UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Psicologia dello Sviluppo e della Socializzazione

Corso di Laurea in Scienze e Tecniche Psicologiche

Elaborato finale

INFERENZA DA CAMPIONI NON PROBABILISTICI: METODI DI STIMA A CONFRONTO

Relatore : Prof. Antonio Calcagnì

Laureanda : Francesca Melegati

Matricola 1174325

Anno accademico 2021/2022

Dedico il presente lavoro a mia figlia, inconsapevole forza motrice di questo percorso. E ringrazio mio marito, per la pazienza ed il supporto.
Un ringraziamento speciale anche al Professor Antonio Calcagnì, per la disponibilità dimostratami.

Indice

Introduzione

Il campionamento non probabilistico è piuttosto diffuso nell'ambito della ricerca sociale. Le ragioni per cui vi si ricorre sono molteplici, dalla necessità di contenere i costi di un'indagine, all'assenza dei presupposti logici per il campionamento probabilistico. Da un punto di vista applicativo questo tipo di campionamento non implica l'impossibilità di stimare le caratteristiche (*parametri*) di interesse nella popolazione target. Tuttavia l'impraticabilità dell'inferenza design– based richiede di fare riferimento a metodi inferenziali opportuni. Ad oggi non esiste ancora un paradigma teorico di riferimento, ma alcuni studiosi si sono occupati di proporre e testare empiricamente differenti metodi di stima.

L'elaborato traccia una breve panoramica sul campionamento non probabilistico, evidenziandone le criticità (capitolo 1), ed illustra alcuni dei principali metodi di stima da campioni non probabilistici presenti in letteratura (capitolo 2). Quindi approfondisce il Propensity Score Adjustment (capitolo 3) e ne mostra alcuni esempi applicativi (capitolo 4).

Capitolo 1

Il campionamento non probabilistico

Il campionamento non probabilistico, introdotto da Anders N.Kiaer nel 1895, è stato spesso utilizzato agli inizi del '900 per indagare caratteri di popolazioni finite. Tuttavia, la scarsa rappresentatività, ovvero le difficoltà nel pervenire a stime accettabili dei parametri in esame, ha portato la comunità scientifica a prediligere il campionamento di tipo probabilistico. A partire dalle trattazioni di Bowley [1906, 1926], di Neyman [1934] e di Horvitz e Thompson[1952], che hanno gettato le basi dell'odierna teoria statistica, i campioni probabilistici hanno costituito lo standard per la ricerca di "buona qualità ". Il coerente impianto di principi matematici sottostanti al disegno campionario e all'applicazione dell'inferenza design – based, giustificano la fiducia nelle stime ricavate.

Le indagini campionarie non probabilistiche non sono però scomparse: hanno continuato e continuano ad essere proposte, soprattutto nella ricerca di mercato e nei sondaggi di opinioni. Nonchè viene comunemente ritenuto lecito il ricorso a campioni di convenienza laddove l'inferenza sia model–

based (ad esempio negli studi sperimentali ed osservazionali) e/o si ipotizzi che gli effetti del fenomeno indagato siano distribuiti in modo omogeneo in tutti i soggetti della popolazione target.

Alla fine del XX secolo, con l'avvento di internet e la possibilità di reclutare numerosi partecipanti tramite panel online, i campioni non casuali hanno suscitato un rinnovato interesse. Il Web consente infatti la raccolta di un gran numero di dati in tempi brevi e con costi "irrisori". Il principale limite delle indagini campionarie probabilistiche stà proprio nel necessitare di risorse cospicue per la realizzazione, sia in termini di tempo che di denaro. A cui si aggiunge, oggigiorno, una decrescita rilevante dei tassi di risposta.

Nell'ultimo decennio si è dunque riacceso il dibattito che vede, da una parte, chi sostiene l'impossibilità di ricavare da un campione non probabilistico informazioni estendibili alla popolazione target e chi, dall'altra, si domanda se e quando, con le dovute assunzioni e le opportune metodologie, si possa giungere a stime accurate dei parametri in esame.

1.1 Le criticità

Alla base del campionamento non probabilistico non c'è la randomizzazione. Non sono note le probabilità di selezione delle unità e tali probabilità potrebbero pure essere nulle.
Le modalità con cui i soggetti vengono inseriti in un campione possono essere molteplici ed anche del tutto arbitrarie. Nel campionamento di convenienza (*convenience sampling*) vengono scelti soggetti "vicini" e disponibili. Il reclutamento di volontari, ad esempio, è piuttosto diffuso negli studi clinici e l'autocandidatura è peculiarità anche nella selezione online di partecipanti

