Operativamente "il numero non eccessivo di passaggi" è misurato in ondate, dove ogni ondata rappresenta un gradino di reclutamento lungo le catene. I semi sono considerati come ondata zero. I rispondenti reclutati dai semi originano la prima ondata. Coloro che sono reclutati dai rispondenti nella prima ondata generano la seconda ondata e così via... In media, dalle quattro alle otto ondate dovrebbero bastare per raggiungere la stabilizzazione delle probabilità di selezione, definita nella terminologia RDS "situazione di equilibrio" (Heckathorn, 1997, 2002; Volz - Heckathorn, 2008).

Una volta che viene raggiunta la situazione di equilibrio, una stima non distorta di una qualsiasi caratteristica (ad es., età o genere) può essere calcolata pesando il valore della variabile per il reciproco del numero, fornito dall'intervistato, di persone della propria rete che fanno parte della popolazione studiata. Questo valore, come già rilevato, è considerato una stima del numero di persone nella popolazione studiata con le quali l'intervistato è in relazione e a partire dalle quali ha reclutato.

L' RDS ha riscosso un notevole successo internazionale ed è stato adoperato per indagini su popolazioni di prostitute, tossicodipendenti, omosessuali in più di una trentina di nazioni (per una rassegna Malekinejad *et al.*, 2008). E' stato anche utilizzato per studiare i musicisti jazz in alcune città americane (Heckathorn - Jeffri, 2001), i lavoratori non in regola (Bernhardt *et al.*, 2006), alcuni gruppi di nativi americani (Walters - Simoni, 2002) e i lavoratori minorenni nel settore diamantifero in Sierra Leone (Bjørkhaug - Hatløy, 2009).

L'entusiasmo con cui è stato accolto questo disegno di campionamento ne evidenzia la capacità di rispondere ad un bisogno diffuso nella comunità scientifica che studia popolazioni di cui non si conosce la lista di campionamento e che per motivi di carattere legale, ideologico o politico tendono ad occultare la loro identità. Questo, però, non deve far dimenticare che il valore del RDS è direttamente legato alla capacità di approssimare il più possibile nelle applicazioni pratiche i principi del modello statistico sottostante. Se questa affermazione è vera per i tipi di campionamento più tradizionali, come ad es. il campionamento casuale semplice, lo è a maggior ragione per quelli che sono ad uno stadio sia teorico che empirico maggiormente "immaturo".

Per valutare la capacità di approssimare i principi del modello statistico e per porre le basi conoscitive indispensabili per la comprensione del lavoro empirico di validazione sviluppato nella seconda parte di questa tesi è necessario approfondire in modo analitico l'applicazione del RDS, che per ora è stata solo abbozzata nelle sue linee principali. I prossimi due paragrafi saranno dedicati a descrivere in dettaglio alcuni aspetti del RDS particolarmente rilevanti per le considerazioni successive e a presentare una sua applicazione, per molti aspetti paradigmatica, per lo studio di una popolazione di tossicodipendenti in Thailandia.

3. Il *respondent-driven sampling*: un approfondimento

Gli aspetti del disegno campionario RDS che verranno considerati in questo paragrafo sono i seguenti:

1. iniziale valutazione dell'applicabilità del RDS;
2. selezione dei semi;
3. numero di persone da reclutare per ogni reclutatore;
4. uso di incentivi per la partecipazione;
5. utilizzo dei coupon;
6. verifica dei requisiti di partecipazione dei reclutati;
7. valutazione del raggiungimento della situazione di equilibrio;
8. stima del numero di persone nella popolazione studiata con le quali l'intervistato è in relazione e a partire dalle quali ha reclutato.

3.1 Iniziale valutazione applicabilità del RDS

Prima di qualsiasi lavoro sul campo si deve decidere se la strategia campionaria RDS è appropriata per l'indagine che si è deciso di intraprendere. La prima domanda da chiedersi è: «La popolazione di interesse è intervistabile presso luoghi definiti e accessibile al ricercatore?». Nel caso la risposta a questa domanda sia affermativa esistono strategie più appropriate del RDS. In particolare diverse forme di *location sampling* (Sudman - Kalton, 1986) possono rivelarsi più adatte del RDS. Diversi disegni campionari che ricadono sotto la più ampia categoria di *location sampling* sono stati utilizzati per campionare una varietà di popolazioni che tendono a concentrarsi in luoghi precisi. Tra questi si può annoverare anche il "campionamento per centri o ambienti di aggregazione" (Blangiardo, 1996; 2004) che è usato nelle indagini campionarie condotte dall' ISMU (Iniziative e Studi sulla Multietnicità)[20] per campionare gli immigrati illegali in Lombardia.

[20] Il sito della fondazione è http://www.ismu.org/

Se la risposta alla domanda circa il fatto che la popolazione di interesse sia intervistabile presso luoghi definiti e accessibili al ricercatore è negativa, allora l'RDS può essere utile. Infatti quando la popolazione target non si raccoglie presso luoghi fisici ben definiti le diverse forme di *location sampling* si rivelano inadatte.

Il fatto che la popolazione sia adatta per l'utilizzo del RDS non è, comunque, sempre chiaro. In ogni caso si possono individuare quattro richieste che devono essere soddisfate (Heckathorn, 1997, 2002; Volz - Heckathorn, 2008).

Le prime due richieste accomunano questo tipo di campionamento con le forme di campionamento (*multiplicity sampling*) che abbiamo incontrato nel terzo capitolo. I rispondenti devono intrattenere relazioni di tipo simmetrico (prima richiesta)[21] e devono riconoscersi l'un l'altro come membri della popolazione di riferimento, poiché altrimenti essi non sono in grado di conoscere chi reclutare (seconda richiesta). Questa richiesta è soddisfatta per popolazioni legate da una partecipazione ad attività comuni (ad es. attori che lavorano insieme in un film). Al contrario le persone che evadono le tasse non vanno bene, poiché due evasori possono essere amici, e non sapere se entrambi fanno parte del gruppo della popolazione oggetto di studio.

In base alla terza richiesta, il numero di legami fra i membri deve essere sufficientemente numeroso da sostenere i processi di reclutamento, poiché altrimenti le catene di reclutamento si arresterebbero dopo poche ondate, e il campionamento non raggiungerebbe la "profondità sociometrica" (lunghe catene di reclutamento che attraversano più ondate) richiesta per assicurare che ogni membro della popolazione abbia una probabilità diversa da zero di entrare a far parte del campione.

In base alla quarta richiesta, la popolazione non deve essere segmentata a tal punto che le catene di reclutamento rimangano intrappolate entro specifici sottogruppi, perché altrimenti il campionamento sarebbe limitato ai sottogruppi a partire dai quali i semi sono stati selezionati, e l'equilibrio non sarebbe mai

[21] Ricordo ancora che con il termine relazione simmetrica si intende che se Tizio, ad una domanda in cui si chiede con chi è in relazione, risponde di essere in relazione con Caio, quest'ultimo alla stessa domanda, deve rispondere di esser in relazione con Tizio.

raggiunto indipendentemente da quante ondate il campione attraversa. Solitamente per valutare quest'ultima condizione viene previsto nei disegni di ricerca RDS uno studio preliminare di tipo esplorativo che comprende, oltre ad interviste a testimoni privilegiati, anche brevi periodi di osservazione della popolazione studiata (Heckathorn 1997, 2002).

3.2 La selezione dei semi

Come nelle tecniche viste nel precedente capitolo i semi costituiscono il primo gruppo di rispondenti, i quali poi troveranno altri individui con alcune caratteristiche desiderate, che saranno inclusi nello studio.

Diversamente dal *multiplicity sampling* i semi non sono, però, membri selezionati in modo casuale da una popolazione: essi sono selezionati sulla base di un campionamento per obiettivi, in particolare in base al criterio della differenziazione che prevede la selezione di soggetti che presentano la massima eterogeneità possibile rispetto ad alcune caratteristiche significative (Bichi, 2002). Questo tipo di selezione dovrebbe accrescere la velocità con la quale la situazione di equilibrio viene raggiunta riducendo il numero di ondate che devono essere attraversate (Gile - Handcock, 2010).

Oltre che in base al criterio di differenziazione i semi sono scelti anche per la loro abilità di reclutare altri nello studio. Tendenzialmente un "buon seme" dovrebbe avere un'ampia rete sociale, dovrebbe avere numerosi legami con gli altri membri della popolazione di interesse. Per questo motivo i semi sono usualmente individuati tra persone ben conosciute all'interno della rete sociale.

Quanti devono essere i semi? In letteratura non c'è un preciso metodo per scegliere il numero corretto di semi: in genere sono scelti fra i sei e i venti semi. È importante notare che avere sia un numero elevato che uno ridotto di semi può essere problematico. Nel caso di pochi semi sono necessarie catene di reclutamento mediamente più lunghe (più ondate) per ottenere ampiezze campionarie della popolazione che garantiscano il raggiungimento dell'equilibrio: questo accresce la probabilità di una loro interruzione. Nel caso

di un numero elevato di semi, le catene risulteranno mediamente più brevi (a parità di ampiezza campionaria), accrescendo la probabilità di stime campionarie distorte perché non si raggiunge la profondità sociometrica attesa per la condizione di equilibrio (Gile - Handcock, 2010).

3.3. Numero di reclutati per reclutatore

Sia i semi che i reclutatori nelle ondate successive alla prima dovrebbero reclutare non più di due o tre amici dalla loro rete sociale. La limitazione del numero di reclutati a tre o quattro mira a sviluppare catene di reclutamento mediamente lunghe per raggiungere la profondità sociometrica richiesta per assicurare che ogni membro della popolazione abbia una probabilità diversa da zero di entrare a far parte del campione e per ridurre l'influenza dei semi/reclutatori con reti personali molto ampie che spesso sono caratterizzate da persone simili rispetto a molte caratteristiche oggetto di studio (McPherson *et al.*, 2001).

3.4 Uso di incentivi per incoraggiare la partecipazione

E' spesso previsto l'uso di forme di incentivo per motivare rispondenti a reclutare i loro pari (Heckathorn, 1997, 2002). Solitamente gli studi RDS prevedono modesti incentivi finanziari (essi possono essere comunque anche in forma di regali o merci e servizi) sia per il completamento dell'intervista che per il reclutamento. Nessun incentivo particolare può essere richiesto per popolazioni che "hanno una storia da raccontare" o che, in generale, sono fortemente interessati all'argomento dello studio. Sono stati rilevati problemi quando gli incentivi erano troppo elevati: persone che non erano membri della popolazione di riferimento fingevano, in accordo con il loro reclutatore, di essere dei membri per ottenere i compensi (Lipovsek - Longfield, 2007) .

Per assicurare che gli incentivi siano distribuiti in modo appropriato, chi recluta deve ricevere il compenso previsto quando coloro che sono reclutati incontrano i criteri di inclusione e quando l'intervista di quest'ultimi è completata con successo. In termini operativi, chi recluta può ricevere l'incentivo sia recandosi al luogo dell'intervista insieme a colui che recluta oppure ritornando in un secondo tempo dai membri dello staff di ricerca e verificando che il reclutato sia stato intervistato.

Un incentivo addizionale nel RDS è l'approvazione sociale all'interno della propria rete di relazioni. L'influenza dei pari è pensata come un fattore in grado di accrescere i tassi di risposte e di cooperazione (Heckathorn, 1997). In ogni caso non esistono evidenze empiriche che confermino questo fenomeno nell'ambito degli studi RDS: per ora si tratta di un'ipotesi, alquanto ragionevole, che richiede però studi mirati per essere avvalorata.

3.5 Il processo di reclutamento: il sistema dei coupon

L'aspetto più complesso degli studi RDS è mantenere traccia delle catene di reclutamento. Sapere "chi ha reclutato chi" è un'informazione indispensabile (come si vedrà nel par. 3.7 e nel par. 4.1) per il calcolo della situazione di equilibrio[22]. Per questo motivo, nell'implementazione del disegno nelle indagini campionarie, è stato disposto un sistema che richiede parecchi controlli incrociati per assicurare che gli individui con le caratteristiche desiderate dai ricercatori siano inclusi nel campione. In questo sistema uno degli elementi centrali è l'uso di coupon di reclutamento (Salganik - Heckathorn, 2004). Per avere un'idea di che cos'è un coupon si può osservare la fig. 2 che riproduce un coupon utilizzato in un'indagine che studia il consumo di droghe (Wang *et al.*, 2007).

I coupon sono usati per controllare il processo di reclutamento e permettere di tracciare chi recluta chi. Solitamente un numero identificativo è scritto su

[22] Si ricorda che nella terminologia RDS l'espressione "situazione di equilibrio" indica che la frequenza di selezione di ogni persona si è stabilizzata ed è proporzionale al suo grado in uscita.

ogni coupon, che mette in relazione i partecipanti ai loro reclutatori permettendo così di ricostruire le catene di reclutamento. Nell'esempio di fig. 2 questo numero è: S002W01-CB-002-A dove S002 indica il reclutatore, in questo caso il seme numero 2; W01 l'ondata di appartenenza (*Wave* 1); CB l'iniziale del luogo di intervista dove è stato distribuito il coupon; 002 il numero identificativo univoco (ID) del reclutato e la lettera A che si tratta del primo dei tre coupon consegnati al reclutatore (A, B, e C indicano rispettivamente i 3 coupons distribuiti). Vista l'importanza dei coupon nell'indagine un ulteriore esempio sarà presentato nel capitolo successivo. Tutti i possibili partecipanti devono avere un coupon quando si presentano al luogo dell'appuntamento per essere intervistati; chi è senza è escluso dallo studio.

Come si può comprendere chiaramente da questa descrizione la gestione dei coupon è uno degli aspetti più delicati della raccolta delle dati, al punto che gli ideatori del campionamento RDS forniscono un programma per facilitarne la gestione: RDSCM (Respondent-driven sampling Coupon Manager).

Fig. 2 - Esempio di coupon di reclutamento

Fonte: Wang *et al..*, *Respondent-driven sampling in the recruitment of illicit stimulant drug users in a rural setting: findings and technical issues*, in "Addictive Behaviors", 2007, 32(5), pp. 924-937

3.6. Verifica dei requisiti di partecipazione dei reclutati

Una volta che si presenta al luogo dell'intervista un possibile intervistato con un coupon regolare lo staff deve verificare:

1. che esista una relazione preesistente fra il reclutato e il reclutatore;
2. che il reclutato sia un membro della popolazione oggetto di studio;
3. che il reclutato non abbia già partecipato all'indagine.

3.6.1.Verifica del tipo di relazione fra reclutato e reclutatore

I ricercatori devono verificare le relazioni preesistenti tra il reclutatore e il reclutato. Questa operazione è molto importante per assicurare la validità della strategia campionaria impiegata: ogni partecipante, esclusi i semi, deve essere invitato da un amico o conoscente. Per indagare la relazione fra reclutato e reclutatore un esempio di domanda utilizzata è la seguente: «Come descriveresti la relazione con la persona che ti ha invitato a partecipare a questa ricerca?». Le modalità di risposta potrebbero essere: "parente, amico o conoscente, collega di lavoro, partner o compagno, estraneo". Se una persona definisce il proprio reclutatore "estraneo" l'intervista viene terminata.

In questo modo si cerca di approssimare il più possibile la condizione del modello matematico che richiede che le relazioni fra i nodi siano di tipo simmetrico.

3.6.2 Verifica che il reclutato sia un membro della popolazione oggetto di studio

Per evitare la presenza di persone che non fanno parte della popolazione di riferimento e che vogliono partecipare solo per ottenere gli incentivi monetari, l'appartenenza del reclutato alla popolazione oggetto di studio deve essere verificata. I membri dello staff di ricerca usualmente pongono alcune domande alle quali solo chi appartiene alla popolazione di riferimento dovrebbe essere in grado di rispondere. Questo punto verrà successivamente chiarito in dettaglio con un esempio nel par. 4.

3.6.3 Verifica che il reclutato non abbia già partecipato all'indagine

I ricercatori devono, inoltre, assicurarsi che i partecipanti non cerchino di essere intervistati più volte sperando di ottenere incentivi multipli. Per facilitare questa operazione sono state suggerite alcune strategie (Lipovsek - Longfield, 2007): ad es., possono essere raccolte ed inserite in un database informazioni riguardo caratteristiche fisiche che permettano di distinguere i partecipanti l'uno dall'altro (colore degli occhi), segni particolari visibili (tatuaggi) o marcatori biometrici (diametro del polso, lunghezza avambraccio, circonferenza della testa). Si possono poi confrontare queste informazioni con quelle raccolte per gli intervistati precedenti e nel caso compaia nel database un altro individuo con le stesse caratteristiche, l'intervistatore cerca di capire se si trova di fronte ad una persona che si è già presentata. Questi suggerimenti seppur utilizzati in alcune ricerche appaiono realisticamente poco praticabili nel complesso.

3.7 Valutazione del raggiungimento della situazione di equilibrio

Nella logica dei disegni RDS il raggiungimento della situazione di equilibrio è condizione necessaria per fare inferenza statistica. Come si è detto in precedenza, gli ideatori del RDS sostengono che è possibile, date certe condizioni, formalizzare matematicamente il processo di reclutamento come una catena markoviana e che, "nel lungo termine", indipendentemente dai semi iniziali scelti, la frequenza di selezione di ogni persona si stabilizza (situazione di equilibrio) ed è proporzionale al numero di persone nella popolazione studiata con le quali è in relazione e a partire dalle quali ha reclutato. Ad es., una persona con venti legami avrà il doppio di probabilità di essere reclutata di una con dieci legami. A partire da queste probabilità è possibile costruire degli stimatori non distorti delle medie, delle percentuali e delle loro varianze campionarie.

Sulla base della teoria statistico-matematica il processo di reclutamento dovrebbe, pertanto, continuare almeno finché il campione raggiunge l'equilibrio (e sia raggiunta l'ampiezza campionaria prevista). Si deve sottolineare l'avverbio "almeno": in base alla logica del RDS, a parità di ampiezza campionaria, più lunghe sono le catene di reclutamento meglio è.

Il raggiungimento della situazione di equilibrio è misurato, nella proposta degli ideatori del RDS, in ondate, dove ogni ondata rappresenta un gradino di reclutamento lungo le catene. I semi sono considerati come ondata zero. I rispondenti reclutati dai semi originano la prima ondata. Coloro che sono reclutati dai rispondenti nella prima ondata generano la seconda ondata e così via...

Il numero di ondate minimo richiesto per raggiungere l'equilibrio è simulato per ogni variabile usando un modello di catene markoviane nelle quali la matrice di transizione è stimata a partire dalla matrice campionaria di reclutamento[23]. (Heckathorn 1997, 2002). L'affermazione appena fatta può risultare oscura. Una piena comprensione del suo significato richiederebbe una descrizione in termini formali-matematici. Verrà presentato, tra breve in questo capitolo (par. 4.1), un esempio della procedura di calcolo dell'equilibrio, applicata in un contesto reale di ricerca. L'esempio dovrebbe chiarirne la logica sottostante, senza addentrarsi nei meandri della sua formalizzazione matematica.

3.8 Stima dell'ampiezza della rete sociale

Una volta raggiunto l'equilibrio nella pratica della ricerca si ritiene di aver a che fare con un campione probabilistico: una stima non distorta di una qualsiasi caratteristica delle persone (età o genere ad es.) può essere calcolata utilizzando

[23] Questo metodo di calcolo del numero di ondate necessarie per raggiungere l'equilibrio è implementato nel programma RDSAT (Volz *et al.*, 2007)

come pesi i valori dell'ampiezza della rete sociale (nella popolazione oggetto di studio) del reclutatore[24].

Operativamente, per stimare questi valori, nelle indagini RDS viene posta una sola domanda (il problema dell'accuratezza di questo tipo di stime verrà considerato successivamente). Particolare attenzione viene prestata nella formulazione di questa domanda per quanto riguarda la descrizione della popolazione oggetto di studio. Come esempio, si può considerare la seguente domanda utilizzata in diverse indagini sui tossicodipendenti che si iniettano le droghe: «Quante persone che si bucano e che hai frequentato almeno una volta nell'ultima settimana conosci per nome o nomignolo?». La sua formulazione, nelle intenzioni del gruppo di ricerca, dovrebbe assicurare che le persone si conoscano l'un l'altra (l'RDS assume relazione simmetriche fra reclutatore e reclutato) e che gli intervistati limitino il conteggio solo alla popolazione di riferimento dei tossicodipendenti che si iniettano sostanze stupefacenti.

Considerando una variabile di tipo categoriale (per un approfondimento dello stimatore RDS nel caso più generale si veda appendice 4) la stima della percentuale di una data modalità, P_A, (es. femmina) è data da:

$$\hat{P}_A = 100\left(\frac{n_A}{n}\right)\left(\frac{D_{totale}}{D_A}\right)$$

dove 100 (n_A/n) è la percentuale di persone del gruppo A nel campione una volta raggiunta la situazione di equilibrio (es. la percentuale di femmine nel campione); D_{totale} è il valore della media armonica del numero di legami dichiarati da ogni persona (d_i) nel campione:

$$D_{totale} = \frac{n}{\sum_{i=1}^{n}\frac{1}{d_i}}$$

e D_A è il valore della media armonica del numero di legami dichiarati da ogni persona (d_i) nel gruppo A:

[24] Tecnicamente pesando ogni valore della variabile dell'unità campionata per il reciproco dell'ampiezza della rete sociale a partire dalla quale recluta (cfr. Appendice 4).

$$D_A = \frac{n_A}{\sum_{i=1}^{n_A} \frac{1}{d_i}}$$

Ad esempio, se la percentuale delle femmine è uguale a 61,7% (dopo che il campione ha raggiunto la condizione di equilibrio) e il rapporto fra le medie armoniche della rete sociale nel complesso e di quella delle femmine è pari a 0,96, la stima RDS della percentuale di femmine sarà pari a 59,1% (61,7% x 0,96).

L'osservazione di questo stimatore RDS può far sorgere spontanea una domanda: «Le stime RDS, che sono pesate, sono da preferire alle stime che non sono pesate? ». Considerando, per semplicità espositiva, una variabile dicotomica (le considerazioni possono essere comunque generalizzate anche a variabili di altro tipo) le stime RDS sono tanto preferibili alle stime semplici quanto maggiore è la differenza fra il numero medio di relazioni nei due gruppi e tendono a coincidere quando i due gruppi hanno, in media, lo stesso numero di relazioni (si veda per una approfondimento l'appendice 9).

4. Uno studio sugli *injection drug users* (IDU) in Thailandia (2005): un esempio di RDS

Per esemplificare nel modo più preciso possibile i diversi aspetti del campionamento RDS analizzerò una ricerca svolta in Thailandia in cui l'RDS fu utilizzato per indagare la popolazione di *IDU* (*drug injection user*) dai 15 ai 45 anni, che viveva e lavorava nell'area urbana di Bangkok (Lipovsek - Longfield, 2007).

I semi furono 10. Per accrescere la velocità con la quale il campione avrebbe dovuto raggiunge l'equilibrio furono selezionati semi con differenti combinazioni delle seguenti caratteristiche: sesso, status occupazionale, lunghezza dell'esperienza di iniezione di droghe e partecipazione a trattamenti a base di metadone. Nella tabella seguente è possibile osservare le caratteristiche dei semi.

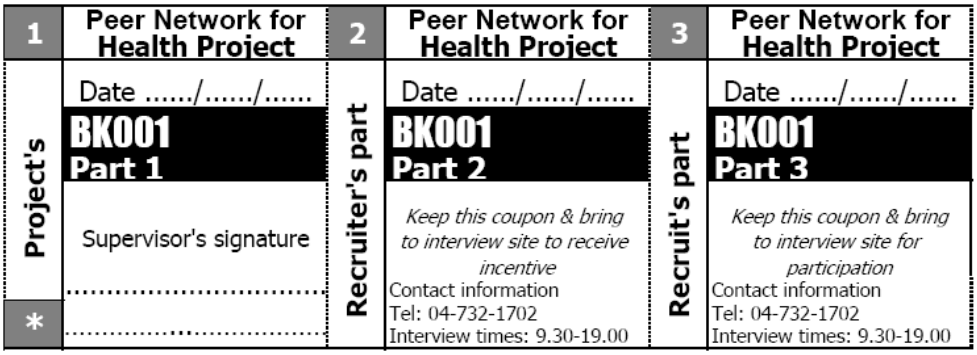
Tab.2 - Le caratteristiche dei semi

No.	Sex	Employment	Experience	Methadone Treatment
1	M	Employed	NEW	NONE
2	F	Employed	NEW	NONE
3	M	Employed	EXP	CURRENT
4	M	Employed	EXP	NONE
5	M	Employed	EXP	CURRENT
6	M	Unemployed	NEW	NONE
7	F	Unemployed	EXP	CURRENT
8	M	Unemployed	EXP	NONE
9	M	Unemployed	EXP	CURRENT
10	F	Unemployed	NEW	NONE

Note: "New" indica IDU che si iniettano droghe da meno di un anno, "Exp" da oltre un anno; "Current" indica gli IDU che stanno seguendo il trattamento a metadone, mentre "None" indica gli IDU che non hanno ricevuto il trattamento oppure lo hanno interrotto.
Fonte: Lipovsek - Longfield, *Sampling Hard-to-Reach Populations-PSI Research & Metrics*, Population Services International, 2007, http://www.psi.org/resources/publications, p. 25

Ognuno dei semi e dei successivi reclutatori ricevette tre coupon (fig. 3) da consegnare a tre membri della propria rete sociale che appartenevano alla popolazione IDU. Tutti coloro che furono reclutati nelle ondate successive ricevettero anch'essi tre coupon.

Fig. 3 - Coupon utilizzato nell'indagine



Fonte: Lipovsek - Longfield, *Sampling Hard-to-Reach Populations-PSI Research & Metrics*, Population Services International, 2007, http://www.psi.org/resources/publications, p. 27

Ogni coupon fu stampato su cartoncino resistente e consisteva di tre parti separabili: Project's, Recruiter's part, Recrutit's part. In tutte e tre le parti furono riportati il nome del indagine: "Peer network for Health Project"(fu dato un nome generico per evitare che un nome più specifico sollevasse sospetti circa lo studio nel caso il coupon fosse finito nelle mani sbagliate, ad es: spacciatori, e per ridurre l'apprensione circa la partecipazione in uno studio che poteva essere considerato stigmatizzante); la data in cui il coupon fu distribuito e il numero identificativo del coupon (BKK001 indicava che questo era il primo coupon distribuito nell'area metropolitana di Bangkok. Il successivo coupon aveva il codice BKK002, il terzo BKK003 e così via….). Il fatto che tutte è tre le parti contenessero lo stesso numero identificativo consentiva di "tracciare" chi reclutava chi: condizione essenziale per le analisi successive.

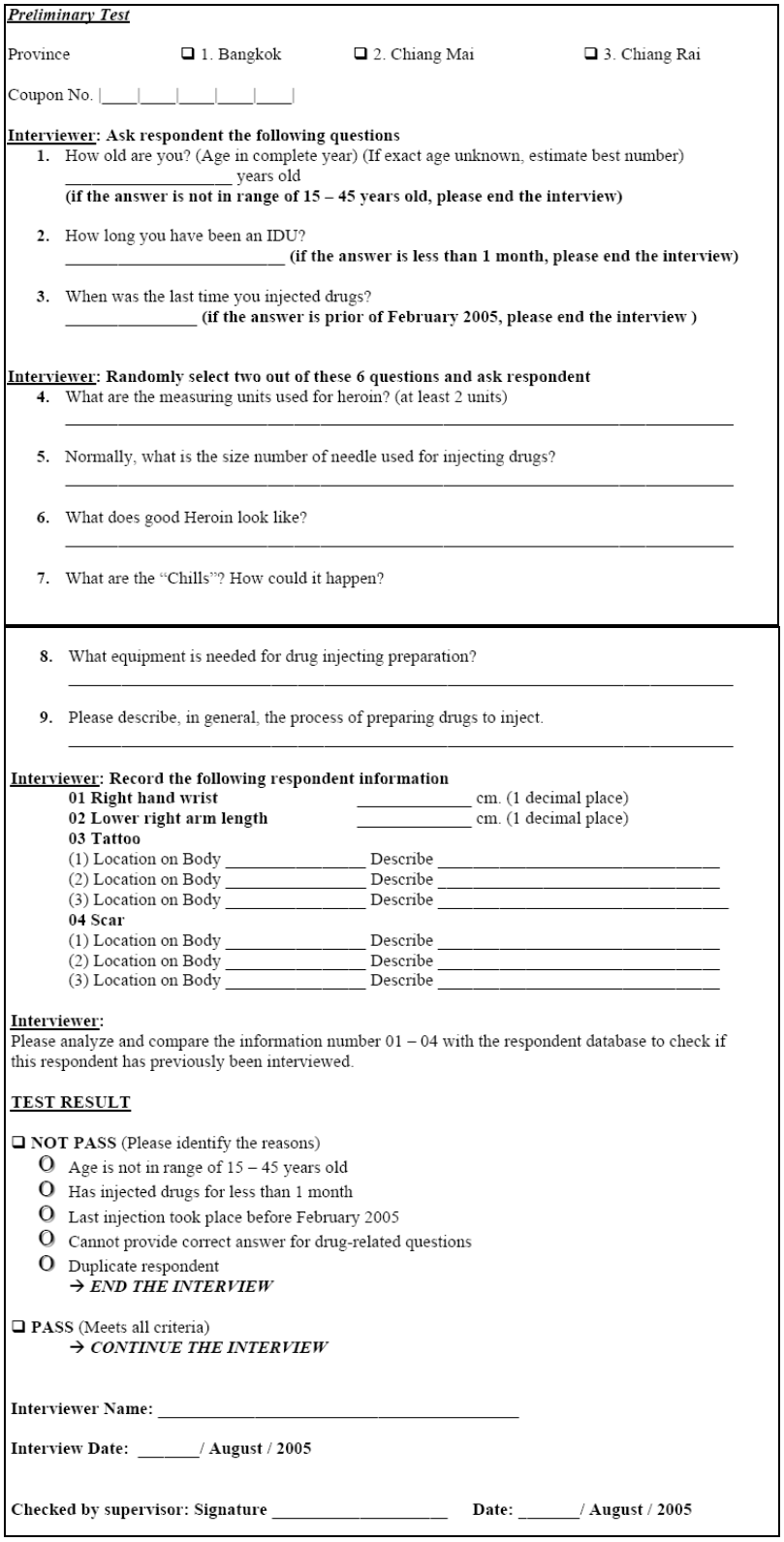
Ogni persona che partecipava all'indagine riceveva tre coupon. Lo staff di ricerca tratteneva la prima parte del coupon. L'intervistato si teneva le due parti restanti. Una volta individuata una persona da reclutare le consegnava la terza parte. Una volta che il reclutato era intervistato riceveva a sua volta tre coupon. Il suo reclutatore si presentava presso la sede dell'intervista con la seconda parte del coupon per ricevere un compenso in denaro. Il possesso della terza parte del

coupon per chi si presentava al luogo dell'intervista era condizione preliminare per partecipare alla ricerca, le persone che non possedevano il coupon erano escluse dalla ricerca.

Gli IDU che ricevevano la terza parte del coupon fissavano un appuntamento con lo staff di ricerca, concordando telefonicamente luogo e ora dell'intervista. Le interviste in quest'indagine furono condotte nei luoghi più diversi: dormitori, appartamenti, case private, parchi, strade, negozi, ospedali, chiese e stazioni di rifornimento per la benzina, tanto per fare qualche esempio.

Prima di iniziare l'intervista gli IDU che si presentavano con la terza parte del coupon dovevano rispondere alle domande di screening per verificare la loro appartenenza alla popolazione oggetto di studio. Venivano loro chieste tre domande selezionate a caso da una lista, alle quali solo un IDU avrebbe potuto rispondere (ad es. descrivere come si prepara un'iniezione di droga) (fig. 4). Veniva anche controllato che non fossero già stati intervistati in precedenza. Lo staff registrava anche dati biometrici per valutare questo secondo aspetto (in particolare, la circonferenza del polso, la lunghezza dell'avambraccio, la circonferenza della testa) e li confrontava con quelli raccolti ed inseriti in un database per gli IDU già intervistati. Nel caso fosse comparso nel database un altro individuo con le stesse misurazioni per le tre caratteristiche, l'intervistatore doveva cercare di capire se si trovava di fronte ad una persona che si era presentata più di una volta.

Fig.4 - Domande di screening per valutare l'appartenenza del reclutato alla popolazione studiata



Preliminary Test

Province ❑ 1. Bangkok ❑ 2. Chiang Mai ❑ 3. Chiang Rai

Coupon No. |___|___|___|___|___|___|

Interviewer: Ask respondent the following questions

1. How old are you? (Age in complete year) (If exact age unknown, estimate best number)
   _____ years old
   (if the answer is not in range of 15 – 45 years old, please end the interview)

2. How long you have been an IDU?
   _____ **(if the answer is less than 1 month, please end the interview)**

3. When was the last time you injected drugs?
   _____ **(if the answer is prior of February 2005, please end the interview)**

Interviewer: Randomly select two out of these 6 questions and ask respondent

4. What are the measuring units used for heroin? (at least 2 units)
   _____

5. Normally, what is the size number of needle used for injecting drugs?
   _____

6. What does good Heroin look like?
   _____

7. What are the "Chills"? How could it happen?

8. What equipment is needed for drug injecting preparation?
   _____

9. Please describe, in general, the process of preparing drugs to inject.
   _____

Interviewer: Record the following respondent information

01 Right hand wrist _____ cm. (1 decimal place)
02 Lower right arm length _____ cm. (1 decimal place)
03 Tattoo
(1) Location on Body _____ Describe _____
(2) Location on Body _____ Describe _____
(3) Location on Body _____ Describe _____
04 Scar
(1) Location on Body _____ Describe _____
(2) Location on Body _____ Describe _____
(3) Location on Body _____ Describe _____

Interviewer:
Please analyze and compare the information number 01 – 04 with the respondent database to check if this respondent has previously been interviewed.

TEST RESULT

❑ **NOT PASS** (Please identify the reasons)
- O Age is not in range of 15 – 45 years old
- O Has injected drugs for less than 1 month
- O Last injection took place before February 2005
- O Cannot provide correct answer for drug-related questions
- O Duplicate respondent

→ ***END THE INTERVIEW***

❑ **PASS** (Meets all criteria)
→ ***CONTINUE THE INTERVIEW***

Interviewer Name: _____

Interview Date: _____/ **August / 2005**

Checked by supervisor: Signature _____ **Date:** _____/ **August / 2005**

Fonte: Lipovsek - Longfield, *Sampling Hard-to-Reach Populations-PSI Research & Metrics*, Population Services International, 2007, http://www.psi.org/resources/publications, pp. 49-50

Una volta superata con successo la fase di screening, l'intervista vera e propria poteva iniziare. Come prima informazione si cercava di identificare il tipo di relazione che legava il reclutato al reclutatore. Tutti coloro che partecipavano dovevano essere reclutati da un amico o conoscente.

Fig. 5 Domanda per verificare la relazione fra reclutato e reclutatore

Q100a	What is the relationship between you and the person who referred you to this study? ***(CHOOSE ONE)***	FRIEND SEXUAL PARTNER RELATIVE INJECTING PARTNER DRUG SELLER DRUG CUSTOMER OTHER_____ (SPECIFY)	1 2 3 4 5 6 7

Fonte: Lipovsek - Longfield, *Sampling Hard-to-Reach Populations-PSI Research & Metrics*, Population Services International, 2007, http://www.psi.org/resources/publications, p. 49

Successivamente si chiedeva all'intervistato di valutare l'ampiezza della sua rete sociale di persone appartenenti alla popolazione oggetto di studio con la seguente domanda (Lipovsek - Longfield, 2007, p. 48):

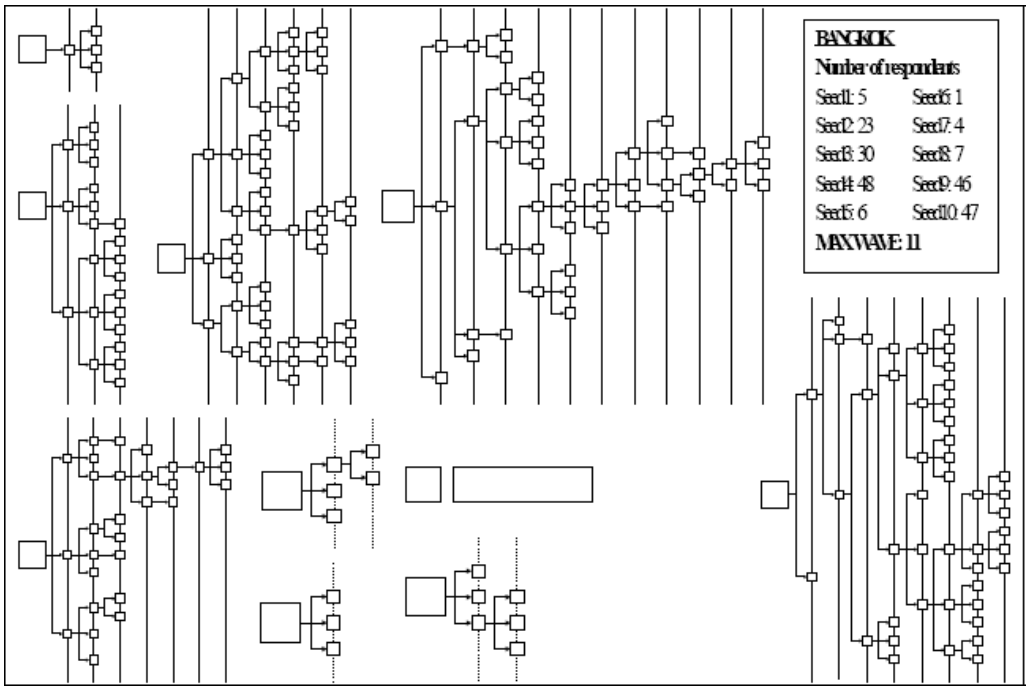
«How many drug users do you personally know, by their nickname or first name, and know how to contact them in any way?

Number: _____ drug users».

Il resto del questionario cercava di raccogliere informazioni da parte degli intervistati sulla pratica dello scambio di siringhe, sulla conoscenza dei meccanismi di trasmissione dell'HIV, sui comportamenti sessuali, sull'uso dei preservativi.

Alla fine del periodo di rilevazione l'indagine coinvolse 217 IDU che vivevano e lavoravano nella città di Bangkok. La fig. 6 mostra le catene e le ondate di reclutamento. Si può notare come un seme non reclutò nessuno, un seme originò una catena di reclutamento di una sola ondata e tutti gli altri semi originarono catene che attraversavano dalle due alle undici ondate.

Fig. 6 - Catene e ondate di reclutamento originate dai dieci semi



Fonte: Lipovsek - Longfield, *Sampling Hard-to-Reach Populations-PSI Research & Metrics*, Population Services International, 2007, http://www.psi.org/resources/publications, p. 24

4.1 Valutazione del raggiungimento della condizione di equilibrio

A questo punto si verificò se l'equilibrio era stato raggiunto per le variabili più rilevanti nella ricerca: in particolare lo scambio di siringhe per iniettarsi la droga, il sesso, l'età e il periodo di tossicodipendenza.

Come scritto nel par. 3.7, in base alla proposta degli ideatori del RDS, il numero di ondate richiesto per raggiungere l'equilibrio è simulato per una variabile usando un modello di catena markoviana nella quale la matrice di transizione è stimata a partire dalla matrice campionaria di reclutamento ("chi recluta chi").

Elemento centrale alla base della procedura è la matrice campionaria di reclutamento delle variabili oggetto di stima. Per le osservazioni successive viene considerata solo la variabile: scambio di siringhe. La matrice di reclutamento associata a questa variabile è mostrata in tab. 3 e fig. 7.

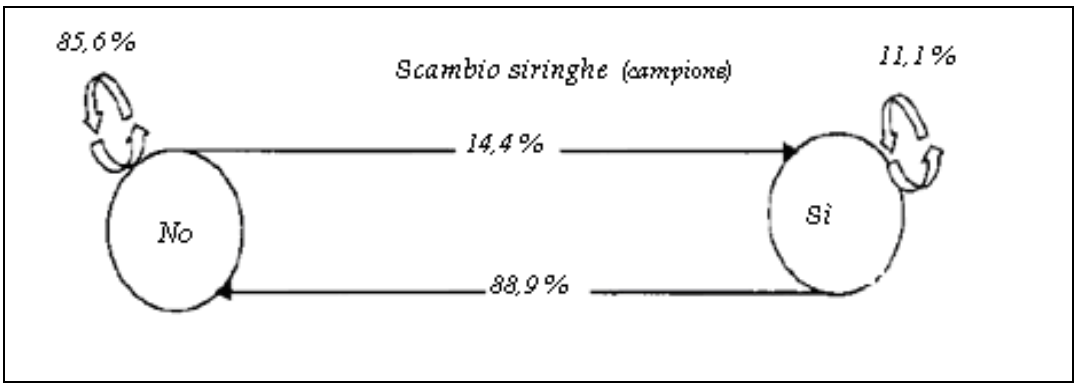
Su un totale di 207 IDU reclutati (sono esclusi i semi) 154 IDU che non scambiavano siringhe reclutarono IDU che non le scambiavano, mentre 26 IDU

Tab. 3 - Matrice di reclutamento della variabile scambio di siringhe (valori assoluti e percentuali)

		Reclutati	
Reclutatori	Scambiano siringhe		
Scambiano siringhe	No	Sì	Totale
No	154 (85,6)	26 (14,4)	180 (100)
Sì	24 (88,9)	3 (11.1)	27 (100)

Fonte: Elaborazione personale da Lipovsek - Longfield., *Sampling Hard-to-Reach Populations-PSI Research & Metrics*, Population Services International, 2007, http://www.psi.org/resources/publications, p. 24

Fig.7 - Rappresentazione grafica della matrice di reclutamento campionaria

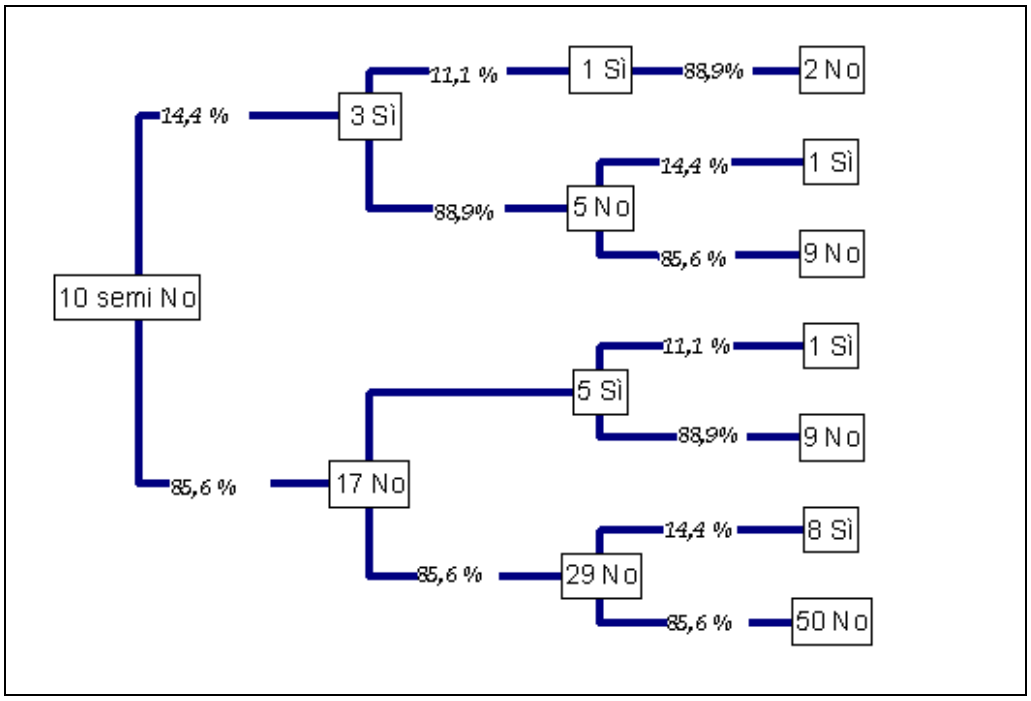


che non le scambiavano reclutarono invece IDU che le scambiavano; 24 IDU che le scambiavano reclutarono IDU che non le scambiavano, mentre 3 IDU che scambiavano siringhe reclutarono IDU che le scambiavano. Le percentuali (85,6; 14,4; 88,9 e 11,1) sono i valori utilizzati nella procedura di calcolo dell'equilibrio.

La procedura di valutazione del raggiungimento dell'equilibrio funziona nelle sue linee essenziali nel modo seguente. Nell'ondata zero si parte dalla situazione in cui tutti i semi appartengono ad un solo gruppo: ad es.10 semi non si scambiano le siringhe. Ogni seme recluta due persone. Nella prima ondata verranno reclutate 17 persone che non condividono le siringhe (85,6% di 20 reclutati) e 3 persone che le scambiano (14,4% delle 20 persone reclutate). Nella seconda ondata verranno reclutate 40 persone di cui 34 non scambiano siringhe

e 6 sì, nella terza ondata verranno reclutate 80 persone di cui 70 non si scambiano siringhe e 10 sì. Il processo simulato di reclutamento si può chiaramente seguire osservando la fig. 8 che mostra l'evolversi delle catene di reclutamento per le ondate.

Fig. 8 - Catene e ondate di reclutamento simulate per il controllo della condizione di equilibrio



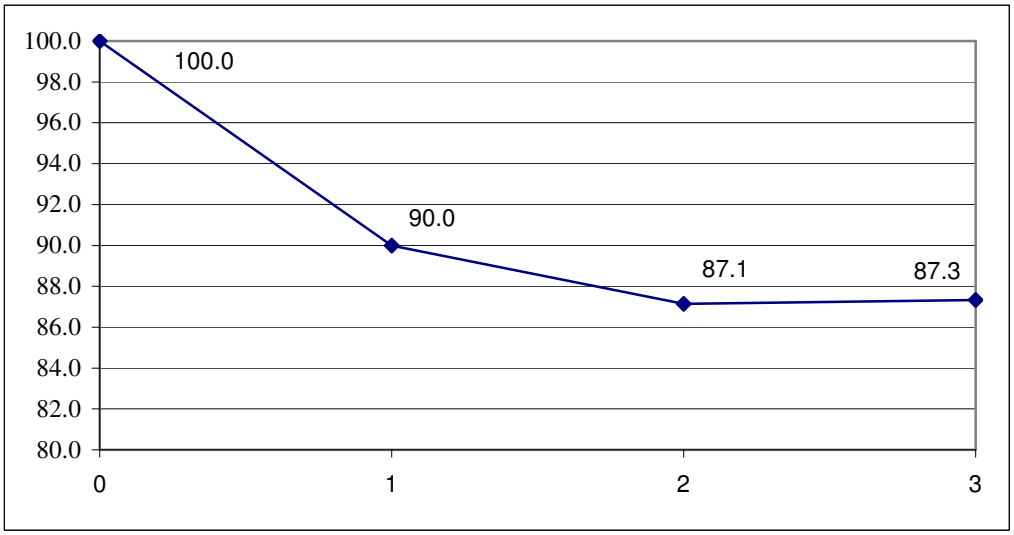
Nel grafico in fig. 9. si vede come, entro le tre ondate, le percentuali si stabilizzano. Nella terza ondata il campione simulato è costituito dall'87,3% di IDU che non scambiano le siringhe e dal 12,7% di IDU che le scambiano. Dal momento che le stime simulate dell'ultima ondata differiscono di poco rispetto a quella precedente (meno del 2% come regola) la simulazione si ferma e si ritiene raggiunta la condizione di equilibrio. Più precisamente secondo la proposta degli ideatori del RDS, il numero di ondate minimo stimato per raggiungere l'equilibrio è pari a tre. I valori 87,3% (IDU che non scambiano le siringhe) e 12,7% (IDU che le scambiano) sono chiamati "valori di equilibrio".

Questi valori di equilibrio risultano coincidere con i valori della distribuzioni nella popolazione nel caso particolare che l'ampiezza media delle reti sia uguale

nei due gruppi (Heckathorn, 2002)[25]. Il numero di ondate stimate e i valori di equilibrio sono successivamente utilizzati per valutare i risultati ottenuti nel campione raccolto con l'RDS.

Si confrontano i valori di equilibrio di una variabile con quelli calcolati nel campione (es. percentuale di IDU che non scambiano le siringhe sul totale degli intervistati). Se i valori stimati nel campione non si discostano eccessivamente da quelli di equilibrio (di solito si ritiene accettabile una differenza di massimo 4 o 5 punti percentuali) e la maggior parte delle catene supera il numero minimo di ondate stimate per raggiungere l'equilibrio si è rassicurati sulla possibilità di formalizzare il processo di reclutamento come una catena markoviana: praticamente si possono iniziare le elaborazioni calcolando le stime RDS (Heckathorn, 2002). Nel caso la differenza fra valori campionari e valori di equilibrio sia troppo elevata, nonostante, in media, il numero di ondate stimate sia stato superato, non è possibile formalizzare il processo di reclutamento come un processo di tipo markoviano e, di conseguenza, risulta molto azzardato il calcolo delle stime RDS.

Fig.9 - Valori percentuali di IDU che non scambiano le siringhe per ondata



[25] Usando un'espressione poco precisa ma che potrebbe aiutare a capire: è come se le informazioni sulla distribuzione della variabile risultassero "incorporate" nella matrice di transizione e che la procedura adottata fosse in grado di "scorporarle".

Tornando all'esempio dello studio sugli IDU in Thailandia, il numero di ondate sufficiente per il raggiungimento dell'equilibrio fu stimato in tre e la "percentuale di equilibrio" di chi non scambiava le siringhe all'87%. La percentuale di chi scambiava siringhe nel campione fu l'85%, mentre oltre il 90% dei reclutati provenivano da catene di almeno tre ondate (fig. 6). Dal momento che la percentuale stimata nel campione non si discostava eccessivamente da quella di equilibrio e che il numero medio di ondate attraversate nel campione era ben al di sopra delle ondate stimate, i ricercatori conclusero che il campione aveva raggiunto l'equilibrio e che si poteva passare alle analisi dei dati.

A questo punto furono analizzati i dati con SPSS (come noto, un software molto utilizzato nelle scienze sociali, non specificatamente pensato per l'analisi di dati provenienti da campioni RDS). Oltre alle usuali analisi mono e bivariate fu presentata una regressione logistica in cui la variabile dipendente era "scambio delle siringhe" e le variabili indipendenti erano tutte quelle risultate significativamente correlate con la variabile indipendente.

Due puntualizzazioni sono doverose a questo punto circa la procedura per il calcolo della condizione di equilibrio proposta dagli ideatori del RDS e descritta in questa sezione.

- Nella descrizione della formalizzazione del processo di reclutamento come processo markoviano e nel grafo di esempio utilizzato all'inizio del capitolo, ogni singolo soggetto selezionava casualmente dalla sua lista personale un conoscente e le singole probabilità di selezione, nel lungo termine, si stabilizzavano. Stante così le cose, ci si sarebbe potuti aspettare che la procedura proposta dagli ideatori del RDS fornisse una regola per decidere quando le singole probabilità dei soggetti si stabilizzano, raggiungono l'equilibrio (e di conseguenza si possono iniziare le analisi). Purtroppo non esistono tutt'ora in letteratura procedure siffatte. La matrice di transizione alla quale si applica una catena di tipo markoviano è, pertanto, diversa da quella considerata nel modello matematico generale presentato all'inizio del capitolo. In quel caso le probabilità inserite nella matrice erano una per ogni unità della popolazione, mentre, nell'esempio mostrato in questa sezione, le

probabilità sono associate a sottogruppi della popolazione, definiti sulla base delle modalità della variabile da stimare. La motivazione di questa scelta è dovuta alla maggiore facilità di trattamento matematico dei dati a livello aggregato. L'utilizzo di probabilità a livello aggregato, invece che individuale, non dovrebbe comportare conseguenze rilevanti in termini di risultati (Heckathorn, 1997; 2002).

- Per la stima della condizione di equilibrio proposta dagli ideatori del RDS, la matrice di transizione alla quale si applica una catena markoviana è calcolata a partire dalla matrice campionaria di reclutamento, cioè è calcolata a partire dai dati raccolti con l'indagine. Quest'ultima fornisce, mediamente, informazioni "non falsate" circa la matrice di transizione nella popolazione, (che il modello matematico mostrato all'inizio del capitolo assume come conosciuta, ma che nella realtà è ignota e si è costretti a stimare a partire dai dati del campione) solo a patto che i reclutatori selezionino casualmente all'interno della loro rete di relazioni o che le diverse scelte individuali, anche se non casuali, si combinino in modo che l'ipotesi di reclutamento casuale sia appropriata almeno a livello di aggregato.