a sondaggi/ricerche. Nella corrispondenza del campione (*sample matching*) si selezionano soggetti in modo da rispecchiare la distribuzione di caratteristiche significative nella popolazione. Come nel campionamento per quote (*quota sampling*), dove, conoscendo le proporzioni (quote) di una certa variabile nella popolazione, per lo più demografica, si vanno ad includere i soggetti in modo che vengano rispettate tali quote anche nel campione. Nel campionamento a rete (*network sampling*) sono gli stessi membri di determinate popolazioni, generalmente rare (come consumatori di droghe) a fornire i nominativi di ulteriori soggetti. Ad esempio reclutandoli tra i loro conoscenti, come nel campionamento a palla di neve (*snowball sampling*).
Al di là della specifica tipologia di selezione, il frame di campionamento tende a non essere ben definito e/o presentare importanti problemi di copertura, con categorie di persone che non hanno la possibilità di essere incluse.
Al fine di poter fare inferenze sulla popolazione target a partire dai dati elaborati dal campione è invece fondamentale che nessuna porzione della popolazione venga sistematicamente esclusa dal campionamento .
È altresì necessario che la composizione del campione rispetto alle caratteristiche osservate e ritenute significative corrisponda (o possa essere adattata) a quella della popolazione e che ci sia una concordanza delle caratteristiche misurate tra soggetti campionati e non campionati [Mercer et al., 2017] .
Di fatto, anche un campione non probabilistico per poter essere funzionale deve essere rappresentativo della popolazione target. E, analogamente che con campioni probabilistici, ci si trova a dover far fronte a:

- problemi di copertura della popolazione(*bias di selezione*)
- mancata partecipazione di alcuni soggetti selezionati o presenza di loro risposte mancanti o incomplete (*nonresponse*)

- errori di misurazione
- perdita di interesse o abbandono dell'indagine di alcune persone (*attrition*). Alcuni soggetti, ad esempio, rispondono in modo sbrigativo e superficiale poichè partecipano solo per ottenere il beneficio economico previsto alla fine dell'indagine [Elliot e Valliant, 2017]

Tali storture rispecchiano i cosiddetti errori non campionari delle indagini probabilistiche. Tuttavia nei campioni non probabilistici il grado con cui si presentano e con cui possono influenzare negativamente le stime è maggiore [Valliant et al., 2013]. L'efficienza dell'inferenza (cioè l'accuratezza delle stime) non viene infatti garantita da una base teorico-matematica considerata universalmente valida. Si fonda esclusivamente sulla corretta specificazione dei modelli da parte dei ricercatori, in base alle loro assunzioni, non testabili, sul fenomeno oggetto di indagine e sulle differenze tra campione e popolazione.

Il campionamento probabilistico è dunque tutt'oggi ritenuto la scelta migliore. Le evidenze empiriche suggeriscono che l'accuratezza delle stime nelle indagini probabilistiche è più elevata che in quelle non probabilistiche [Cornesse et al., 2020]. Nonostante ciò gli ingenti costi e le tempistiche portano spesso a preferire campioni non casuali. E per questo in letteratura sono presenti diversi metodi che consentono di correggere gli stimatori e migliorare l'accuratezza delle stime.

Capitolo 2

Metodi di stima

I bias di selezione legati all'autocandidatura e i problemi di copertura sono indubbiamente le sfide principali per le indagini che impiegano campioni non probabilistici. Nei sondaggi-indagini-studi che ricercano volontari tramite il web, ad esempio, oltre all'esclusione delle persone senza accesso ad internet, si registra frequentemente una prevalente presenza di giovani con medio/alto livello di istruzione.

La situazione che si presenta è la seguente. Sia S un campione non probabilistico di n_s unità da una popolazione U, siano x_i e y_i rispettivamente i valori delle informazioni ausiliarie e delle variabili d'analisi rilevate sull'i-esima unità di S. Poichè S non è rappresentativo di U, l'applicazione della statistica inferenziale ai dati raccolti, seppur possibile, conduce a stime dei parametri di Y in U (come medie, proporzioni, totali, coefficienti di regressione, etc.) affette da bias e ne indebolisce le proprietà asintotiche. Senza opportune metodologie ed aggiustamenti sarebbe impossibile trarre generalizzazioni dal campione S alla popolazione U.

Come viene affrontata l'inferenza? Come vengono analizzati ed elaborati i dati e come vengono eventualmente corretti i risultati?

In alcuni casi i ricercatori adottano particolari strategie già durante la fase di campionamento con l'intento di attenuare i potenziali bias, rendendo il campione simile alla popolazione. Una strategia è quella di ricorrere al campionamento per quote. Un'altra è di estrarre il campione non random da un ampio panel di volontari facendo in modo di creare una corrispondenza di caratteristiche tra gli individui che lo compongono e quelli appartenenti ad un campione probabilistico, estratto da un frame che copre la stessa popolazione di riferimento U. Mentre il campione casuale ha la sola funzione di consentire il matching, il campione non probabilistico viene sottoposto all'indagine sull'argomento di studio (cioè vengono rilevati e analizzati i valori delle variabili di interesse) [Rivers, 2007]. Un'ulteriore strategia consiste nell'individuare i partecipanti facendo riferimento a fonti presumibilmente affette da distorsioni opposte [Comer, 2019].