5. Il *respondent-driven sampling*: aspetti problematici

In precedenza si è scritto che il valore del RDS è direttamente legato alla capacità di approssimare il più possibile nelle applicazioni pratiche i principi del modello statistico sottostante. Dopo aver chiarito in cosa consista un disegno RDS è ora possibile valutare questo punto.

In particolare verranno considerati alcuni aspetti dei disegni RDS che in letteratura sono oggetto di studi, riflessioni e talvolta di critiche serrate. In particolare:

1. la modalità con cui viene stimata la probabilità di un individuo di entrare a far parte del campione;
2. la modalità di reclutamento dei soggetti che fanno parte del campione;
3. il numero di ondate minimo che le catene di reclutamento devono attraversare per raggiungere l'equilibrio;
4. la possibilità di salvaguardare la logica delle catene markoviane nel caso di un campionamento senza reimmissione;
5. la precisione delle stime RDS.

5.1 La modalità con cui viene stimata la probabilità di un individuo di entrare a far parte del campione

Nelle indagini RDS l'intervistatore, oltre alle domande sostanziali, pone agli intervistati delle domande per raccogliere l' informazione relativa al numero di conoscenti della loro rete sociale che fanno parte della popolazione studiata. Ad esempio nelle indagini su popolazioni che fanno uso di droga si utilizza spesso la seguente domanda: «Quante persone che fanno uso di droga e che hai frequentato almeno una volta nell'ultima settimana conosci per nome?». In un'altra indagine, sui comportamenti omosessuali, è stata utilizzata una domanda articolata in due tempi: in un primo momento si è chiesto «Approssimativamente, quante persone conosci per nome? (Considera solo le persone che tu conosci e che conoscono te, che sei in grado di contattare e che

hai incontrato almeno una volta negli ultimi sei mesi)» e successivamente «Di queste persone che tu conosci per nome, quante diresti che sono di origine messicana e contemporaneamente omosessuali o bisessuali?» (Ramirez-Valles *et al.*, 2005).

Domande dirette di questo tipo danno per scontato che i rispondenti, oltre ad essere disponibili, siano in grado di riportare accuratamente l'ampiezza della loro rete personale. Nonostante si incontrino comunemente negli studi delle reti sociali e nonostante esistano riscontri empirici sulla loro affidabilità (almeno in termini relativi rispetto ad altri indicatori di rete) (Marsden, 1990), gli indicatori dell'ampiezza totale della rete di conoscenze pongono problemi di misura estremamente delicati (Brewer, 2000; Bell *et al.*, 2007), che non sono risolvibili semplicemente specificando con estrema precisione le caratteristiche di chi considerare nel computo. Senza giungere a posizioni radicali che attribuiscono nessun valore alle conclusioni tratte dai dati raccolti con domande di questo tipo, nondimeno la scarsa accuratezza nel valutare l'ampiezza della propria rete sociale da parte dei rispondenti rappresenta una seria fonte di errore, introducendo una distorsione di entità sconosciuta nelle stime RDS.

Nel contesto specifico dei disegni di ricerca RDS non esiste ancora una consolidata letteratura che individui quali siano i metodi di stima più appropriati. In questo campo va comunque segnalato il contributo pionieristico di Wejnert (2009) che testa, in una websurvey su una popolazione di studenti di college, differenti domande per rilevare l'ampiezza del rete di conoscenti e conclude che le stime non sono influenzate in modo marcato dalle diverse domande.

5.2 La modalità di reclutamento dei soggetti che fanno parte del campione

Nel RDS vengono lasciate all'intervistato le operazioni di selezione e di reclutamento di un altro intervistato. Secondo la teoria statistica si assume che il soggetto scelga in modo casuale fra i suoi conoscenti. Assunzione difficilmente plausibile nel caso del RDS, anche se alcune evidenze empiriche sembrano

essere state raccolte sul fatto che l'assunto della scelta casuale si sia rivelato adeguato in alcune indagini che hanno utilizzato l'RDS (Wang *et al.*, 2005; Ramirez-Valles *et al.*, 2005). Non sono comunque state considerate in modo sistematico le condizioni che identificano in quali casi l'assunto ha maggiori probabilità di essere rispettato e in quali no. La violazione di questo assunto introduce negli stimatori una distorsione, il cui segno e forza sono sconosciuti.
In ogni caso sebbene non sia plausibile aspettarsi a livello individuale la scelta di tipo casuale, sembra più ragionevole ipotizzare che le diverse scelte individuali, anche se non casuali, si combinino in modo che l'ipotesi di reclutamento casuale sia appropriata almeno a livello di aggregato (Volz - Heckathorn, 2008, p. 83).

5.3 Il numero di ondate minimo che le catene di reclutamento devono attraversare per raggiungere l'equilibrio

Si è scritto più volte che, da un punto di vista matematico, in un processo di selezione dei soggetti di tipo markoviano è possibile dimostrare che nel lungo termine, indipendentemente dai semi iniziali scelti, la frequenza di selezione di ogni persona si stabilizza. Questa affermazione implica che un certo numero di ondate è necessario per permettere alle catene di reclutamento di esplorare le diverse parti della struttura della rete e di raggiungere membri che potrebbero avere bassissime probabilità di essere inclusi come semi iniziali. Il problema pratico consiste nello stabilire quante ondate dovrebbe attraversare una catena di reclutamento.

Ricerche su reti sociali sia simulate che reali (Watts, 2003) hanno mostrato che il percorso che lega due persone è caratterizzato, nella stragrande maggioranza dei casi, da sentieri di lunghezza media abbastanza brevi: cioè prese due persone a caso nella rete sono necessari pochi intermediari per passare dall'una all'altra. Se così fosse, la probabilità di selezione per un nodo dovrebbe stabilizzarsi dopo un numero non eccessivo di passaggi (Volz - Heckathorn, 2008, p. 83). Questa considerazione si traduce nei disegni RDS nel

suggerimento di continuare il processo di selezione dei casi finché, siano state attraversate, per il maggior numero dei semi, dalle sei alle otto ondate (Salganik - Heckathorn, 2004).

A conclusioni del tutto opposte giungono invece altri autori i quali, sulla base di alcuni studi su reti simulate, affermano che nelle reti sociali è irragionevole aspettarsi che la probabilità di selezione dei nodi (cioè a livello individuale) si stabilizzi nel numero di ondate che tipicamente caratterizzano gli studi RDS (Gile - Handcock, 2010).

La questione è tutt'ora aperta. In ogni caso, se anche le conclusioni di Gile e Handcock venissero confermate, la logica del raggiungimento dell'equilibrio in un numero di ondate non troppo elevato potrebbe essere ragionevole a livello aggregato: cioè a livello di gruppi. In altri termini, focalizzando l'attenzione sulla stima di percentuali e nonostante non si giunga alla stabilizzazione completa delle probabilità di selezioni individuali, non si può escludere che le percentuali convergano, entro un numero di ondate non troppo elevate, a valori prossimi a quelli della popolazione. Alcune dimostrazioni matematiche e simulazioni sono state proposte a sostegno di questo assunto (Heckathorn 1997, 2002; Salganik - Heckathorn, 2004).

5.4 Campionamento senza reinserimento

Il modello matematico alla base del RDS implica che il campionamento sia con reinserimento (par. 1), cioè comporta il ritorno di ogni elemento campionato nella popolazione, che rimane quindi costante. Nelle applicazioni RDS il campionamento è invece senza reinserimento: quando un nodo è selezionato per far parte del campione non può più essere selezionato una seconda volta. Vengono, inoltre, rimossi tutti i percorsi di reclutamento che passano attraverso di esso.

Se la frazione campionaria è bassa e la popolazione ampia si può considerare a tutti gli effetti pratici il campionamento senza reinserimento equivalente a quello con reinserimento (Volz - Heckathorn, 2008). Questa conclusione trova

riscontro in alcune simulazioni che mostrano che l'effetto sulle stime risulta estremamente contenuto se si adotta un campionamento senza reinserimento in popolazioni ampie (Salganik - Heckathorn, 2004; Lu *et al*., 2010).

Va comunque sottolineato che, nel caso di una popolazione di ampiezza limitata, il fatto di campionare senza reinserimento può avere pesanti effetti sulla distorsione delle stime (Gile - Handcock, 2010).

Il perché di questa distorsione può essere compreso richiamando il fatto, più volte ripetuto nel capitolo, che il processo di reclutamento RDS è formalizzato come una catena markoviana e, di conseguenza, a patto di campionare con reinserimento, nel "lungo termine", indipendentemente dai semi iniziali scelti, la frequenza di selezione di ogni persona dovrebbe stabilizzarsi ed essere proporzionale al suo grado in uscita. Nel caso di campionamento senza reinserimento si verifica un allontanamento da questa situazione: le probabilità di selezione degli intervistati non sono più proporzionali ai rispettivi gradi in uscita, ma tendono ad essere simili fra loro, di mano in mano che la frazione campionaria diventa maggiore (nella situazione limite di una frazione campionaria pari ad uno, verranno campionate tutte le persone nella popolazione e di conseguenza la loro probabilità di esser incluse nel campione sarà uguale ad uno). Dal momento che lo stimatore RDS (si veda appendice 4) pesa le risposte delle persone per il reciproco del grado in uscita, quando le reali probabilità di inclusione sono simili vengono "sovrapesate" le risposte delle persone con pochi legami e "sottopesate" quelle delle persone con molti legami.

5.5 L'e*ffetto del disegno*

Ammettendo che sia possibile calcolare degli stimatori RDS non distorti sulla base dei dati campionari tuttavia, come abbiamo sottolineato anche nel precedente capitolo, la non distorsione di una stimatore da sola non è decisiva per il suo utilizzo; da un punto di vista pratico risulta più importante la variabilità dello stimatore. Le stime calcolate a partire da stimatori corretti con

un'elevata variabilità ci daranno stime poco precise (o molto disperse) attorno al valore della popolazione; al limite sono preferibili ad essi stimatori leggermente distorti ma poco variabili.

Così come nel caso del *multiplicity sampling*, la variabilità degli stimatori RDS può essere quantificata in termini dell'effetto del disegno (Kish, 1965), il quale valuta la precisione delle stime di un tipo di campionamento, in questo caso l'RDS, confrontandola con quella che si ottiene con un campione casuale semplice di pari ampiezza campionaria.

Alcuni studi che valutano l'effetto del disegno RDS sono stati condotti negli ultimi anni (Salganik, 2006; Wejnert - Heckathorn, 2008; Wejnert, 2009; Goel - Salganik, 2009, 2010; Gile - Handcock, 2010). Il contributo più rilevante fra questi è quello di Goel - Salganik (2010), i quali hanno valutato la performance del RDS simulando il campionamento e i processi di stima RDS su 85 popolazioni reali di cui disponevano sia della struttura della rete sociale che dei valori individuali di alcune variabili demografiche. I due studiosi con queste informazioni hanno potuto confrontare empiricamente le stime RDS con i valori veri dei parametri delle popolazioni e misurare la loro variabilità. Essi hanno trovato che la varianza delle stime RDS era dalla 5 alle 10 volte più elevata di quella delle stime calcolate con un campione casuale semplice di pari numerosità. Questo risultato tradotto in termini pratici significa che un campione RDS richiede numerosità campionarie cinque, dieci volte più grandi di un campione casuale semplice per ottenere lo stesso margine di errore delle stime: ad es. se con un campione casuale semplice potrebbero bastare 400 casi per calcolare attorno alle stime percentuali margini di errore del 5%, si dovrebbero estrarre dai 2.000 ai 4000 casi con un campione RDS. Questi risultati, se confermati da ulteriori studi, suggeriscono una scarsa applicabilità delle forme di campionamento RDS. Anche su questo punto la riflessione è comunque all'inizio: lo studio della variabilità degli stimatori RDS resta ancora un aspetto poco approfondito ed è stato descritto da alcuni come la "nuova frontiera" dei disegni RDS (Wejnert, 2009).

5. LA VALUTAZIONE DEL RESPONDENT-DRIVEN SAMPLING: LA RICERCA

1. La valutazione delle stime *respondent-driven sampling*: inquadramento teorico della ricerca

La seconda parte di questa tesi si caratterizza come un contributo empirico al dibattito in generale sulla validità del RDS e, in particolare, sull'accuratezza delle stime RDS cioè il grado di corrispondenza tra la stima ottenuta dall'indagine e il vero (ma ignoto) valore della caratteristica studiata nella popolazione di riferimento.

Alcune brevi considerazioni sui lavori che valutano l'accuratezza delle stime RDS[26] servirà ad inquadrare meglio la logica e le scelte della presente ricerca. Innanzitutto il loro numero è limitato: ad oggi vanno segnalati, per il rigore e la rilevanza, i seguenti contributi: Salganik - Heckathorn (2004); Salganik (2006); Volz - Heckathorn (2008); Wejnert - Heckathorn (2008); Goel - Salganik (2009, 2010); Neely (2009); Wejnert (2009); Gile - Handcock (2010) e Lu *et al.* (2010).

Sono stati esclusi dal computo studi di tipo metodologico che confrontano i risultati di campioni RDS con quelli derivati da altre tecniche alternative di campionamento, di tipo non probabilistico, applicate sulla stessa (o simile) popolazione (Ramirez-Valles *et al.*, 2005; Abdul-Quader *et al.*, 2006; Platt *et al.* 2006; Robinson *et al.*, 2006; Johnston *et al.*, 2008, 2009; Kendall *et al.*, 2008; Burt *et*

[26] I risultati più rilevanti di questi lavori sono già stati descritti e commentati nel par. 5 del cap.4 – "Aspetti problematici del RDS". In questo contesto sono ripresi da una diversa prospettiva che ne mette in luce alcune caratteristiche "formali".

al., 2010). Il motivo della loro esclusione è semplice: in tutti questi contributi i veri valori delle caratteristiche studiate risultano ignoti, pertanto non è possibile valutare, se non con larghi margini di insicurezza, l'accuratezza delle stime.

I lavori considerati sono riconducibili a due strategie di validazione: la prima di tipo analitico matematico e la seconda di tipo empirico.

La strategia di validazione analitica (Goel - Salganik, 2009; Neely, 2009) cerca di formalizzare matematicamente la relazione fra il grado di accuratezza delle stime, calcolate a partire da campioni con disegno RDS, con la variazione di alcune condizioni ed assunzioni. Purtroppo questa strategia si scontra con un considerevole problema: la procedura di campionamento RDS è una procedura molto complessa da formalizzare matematicamente. I tentativi di formalizzazione risultano, di conseguenza, spesso incapaci di incorporare caratteristiche critiche dei disegni campionari RDS che hanno un impatto rilevante sull'accuratezza delle stime.

La strategia di validazione empirica consiste, in termini generali, nell'esaminare il grado di accuratezza delle stime calcolate a partire da campioni con disegno RDS, indipendentemente da ogni formula matematica; cioè come se tali formule non esistessero. Se i risultati si discostano in modo non rilevante da quelli previsti dalla teoria statistico-matematica il metodo risulta empiricamente validato. I lavori appartenenti alla seconda strategia sono divisibili a loro volta in simulazioni (Salganik, 2006; Salganik - Heckathorn, 2004; Volz - Heckathorn, 2008; Gile - Handcock, 2010; Goel - Salganik, 2010; Lu *et al.*, 2010) o test su popolazioni reali (Wejnert - Heckathorn, 2008; Wejnert, 2009).

1.1 Validazioni di tipo empirico tramite simulazioni

L'idea che sottende l'uso delle simulazioni è, *in nuce*, molto semplice: si costruiscono popolazioni artificiali con determinate caratteristiche (Salganik - Heckathorn, 2004; Salganik, 2006; Volz - Heckathorn, 2008; Gile - Handcock, 2010) oppure si considerano popolazioni reali di cui è possibile conoscere il grafo della rete sociale e alcune caratteristiche dei singoli individui (Goel -

Salganik, 2010; Lu *et al.*, 2010) e si campiona ripetutamente da queste popolazioni. L'accuratezza delle stime è valutata calcolando la media e la varianza delle stime che provengono da questo elevato numero di campioni. Se, per esempio, la media di queste stime è vicina a quella della popolazione, artificiale o reale che sia, e presenta al contempo una variabilità contenuta, si hanno dati di supporto all'ipotesi che le procedure RDS forniscono stime accurate.

La strategia presenta notevoli vantaggi. In primo luogo permette di applicare la logica dell'esperimento. Si possono variare secondo un piano sistematico alcune caratteristiche della popolazione, tenendo sotto controllo tutte le altre, al fine di comprenderne gli effetti sulla qualità dei dati raccolti. In secondo luogo, presenta il vantaggio di rispondere a domande che non sono trattabili matematicamente.

I lavori che sfruttano la simulazione hanno permesso di approfondire la conoscenza della relazione fra accuratezza delle stime, da una parte, e struttura delle reti sociali sottostanti, distribuzione delle caratteristiche studiate nella rete e dinamiche di reclutamento, dall'altra. Nonostante questo, la strategia non è esente da problemi: le popolazioni artificiali sono spesso semplificazioni della complessità delle relazioni che caratterizzano le popolazioni reali; inoltre, le stesse procedure di reclutamento, anche nel caso di applicazione a popolazioni reali, sono semplificazioni ben lontane dalle complesse dinamiche di reclutamento che si generano durante un reale processo RDS; infine, i risultati fra le diverse simulazioni non sono facilmente comparabili dal momento che i risultati dipendono fortemente dagli assunti adottati.

1.2 Validazioni di tipo empirico con indagini-test su popolazioni reali

Questa strategia è molto articolata e implica i seguenti passaggi: l'individuazione di una popolazione di cui si conosce la distribuzione di alcune variabili; la costruzione di un campione secondo la logica RDS; la raccolta tramite questionario dei dati di interesse; il confronto, per le stesse variabili di

cui si conosce già la distribuzione, fra le stime campionarie e i valori reali noti della popolazione.

Purtroppo a causa del fatto che solo un singolo campione RDS è estratto in ogni indagine-test, è molto difficile valutare l'accuratezza delle stime. Quest'ultima nel caso di una singola rilevazione deve essere, infatti, dedotta a priori all'interno del campione dando per scontata l'applicabilità del modello matematico sottostante al metodo con cui il campione è stato costruito. Solo avendo a disposizione un congruo numero di stime calcolate a partire da diversi campioni estratti della stessa popolazione (teoricamente dovrei ripetere più volte la stessa indagine sulla stessa popolazione) è possibile controllare empiricamente l'accuratezza del metodo di campionamento. Ovviamente questo è difficilmente praticabile: la ripetizione della stessa indagine sulla stessa popolazione, oltre ad essere proibitivamente onerosa, comporterebbe un disturbo difficilmente sopportabile dalla popolazione oggetto di studio (molte delle persone dovrebbero partecipare più volte all'indagine).

Nonostante questi problemi, l'utilizzo di indagini-test risulta indispensabile per una corretta valutazione metodologica delle stime RDS, perché permette di gettare una luce sulle complesse dinamiche sociali che giocano durante il processo di reclutamento RDS e evidenzia le difficoltà dell'applicabilità in contesti reali di molti degli assunti considerati nella letteratura RDS di tipo teorico-matematico.

Il presente lavoro si inserisce all'interno della strategia empirica, ma contiene un elemento di novità rispetto ai precedenti studi: le performance delle stime RDS saranno valutate sia con l'uso di simulazioni sia svolgendo un'indagine RDS su una popolazione reale di cui è stata ricostruita la struttura completa delle relazioni che legano le persone e alcune caratteristiche di ogni persona (es. genere, età, ecc...). L'utilizzo combinato di simulazione e indagine-test dovrebbe consentire di superare alcune delle limitazioni associate alle due strategie considerate singolarmente, permettendo di sviluppare una comprensione qualitativamente più profonda di questa forma di campionamento.
La presentazione della ricerca, che occuperà questo capitolo, si articola nei seguenti punti:

1. dati: descrizione della popolazione oggetto di studio;
2. valutazione dell'applicabilità del RDS;
3. descrizione delle caratteristiche della simulazione e dell'indagine-test;
4. risultati della simulazione e dell'indagine test;
5. considerazioni conclusive.

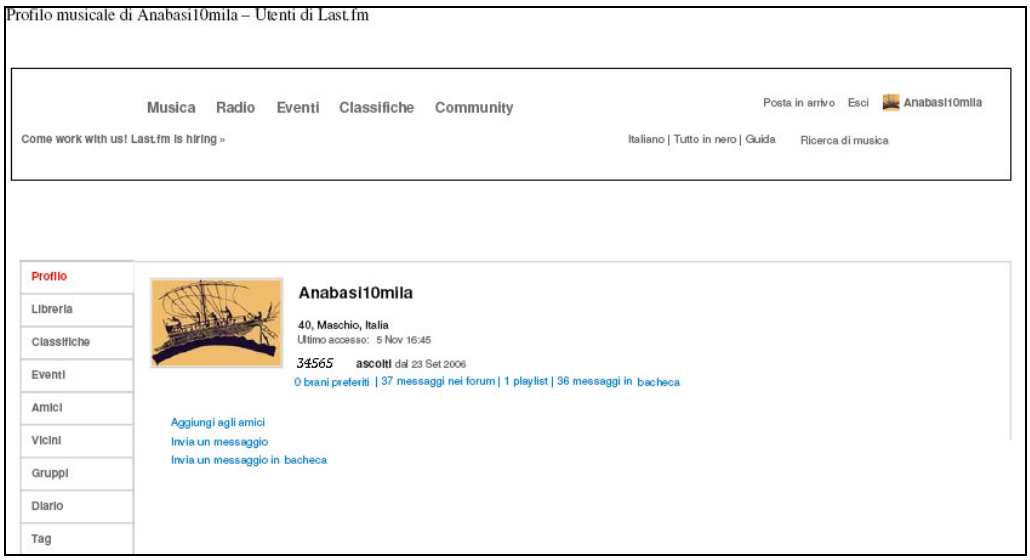
2. Dati: descrizione della popolazione oggetto di studio

La popolazione considerata è il frutto di una ricostruzione dei legami di oltre mille persone che possiedono un account su Last.Fm e che sono iscritti ad un gruppo di discussione, "Last.fm aNobii", rivolto a persone «che divorano con la stessa voracità libri e dischi», come recita lo slogan del gruppo.
Last.fm[27] è una radio su Internet e anche una rete sociale su Internet, fondata nel regno Unito nel 2002. Ad ogni utente viene attribuita, al momento dell'adesione, una pagina personale che include alcune informazioni come nome utente, genere, età, cittadinanza, data di registrazione, numero totale di brani ascoltati e lista di amici (fig. 1). Alcune informazioni come genere, età, nazionalità sono facoltative, pertanto non sono sempre presenti. Le pagine personali sono pubbliche e visibili a tutti gli altri membri. Gli iscritti possono,

[27] L'indirizzo del sito, in italiano, è http://www.lastfm.it.

inoltre, mandare comunicazioni in una bacheca virtuale, ricevere e mandare messaggi privati. Last.fm permette, infine, la formazione di gruppi di utenti con interessi in comune: per esempio fan di un cantante, stimatori di un genere musicale.

Fig. 1 - Immagine semplificata della home page del profilo di un utente

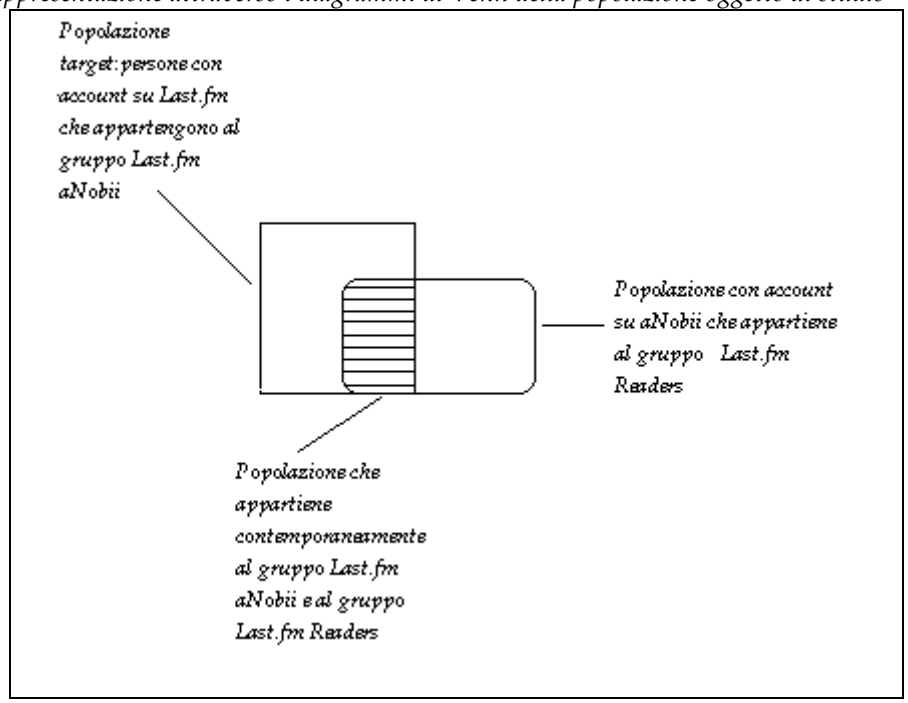


La popolazione oggetto di studio è costituita dai membri del gruppo di discussione "Last.fm aNobii" nella rete sociale Last.fm (fig. 2, indicata dal rettangolo grande).

Per comprendere i risultati di questo studio è necessario capire come è nato questo gruppo di discussione, anche se la descrizione può risultare cervellotica. Inizialmente, nel 2007, l'apertura del gruppo di discussione "Last.fm aNobii" è stata proposta da alcuni membri di Last.fm che facevano parte di un gruppo di discussione, chiamato "Last.Fm Readers", in un'altra rete sociale: aNobii[28] (indicato in fig. 2 dal rettangolo arrotondato). Essi si proponevano di costituire un gruppo "gemello" di discussione, all'interno di Last.fm, in grado di raccogliere il maggior numero di utenti che erano iscritti al gruppo "Last.Fm Readers" in aNobii (l'intersezione tratteggiata in fig. 2). Come era logico

[28] aNobii (http://www.anobii.com) è una rete sociale in Internet dedicata ai libri. Gli utenti iscritti possono mettere in linea la propria libreria attraverso i codici ISBN o un motore di ricerca interno, condividendo recensioni, commenti, votazioni, dati sull'acquisto e sulla lettura, lista dei desideri e suggerimenti con altri utenti, direttamente o attraverso gruppi.

Fig. 2 - Rappresentazione attraverso i diagrammi di Venn della popolazione oggetto di studio



Nota: le aree delle figure sono state costruite per aiutare a capire le relazioni fra le popolazioni coinvolte, non sono proporzionali alle grandezze dei rispettivi gruppi.

aspettarsi il gruppo di discussione "Last.fm aNobii" era costituito, inizialmente, solo da persone che possedevano contemporaneamente un account su Last.fm ed aNobii. In seguito il gruppo "Last.fm aNobii" si è sviluppato, secondo logiche autonome, all'interno della rete sociale Last.fm, ed ora, la maggioranza degli appartenenti a questo gruppo non è iscritta alla rete sociale aNobii.

Alla data del 15 aprile 2010, i nomi utente degli iscritti (in totale 1.451) sono stati scaricati manualmente e archiviati in un file. I diversi profili utente sono stati, successivamente, valutati sulla base della data di ultimo accesso: utenti il cui ultimo accesso risaliva a più di quattro mesi precedenti sono stati scartati (114 membri). Per ognuno dei 1.337 utenti validi iscritti al gruppo sono state scaricate le liste di amici, utilizzando un programma creato appositamente allo scopo[29]. In totale sono stati scaricati 57.777 amici; da questi sono stati eliminati tutti gli amici che non appartenevano alla popolazione obiettivo: cioè il gruppo

[29] Lfm COL - last.fm Data Collector for Windows è un programma gratuito scritto da Klaus Tockloth (http://www.easyclasspage.de/lastfm/seite-12.html). Esso permette di scaricare alcuni dati specifici di ogni utente e di archiviarli in un file XML, che può essere successivamente caricato in Excel.

di discussione "Last.Fm aNobii". Dei 1.337 utenti validi 1.080 sono risultati avere almeno una relazione all'interno del gruppo, mentre i restanti 257 non hanno nessuna relazione con altri elementi del gruppo.

A questo punto è necessaria una precisazione: la generalizzazione dei risultati delle procedure RDS si dovrà limitare solo alla persone con almeno una relazione, dal momento che i 257 casi senza relazioni all'interno del gruppo hanno una probabilità pari a zero di essere estratti in un campione RDS (cfr. cap. 4, par. 1). Per essere rigorosi la definizione della popolazione oggetto di studio necessita di una rifinitura: essa è costituita dai membri del gruppo di discussione "Last.fm aNobii" nella rete sociale Last.fm che hanno almeno un amico che fa parte anch'esso del gruppo.

Le informazioni scaricate da ogni profilo utente hanno permesso di costruire due matrici:

- la matrice di adiacenza binaria (cfr. cap. 2, par. 2) che rappresenta le relazioni del grafo della rete sociale (chi è amico di chi) e
- una matrice casi per variabili. Le variabili considerate sono "genere", "età in anni" (variabile continua), "numero di relazioni" e "iscrizione al gruppo Last.Fm Readers nella rete sociale aNobii". Le informazioni sul genere e sull'età sono state estratte, manualmente, da ogni pagina utente; l'informazione sul numero di relazioni è stata estratta da un'elaborazione della matrice di adiacenza binaria e quella sull'iscrizione al gruppo "Last.Fm Readers" nella rete sociale aNobii (d'ora in avanti per semplicità sarà utilizzata l'espressione "iscrizione ad aNobii") è stata estratta manualmente a partire dalle mail di adesione inviate alla bacheca del gruppo.

Le informazioni che hanno permesso di ricostruire i legami delle 1.080 persone e di costruire le variabili considerate sono pubbliche; pertanto, non sono state necessarie particolari autorizzazioni per la loro raccolta e non si è violata la riservatezza dei soggetti studiati.

Nei grafici in figg. 3-6 sono mostrate le distribuzioni nella popolazione di ogni variabile considerata: il diagramma a torta per le variabili "genere" e "iscrizione ad aNobii", l'istogramma per la variabile "età" e la curva di

frequenza per la variabile "numero di relazioni". Considerando la curva di frequenza del numero di relazioni si può notare come la distribuzione si caratterizzi per la presenza di un elevato numero di persone con poche relazioni e di un piccolo numero con tantissime relazioni. Questa distribuzione è abbastanza tipica: infatti è stata riscontrata in un elevato numero di reti sociali (Albert – Barabasi, 2002).

Fig. 3 - Distribuzione percentuale del genere (N=1.019)

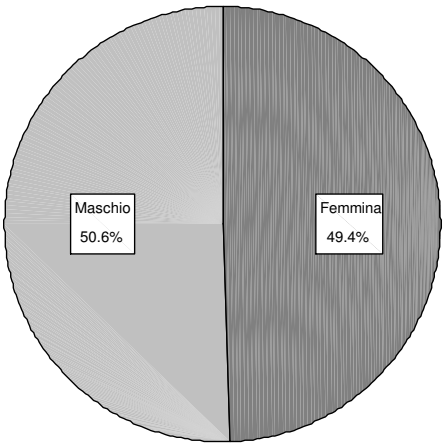


Fig. 4 - Distribuzione percentuale degli iscritti ad aNobii (N=1.080)

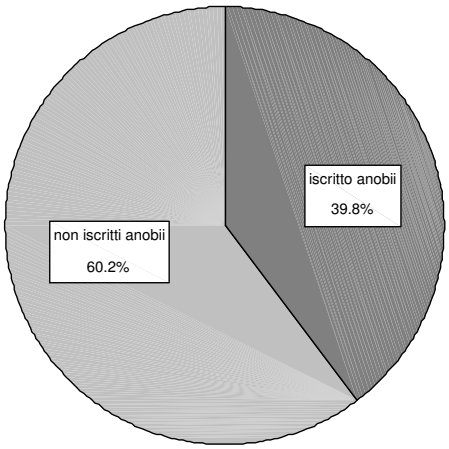


Fig. 5 - Distribuzione dell' età

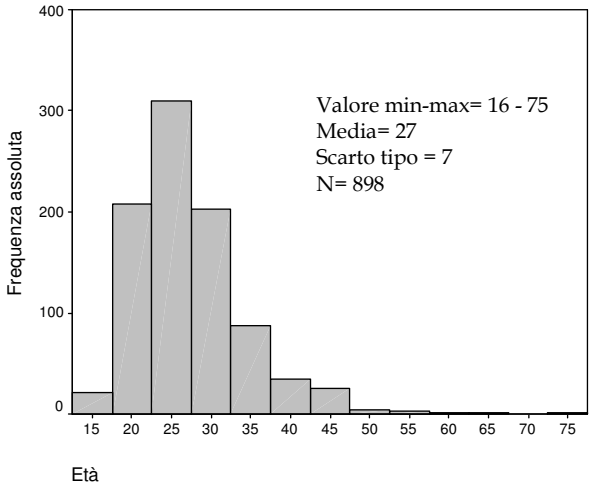
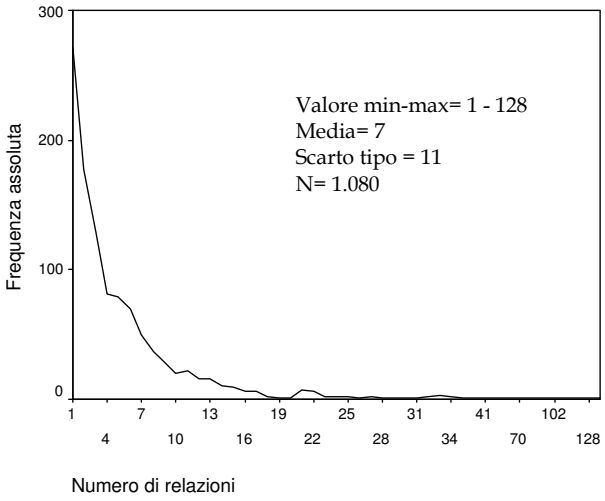


Fig. 6 - Distribuzione del numero di relazioni

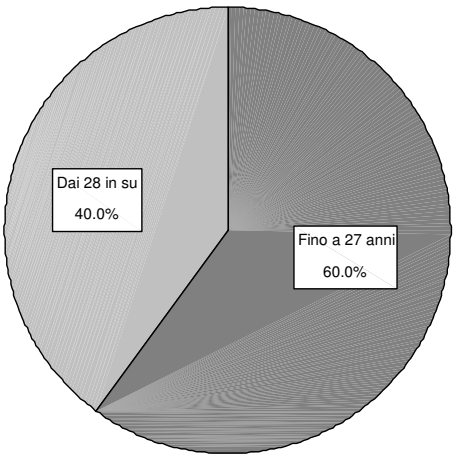


Nelle analisi delle performance del campionamento RDS le variabili considerate per la valutazione dell'accuratezza delle stime (d'ora in avanti variabili test) sono "l'età" e "l'iscrizione ad aNobii". La variabile "genere" è stata scartata: la differenza, nella popolazione, fra le frequenze percentuali di

maschi e femmine risulta così piccola da non esser rilevabile con un campione di ampiezza ragionevole.

Per semplificare le analisi e la loro successiva presentazione la variabile "età" è stata ricodificata in due categorie utilizzando, come punto di taglio, il valore della media. Il risultato è la variabile "classe di età" divisa in due modalità: "fino a 27 anni" la prima (60%) e "dai 28 anni in su" la seconda (40%) (fig. 7)

Fig. 7 - Distribuzione percentuale della variabile "classe di età" (N=898)



3. Valutazione dell'applicabilità del disegno RDS per la popolazione di riferimento

La verifica dell'applicabilità del RDS per la popolazione di riferimento è un passo necessario della ricerca. E' evidente che nel caso la popolazione non si presti all'applicazione del disegno campionario oggetto di valutazione, qualsiasi considerazione basata sui risultati della analisi sarebbe infondata.
La popolazione oggetto di studio incontra i quattro requisiti di applicabilità considerati nel precedente capitolo (cap. 4, par. 3.1).
In primo luogo, la relazione di amicizia è una relazione simmetrica: se A è nella lista di B, B sarà nella lista di A. Infatti per essere inseriti nella lista di amici di un utente è necessario inviargli una mail con la richiesta di amicizia e solo dopo

che la richiesta di amicizia viene accettata le due persone entrano a far parte delle rispettive liste di amici.

In secondo luogo, i rispondenti sono in grado di riconoscersi l'un l'altro come membri della popolazione di riferimento. Ogni membro della popolazione può sapere se un amico appartiene alla popolazione di riferimento (se cioè appartiene al gruppo di discussione "Last.fm aNobii") consultandone la lista dei gruppi ai quali è iscritto.

In terzo luogo, la rete risulta abbastanza "densa" da sostenere i processi di reclutamento e garantire che il campionamento raggiunga la "profondità sociometrica" (lunghe catene di reclutamento che attraversano ondate) per assicurare che ogni membro della popolazione abbia una probabilità diversa da zero di entrare a far parte del campione. La popolazione è infatti un esempio di *small word network* cioè un tipo di grafo nel quale la maggioranza dei nodi può essere raggiunta da qualsiasi altro nodo in un numero non troppo elevato di passaggi (Watts, 2003). Nella popolazione studiata il percorso medio che lega due persone prese a caso è di circa quattro passaggi. Secondi gli ideatori del RDS, popolazioni del tipo *small word* presentano una densità della rete adatta per l'applicazione del RDS (Volz - Heckathorn, 2008, p. 83)[30].

In quarto e ultimo luogo, la popolazione non presenta livelli di segmentazione tali da intrappolare le catene di reclutamento entro specifici sottogruppi. Infatti, considerando la popolazione divisa sulla base delle modalità delle variabili test, il 25% dei legami unisce persone che appartengono ai sottogruppi "fino a 27 anni" e "dai 28 anni in su", mentre il 29% dei legami unisce persone che appartengono ai sottogruppi "iscritto ad aNobii" e " non iscritto ad aNobii" [31].

---

[30] La distribuzione del numero di relazioni come quella della popolazione target non è come potrebbe sembrare a prima vista in conflitto con il fatto che esita una distanza relativamente contenuta fra due persone prese a caso; anzi è il contrario. Una distribuzione simili a quella della popolazione target è considerata come un indicatore di una rete del tipo *small world* (Watts, 2003).

[31] La forza della segmentazione è stata valutata anche utilizzando l'indice di segregazione proposto da Heckathorn (1997, 2002) che varia da 100% per sottogruppi con relazioni solo al proprio interno a 0% (nel caso in cui i legami fra persone siano distribuiti casualmente fra e dentro i sottogruppi) a –100% (nel caso i membri di un sottogruppo abbiano relazioni solo con i membri di un altro sottogruppo). L'indice di segmentazione per le variabili test: "classe di età" e "iscrizione ad aNobii" è rispettivamente uguale a 30% e 17%. Generalmente valori dell'indice inferiori al 50% sono considerati accettabili per l'utilizzo del RDS (Heckathorn, 1997, 2002). Per

Alcune ulteriori precisazioni sono necessarie per inquadrare correttamente le considerazioni successive. La popolazione è costituita da un numero non elevato di casi. La scelta è dovuta al desiderio, in primo luogo, di testare il disegno RDS in una situazione diffusa della ricerca (in molti casi l'RDS è utilizzato su popolazioni di ampiezza contenuta [Malekinejad et al., 2008]) ed, in secondo luogo, di valutare la problematicità del campionamento senza reinserimento nei disegni RDS nel caso di una frazione campionaria apprezzabile. Si ricorda che la teoria statistica dalla base del RDS prevede come condizione necessaria il campionamento con reintroduzione, mentre nella pratica della ricerca si campiona invece senza reintroduzione. Ai partecipanti è, infatti, proibito partecipare più volte all'indagine, in modo da scoraggiare comportamenti fraudolenti volti ad ottenere i compensi che sono quasi sempre previsti nelle indagini (cfr. cap.4, par. 3.6.3)[32].

4. Il disegno della ricerca

4.1 Le simulazioni

La simulazione è partita da otto semi che sono stati estratti casualmente dalla lista dei membri che erano nel decimo percentile più elevato per numero di amici (tab. 1). Questa scelta vorrebbe concretizzare il suggerimento da parte degli ideatori del RDS che un "buon seme" dovrebbe avere un'ampia rete sociale, numerosi legami con gli altri membri della popolazione di interesse (cfr. cap. 4, par 3.2).

---

una descrizione del calcolo dell'indice di segmentazione si rimanda a Heckathorn (2002, pp. 20-21).

[32] Altri aspetti marginali, ma utili per comprendere la scelta di questa popolazione per l'applicazione del RDS sono: gli accessi regolari e frequenti dei membri del gruppo Last.Fm aNobii (in molti gruppi i membri si iscrivono ma poi non si presentano più); sono io stesso iscritto al gruppo Last.Fm (l'effetto familiarizzazione dovrebbe attenuare, nella fase dell'indagine-test, la sensazione di sfruttamento data da un ricercatore che arriva, chiede e se ne va), l'atteggiamento di apertura che contraddistingue i membri del gruppo evidenziato dalla disponibilità a condividere e discutere nei forum ciò che amano: cioè la musica e i libri.

Tab. 1 - Caratteristiche dei semi in termini di genere, età, iscrizione ad anobii e numero di relazioni

Seme ID	*Genere*	*Età*	*Iscrizione aNobii*	*Numero relazioni*
001	2	27	1	34
002	1	30	1	30
003	2	19	1	21
004	1	31	1	13
005	1	39	2	24
006	2	33	1	15
007	2	37	2	127
008	2	38	1	42

Per escludere, ragionevolmente, che i risultati siano dovuti alle caratteristiche dei semi iniziali (appartenere al decimo percentile più elevato per numero di amici), sono stati estratti altri cento campioni e calcolate le stime RDS (e la loro variabilità) a partire da altri otto semi estratti casualmente dall'intera popolazione.

Le simulazioni fatte a partire dai semi estratti casualmente nel decimo percentile più elevato per numero di amici saranno chiamate "simulazioni base", mentre quelle che utilizzano gli otto semi estratti casualmente dall'intera popolazione saranno chiamate "simulazioni di controllo".

In entrambe i tipi di simulazioni, base e di controllo, piuttosto che tentare di riprodurre le complesse dinamiche sociali che giocano nei processi di reclutamento RDS, sono state assunte le condizioni ideali così come specificate dagli ideatori del RDS (Salganik – Heckathorn, 2004; Heckathorn, 2007; Volz – Heckathorn, 2008), ad eccezione di una. In particolare: le relazioni fra gli attori sono simmetriche; i partecipanti reclutano in modo casuale semplice a partire dalla loro lista di amici; i rispondenti riportano in modo accurato il numero di persone che fanno parte della loro lista di amici e partire dai quali hanno reclutato; il soggetto che partecipa all'indagine è estremamente motivato nel reclutare (se possibile, individuerà tra i suoi amici una persona che non è già stata intervistata); coloro che sono reclutati partecipano sempre all'indagine.

La condizione che non viene considerata è che gli individui possono essere reclutati più di una volta (condizione di campionamento con ripetizione): nella simulazione ogni individuo può partecipare una volta sola, così come nelle reali ricerche RDS. Di conseguenza, la lista da cui si estrae, ad ogni ondata, considera solo gli amici che non sono già stati estratti nelle ondate precedenti. Ogni intervistato può reclutare fino ad un massimo di due amici e le catene di reclutamento si interrompono quando un soggetto non ha alcun amico nella sua lista che non sia stato già selezionato (per una definizione formale del grafo e delle procedura di selezione si rimanda all'appendice 5).

Il processo viene ripetuto per quattro ondate: cioè il numero minimo di ondate stimato sufficiente per raggiungere l'equilibrio per livelli di segmentazione della popolazione come quelli evidenziati dalla popolazione studiata (Heckathorn, 1997, 2002)[33].

Sono stati estratti cento campioni sia per le simulazioni di base che per quella di controllo. In entrambe i casi, il processo ha generato, considerando ogni singola ondata, cento campioni parziali di prima ondata, cento di seconda ondata, cento campioni di terza ondata e cento campioni di quarta ondata.

Teoricamente se tutti le persone coinvolte in ogni ondata reclutassero due persone sarebbero selezionati, in ognuno dei cento campioni, 24 rispondenti entro la prima ondata (gli 8 semi più 16 casi della prima ondata), 56 entro la seconda (i 24 casi delle ondate precedenti più i 32 della seconda), 120 entro la terza (i 56 casi delle precedenti più i 64 della terza ondata) e 248 entro la quarta ondata (i 120 casi delle ondate precedenti più i 128 della quarta ondata).

In realtà, come vedremo nella sezione in cui si discuteranno i risultati delle simulazioni, le ampiezze campionarie nelle diverse ondate risultano inferiori a quelle teoriche (non tutti sono sempre in grado di reclutare in ogni ondata due casi) e variano da una simulazione all'altra (per il normale evolversi delle catene di reclutamento all'interno della popolazione studiata) e a seconda della variabile test considerata (per la presenza di mancate risposte).

[33] Il numero di ondate è limitato a quattro, anche per mantenere la frazione campionaria non troppo elevata (solo in questo caso è possibile considerare a livello pratico il campionamento senza reinserimento equivalente a quello con reinserimento, condizione quest'ultima necessaria per l'utilizzo del RDS – cfr. 4, par. 5.4).

Si sarebbe potuto ragionare in termini di ampiezze campionarie costanti (fermandosi, ad es., indipendentemente dal numero di ondate solo quando in ogni campione fossero stati selezionati 200 casi) e non di ondate; questo avrebbe permesso di fare confronti fra campioni simulati di ampiezza uguale. E' stata, ritenuta migliore la scelta di ragionare in termini di ondate, lavorando su campioni di ampiezza variabile per due motivi. Il primo motivo è di ordine teorico: l'importanza che, nella letteratura RDS, viene posta sul concetto di ondata, di iterazione dei processi di selezione (basti pensare, ad esempio, alle procedure di calcolo dell'equilibrio) e non tanto di ampiezza campionaria. Il secondo è di ordine pratico: a parità di ondata le variazioni dell'ampiezza campionaria non sono così rilevanti da preoccupare per il loro impatto sulle considerazioni sostantive circa le performance del RDS.

Per poter operare il processo di selezione dei casi a partire dal grafo della popolazione è stato utilizzato un programma[34], chiamato *Snowball*, scritto appositamente per questa ricerca e il cui algoritmo è mostrato in appendice 5.

4.2 L'indagine-test

L'indagine–test è consistita in una web-survey[35]. Per una popolazione elettronicamente connessa la scelta di somministrare il questionario per via telematica è apparsa del tutto naturale. Inoltre, questa modalità è risultata particolarmente attraente considerate anche le limitate risorse, sia monetarie che di personale, a disposizione. Infine, l'indagine web garantisce le condizioni ideali per una crescita esponenziale del campione così come ipotizzata nelle riflessioni teoriche del RDS.

La procedura RDS è iniziata con l'invio di una mail di partecipazione alle stesse persone che sono state utilizzate come semi nella simulazione. Questa

[34]Per avere gratuitamente il programma corredato di una breve documentazione inviare una mail a: *alberto.vitalini@unicatt.it*

[35]Il disegno dell'indagine web ricalca quella di Wejnert - Heckathorn (2008) ai quali sono debitore di numerosi suggerimenti. Per creare il questionario, raccogliere le risposte e scaricare i dati sono stati utilizzati gli strumenti gratuiti forniti nel sito *Kwik Survey* (http://www.kwiksurveys.com/).

scelta, almeno in linea di principio, permette di confrontare i risultati dell'indagine-test con quelli delle simulazioni.

La mail di invito informava brevemente dei propositi del progetto rimandando ad un pagina Web per una descrizione più articolata della ricerca e per iniziare la compilazione del questionario (si veda appendice 6).

Le informazioni richieste nel questionario erano alcune di quelle già presenti nel pagina del profilo utente: in particolare "genere"; "età"; "nome utente"; "artista, album e brano ascoltato maggiormente negli ultimi tre mesi"; "numero di amici nella rete sociale Last.fm" e "numero di amici nel gruppo di discussione Last.fm aNobii".

Il rispondente non doveva rendere manifesto nulla di più di quanto aveva già deciso di fare nella rete sociale Last.fm. Questa scelta mirava a rassicurare l'intervistato, velocizzare l'intervista [36] e limitare, nel rispetto delle norme etiche della ricerca, i rischi connessi alla partecipazione all'indagine e dovuti alla rivelazione di informazioni di tipo personale.

Alla fine del questionario si chiedeva all'intervistato il permesso di inoltrare, con l'indicazione che egli aveva già preso parte all'indagine, una mail di invito alla ricerca ad alcuni dei suoi amici.

Successivamente, venivano scelti, casualmente dalla lista di amici degli intervistati che avevano espressamente fornito il consenso, tre nominativi ai quali veniva inviata la mail di invito all'indagine. Gli intervistati che non fornivano il consenso terminavano la catena di reclutamento Come nella simulazione ogni individuo partecipava una volta sola all'indagine. Di conseguenza, la lista di estrazione, ad ogni ondata, comprendeva solo gli amici che non erano già stati estratti ed invitati nelle ondate precedenti.

Questa procedura, a prima vista macchinosa, è stata pensata per salvaguardare la logica, precipua del RDS, del reclutamento basato su legami simmetrici e contemporaneamente garantire la condizione, altamente problematica nella ricerca reale, che il reclutamento dei soggetti che fanno parte

[36]L'intervista, pur nella sua semplicità, ha una durata media dai dieci ai quindici minuti. E' stato possibile calcolare il tempo di intervista dal momento che nel file di dati è contenuta l'informazione circa l'ora di inizio e quella di fine dell'intervista.

del campione sia il frutto di una selezione casuale fra i componenti della propria rete di conoscenze.

Il processo è stato ripetuto per chi era stato invitato e partecipava all'indagine fino a che le catene di reclutamento non si interrompevano o per mancanza di amici che non erano già stati invitati o per non risposta o per non concessione dell'utilizzo del nome. E' stata prevista una seconda mail di invito a distanza di circa un mese dall'invio della prima per i membri contattati che non avevano ancora risposto. L'indagine è durata due mesi: è iniziata a mezzanotte del 17 aprile 2010 ed è terminata a mezzanotte del 17 giugno 2010.

5. I risultati

5.1 Le simulazioni

Nelle tabelle 2 e 3 sono mostrate, per ondata, alcune informazioni circa le numerosità dei cento campioni estratti nelle simulazioni (non viene mostrata l'ondata 0 che contiene i semi). Si può osservare che la numerosità dei campioni varia a seconda della variabile: considerando i casi selezionati in quattro ondate i casi validi per il calcolo delle risposte sono, in media, 158 per la variabile "classe di età" e 188 per la variabile "iscrizione ad aNobii".

L'ampiezza campionaria varia anche all'interno della stessa ondata. Questo fatto, come già anticipato, potrebbe essere visto come un problema quando si confronteranno le distribuzioni delle stime all'interno di una stessa ondata, dal momento che le stime sono calcolate a partire da numerosità campionarie non costanti. In realtà da un punto di vista pratico le variazioni dell'ampiezza campionaria non sono così rilevanti da preoccupare per il loro impatto sulle considerazioni sostantive circa le performance del RDS. Considerando i casi selezionati in quattro ondate la distanza interquartilica è di 15 casi per la variabile "classe di età" e di 12 casi per la variabile "iscrizione ad aNobii". Lo svantaggio di utilizzare campioni di ampiezza non costante viene più che

compensato dal fatto che i ragionamenti potranno essere svolti in termini di ondate, concetto più aderente alla riflessione teorica RDS.

Tab. 2 – Alcune informazioni statistiche relative alla numerosità dei campioni estratti per la variabile "classe di età", a seconda dell'ondata di appartenenza

	ondata1	ondata2	ondata3	ondata4
Valore minimo	10	33	73	134
1° quartile	13	38	80	150
Mediana	14	40	84	158
3° quartile	15	41	87	165
Valore massimo	16	46	97	180
Intervallo valore max- min	6	13	24	46
Media	14	39	84	158
Scarto tipo	1	3	5	10

Nota: L'ondata1 comprende i casi della prima ondata più i semi, l'ondata2 comprende i casi delle prime due ondate più i semi; l'ondata3 comprende i casi delle prime tre ondate più i semi, l'ondata4 comprende tutti i casi selezionati nelle quattro ondate più i semi

Tab. 3 Alcune informazioni statistiche relative alla numerosità dei campioni estratti per la variabile "iscrizione ad aNobii", a seconda dell'ondata di appartenenza

	Ondata1	Ondata2	Ondata3	Ondata4
Valore minimo	15	42	84	155
1° quartile	16	45	96	183
Mediana	16	46	98	189
3° quartile	16	47	101	195
Valore massimo	16	48	108	211
Intervallo valore max- min	1	6	24	56
Media	16	46	98	188
Scarto tipo	0	2	4	10

Nota: L'ondata1 comprende i casi della prima ondata più i semi, l'ondata2 comprende i casi delle prime due ondate più i semi; l'ondata3 comprende i casi delle prime tre ondate più i semi, l'ondata4 comprende tutti i casi selezionati nelle quattro ondate più i semi

A partire dai dati delle simulazioni sono state calcolate, per ogni ondata, le stime RDS e le stime semplici delle variabili "classe di età" e "iscrizione ad aNobii". Per mostrare le diverse distribuzioni delle stime si farà uso dei grafici a scatola (*box-plot*) in figg. 8-9. Ogni grafico a scatola fornisce una concisa indicazione dell'andamento della distribuzione delle stime nelle diverse ondate (non viene mostrata l'ondata 0 che contiene i semi). La prima scatola mostra la distribuzione delle stime per i cento campioni che comprendono i casi

selezionati entro la prima ondata (comprende anche i semi), la seconda scatola la distribuzione delle stime per i cento campioni che comprendono i casi selezionati entro la seconda ondata, la terza scatola la distribuzione delle stime per i cento campioni che comprendono i casi selezionati entro la terza ondata e la quarta scatola mostra la distribuzione delle stime per i cento campioni che comprendono i casi selezionati in tutte le ondate considerate.

La scatola vera e propria fornisce, contemporaneamente, due informazioni chiave per valutare l'accuratezza delle stime: il valore della media di tutte le stime, indicato dalla linea trasversale, e il valore della variabilità delle stime indicato dalla lunghezza della scatola. Quest'ultima non è altro che la distanza interquartilica: infatti la scatola contiene il 50% dei casi (cioè quelli che cadono fra il 25mo e il 75mo percentile) e, di conseguenza, più la scatola è allungata maggiore è la variabilità delle stime[37].

Osservando i grafici a sinistra (figg. 8-9) risulta immediato che, all'aumentare delle ondate, consistentemente con le ipotesi teoriche avanzate dagli ideatori del RDS, la distorsione delle stime RDS diventa minore: infatti, passando dalla prima alla quarta ondata, la linea trasversale che indica il valore medio si avvicina, per entrambe le variabili test, al valore reale della popolazione indicato dalla linea che attraversa il grafico. Considerando le stime RDS calcolate a partire dai dati raccolti in quattro ondate, la distanza fra la media delle stime e il valore reale nella popolazione è inferiore di 4 punti per la percentuale di persone fino a 27 anni di età e di 2 punti per la percentuale di persone iscritte ad aNobii.

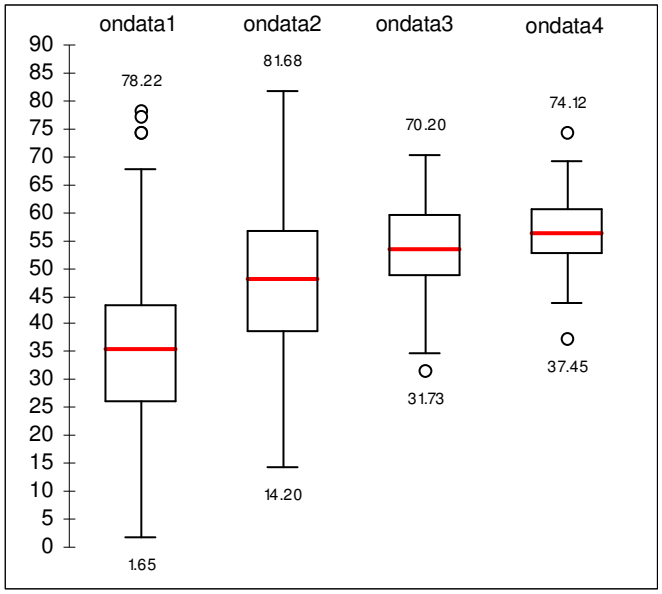
Inoltre, sempre coerentemente con le ipotesi teoriche avanzate dagli ideatori del RDS, le stime RDS (che pesano inversamente le risposte di ogni singolo soggetto per il reciproco del numero di amici- cfr. appendice 4 e cap. 4, par. 3.8) evidenziano migliori risultati rispetto alle stime campionarie semplici (che non sono pesate). Il confronto fra i grafici delle stime RDS e di quelle semplici (figg.

[37] La scatola presenta due altri appendici, chiamate "baffi". Il baffo superiore ed inferiore indicano rispettivamente il valore più elevato e più piccolo della distribuzione non considerati fuori dalla distribuzione (outlier), cioè inferiori o uguali a 1.5 volte la lunghezza della scatola. I valori al di là dei baffi, rappresentati ognuno da un cerchietto, sono considerati degli outlier.

8-9) mette in evidenza livelli di distorsione delle prime, decisamente minori per entrambi le variabili test.

Fig. 8 - Distribuzione per ondata delle stime RDS della frequenza percentuale di persone che dichiarano di avere fino a 27 anni. Stime calcolate a partire da cento campioni estratti dalla popolazione oggetto di studio (N=1.080). La linea trasversale indica la percentuale, nella popolazione, di chi dichiara di fino a 27 anni (60%)*

Stime RDS (pesate)



Stime semplici (non ponderate)

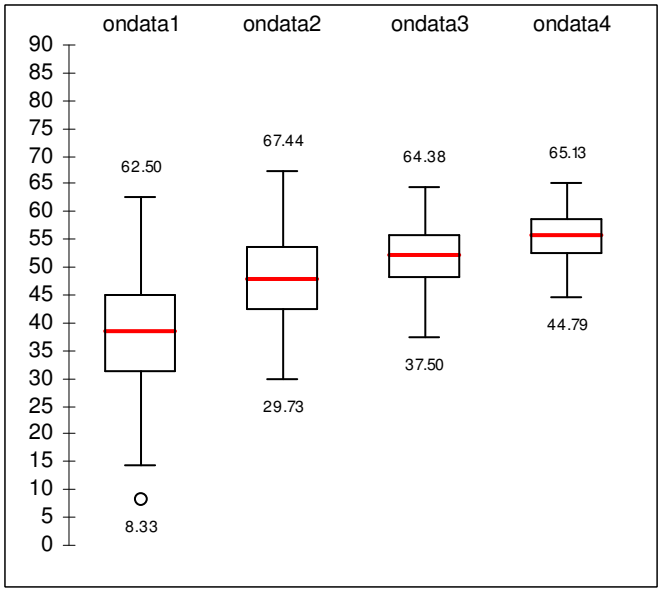
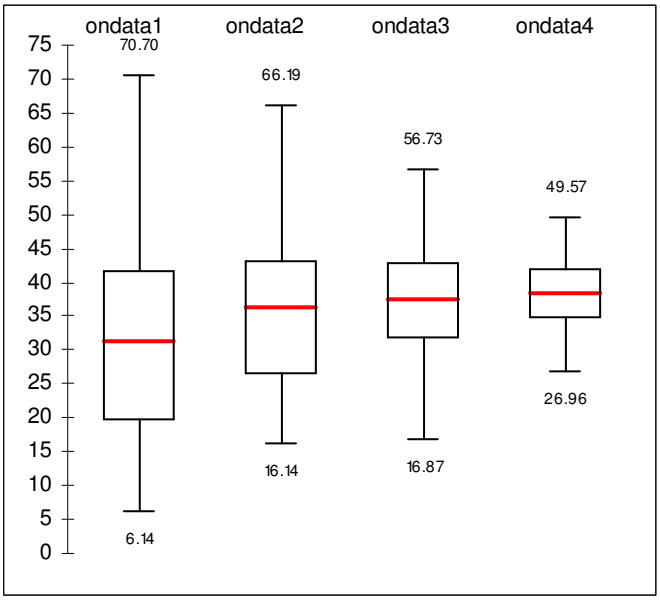
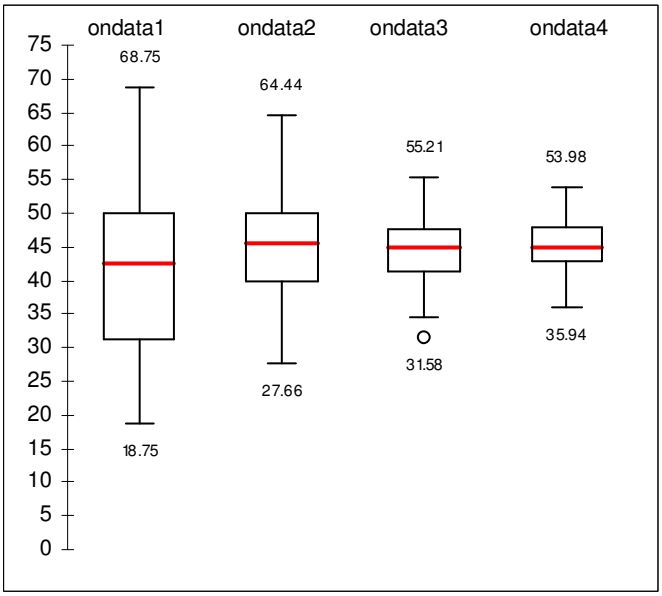


Fig 9 - Distribuzione per ondata delle stime RDS della frequenza percentuale di persone che dichiarano di essere iscritte ad aNobii. Stime calcolate a partire da cento campioni estratti dalla popolazione oggetto di studio (N=1.080). La linea trasversale indica la percentuale, nella popolazione, di chi dichiara di essere iscritto ad aNobii (39,8%).*

Stime RDS (pesate)



Stime semplici (non pesate)



** Nota: L'ondata1 comprende i casi della prima ondata più i semi, l'ondata2 comprende i casi delle prime due ondate più i semi; l'ondata3 comprende i casi delle prime tre ondate più i semi, l'ondata4 comprende tutti i casi selezionati nelle quattro ondate più i semi.*

Le simulazioni di controllo rimandano un medesimo quadro (si veda appendice 7): al crescere delle ondate la distorsione delle stime diventa minore ed, inoltre, il confronto fra le stime RDS (che pesano inversamente le risposte al numero di amici) rispetto a stime non pesate evidenziano una minore distorsione fra le prime.

A questo punto i risultati delle simulazioni concordano sostanzialmente con le ipotesi degli ideatori del RDS (cioè date le condizioni poste nei modelli matematici, i risultati del RDS sono in linea con le aspettative teoriche) e questo, pur non essendo una dimostrazione della validità del RDS, ne avvalora l'impianto concettuale.

Rassicurati su questo punto possiamo cercare di valutare l'accuratezza delle stime RDS. Considerando la distribuzione delle stime RDS calcolate a partire dai dati selezionati nelle quattro ondate, il 70% di esse non si discosta dal valore reale nella popolazione per più di 7 punti percentuali per la variabile "classe di età" (le stime sono, cioè, comprese nell'intervallo 60,0% ± 7% il 70% delle volte) e per più di 6,3 punti percentuali per la variabile "iscrizione ad aNobii" (le stime RDS sono, cioè, comprese nell'intervallo 39,8% ± 6,3% il 70% delle volte).

Ci si può chiedere se questi sono dei buoni risultati? Non è possibile dare, in assoluto, una riposta affermativa o negativa; essa dipende dagli obiettivi teorici che i ricercatori si pongono, dai vincoli di tempo e di denaro che pesano su un'indagine. Nonostante questa precisazione, si può ragionevolmente affermare che, con campioni di ampiezza contenuta (campioni di circa 160 per la variabile "classe di età" e di 190 casi per la variabile "iscrizione ad aNobii"), si hanno buone probabilità che la stima fornisca un'informazione corretta sulla popolazione: cioè che i giovani fino a 27 anni e i non iscritti ad aNobii sono la maggioranza nella popolazione.

Un'ulteriore possibilità di valutazione dei risultati consiste nel confrontare le stime RDS rispetto ad uno standard di riferimento: i risultati che si otterrebbero utilizzando il campionamento casuale semplice. Estraendo dalla popolazione studiata, con un campionamento casuale semplice, un centinaio di campioni di 160 casi ciascuno ci aspetteremmo che il 95% delle stime delle frequenze relative

delle persone fino a 27 anni sia compreso nell'intervallo 60,0% ± 7%, mentre estraendo un centinaio di campioni di 190 casi ciascuno ci aspetteremmo che il 95% delle stime delle frequenze relative di iscritti ad aNobii cada nell'intervallo 39,8% ± 6,3%.

Coerentemente con le riflessioni fatte nei precedenti capitoli (cfr. cap. 4, par. 5.5) le stime provenienti da campioni RDS risultano meno precise rispetto a quelle provenienti da campioni casuali semplici: la loro distribuzione attorno al dato reale risulta più dispersa. Nel caso specifico, l'effetto del disegno dovuto al RDS è circa 2: il campione RDS richiede, cioè, una numerosità campionaria due volte più grande di un campione casuale semplice per ottenere lo stesso margine di errore delle stime delle variabili test.

5.2 Indagine-test

Il numero di persone alle quali è stata inoltrata la mail di partecipazione erano 295, di queste 154 (semi compresi) hanno preso parte all'indagine. Considerando le ondate generate dai singoli semi (tab. 4), si può notare che il seme 006 origina la più lunga catena nel campione che attraversa 13 ondate e comprende il 59% dei reclutati. La presenza di un "super" seme è coerente con altri ricerche RDS (Weijnert - Heckathorn, 2008). Le catene costituite da almeno quattro ondate comprendono il 95% dei casi. Un seme non ha prodotto nessun reclutamento.

Tab. 4 - Numero di ondate e numero di persone reclutate per seme

Seme ID	Numero di ondate generate	Numero di persone reclutate
001	3	6
002	4	9
003	1	1
004	5	16
005	4	15
006	13	86

007	6	13
008	0	0

Nota: l'ondata che contiene i semi è considerata come ondata 0. I semi non sono considerati in questa tabella.

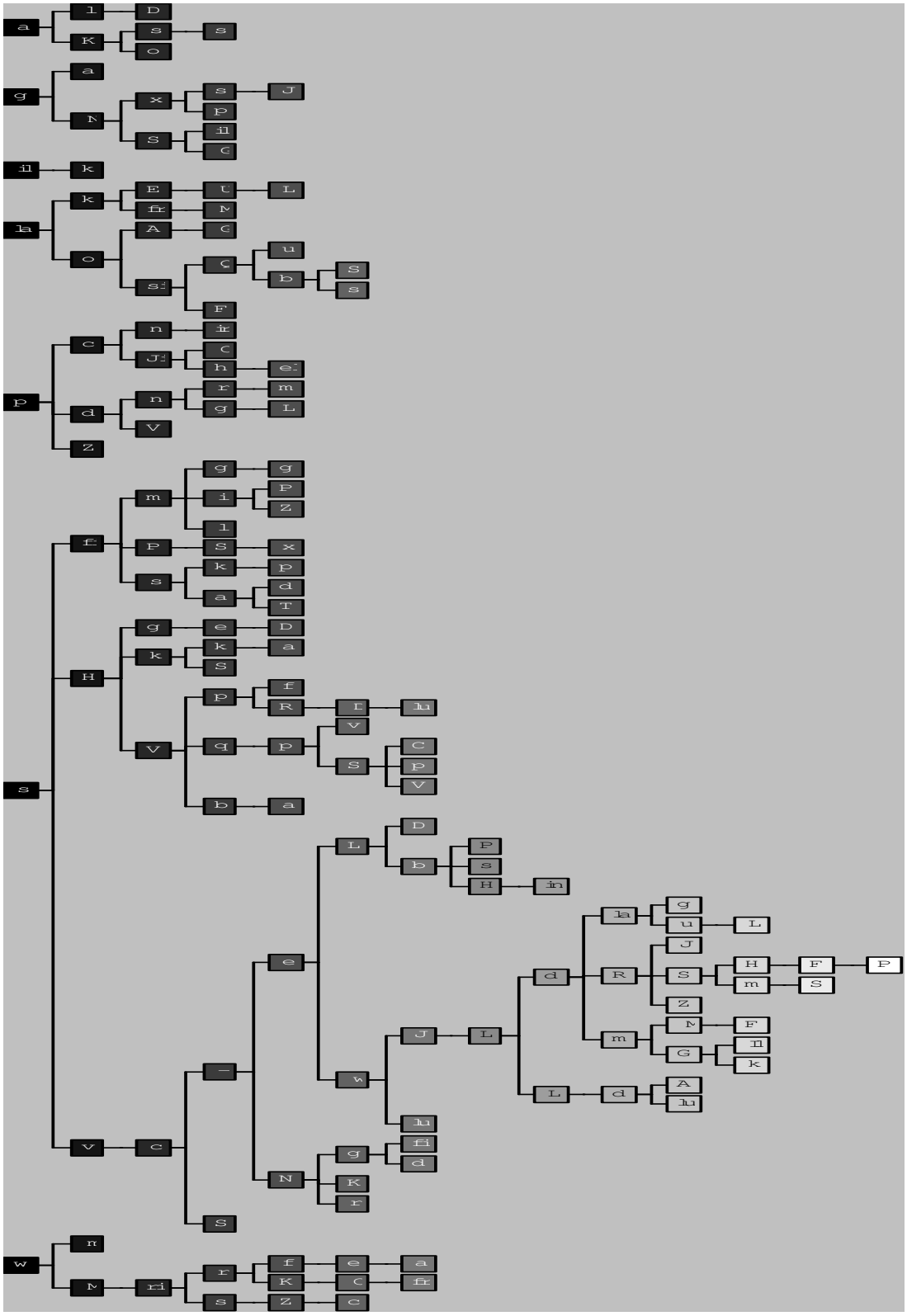
In fig. 10 si possono osservare le diverse catene e ondate di reclutamento a partire dai sette semi "fertili".

Prima di procedere con le analisi è necessario valutare il raggiungimento della situazione di equilibrio (cfr. cap. 4, par. 3.7). Qualsiasi riflessione sulle performance del RDS sarebbe priva di sostegno nel caso l'equilibrio non fosse stato raggiunto. Questa operazione, come abbiamo visto nell'esempio del capitolo precedente (cfr. cap. 4, par. 4.1), viene fatta simulando il numero minimo di ondate da raggiungere per ogni variabile a partire dalle matrici di reclutamento campionarie[38] e comparando, successivamente, il numero di ondate stimato sufficiente per raggiungere l'equilibrio con quello raggiunto nel campione. Applicando la procedura nel caso delle due variabili test[39], il numero di ondate simulate sufficiente per raggiungere l'equilibrio è risultato rispettivamente uguale a quattro per la variabile "classe di età" (valore di equilibrio della percentuale di persone fino ai 27 anni pari a 61,4%) e a tre per la variabile "iscrizione ad aNobii" (valore di equilibrio per la percentuale di iscritti ad aNobii pari a 61,3%).

[38] Le matrici di reclutamento delle variabili test sono presentate in appendice 8.

[39] Per il calcolo si è utilizzata la funzione interna al programma RDSAT 6.01. Il programma si può scaricare gratuitamente dal sito: http://www.respondentdrivensampling.org/

Fig. 10 - Catene e ondate di reclutamento a partire dai sette semi che hanno reclutato



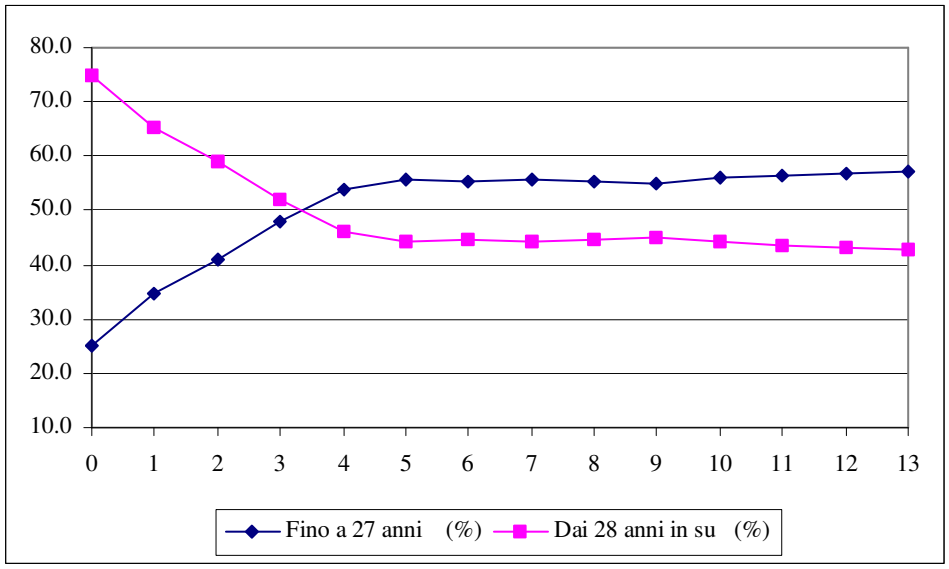
Considerando i dati campionari raccolti (tab. 5), all'aumentare delle ondate, le percentuali delle variabili test si stabilizzano, cioè l'aggiunta di una maggior numero di casi non modifica praticamente i risultati. I grafici (fig. 11-12) illustrano con immediatezza questo processo, mostrando un graduale assestamento della curva verso una posizione più o meno statica, di mano in mano che aumentano il numero di ondate e le persone intervistate.

Dal momento che i valori stimati nel campione per le variabili test non si discostano eccessivamente da quelli di equilibrio (57,1% vs 61,4% per la variabile "classe di età" e 61,7% vs 61,3% per la variabile "iscrizione ad aNobii") e che le ondate con almeno quattro iterazioni comprendono il 95% dei casi si può tranquillamente ritenere, sulla base delle procedure proposte dagli ideatori del RDS, che il processo di reclutamento è penetrato profondamente nella popolazione target e che la condizione di equilibrio è stata raggiunta.

Tab. 5 - Composizione del campione per variabile e ondata di reclutamento

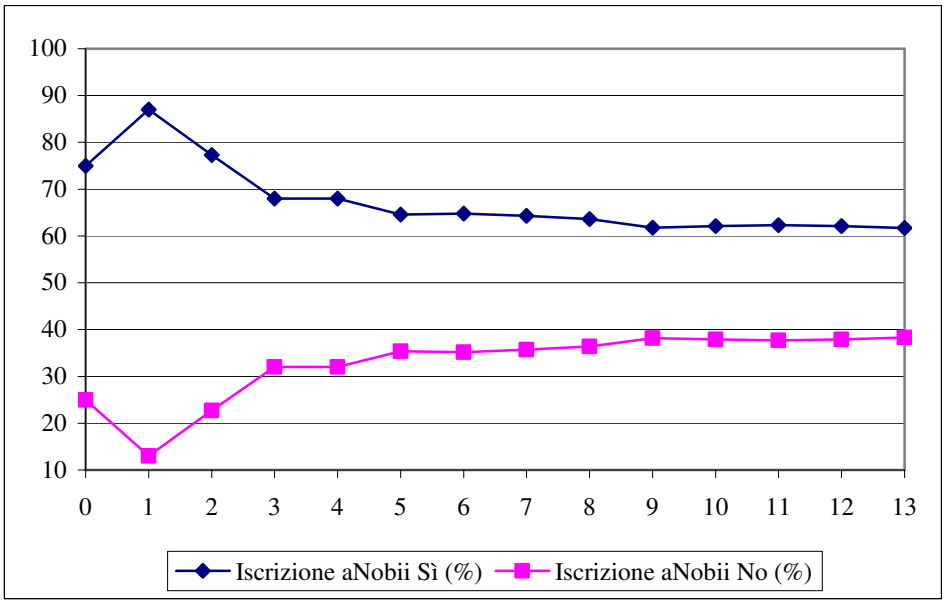
Ondate	n	Differenza in n	Classe di età – Fino a 27 anni (%)	Iscrizione aNobii Sì (%)i
0	8	-	25.0	75.0
0-1	23	15	34.8	87.0
0-2	44	21	40.9	77.3
0-3	75	31	48.0	68.0
0-4	100	25	54.0	68.0
0-5	113	13	55.8	64.6
0-6	125	12	55.2	64.8
0-7	129	4	55.8	64.3
0-8	132	3	55.3	63.6
0-9	136	4	55.1	61.8
0-10	145	9	55.9	62.1
0-11	151	6	56.3	62.3
0-12	153	2	56.9	62.1
0-13	154	1	57.1	61.7

Fig. 11 - Spezzata che unisce le frequenze percentuali dei rispondenti che dichiarano di avere fino a 27 o dai 28 anni in su, per ondata



Nota: L'ondata0 comprende i semi, l'ondata1 comprende i casi della prima ondata più i semi, l'ondata2 comprende i semi più i casi delle prime due ondate; l'ondata3 comprende i semi più i casi delle prime tre ondate e così via fino all'ondata13 che comprende tutti casi del campione.

Fig. 12 - Spezzata che unisce le frequenze percentuali di rispondenti che dichiarano di essere iscritti o meno ad aNobii, per ondata



Nota: L'ondata0 comprende i semi, l'ondata1 comprende i casi della prima ondata più i semi, l'ondata2 comprende i semi più i casi delle prime due ondate; l'ondata3 comprende i semi più i casi delle prime tre ondate e così via fino all'ondata13 che comprende tutti casi del campione.

Dopo aver verificato che la condizione di equilibrio è stata raggiunta, le stime campionarie si devono ponderare per il reciproco del grado in uscita ottenendo, in questo modo, le stime RDS (cfr. cap. 4, par. 3.8 e appendice 4). Come descritto in precedenza, gli ideatori del RDS suppongono di poter conoscere il valore del grado in uscita (il cui reale valore è ignoto) utilizzando la risposta dell'intervistato ad una domanda sul numero di persone della propria rete sociale che fanno parte della popolazione studiata (cfr. capitolo 4, par. 3.8). Questo presuppone che i rispondenti riportino in modo accurato il numero di persone che fanno parte della loro lista di amici e a partire dalle quali hanno reclutato. Si è già sollevata la problematicità di questo presupposto e delle conseguenze circa una sua violazione nella ricerca (cfr. capitolo 4, par. 5.1) .

Per tenere sottocontrollo questa possibile fonte di distorsione è stato utilizzato, come peso di ponderazione, il numero reale degli amici calcolato a partire dalla matrice di adiacenza binaria della popolazione. In ogni caso nel questionario era inserita, con funzioni di controllo, la domanda circa l'ampiezza della rete al fine di valutare l'attendibilità delle possibili risposte. Il risultato non è dei più rassicuranti: il 60% delle risposte è inutilizzabile (non risponde o non so).

Ci si potrebbe chiedere la motivazione di questo comportamento da parte di persone che hanno dimostrato la loro disponibilità a collaborare all'indagine. Una possibile spiegazione è legata al fatto che, per conoscere chi dei propri amici faceva parte del gruppo Last.fm aNobii, il rispondente doveva entrare nella pagina di ognuno dei suoi amici e controllare l'elenco dei gruppi ai quali era iscritto. E' facile immaginare che lo sforzo richiesto, soprattutto nel caso di un elevato numero di amici, ha scoraggiato molti intervistati che hanno risposto in modo sbrigativo.

Il fatto che l'informazione fornita dai rispondenti sia inutilizzabile non è un problema in questa indagine-test, dal momento che si dispone del dato reale. In ogni caso questo risultato ribadisce la problematicità, già peraltro evidenziata abbondantemente in letteratura, della raccolta di un'informazione critica per il funzionamento del RDS.

Le stime RDS delle variabili "classe di età" e "iscrizione ad Anobii" sono presentate in figg. 13-14 (lato destro). Accanto ad ognuna è mostrato anche il corrispondente valore del parametro della popolazione (lato sinistro)[40].

Considerando la variabile "classe di età" il risultato campionario rimanda un'immagine non distorta del valore del parametro della popolazione: esso differisce per meno di 2 punti percentuali dal valore reale nella popolazione (con un intervallo di confidenza al 95% che va da 49,9% al 71,3[41]).

Senza voler fare affermazioni definitive se ci fermassimo a questo punto la nostra fiducia nelle possibilità di utilizzo del RDS sarebbe ragionevolmente elevata a patto di poter ottenere stime accurate dell'ampiezza delle reti.

Purtroppo questa sicurezza, come vedremo tra breve, verrà messa in crisi dai risultati relativi alla seconda variabile test: "iscrizione ad aNobii". Come si può osservare nella fig. 14, mentre la percentuale stimata nel campione di chi dichiara di essere iscritto ad aNobii è pari al 58,5% (con un intervallo di confidenza al 95% che va da 48,8% al 68,1%), nella popolazione essa è pari al 39,8%: il campione rimanda un'immagine fortemente distorta della distribuzione nella popolazione e, inoltre, l'intervallo di confidenza non riesce a comprendere il valore del parametro della popolazione.

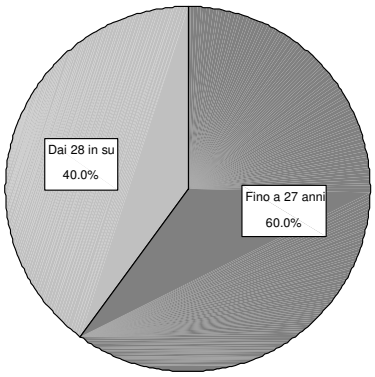
Questo risultato non sarebbe di per sé un problema se l'applicazione delle procedure e l'utilizzo dei criteri standard, proposti dagli ideatori del RDS, per la valutazione delle stime (numero ondate di convergenza, percentuale di equilibrio e fiducia statistica) avessero rivelato la presenza di problemi; se, usando un'immagine, si fosse accesa "una spia rossa di pericolo" in uno o più passaggi dell'analisi. Non solo questo non è avvenuto, ma sulla base dei criteri di analisi standard in campioni RDS, un ricercatore sarebbe stato abbastanza fiducioso circa la conclusione che, al momento della rilevazione, la percentuale

[40] In appendice 9 sono mostrati i grafici che confrontano le stime RDS pesate e le stime semplici non pesate. Una loro osservazione permette di comprendere più approfonditamente le considerazioni, svolte alla fine del par. 3.8, cap. 4, circa il tipo di stime da preferire in una ricerca.

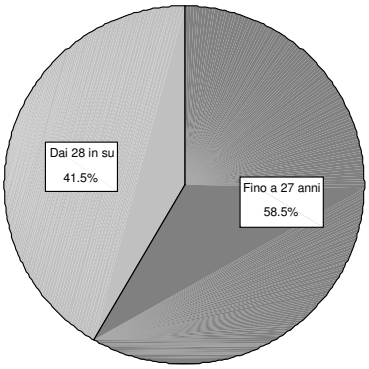
[41] Per valutare la grandezza di questo intervallo si può considerare l'intervallo al 95% che si calcolerebbe in un campione casuale semplice di pari numerosità: esso andrebbe da 51,3% a 65,7%.

di iscritti al sito aNobii era, se non la maggioranza nella popolazione, almeno attorno alla metà.

Fig. 13 - Grafici a torta delle frequenze percentuali dei rispondenti che dichiarano di avere fino a 27 o dai 28 anni in su

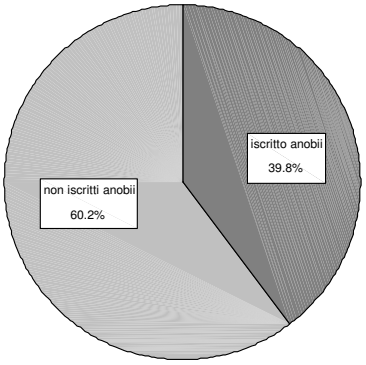


Valore reale nella popolazione

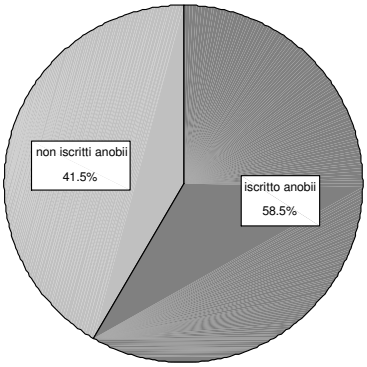


Stima RDS

Fig. 14 - Grafici a torta delle frequenze percentuali dei rispondenti che dichiarano di essere iscritti o meno ad aNobii



Valore reale nella popolazione



Stima RDS

Una possibile obiezione può essere opposta a questa affermazione. I risultati ottenuti nell'indagine-test potrebbero essere un'eccezione, un caso sfortunato, un outlier: se potessi ripetere più volte la stessa indagine utilizzando le stesse procedure sulla popolazione studiata, nella maggioranza dei casi le stime RDS si rivelerebbero accurate come nel caso delle simulazioni.

Nonostante non si possa negare, in generale, la ragionevolezza di questa obiezione, in questo caso particolare si può escludere che la performance

negativa della stima RDS per la variabile "iscrizione ad aNobii" sia imputabile solo a condizioni accidentali. In questa specifica circostanza la performance negativa è imputabile a due fattori:

- la violazione dell'assunto, critico per l'applicazione del disegno RDS, che prevede che le diverse scelte individuali, anche se non casuali, si combinino in modo che l'ipotesi di reclutamento casuale sia appropriata almeno a livello di aggregato (cfr. cap. 4, par. 5.2) e
- l'incapacità delle stime RDS, di tener conto, di compensare gli effetti di questa violazione.

Per valutare la violazione dell'assunto sono state confrontate le distribuzioni delle persone nelle matrici campionarie di reclutamento delle variabili test con quelle attese nell'ipotesi di reclutamento casuale. E' stato possibile fare questa operazione dal momento che sono note le distribuzioni delle relazioni fra i sottogruppi nella popolazione: così, ad esempio, considerando la variabile "classe di età", se il 32% delle relazioni presenti nella popolazione sono fra persone che appartengono al gruppo fino a 27 anni di età, ci si può aspettare che, in caso di reclutamento casuale, il numero di persone reclutate, nel campione, da persone fino a 27 anni di età sia il 32% di tutti i reclutati cioè 47 persone circa [(32 x 146)/100]; lo stesso ragionamento vale per gli altri gruppi (tabb. 6-7).

Tab. 6 - *Matrice di reclutamento per la variabile "classe di età" con i valori delle frequenze osservate e di quelle attese (fra parentesi)*

	Reclutati	
Reclutatori	*Classe di età*	
Classe di età	Fino a 27 anni	Dai 28 anni in su
Fino a 27 anni	50 (47)	22 (29)
Dai 28 anni in su	36 (29)	38 (40)

Tab. 7 *Matrice di reclutamento per la variabile "iscrizione ad aNobii" con i valori delle frequenze osservate (in corsivo) e di quelle attese (fra parentesi)*

	Reclutati	
Reclutatori	*Iscritto ad Anobii*	
Iscritto ad Anobii	Sì	No
Sì	55 (34)	40 (33)
No	34 (33)	17 (46)

Nel caso della variabile "classe di età" lo scostamento fra valori osservati ed attesi non è marcato (tab. 6), possiamo dunque accettare l'ipotesi che almeno a livello aggregato il reclutamento sia casuale. Invece nel caso della variabile "iscrizione ad aNobii" lo scostamento fra valori è elevato (tab. 7): coloro che sono iscritti ad aNobii sono sovrarappresentati nella matrice di reclutamento: non solo gli iscritti ad aNobii, ma anche i non iscritti, reclutano più iscritti ad aNobii di quanto ci si aspetterebbe nell'ipotesi di casualità[42].

La violazione dell'assunto può, a livello generale, essere attribuito a due fattori che possono agire sia singolarmente che in combinazione: comportamenti selettivi di reclutamento oppure diversa propensione a partecipare da parte di coloro che sono contatti.

Dal momento che, nell'indagine–test, la scelta delle persone dalla lista degli amici era casuale (la scelta delle persone a cui inviare la mail era rigorosamente controllata) si può escludere che la sovrarappresentazione sia frutto di una

[42] L'ipotesi è stata sottoposta anche al test χ^2 della bontà di adattamento fra distribuzione osservata e attesa (cfr. appendice 10). Per quanto riguarda la variabile "classe di età" il valore del χ^2 è uguale 3,67 ($p > 0,05$), mentre per la variabile "iscrizione ad aNobii" il valore del χ^2 è uguale 32,76 ($p < 0,01$).

distorsione nella selezione. Essa è, dunque, imputabile alla distorsione indotta dalla maggiore partecipazione, una volta invitati, di coloro che erano iscritti ad aNobii.

Le ragioni di questa maggiore partecipazione non sono difficili da immaginare, ricordando la vicende di creazione del gruppo "Last.fm aNobii". Coloro che sono contemporaneamente membri delle due reti sociali, Last.fm ed aNobii, sentono molto probabilmente un senso di appartenenza molto forte al gruppo: in fondo è il "loro" gruppo. Per gli "altri", che non sono iscritti ad aNobii, il gruppo "Last.fm aNobii" può risultare uno dei tanti ai quali si sono iscritti e forse neanche il più significativo.

Stabilito che la sovrarappresentazione degli iscritti ad aNobii è imputabile ad una violazione della condizione di reclutamento casuale, si può sottolineare che lo stimatore RDS non è in grado di compensare automaticamente questo tipo di distorsione: esso, infatti, corregge la stima percentuale semplice con un fattore di correzione che tiene in considerazione solo la distorsione indotta dal fatto che persone con un maggior numero di relazioni hanno una maggiore probabilità di essere estratte nel campione (cfr. cap. 4, par. 3.8 per la versione semplificata).

6. RDS: alcune considerazioni e possibili sviluppi

La ricerca presentata nella tesi non vuole, e non può, porsi come una risposta definitiva circa la valutazione dell'accuratezza delle stime RDS. Essa è da considerarsi, invece, come un contributo parziale alla riflessione, in seno ad un dibattito metodologico internazionale che è tutt'ora in corso.

Una delle principali indicazioni che emergono dal lavoro (e in particolare dai risultati dell'indagine-test descritti in precedenza) è l'opportunità di accostarsi con una certa dose di scetticismo (o quantomeno di prudenza) nei confronti della validità delle procedure RDS.

Nella nostra ricerca è vero che, date le condizioni richieste dal modello matematico alla base del RDS, i risultati sono in linea con le aspettative teoriche: le simulazioni, nelle quali sono state assunte le condizioni ideali, hanno mostrato, infatti, performance positive delle stime RDS.

Nonostante ciò, l'analisi dei dati provenienti dall'indagine-test ha rimandato un quadro molto meno ottimista rispetto a quello emerso nelle simulazioni. La violazione di un assunto centrale nell'applicazione del RDS ha indotto una distorsione in una variabile test, che lo stimatore RDS non è stato in grado di compensare automaticamente. A peggiorare questo risultato va aggiunto che l'applicazione delle procedure e l'utilizzo dei criteri standard, proposti dagli ideatori del RDS, per la valutazione delle stime (numero ondate di convergenza, percentuale di equilibrio e fiducia statistica) non hanno "rivelato" la presenza di problemi.

I risultati di questa ricerca sono soggetti ad una potenziale obiezione: la popolazione considerata in questo studio (i membri di una comunità virtuale) non può essere considerata una popolazione "tipo" delle popolazioni che usualmente si studiano con l'RDS (cfr. cap. 4, par. 2). Non si può negare la ragionevolezza di questa obiezione, almeno a livello generale. Proprio per questo, anche mettendo in luce le limitazioni riscontrabili in una singola e parziale applicazione, si rafforza la convinzione che i risultati di questa ricerca evidenziano la necessità, in primo luogo, di ulteriori e sistematiche valutazioni

delle procedure RDS (testandole ad esempio su altre popolazioni di cui si conoscono i valori reali di alcune variabili); in secondo luogo, di migliorare la comprensione del processo di reclutamento (ad es. con interviste follow-up quando i partecipanti tornano per ricevere gli incentivi oppure con tecniche non standard come l'osservazione partecipante); in terzo luogo, di individuare strumenti, test per identificare e misurare le distorsioni indotte dalle violazioni degli assunti e, in quarto e ultimo luogo, di sviluppare stimatori in grado di compensare le distorsioni, una volta individuate.

Il consiglio relativo ad un cauto utilizzo del RDS va dato per obiettività e per non coltivare illusioni apportatrici di discredito e di sfiducia in un metodo che può essere utile (a patto che si risolvano alcuni problemi) per comprendere le caratteristiche di alcune popolazioni delle quali altrimenti si rischierebbe di non poter dire nulla.

Si potrebbe pensare che sia fatica inutile studiare, perfezionare questa tecnica o tecniche affini. Oltre agli errori di misurazione indotti dalla deviazione dell'applicazione dalla teoria statistica rimangono anche quelli che affliggono le indagini sociali tradizionali, in particolare gli errori di osservazione legati allo strumento, all'intervistatore e all'intervistato.

Una risposta a questo dubbio legittimo non può essere definitiva e rimanda, in ultima istanza, alla visione del proprio lavoro posseduta da ogni singolo ricercatore. La posizione adottata in questa tesi è ben espressa dalle parole di Tagliacarne che, circa mezzo secolo fa, in risposta alle critiche di inutilità mosse al campionamento di tipo tradizionale, scrisse "Noi non seguiremo mai questo modo di pensare e riteniamo che il progresso dei nostri studi si debba conseguire per due vie convergenti: affinare la tecnica scientifica del campione e curare al massimo quella delle interviste ed ogni altra forma di indagine che possa essere utile" (Tagliacarne, 1964, p. 297).

Appendici

Appendice 1 - Concetti base di teoria della stima

Si immagini di aver svolto un'ipotetica indagine campionaria utilizzando un qualsiasi disegno campionario di tipo probabilistico e di voler stimare il valore di alcune variabili. Limiteremo, per semplicità, le considerazioni soltanto a stime di totale (ad esempio, il numero di persone che leggono almeno un libro durante l'anno). Altri tipi di stima molto comuni nelle pubblicazioni e nelle ricerche sociali, quali le stime di frequenze percentuali e di medie si possono ricavare da quella di un totale dividendo il valore del totale per la numerosità della popolazione studiata (N). Una volta effettuata l'indagine si dispone dei risultati relativi al campione selezionato di numerosità n. Per calcolare la stima del totale si usa la seguente formula:

$$\hat{\tau} = \sum_{i=1}^{n} w_i y_i \qquad (1)$$

La formula richiede tre passaggi: si determina un peso da attribuire a ciascuna unità inclusa nel campione che dipende dal disegno campionario e dal metodo di stima utilizzato (w_i). Il valore del peso viene calcolato a partire dalle probabilità di selezione dell' unità nel campione e serve per compensare l'effetto distorsivo sulle stime dovuto ad eventuali differenti probabilità di selezione fra le diverse unità; si moltiplica il valore relativo ad una data variabile oggetto di indagine (y_i), rilevata sulla generica unità (i) inclusa nel campione, per il peso attribuito alla medesima unità (w_i): $w_i\, y_i$ (nell'esempio il

numero di libri letti da ciascun individuo campionato moltiplicato per il corrispondente peso). Si effettua la somma, per ogni unità campionata, dei prodotti del punto precedente ($\sum_{i=1}^{n} w_i y_i$), la quale fornisce una stima del totale ($\hat{\tau}$) dei libri letti in un anno nella popolazione. Se vogliamo calcolare una stima del numero medio dei libri letti in un anno ci basta dividere il totale per la numerosità della popolazione.

Probabilità di selezione e di inclusione

Esistono due tipi di probabilità la cui determinazione sta alla base del calcolo del peso: la probabilità di selezione e quella di inclusione. La prima è la probabilità di estrarre un unità in ogni singola estrazione. Se da un'urna che contiene 4 palline di colore diverso: verde, bianca, rossa e blu (N=4) si estraggono 2 palline (n=2) la probabilità di selezione della prima pallina estratta sarà ¼ cioè 1/(N); la probabilità della seconda pallina sarà ancora ¼ se la prima pallina estratta viene reinserita nell'urna prima di estrarre la seconda pallina (estrazione con reimmissione) e 1/3 [1/(N-1)] se la prima pallina estratta *non* viene reinserita nell'urna prima di estrarre la seconda pallina (estrazione senza reimmissione o in blocco). La probabilità di inclusione π_i è la probabilità di includere nel campione l'i-esima unità; per una qualsiasi unità è data dalla somma delle probabilità di tutti i campioni che contengono quell'unità. Sempre considerando l'esempio campionario precedente si possono estrarre (senza reimmissione) 6 possibili campioni (c_i) di due unità: c_1(verde, bianca); c_2 (verde, rossa); c_3 (verde, blu); c_4 (bianca,rosso); c_5 (bianco,blu); c_6 (blu, rossa). Ognuno di questi campioni ha una probabilità di selezione uguale a $p(c_i)=1/6$. Qual è la probabilità di inclusione? Considerando la pallina verde la probabilità di inclusione è data dalla somma delle probabilità dei campioni che contengono la pallina verde che sono tre: $\pi_{verde}=p(c_1)+p(c_2)+p(c_3)=1/6+1/6+1/6=1/2$. La probabilità di inclusione è uguale anche per la pallina bianca, per quella rossa e per quella blu ($\pi_{rossa}=\pi_{bianca}=\pi_{blu}=1/2$). La somma delle probabilità di inclusione di tutte le palline coincide con la numerosità campionaria: nell'esempio è

uguale a 2. Nel caso di un campione casuale semplice la probabilità di selezione di un'unità è data da $p_i=1/N$ dove N è l'ampiezza della popolazione, mentre la probabilità di inclusione è data da $\pi_i = \frac{n}{N}$ dove n è l'ampiezza campionaria.

Stimatore con probabilità di selezione

Supponendo che il campione sia con rimpiazzo e che la probabilità di selezione di ogni singola unità della popolazione p_i per i=1,2,….,N. Allora uno stimatore non distorto del totale della popolazione τ è dato da (Thompson, 2002, pp. 51-53):

$$\hat{\tau}_{HH} = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \frac{y_i}{p_i} \tag{2}$$

Questo stimatore fu introdotto da Hansen e Hurwitz nel 1943. Il fatto che il campionamento sia con rimpiazzo può portare a selezionare la stessa unità più di una volta nel campione. Nel qual caso il valore dell'unità compare nello stimatore tante volte quante essa è selezionata.

Uno stimatore non distorto della varianza campionaria di $\hat{\tau}_{HH}$ è dato da:

$$\hat{\mathrm{var}}(\hat{\tau}_{HH}) = \frac{1}{\mathrm{n(n-1)}} \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \hat{\tau} \right)^2 \tag{3}$$

Uno stimatore non distorto della media della popolazione μ è $\hat{\mu} = (1/N)\hat{\tau}_{HH}$ e della varianza campionaria $\hat{\mathrm{var}}(\hat{\mu}) = (1/N^2)\hat{\mathrm{var}}(\hat{\tau}_{HH})$

Stimatore con probabilità di inclusione

Con qualsiasi disegno campionario data la probabilità di inclusione π_i per una qualsiasi unità inclusa nel campione, uno stimatore non distorto del totale della popolazione τ è dato da:

$$\hat{\tau}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} \tag{4}$$

e uno stimatore non distorto della sua varianza campionaria è dato da:

$$\hat{\text{var}}(\hat{\tau}_{HT}) = \sum_{i=1}^{n}\sum_{l>i}\left(\frac{y_i}{\pi_i} - \frac{y_l}{\pi_l}\right)^2 \frac{\pi_i\pi_l - \pi_{il}}{\pi_{il}} \qquad (5)$$

dove π_{il} è la probabilità di inclusione di secondo ordine che le unità i.ma e l.ma siano entrambe comprese nel campione. Questo stimatore fu introdotto da Horvitz e Thompson nel 1952.

Uno stimatore della varianza campionaria che non richiede il calcolo delle probabilità di inclusione di secondo ordine è stato proposto da Brewer e Hanif nel 1983 ed è dato da:

$$\tilde{\text{var}}(\hat{\tau}_{HT}) = \left(\frac{N-n}{Nn(n-1)}\right)\sum_{i=1}^{n}\left(\frac{ny_i}{\pi_i} - \hat{\tau}_{HT}\right)^2 \qquad (6)$$

Questo stimatore pur essendo distorto (sovrastima tendenzialmente il valore della varianza campionaria) è preferibile per la sua semplicità di calcolo.

Uno stimatore non distorto della media della popolazione μ è $\hat{\mu} = (1/N)\hat{\tau}_{HT}$ e della varianza campionaria $\hat{\text{var}}(\hat{\mu}) = (1/N^2)\hat{\text{var}}(\hat{\tau}_{HT})$

L'intervallo di confidenza

Gli intervalli di confidenza al 95% per gli stimatori $\hat{\tau}_{HH}$ e $\hat{\tau}_{HT}$ sono dati rispettivamente da:

$$\hat{\tau}_{HH} \pm z\sqrt{\hat{\text{var}}(\hat{\tau}_{HH})} \qquad (7)$$

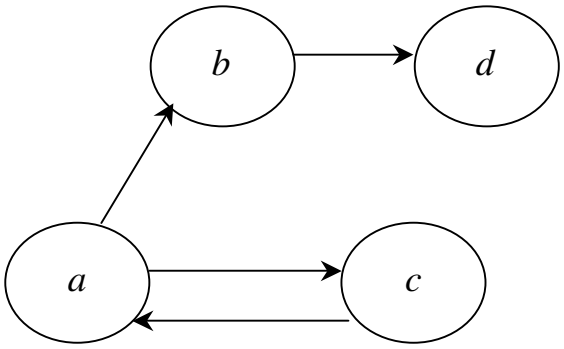
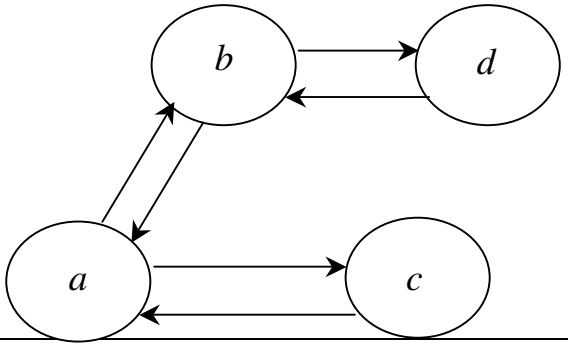
e da $$\hat{\tau}_{HT} \pm z\sqrt{\hat{\text{var}}(\hat{\tau}_{HT})} \qquad (8)$$

dove z è il coefficiente che dipende dal livello di fiducia nella stima che si desidera avere. In un campione con n >30 è uguale a 1,96 per un livello di confidenza del 95% (il più comunemente utilizzato). Il livello di confidenza descrive l'incertezza associata col metodo di campionamento. Supponiamo di usare lo stesso metodo di campionamento per selezionare 100.000 differenti campioni a partire dalla stessa popolazione e di calcolare 100.000 intervalli di

confidenza al 95%: uno per ogni campione. Alcuni intervalli includeranno il parametro della popolazione (il valore reale nella popolazione) ed alcuni non lo comprenderanno. Un intervallo di confidenza al 95% significa che 950.000 dei 100.0000 intervalli calcolati dovrebbero includere il parametro. Di conseguenza, l'intervallo del 95% calcolato sui dati del campione realmente selezionato in una ricerca sociale (che è solo uno dei possibili campioni che possono essere generati dal disegno campionario adottato) fa ben sperare che sia uno dei 950.000 che comprendono il parametro della popolazione, anche se a rigore potrebbe essere uno dei 50.000 che non lo contengono.

Appendice 2- Il Grafo

Un *grafo G=(V,E)* è una coppia di insiemi disgiunti dove *V* è un insieme non vuoto, discreto e finito, i cui elementi sono i *nodi* del grafo: $V = \{v_1, v_1, ..., v_N\}$. L'insieme E è un sottoinsieme del prodotto cartesiano VxV ($\{E \subseteq V * V\}$) costituito da un insieme di coppie ordinate di nodi del grafo: $E = \{(u,v) : u,v \in V\}$.Una coppia $(u,v) \in E$ è definita *arco* (figura 1). Il *grafo* G è *orientato* se gli archi sono considerati coppie ordinate di nodi: $(u,v) \neq (v,u)$. Se invece tutti gli archi del grafo sono considerati come coppie non ordinate, cioè $(u,v) = (v,u)$ il *grafo* G sarà *non orientato*. Se (u,v) è un arco del grafo orientato G, diremo che l'arco è uscente da *u* ed è entrante in *v*. Nel caso dei grafi orientati possiamo distinguere il *grado entrante di un nodo u* dal suo *grado uscente* (rispettivamente *de(u)* e *du(u)*): il primo è dato dal numero di archi entranti in *u*, mentre il secondo è dato dal numero di archi uscenti. Nel caso di un grafo non orientato, *de(u)* è uguale a *du(u*). I grafi possono essere disegnati rappresentando i nodi come punti e gli archi come frecce (orientate o non orientate) che uniscono coppie di nodi. A rigore non si dovrebbe mai confondere un grafo con la sua rappresentazione attraverso un diagramma a frecce, dal momento che la collocazione in un piano dei nodi è del tutto arbitraria come lo è anche il disegno dei segmenti che rappresentano gli archi. In altre parole uno stesso grafo può essere disegnato in modi molto diversi.

Grafo orientato (o diretto)=(V,E) V = nodi o nodi; E = archi $V = \{a, b, c, d\}$ $E = \{(a, b), (a, c), (c, a), (b, d)\}$	Grafo non orientato =(V,E) V = nodi o nodi; E = archi $V = \{a, b, c, d\}$ $E = \{(a, b), (b, a), (a, c), (c, a), (b, d), (d, b)\}$
	

Appendice 3 - Stimatore *multiplicity sampling*[♣]

Nel caso del *multiplicity sampling (*e dell'*adaptative cluster sampling)*, la probabilità di inclusione di un'unità della popolazione oggetto di studio è data dalla probabilità che il grappolo sia intersecato dal campione iniziale. Per un intervistato inserito in un grappolo di una persona, la probabilità di intersezione è semplicemente la probabilità di essere estratto nel campione iniziale. Per grappoli più ampi la probabilità di intersezione è data dalla probabilità che uno o più delle unità del grappolo siano incluse nel campione iniziale .

La popolazione può essere divisa in K grappoli, etichettati da 1,…,K; y_k^* è la somma dei valori della variabile y per tutti i membri del grappolo *k*-esimo e n_k è l'ampiezza del grappolo k-esimo. La probabilità di inclusione di ogni unità che appartiene al k-esimo grappolo è calcolabile come uno meno la probabilità che il campione iniziale *N* non intersechi mai il grappolo *k* ed è data dalla formula:

$$\pi_k = 1 - \binom{N - n_k}{n} \Big/ \binom{N}{n} \qquad (1)$$

dove *n* indica il numero di unità estratte nel campione iniziale (semi).
Questa probabilità nel caso di un campione casuale semplice con reinserimento si semplifica:

$$\pi_k = 1 - (1 - n_k / N)^n \qquad (2)$$

Lo stimatore di Horvitz-Thompson del totale della popolazione è dato pertanto da:

$$\hat{\tau} = \sum_{k=1}^{K} \frac{y_k^*}{\pi_k} \qquad (3)$$

Uno stimatore non distorto della varianza campionaria di $\hat{\tau}$ è dato da:

[♣] Nell'appendice verrà seguita la trattazione di questo stimatore presentata da Thompson (2002, pp. 176-179 e 295-297).

$$\hat{\text{var}}(\hat{\tau}) = \sum_{k=1}^{K}\left(\frac{1}{\pi_k^2} - \frac{1}{\pi_k}\right) y_k^{*2} + \sum_{k=1}^{K}\sum_{l\neq k}\left(\frac{1}{\pi_k \pi_l} - \frac{1}{\pi_{kl}}\right) y_k^{*} y_l^{*} \qquad (4)$$

$$\text{dove } \pi_{kl} = 1 - \left[\binom{N-n_k}{n} + \binom{N-n_l}{n} - \binom{N-n_k-n_l}{n} \Big/ \binom{N}{n}\right] \qquad (5)$$

Uno stimatore non distorto della media della popolazione μ sarà dato da $\hat{\mu} = (1/N)\hat{\tau}$ e della varianza campionaria $\hat{\text{var}}(\hat{\mu}) = (1/N^2)\hat{\text{var}}(\hat{\tau})$.

Appendice 4 - Stimatore *respondent-driven sampling*

Per comprendere lo stimatore RDS proposto da Volz - Heckathorn (2008) inizieremo il ragionamento riconsiderando (cfr. appendice 2) lo stimatore con probabilità di selezione introdotto da Hansen-Hurwitz (1943). Data la probabilità di selezione di ogni singola unità della popolazione p_i per i=1,2,….,N, uno stimatore non distorto del totale della popolazione τ è dato da:

$$\hat{\tau}_{HH} = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \frac{y_i}{p_i} . \qquad (1)$$

Gli ideatori del RDS sostengono che il processo di reclutamento RDS può essere formalizzato come una catena markoviana e che le probabilità di selezione delle singole unità, p_i , convergono in un numero finito di iterazioni a valori proporzionali al grado di uscita di ogni unità. Le probabilità p_i da inserire nella formula possono essere stimate a partire dai dati campionari come

$$\hat{p}_i = \frac{\delta_i}{N \hat{\delta}_U} \qquad (2)$$

dove δ_i è il grado in uscita rilevato sull'unità campionata e $\hat{\delta}_U$ è una stima del valore medio del grado in uscita di tutte le unità campionate, calcolata a partire dal rapporto di due stimatori Hansen-Hurwitz (Volz - Heckathorn, 2008, p. 85 e Salganik - Heckathorn 2004, pp. 216-218). La stima di $\hat{\delta}_U$ è data da:

$$\hat{\delta}_U = \frac{\sum_{i=1}^{n} \delta_i / np_i}{\sum_{i=1}^{n} 1/np_i} = \frac{n}{\sum_{i=1}^{n} \delta_i^{-1}} \qquad (3)$$

Dal momento che il numeratore e il denominatore sono entrambi stimatori Hansen-Hurwitz non distorti, si può dimostrare che anche il loro rapporto è asintoticamente non distorto con valori della distorsione che diminuiscono all'aumentare dell'ampiezza campionaria.

A questo punto il normale stimatore del totale della popolazione HH si può scrivere come:

$$\hat{\tau}_{HH} = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \frac{y_i}{p_i} = \frac{1}{n} \sum_{i=1}^{n} \frac{N \hat{\delta}_U \; y_i}{\delta_i} = \frac{N \hat{\delta}_U}{n} \sum_{i=1}^{n} \delta_i^{-1} y_i \tag{4}$$

Dal momento che tipicamente l'RDS si applica in situazioni in cui non conosce l'ampiezza della popolazione totale (*N* è sconosciuto), possiamo eliminare dalla (4) *N* dividendo entrambi i membri dell'equazione per *N*: si ottiene così lo stimatore della media ed eliminando la necessità di conoscere *N*:

$$\hat{\tau}_{HH} \Big/ N = \hat{\mu} = \frac{\hat{\delta}_U}{n} \sum_{i=1}^{n} \delta_i^{-1} y_i \tag{5}$$

Sostituendo in questa $\hat{\delta}_U$ con la (3) otteniamo lo stimatore RDS di una media o una percentuale proposto da Volz - Heckathorn (2008):

$$\hat{\mu}_{RDS} = \frac{\sum_{i=1}^{n} \delta_i^{-1} y_i}{\sum_{i=1}^{n} \delta_i^{-1}} \tag{6}$$

Questa equazione può essere riscritta (Volz - Heckathorn, 2008, p. 86) nel caso di stima della percentuale di una variabile di tipo categoriale come:

$$\hat{P}_A = 100 \left(\frac{n_A}{n} \right) \left(\frac{\hat{\delta}_U}{\hat{\delta}_A} \right) \tag{7}$$

dove 100 (n_A/n) è la percentuale di persone del gruppo A nel campione una volta raggiunta la situazione di equilibrio (es. la percentuale di femmine nel campione) e $\hat{\delta}_A$ è il valore della media armonica del numero di legami dichiarati da ogni persona (δ_i) nel gruppo A, cioè: $\hat{\delta}_A = \frac{n_A}{\sum_{i=1}^{n_A} \frac{1}{\delta_i}}$

Appendice 5 - Algoritmo della procedura di simulazione del *respondent-driven sampling*

GRAFO

$$V = \{v_1, v_1, ..., v_N\}$$

$$N = |V|$$

$$E \subseteq V * V$$

$$M = |E|$$

$$\forall u \in V, (u,u) \notin E \;\; \forall (u,v) \in E, (u,v) \in E \rightarrow (v,u) \in E$$

ONDATE

$$W_1, W_2, W_3, ..., W_t \subseteq V, t \geq 1$$

$$W_i \neq \{\}, 1 \leq i \leq t$$

$$W_i \cap W_j = \{\}, 1 \leq i, j \leq t \wedge i \neq j$$

$$W_1 \cup W_2 \cup W_3 \cup ... \cup W_t = X, t \geq 1$$

$$X \subseteq V$$

$$|W_1| = C, 1 < C \leq N$$

$$\forall u \in W_{t-1}, \exists v \in W_t : (u,v) \in E, t = 2, ..., t-1$$

Input1: V e C **Output1**: Calcolo prima ondata (W_1)

1 $S \leftarrow \{\}$

2 Per i =1,2,…,C ripeti

3 Seleziona $u \in V \setminus S$

4 $S \leftarrow S \cup \{u\}$

5 Fine–ciclo

6 Restituisci S

Input2: V,E,X,k **Output2**: Calcolo nuova ondata (W_i)

1 $S \leftarrow \{\ \}$

2 Per ogni $u \in W_{i-1}$ ripeti

3 Per *i* =1,2,…,k ripeti

4 $T \leftarrow (\{v \in V \mid (u,v) \in E\} \setminus (X \cup S)$

5 Se $\{\ \} \neq T$ allora

6 Seleziona $v \in T$

7 $S \leftarrow S \cup \{v\}$

8 Fine-condizione

9 Fine–ciclo

A Fine–ciclo

B Restituisci S

Input2: V,E,C,t **Output2**: Esporta campione SNOWBALL

1 $W_1 \leftarrow$ Output1

2 $X \leftarrow W_1$

3 Per *i* =2,…,t ripeti

4 $W_i \leftarrow$ Output2

5 $X \leftarrow X \cup W_i$

6 Fine–ciclo

7 Restituisci *X*

Appendice 6 – Documenti dell'indagine-test

Fac-simile mail di invito Indagine "Music Lab - Anobii Italia" per i semi

Ciao [nome utente Last.fm],
Il mio nome è Alberto Vitalini e lavoro presso il Dipartimento di Sociologia dell'Università Cattolica di Milano (in fondo alla mail troverai tutti gli indirizzi nel caso tu voglia controllare/ avere delle referenze). Al momento sto conducendo una ricerca per testare una nuova procedura per lo studio di popolazioni presenti in Internet e in particolare il gruppo Anobii-Italia.
Con questa mail voglio invitarti a partecipare alla mia ricerca (nei prossimi giorni anche altri membri del gruppo Anobii-Italia saranno contattati). Ti chiedo solo di compilare un breve questionario on line sui tuoi gusti musicali. Non ti verrà fatta alcuna domanda personale né ti verranno sottratte con inganno informazioni che non vuoi cedere. Il questionario è anonimo e confidenziale e svolto per soli fini di ricerca. Per compilarlo ti basteranno pochi minuti.

Per partecipare all'indagine e per rispondere alle domande clicca sul seguente link:

Indagine "Music Lab - Anobii Italia" http://www.kwiksurveys.com/online-survey.php?surveyID=KJNHMO_60e885db&UID=2349958492

Ti ringrazio per la collaborazione e l'aiuto che vorrai darmi

Alberto Vitalini
P.S.
Se hai qualsiasi domanda invia una mail a: alberto.vitalini@unicatt.it.
Se hai qualche dubbio sulla serietà della mail puoi consultare, nel sito dell'Università Cattolica di Milano (http://milano.unicatt.it/), la pagina del prof. Marco Caselli, nella sezione Ricerca.

Fac-simile mail di invito Indagine "Music Lab - Anobii Italia" (per reclutandi)

Ciao [*nome utente Last.fm*],
Il mio nome è Alberto Vitalini e lavoro presso il Dipartimento di Sociologia dell'Università Cattolica di Milano (in fondo alla mail troverai tutti gli indirizzi nel caso tu voglia controllare/ avere delle referenze). Al momento sto conducendo una ricerca per testare una nuova procedura per lo studio di popolazioni presenti in Internet e in particolare il gruppo Anobii-Italia.

Con questa mail voglio invitarti a partecipare alla ricerca, che consiste nel compilare un breve questionario on line sui tuoi gusti musicali. Sei stata scelta perché fai parte della lista di amici di [*nome utente Last.fm che ha già partecipato all'indagine*], che ha già preso parte alla ricerca.
Non ti verrà fatta alcuna domanda personale né ti verranno sottratte con inganno informazioni che non vuoi cedere. Il questionario è anonimo e confidenziale e svolto per soli fini di ricerca. Per compilarlo ti basteranno pochi minuti.

Per partecipare all'indagine e per rispondere alle domande clicca sul seguente link:

Indagine "Music Lab - Anobii Italia" http://www.kwiksurveys.com/online-survey.php?surveyID=KJNHMO_60e885db&UID=2349958492

Ti ringrazio per la collaborazione e l'aiuto che vorrai darmi

Alberto Vitalini
P.S.
Se hai qualsiasi domanda invia una mail a: alberto.vitalini@unicatt.it.
Se hai qualche dubbio sulla serietà della mail puoi consultare, nel sito dell'Università Cattolica di Milano (http://milano.unicatt.it/), la pagina del prof. Marco Caselli, nella sezione Ricerca.

Presentazione indagine e questionario on line

Indagine "Music Lab - Anobii Italia"

Grazie per aver deciso di partecipare a questa indagine, condotta dal dott. Alberto Vitalini sotto la supervisione del prof. Clemente Lanzetti e del prof. Marco Caselli (marco.caselli@unicatt.it) presso il Dipartimento di Sociologia dell'Università Cattolica di Milano.

L'obiettivo di questa ricerca è capire come si possano raccogliere informazioni accurate circa i membri di una popolazione elettronicamente connessa utilizzando il "passaparola". Per questo ti chiedo di rispondere ad alcune brevi domande sui tuoi gusti musicali e poi di indicarmi se

posso scrivere, nelle mail di invito che inoltrerò ad alcuni tuoi amici, che hai accettato di partecipare alla ricerca.

Il questionario è anonimo e confidenziale e svolto per soli fini di ricerca. Le informazioni che ti verranno richieste sono alcune di quelle che hai già reso accessibili nel tuo profilo: non dovrai svelare nulla di più di quanto hai già deciso di fare.

Pochi minuti del tuo tempo contribuiranno al buon esito della ricerca!

La partecipazione a questo studio è completamente volontaria e puoi ritirarti in qualsiasi momento. Le risposte saranno usate solo per fini di ricerca scientifica e in modo confidenziale. Il tuo nome utente su Last.fm sarà rimpiazzato da un numero e non sarà mai mostrato in nessuna presentazione o pubblicazione legata a questa ricerca. I dati raccolti per questo studio saranno immagazzinati su un server per un anno e poi distrutti. Nel caso fossi interessato/a ti verranno inviate gratuitamente, su richiesta, le copie delle pubblicazioni e degli articoli relativi alla ricerca.
Se desideri ulteriori informazioni riguardo allo studio, scrivimi: alberto.vitalini@unicatt.it. Puoi contattare anche il mio supervisore: Marco Caselli (marco.caselli@unicatt.it).

Avanti per rispondere al questionario

Sesso

- ☐ Maschio
- ☐ Femmina

Età in anni compiuti

Qual è il tuo nominativo Last.fm?

Sei iscritto anche nel gruppo gemello di aNobii.com:" Last.Fm Readers"?

- ☐ SÌ
- ☐ No

Qual è l'artista che, negli ultimi tre mesi, hai ascoltato maggiormente?

Qual è l'album che, negli ultimi tre mesi, hai ascoltato maggiormente?

Qual è il brano che, negli ultimi tre mesi, hai ascoltato maggiormente?

Quanti amici Last.fm hai?

Se hai amici Last.fm, quanti di questi sono iscritti, come te, al gruppo Anobii - Italia?

Nei prossimi giorni inviterò a partecipare a questa ricerca alcuni membri del gruppo Anobii-Italia, che appartengono alla tua lista di amici. Posso dire loro che hai accettato di partecipare alla ricerca?

- SÌ
- No

Indietro	**Fine dell'indagine**

Grazie mille per aver preso parte a questa ricerca.

Appendice 7 – Risultati delle simulazioni di controllo

Tab.1 - Alcune informazioni statistiche relative alla numerosità dei campioni estratti per la variabile classe di età, a seconda dell'ondata di appartenenza. Simulazioni a partire da otto semi estratti casualmente dalla lista della popolazione

	ondata1	ondata2	ondata3	ondata4
Valore minimo	8	26	52	102
1° quartile	10	29	63	123
Mediana	11	30	68	132
3° quartile	12	32	71	139
Valore massimo	13	36	78	158
Intervallo valore max- min	5	10	26	56
Media	11	30	67	130
Scarto tipo	1	2	6	12

Nota: L'ondata1 comprende i casi della prima ondata più i semi, l'ondata2 comprende i casi delle prime due ondate più i semi; l'ondata3 comprende i casi delle prime tre ondate più i semi, l'ondata4 comprende tutti i casi selezionati nelle quattro ondate più i semi

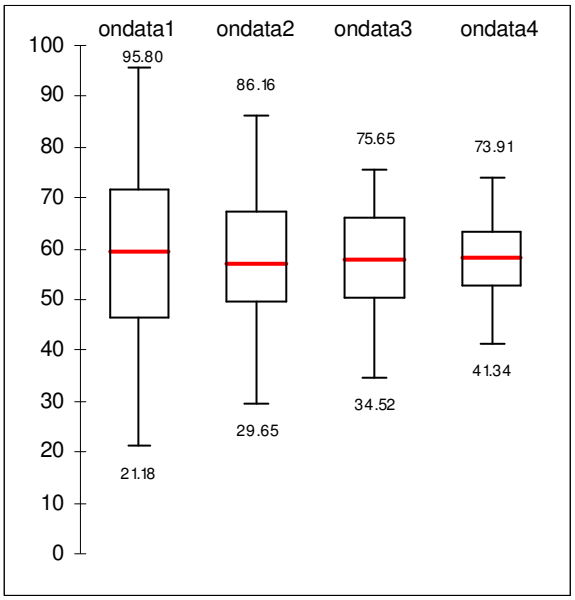
Tab.2 - Alcune informazioni statistiche relative alla numerosità dei campioni estratti per la variabile iscrizione ad aNobii, a seconda dell'ondata di appartenenza. Simulazioni a partire da otto semi estratti casualmente dalla lista della popolazione

	Ondata1	Ondata2	Ondata3	Ondata4
Valore minimo	14	33	67	123
1° quartile	14	37	77	146
Mediana	14	38	81	157
3° quartile	14	39	84	162
Valore massimo	14	40	92	183
Intervallo valore max- min	0	7	25	60
Media	14	37	80	154
Scarto tipo	0	2	6	13

Nota: L'ondata1 comprende i casi della prima ondata più i semi, l'ondata2 comprende i casi delle prime due ondate più i semi; l'ondata3 comprende i casi delle prime tre ondate più i semi, l'ondata4 comprende tutti i casi selezionati nelle quattro ondate più i semi

Fig. 1 - Distribuzione per ondata delle stime RDS della frequenza percentuale di persone che dichiarano di avere fino a 27 anni. Simulazioni a partire da otto semi estratti casualmente dalla lista della popolazione. La linea trasversale indica la percentuale, nella popolazione, di chi dichiara di fino a 27 anni (60%)*

Stime RDS (pesate)



Stime semplici (non ponderate)

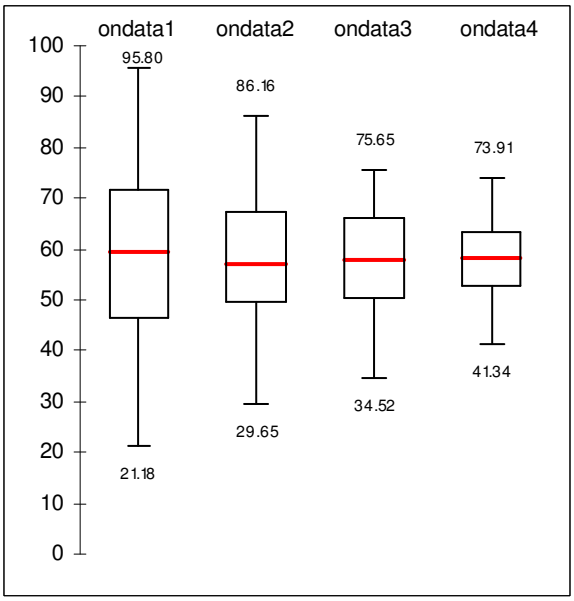
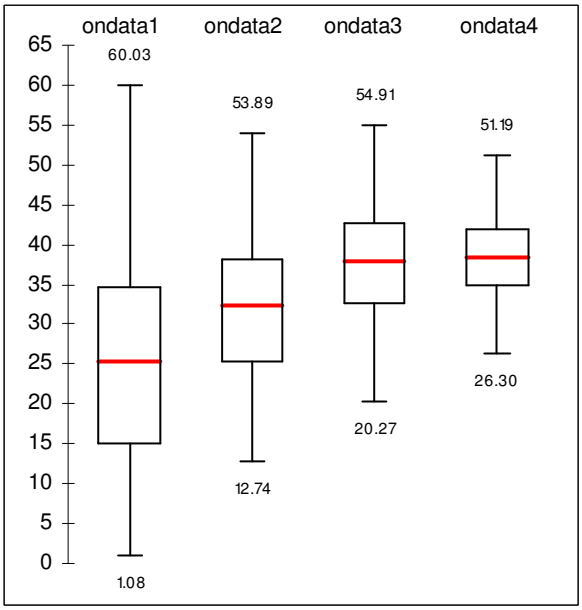
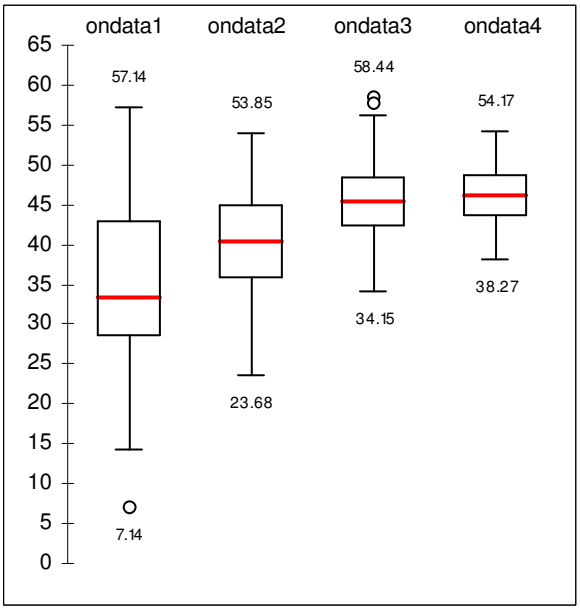


Fig.2 - Distribuzione per ondata delle stime RDS della frequenza percentuale di persone che dichiarano di essere iscritte ad aNobii. Simulazioni a partire da otto semi estratti casualmente a partire dalla lista della popolazione. La linea trasversale indica la percentuale, nella popolazione, di chi dichiara di essere iscritto ad aNobii nella popolazione (39,8%).*

Stime RDS (pesate)



Stime semplici (non pesate)



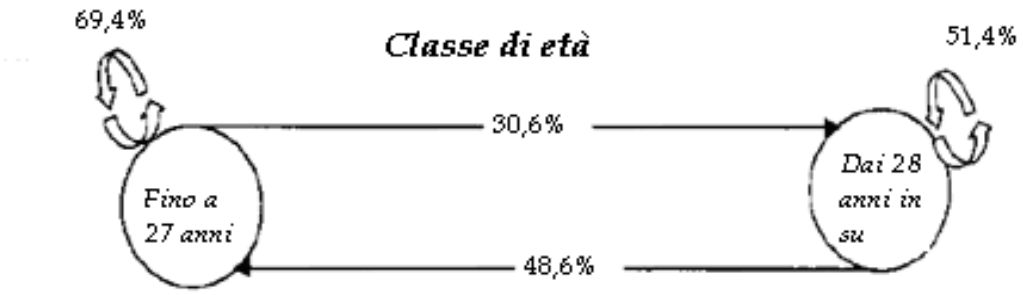
**Nota: L'ondata1 comprende i casi della prima ondata più i semi, l'ondata2 comprende i casi delle prime due ondate più i semi; l'ondata3 comprende i casi delle prime tre ondate più i semi, l'ondata4 comprende tutti i casi selezionati nelle quattro ondate più i semi*

Appendice 8 - Matrici di reclutamento utilizzate per il calcolo del numero di ondate di equilibrio

Tab.1 - Matrice di reclutamento della variabile classe di età (valori assoluti e percentuali)

	Reclutati			
Reclutatori	Classe di età			
Classe di età	Fino a 27 anni	Dai 28 anni in su	Totale	v.a.
Fino a 27 anni	69,4	30,6	100	(72)
Dai 28 anni in su	51,4	48,6	100	(74)

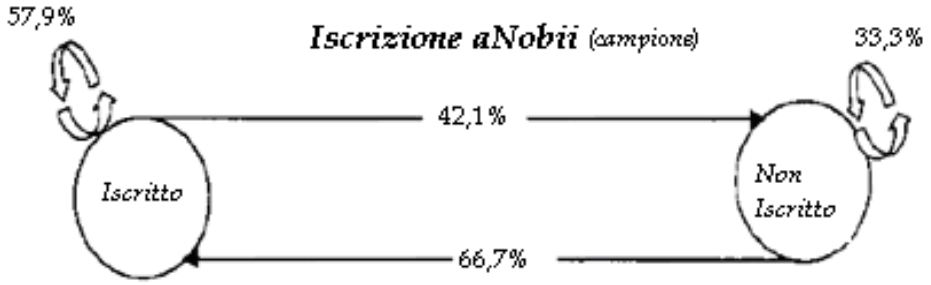
Fig. 1 - Rappresentazione grafica della matrice di reclutamento campionaria della tab. 1



Tab.2 - Matrice di reclutamento della variabile iscrizione ad aNobii (valori assoluti e percentuali)

	Reclutati			
Reclutatori	*Iscritto ad Anobii*			
Iscritto ad Anobii	Sì	No	Totale	v.a.
Sì	57,9	42,1	100	(95)
No	66,7	33,3	100	(51)

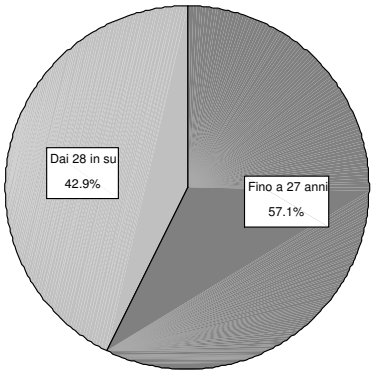
Fig. 2 - Rappresentazione grafica della matrice di reclutamento campionaria della tab. 2



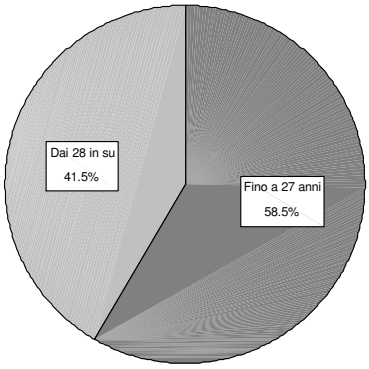
Appendice 9 – Confronto fra le stime *respondent-driven sampling* e le stime semplici delle variabili test

Nelle tabelle 1 e 2 sono mostrate le stime RDS pesate e le stime semplici non pesate. Si può notare che tra i due tipi di stime non c'è molta differenza. La ponderazione non ha cambiato in modo sostanziale i risultati. Può sorgere spontanea una domanda : «Quando le stime RDS, che sono pesate, sono da preferire alle stime semplici che non sono pesate? ».

Fig. 1 - Grafici a torta delle frequenze percentuali dei rispondenti che dichiarano di avere fino a 27 o dai 28 anni in su

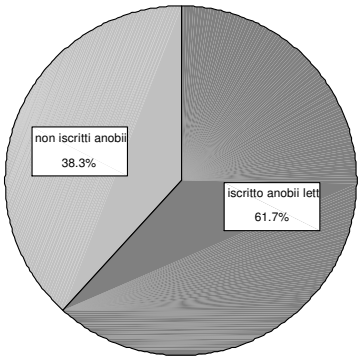


Stima campionaria semplice (non pesata)

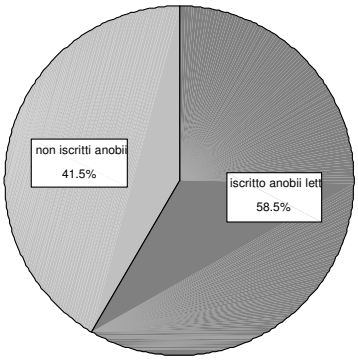


Stima RDS (pesata)

Fig. 2 - Grafici a torta delle frequenze percentuali dei rispondenti che dichiarano di essere iscritti o meno ad aNobii



Stima campionaria semplice (non pesata)



Stima RDS (pesata)

Una risposta a questa domanda richiede di considerare nuovamente lo stimatore semplificato 7 in appendice 4. Nel caso di stima di una variabile dicotomica la percentuale di una data modalità, P_A , risulta:

$$\hat{P}_A = 100\left(\frac{n_A}{n}\right)\left(\frac{D_{totale}}{D_A}\right)$$

dove 100 (n_A/n) è la percentuale di persone del gruppo A nel campione una volta raggiunta la situazione di equilibrio (es. la percentuale di femmine nel campione); D_{totale} è il valore della media armonica del numero di legami dichiarati da ogni persona (d_i) nel campione:

$$D_{totale} = \frac{n}{\sum_{i=1}^{n}\frac{1}{d_i}}$$

e D_A è il valore della media armonica del numero di legami dichiarati da ogni persona (d_i) nel gruppo A:

$$D_A = \frac{n_A}{\sum_{i=1}^{n_A}\frac{1}{d_i}}$$

Dal momento che la stima RDS è pesata per il fattore

$$\left(\frac{D_{totale}}{D_A}\right)$$

le stime RDS sono tanto preferibili alle stime semplici quanto maggiore è la differenza fra il numero medio di relazioni nei due gruppi e tendono a coincidere quando i due gruppi hanno, in media, lo stesso numero di relazioni. In quest'ultimo caso il valore del fattore di ponderazione sarà uguale nei due gruppi ad 1 e non modificherà minimamente le stime percentuali semplici.

Tornando alle stime delle variabili-test della ricerca mostrata in tesi, dovrebbe essere chiaro, a questo punto, il motivo di una differenza così limitata fra le stime RDS e le stime semplici: i sottogruppi "fino a 27 anni" e "dai 28 anni in su" e "iscritti ad aNobii" e "non iscritti ad aNobii" hanno un numero medio di legami simili.

Appendice 10 - Calcolo del test di adattamento per la valutazione dell'assunto di reclutamento casuale

Tab.1 - Matrice di reclutamento RDS (valori assoluti osservati)

	Reclutati	
Reclutatori	Classe di età	
Classe di età	Fino a 27 anni	Dai 28 anni in su
Fino a 27 anni	*50*	*22*
Dai 28 anni in su	*36*	*38*

Tab.2 - Distribuzione delle relazioni fra gruppi nella popolazione (valori percentuali sul totale)

	Classe di età	
Classe di età	Fino a 27 anni	Dai 28 anni in su
Fino a 27 anni	*32.2*	*20.0*
Dai 28 anni in su	*20.0*	*27.8*

Nota: Ad es: il valore 32.3% indica che il 32.3% delle relazioni presenti nella popolazione sono fra persone fino a 27 anni di età, il 27.8 fra persone dai 28 anni in su e il 20.0% fra persone appartenenti a classi di età diverse.

Tab.3 - Matrice di reclutamento RDS (valori assoluti attesi)

	Reclutati	
Reclutatori	Classe di età	
Classe di età	Fino a 27 anni	Dai 28 anni in su
Fino a 27 anni	*47*	*29*
Dai 28 anni in su	*29*	*40*

*Tab.4 - Matrice residui standardizzati**

	Reclutati	
Reclutatori	Classe di età	
Classe di età	Fino a 27 anni	Dai 28 anni in su
Fino a 27 anni	*0.19*	*1.69*
Dai 28 anni in su	*1.69*	*0.1*

$\chi^2 = 0.19 + 1.69 + 1.69 + 0.1 = 3.67$ ***con 3 gradi di libertà p-value > 0.05***

*Nota: *I valori delle celle sono dati dalla formula (frequenze osservate-frequenze attese)^2/frequenze attese. Ad es. nella prima cella in alto a sinistra il valore 0.19=(50-47)^2/47.*

Tab.5 - Matrice di reclutamento RDS (valori assoluti osservati)

	Reclutati	
Reclutatori	Iscritto ad Anobii	
Iscritto ad Anobii	Sì	No
Sì	*55*	*40*
No	*34*	*17*

Tab.6 Distribuzione delle relazioni fra gruppi di relazioni nella popolazione (valori percentuali sul totale)

	Reclutati	
Reclutatori	Iscritto ad Anobii	
Iscritto ad Anobii	Sì	Sì
Sì	*23.0*	*22.8*
No	*22.8*	*31.4*

Nota: Ad es: il valore 23.0% indica che il 23.0% delle relazioni presenti nella popolazione sono fra persone iscritte ad aNobii, il 31.4% fra persone non iscritte ad aNobii e il restante 22.8% fra persone appartenenti a classi di età diverse.

Tab.7 Matrice di reclutamento RDS (valori assoluti attesi)

	Reclutati	
Reclutatori	Iscritto ad Anobii	
Iscritto ad Anobii	Sì	No
Sì	*34*	*33*
No	*33*	*46*

*Tab.8 Matrice residui standardizzati**

	Reclutati	
Reclutatori	Iscritto ad Anobii	
Iscritto ad Anobii	Sì	No
Sì	*12.97*	*1.48*
No	*0.03*	*18.28*

$\chi^2 = 12.97 + 1.48 + 0.03 + 18.28 = 32.76$ ***con 3 gradi di libertà p-value < 0.01***

*Nota: *I valori delle celle sono dati dalla formula (frequenze osservate-frequenze attese)^2/frequenze attese. Ad es. nella prima cella in alto a sinistra il valore 12.97=(55-34)^2/34*

BIBLIOGRAFIA*

ABDUL-QUADER A.S. – HECKATHORN D.D. – MCKNIGHT C. – BRAMSON H. – NEMETH C. – SABIN K. – GALLAGHER K. – DES JARLAIS D.C.
2006 *Effectiveness of respondent-driven sampling for recruiting drug users in New York City: Findings from a pilot study*, in "Journal Urban Health", 83, pp. 459–476.

ALBERT R. - BARABÁSI A.-L.
2002 *Statistical mechanics of complex networks*, in "Reviews of Modern Physics", 74, pp. 47–97.

ATKINSON R.- FLINT J.
2001 *Accessing hidden and hard-to-reach populations: Snowball research strategies*, in "Social Research Update", 33. http://sru.soc.surrey.ac.uk/SRU33.html

BABBIE E.
2010 *Ricerca sociale*, Apogeo, Milano.

BARBAGLI M.
1995 *L' occasione e l'uomo ladro. Furti e rapine in Italia*, Il Mulino, Bologna.

BELL D.C. – BELLI MCQUEEN B. - HAIDER A.
2007 *Partner naming and forgetting: Recall of network members*, in "Social Networks", 29, pp. 279-299.

BERGSTEN J.W. – PIERSON S.A.
1982 *Telephone Screening for Rare Characteristics Using Multiplicity Counting Rules*, p. 145-150 in Proceedings of the Survey Research Methods Section, American Statistical Association (1982). http://www.amstat.org/sections/srms/Proceedings/

BERNHARDT A. - HECKATHORN D. D.- MILKMAN R. - THEODORE N.
2006. *Documenting unregulated work: A survey of workplace violations in New York City. The Future of Work*. http://www.russellsage.org.

BICHI R.
2002 *L'intervista biografica. Una proposta metodologica*, Vita e Pensiero, Milano.
2007 *La conduzione delle interviste nella ricerca sociale*, Carocci, Roma.

BIRNBAUM Z.W. – SIRKEN M.G.
1965 *Design of sample surveys to estimate the prevalence of rare diseases: Three unbiased estimates*, in "Vital Health Statistics", 2 (11), pp. 1-8.

*La sitografia si intende aggiornata al 10 dicembre 2010.

BJØRKHAUG I. - HATLØY A.
2009 *Utilization of respondent-driven sampling among a population of child workers in the diamond-mining sector of Sierra Leone*, in "Global Public Health", 4 (1), pp.96-109.

BLANGIARDO G. C.
1996 "Il campionamento per centri o ambienti di aggregazione nelle indagini sulla presenza straniera", in *Studi in onore di Giampiero Landenna*, Giuffrè, Milano.
2004 "Campionamento per centri nelle indagini sulla presenza straniera in Lombardia: una nota metodologica", in M.M. Pelegatti (a cura di), *Studi in ricordo di Marco Martini*, Giuffrè, Milano.

BOLLING K. - GRANT C. - SINCLAIR P.
2008 *2006-07 British Crime Survey (England and Wales). Technical Report. Volume I.* http://www.homeoffice.gov.uk/rds/pdfs07/bcs0607tech1.pdf

BRACKERTZ N.
2007 *Who is hard to reach and why?*, ISR Working Paper, January 2007. www.sisr.net/publications/0701brackertz.pdf

BREWER D.D.
2000 *Forgetting in the recall-based elicitation of personal and social networks*, in "Social Networks", 22, pp. 29–43.

BRICK J.M.
1990 *Multiplicity sampling in an RDD Telephone Survey*, pp. 296 –301 in Proceedings of the Survey Research Methods Section, American Statistical Association. http://www.amstat.org/sections/srms/Proceedings/

BRUSCHI A.
2005 *La metodologia della ricerca sociale*, Editori Laterza, Roma.

BURT R.D. – HAGAN H. – SABIN K. – THIEDE H.
2010 *Evaluating respondent-driven sampling in a major metropolitan area: Comparing injection drug users in the 2005 Seattle area National HIV Behavioral Surveillance System with injectors in the RAVEN and Kiwi studies*, in "Annals of Epidemiology", 20, pp. 159–167.

CASELLI M.
2005 *Indagare col questionario. Introduzione alla ricerca sociale di tipo standard*, Vita e Pensiero, Milano.

CERVANTES I. F. - KALTON G.
2008 "Methods for Sampling Rare Populations in Telephone Surveys", pp. 113-132 in Lepkowski J. M., Tucker C., Brick J.M., de Leeuw E.D., Japec L., Lavrakas P.J. , Link M. W., Sangster R. L. (a cura di), *Advances in Telephone Survey Methodology*, J.W. Wiley and Sons, New York.

CHIESI A. M.
1996 *Attori e relazioni tra attori mediante l'analisi dei reticoli multipli*, in "Rassegna Italiana di Sociologia", XXXVII, n. 1, pp. 57-81.
1999 *L'analisi dei reticoli*, Franco Angeli, Milano.

COLEMAN J.S.
1958 *Relational analysis: the study of social organization with survey methods*, in "Human Organisation", 17, pp. 28–36.

CORBETTA P.
1999 *Metodologia e tecniche della ricerca sociale*, Il Mulino, Bologna.

CZAJA R. - BLAIR J.
1990 *Using Network Sampling in Crime Victimization Surveys*, in "Journal of Quantitative Criminology", 6, pp. 185-206.

ERENS B. - PRIOR G. - KOROVESSIS C. - CALDERWOOD L. - BROOKES M. - PRIMATESTA P.
2001 "Survey methodology and response", cap. 4 in B. Erens, P. Primatesta and G. Prior (a cura di), *Health Survey for England – The Health of Minority Ethnic Groups '99. Volume 2: Methodology and Documentation*, The Stationery Office, London. http://www.archive.official-documents.co.uk/document/doh/survey99/hse99-00.htm

ERICKSON B. H.
1979 *Some Problems of Inference from Chain Data*, in "Sociological Methodology", 10, pp. 276–302.

FRANK O.
1977 *Survey sampling in graphs*, in "Journal of Statistical Planning and Inference", 1, pp. 235-264.

FRANK O. - SNIJDERS T.
1994 *Estimating the size of hidden populations using snowball sampling*, in "Journal of Official Statistics", 10, pp. 53-67.

FROST S.D.W. - BROUWER K.C. - FIRESTONE CRUZ M.A. - RAMOS R. - RAMOS M.E. - LOZADA R.M. - MAGIS-RODRIGUEZ C. - STRATHDEE S.A.
2006 *Respondent-driven sampling of injection drug users in two US.-Mexico border cities: Recruitment dynamics and impact on estimate of HIV and Syphilis prevalence*, in "J. Urban Health", 83, pp. 83–97.

GILBERT N.
2008 *Researching Social Life*, Sage Publications, London.

GILE K.J - HANDCOCK M.S.
2010 *Respondent-Driven Sampling: An Assessment of Current Methodology*, in "Sociological Methodology", 40 (1), pp. 285–327.

GOEL S. - SALGANIK M.J.
2009 *Respondent-driven sampling as Markov-chain Monte Carlo*, in "Statistics in medicine", 28 (17), pp. 2202-2229.
2010 *Assessing respondent-driven sampling*, in "Proceedings of the National Academy of Sciences", 107 (15), pp. 6743-6747.

GOODMAN L.A.
1961 *Snowball Sampling*, in "Annals of Mathematical Statistics", 32 (1), pp.148-170.

GRANOVETTER M.
1976 *Network Sampling: Some First Steps* in "The American Journal of Sociology", 81 (6), pp. 1287-1303.

GRIBAUDI M.
1996 *L'analisi di rete: tra struttura e configurazione*, in "Rassegna italiana di Sociologia", XXXVII, 1, pp. 31-55.

GRINSTEAD C. M. - SNELL J. L.
2006 *Introduction to Probability*, American Mathematical Society. http://www.math.dartmouth.edu/~prob/prob/prob.pdf

HECKATHORN D.D.
1997 *Respondent driven-sampling: A new approach to the study of hidden populations*, in "Social Problems", 44, pp. 174-199.
2002 *Respondent-driven sampling II: Deriving valid population estimates from chain referral samples of hidden populations*, in "Social Problems", 49, pp. 11-34.
2007 *Extensions of Respondent-Driven Sampling: Analyzing Continuous Variables and Controlling for Differential Recruitment*, in "Sociological Methodology", 37 (1), pp. 151-207.

HECKATHORN D.D. - JEFFRI J.
2001 *Finding the beat: Using respondent-driven sampling to study jazz musicians*, in "Poetics", 28, pp. 307–329.

HERZEL A.
1991 "Teoria e tecniche dei campioni" pp. 626-630, in *Enciclopedia delle Scienze Sociali*, vol. I, Istituto dell'Enciclopedia Italiana Treccani, Roma.

ISRAEL G.
2002 *Modelli matematici. Introduzione alla matematica applicata*, Franco Muzzio Editore, Padova.

ISTAT
2001 *Indagini sociali telefoniche. Metodologia ed esperienze della statistica ufficiale*, Istituto Nazionale di Statistica, Roma.
2006 *Il sistema delle indagini sociali multiscopo – Contenuti e metodologia dell'indagine*, Istituto Nazionale di Statistica, Roma.

JOHNSTON L.G. - KHANAM R. - REZA M. - KHAN I. S. - BANU S. - ALAM MD. S. - RAHMAN M. - AZIM T.
2008 *The effectiveness of respondent-driven sampling for recruiting males who have sex with males in Dhaka, Bangladesh*, in "AIDS and Behavior", 12 (2), pp. 294–304.

JOHNSTON L.G. - TRUMMAL A. - LOHMUS L. - RAVALEPIK A.
2009 *Efficacy of convenience sampling through the internet versus respondent-driven sampling among males who have sex with males in Tallin and Harju County, Estonia: Challenges reaching a hidden population*, in "AIDS Care", 21 (9), pp.1195–1202

KALTON G.
2009 *Methods for oversampling rare subpopulations in social surveys*, in "Survey Methodology", 35 (2), pp. 125-141.

KALTON G. - ANDERSON D.W.,
1986 *Sampling rare populations*, in " Journal of the Royal Statistical Society" Ser. A 149, pp. 65-82.

KENDALL C. - KERR L.R.F.S. - GONDIM R.C. - WERNECK G. L. - MAIA MACENA R.H. - PONTES M. K. - JOHNSTON L.G. - SABIN K. - MCFARLAND W.
2008 *An empirical comparison of respondent-driven sampling, time location sampling, and snowball sampling for behavioral surveillance in men who have sex with men, Fortaleza, Brazil*, in "AIDS and Behavior ", 12 (Supplement 1), pp. 97–104.

KISH L.
1965 *Survey Sampling*, Wiley, New York.
1991 *Taxonomy of Elusive Populations*, in "Journal of Official Statistics", 7, pp. 339-347.
1992 *Weighting for unequal* p_i, in "Journal of Official Statistics", 8, pp. 183–200.

KLOVDAHL A.
1989 "Urban social networks: some methodological problems and possibilities", pp. 176 –210 in Kochen, M. (a cura di), *The Small World*, Ablex Publishing, Norwood, NJ.

KNOKE D. - YANG S.
2008 *Social network analysis* (2n ed.), Sage, Beverly Hills and London.

LANZETTI C.
2004 *La qualità del servizio in ospedale - Una ricerca sull'esperienza dei malati*, Franco Angeli, Milano.

LANZETTI C. – LOMBI L. - MARZULLI M. (a cura di)
2008 *Metodi qualitativi e quantitativi per la ricerca sociale in sanità*, Franco Angeli, Milano.

LIPOVSEK V. - LONGFIELD K.
2007 *Sampling Hard-to-Reach Populations-PSI Research & Metrics*, Population Services International. http://www.psi.org/resources/publications

LU X. - BENGTSSON L. - BRITTON T. - CAMITZ M. - KIM B.J. - THORSON A. - LILJEROS F.
2010 *The Sensitivity of Respondent-driven Sampling Method.* http://arxiv.org/abs/1002.2426v3

MALEKINEJAD M. - JOHNSTON L. - KENDALL C. - KERR L. - RIFKIN M. - RUTHERFORD G.
2008 *Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance in International Settings: A Systematic Review*, in "AIDS and Behavior ", 12 (Supplement 1), pp. 105–130.

MARRADI A.
1997 "Casuale e rappresentativo: ma cosa vuole dire?" pp. 9-52 in Paolo Ceri (a cura di), *La politica e i sondaggi*, Rosenberg & Sellier, Torino.
2007 *Metodologia delle scienze sociali*, (a cura di Pavsic R.; Pitrone M. C.), Il Mulino, Bologna.

MARSDEN P.
1990 *Network data and measurement*, in "Annual Review of Sociology", 16, pp. 435-463.

MATTIOLI F.
1995 *Sociometria*, la Goliardica, Roma.

MCPHERSON M. - SMITH-LOVIN L. - COOK J. M.
2001 *Birds of a Feather: Homophily in Social Networks*, in "Annual Review of Sociology", 27, pp. 415–44.

MUTTI A.
1996 *Reti sociali tra metafore e programmi teorici*, in "Rassegna italiana di Sociologia", XXXVII, 1, pp. 5-30.

NEELY W. W.
2009 *Statistical Theory for Respondent-Driven Sampling*. Ph.D. thesis, University of Wisconsin-Madison, Madison, Wisconsin.

PALUMBO M. - GARBARINO E.
2004 *Strumenti e strategie della ricerca sociale. Dall'interrogazione alla relazione,* FrancoAngeli, Milano.

PATRICK J.
1973 *A Glasgow gang observed*, Eyre Methuen, London.

PISELLI F.
1995 *Reti. L'analisi di network nelle scienze sociali*, Donzelli, Roma.

PLATT L. - WALL M. - RHODES T. - JUDD A. - HICKMAN M. - JOHNSTON L.G. - RENTON A. - BOBROVA N. - SARANG A.
2006 *Methods to recruit hard-to-reach groups: Comparing two chain referral sampling methods of recruiting injection drug users across nine studies in Russia and Estonia*, in "Journal of Urban Health", 83, pp. 39–53.

RAMIREZ-VALLES J. - HECKATHORN D.D. - VASZQUEZ R. - DIAZ R.M. - CAMPBELL R.T.
2005 *From networks to populations: the development and application of respondent-driven sampling among IDUs and Latino Gay Men*, in "AIDS Behavior", 9 (4), pp. 387–402.

ROBINSON W.T., RISSER J.M.H., MCGOY S., BECKER A.B., REHMAN H., JEFFERSON M., GRIFFIN V., WOLVERTON M., TORTU S.
2006 *Recruiting injection drug users: A three-site comparison of results and experiences with respondent-driven and targeted sampling procedures*, in "Journal of Urban Health", 83, pp. 29–38.

ROTHBART G. S. - FINE M. - SUDMAN S.
1982 *On Finding and Interviewing the Needles in the Haystack: The Use of Multiplicity Sampling*, in "Public Opinion Quarterly", 46 (3), pp. 408-421.

SALGANIK M.J.
2006 *Variance estimation; design effects, and sample size calculations for Respondent-Driven Sampling*, in "Journal of Urban Health", 83, pp-98-112.

SALGANIK M.J. - HECKATHORN D.D.
2004 *Sampling and estimation in hidden populations using respondent-driven sampling*, in "Sociological Methodology", 34, pp.193–239.

SCHWARTZ N. - WELLENS T.
1997 *Cognitive Dynamics of Proxy Reporting: The Diverging Perspectives of Actor and Observer*, in " Journal of Official Statistics", 13 (2), pp. 159–179.

SCLAVI M.
1994 *La signora va nel Bronx*, Anabasi, Milano.

SCOTT G.
2008 *They got their program, and I got mine": A cautionary tale concerning the ethical implications of using respondent-driven sampling to study injection drug users*, in "International Journal of Drug Policy", 19, pp. 42–51.

SIRKEN M.G.
1970 *Household Surveys with Multiplicity*, in "Journal of the American Statistical Association", 65, pp. 257-266.
2006 *Network sampling* , in "Encyclopedia of Biostatistics", Vol. 4, John Wiley and Sons, pp. 2977-2986.

SNIJDERS T.A.B.
1992 *Estimation on the basis of snowball samples: how to weight*, in "Bulletin of Sociological Methodology", 36, 59-70.

STUART A.
1996 *I sondaggi di opinione - Idee per il campionamento*, Newton & Compton, Roma.

SUDMAN S. - FREEMAN H.
1988 *The Use of Network Sampling for Locating the Seriously Ill*, in "Medical Care", 26 (10), pp. 992-999.

SUDMAN S. - SIRKEN M.G. - COWAN C.D.
1988 *Sampling Rare and Elusive Populations*, in "Science", 240 (4855), pp. 991-996.

TAGLIACARNE G.
1964 *Tecniche e pratica delle ricerche di mercato*, Dott. A. Giuffrè, Milano.

THOMSON S.K.
1997 *Adaptive sampling in behavioural surveys*, in " NIDA Research Monograph", pp. 296-319.
2002 *Sampling,* John Wiley and Sons, New York.
2006 *Targeted random walk designs*, in "Survey Methodology" 32, pp. 11-24.

THOMPSON S.K. - COLLINS L.M.
2002 *Adaptive sampling in research on risk-related behaviors*, in "Drug and Alcohol Dependence", 68, pp. 557–567.

THOMPSON S. K. - SEBER G.A.F.
1996 *Adaptive Sampling*, John Wiley & Sons, New York.

TORTORA R. - GROVES R. M. - PEYTCHEVA E.
2008 "Multiplicity-Based Sampling for the Mobile Telephone Population: Coverage, Nonresponse, and Measurement Issues", pp. 133–148 in Lepkowski J. M., Tucker C., Brick J.M., de Leeuw E.D., Japec L., Lavrakas P.J. , Link M. W., Sangster R. L. (a cura di), *Advances in Telephone Survey Methodology*, J.W. Wiley and Sons, New York.

VITALINI A.
2010a "*L'uso delle reti sociali per la costruzione di campioni probabilistici: possibilità e limiti per lo studio di popolazioni nascoste*", in "Studi di Sociologia", 3, pp. 251-266.
2010b *Il campionamento casuale – Manuale pratico per ricercatori sociali*, Franco Angeli, Milano.

VOLZ E. - HECKATHORN D.D.
2008 *Probability-based estimation theory for respondent-driven sampling,* in "Journal of Official Statistics", 24 (1), pp. 79-87.

VOLZ E. - WEJNERT C. - DEGANI I. - HECKATHORN D.D.
2007 *Respondent-Driven Sampling Analysis Tool* (RDSAT) Version 6.0.1., Ithaca, NY. http://www.respondentdrivensampling.org/

WALTERS K. - SIMONI J.
2002 *Health survey of two-spirited native americans*, Grant no. 1R01MH065871-01, National Institute of Mental Health.

WANG J. - ROBERT G. C. - RUSSEL S. F. - HARVEY A. S. - AHMMED R. - LINNA L.
2005 *Respondent-Driven Sampling to Recruit MDMA Users: A Methodological Assessment* in "Drug and Alcohol Dependence", 78, pp.147–157.

WASSERMAN S. - FAUST, K.
1994 *Social Network Analysis: Methods and Application*, Cambridge University Press, New York.

WATTERS J.K. - BIERNACKI P.
1989 *Targeted sampling: options for the study of hidden populations* in "Social Problems", 36, pp. 416-430.

WATTS D.J.
2003 *Small Worlds: The Dynamics of Networks Between Order and Randomness,* Princeton University Press, Princeton, NJ.

WEJNERT C.
2009 *An empirical test of respondent –Driven Sampling: point estimates, variance, degree measures, and out-of-equilibrium data*, in "Sociological Methodology", 39 (1), pp. 73-116.

WEJNERT C. - HECKATHORN D.D.
2008 *Web-based network sampling: Efficiency and efficacy of respondent-driven sampling for on-line research*, in "Sociological Methods Research", 37, pp. 105–134.

WHYTE W. F.
1955 *Street corner society: the social structure of an Italian slum*, University of Chicago Press, Chicago.