Laddove invece si intervenga dopo la raccolta dei dati, esistono differenti metodi per correggere gli stimatori in modo da ridurre gli effetti delle distorsioni nelle stime dei parametri d'interesse, rendendo più attendibile il processo inferenziale ed aumentando la validità esterna di un'indagine. Fondamentale in questo senso si è rivelato l'adattamento di metodologie già utilizzate per ridurre gli errori di copertura e di mancata risposta nelle indagini campionarie probabilistiche.

Le diverse prospettive sono accomunate dal fare ricorso ad informazioni ausiliarie (le covariate x), osservate nel campione non casuale e di cui devono esserne almeno noti:

- i totali (o le medie) nella popolazione target U

- i valori per ciascuna unità nella popolazione target U

- i valori per ciascuna unità di un campione probabilistico di riferimento S_r (*reference sample*)

Per questioni di reperibilità ci si affida prevalentemente alle variabili socio-demografiche (età, genere, etnia, livello di istruzione, reddito etc.). I dati possono essere ricavati da censimenti, da fonti amministrative o demografiche, da statistiche ufficiali oppure da indagini probabilistiche di elevata qualità. Talvolta vengono condotte indagini probabilistiche parallele con lo scopo di rilevare variabili più pertinenti ed in linea con l'indagine.

In generale è possibile suddividere i metodi di stima da campioni non probabilistici in approcci di ponderazione (*weighting*) e approcci predittivi.

2.1 Approcci di Ponderazione

Gli approcci di ponderazione cercano di ridurre i bias di selezione e i problemi di copertura manipolando i dati affichè il campione arrivi a rispecchiare la popolazione nella distribuzione delle covariate. Proprio dal loro impiego in differenti modelli si ricavano i pesi correttivi, w_i ($\forall i \in S$), con cui le unità campionarie vengono bilanciate e/o "gonfiate" alla popolazione.

Gli approcci di ponderazione producono un aggiustamento globale, che può essere indistintamente applicato per tutte le variabili Y in esame [Cornesse et al., 2020]. Ovvero, presupponendo una certa interscambiabilità tra le unità di S e quelle di U-S, i coefficienti correttivi ottenuti sono adoperabili nelle analisi inferenziali che coinvolgono tutte le variabili oggetto di studio: lo stesso set di pesi può essere inserito negli stimatori di medie, proporzioni e totali di ciascuna.

La ponderazione vincolata o *calibrazione* [Deville and Särndal, 1992] individua i pesi imponendo un'equivalenza tra i totali delle variabili ausiliarie nel campione e i totali noti delle stesse variabili nella popolazione. Il set di pesi ottimale è quello che minimizza la distanza dal set di pesi iniziali (si considera convenzionalmente il valore unitario come peso base delle unità campionarie nei campioni non random) e, al tempo stesso, rispetta il vincolo dell'uguaglianza, riproducendo i totali noti delle covariate nella popolazione

$$\overset{min}{i \in S} \sum_{i \in S} G_i(w_i^{cal}, d_i) \quad t.c. \sum_{i \in S} w_i x_i = X \tag{2.1}$$

con d_i peso base di una unità i del campione non probabilistico (in genere 1 poichè non è nota la probabilità di selezione), $G(.)$ funzione di distanza prescelta e X vettore dei totali noti delle covariate nella popolazione. La scelta della funzione di distanza, che deve essere positiva e strettamente convessa, può ricadere su di molteplici alternative (lineare, logaritmica, chi-quadrato, etc.). La maggiormente adoperata è quella lineare.

Altre tecniche di ponderazione coinvolgono un campione probabilistico di riferimento, con maggiore copertura di U. Elliot e Valliant [2017] utilizzano il termine quasi – randomizzazione per indicare quei metodi che, avvalendosi del campione di riferimento, trattano il campione non probabilistico come se fosse stato generato da un meccanismo di probabilità sconosciuto quindi individuano modelli per stimare le probabilità di "pseudo-inclusione" delle unità campionarie.

Il principale di questi metodi è la ponderazione del punteggio di propensione (*Propensity Score Adjustment*), che ricava i coefficienti di ponderazione

dagli inversi delle stime della probabilità di pseudo – inclusione [Lee e Valliant, 2009; Valliant e Dever, 2011; Schonlau and Couper, 2017; Valliant, 2019]. Le stime si ottengono combinando insieme i due campioni, non casuale e di riferimento, ed utilizzando le informazioni ausiliarie, misurate in entrambi, in un modello di regressione binaria (logit o probit) volto a predire la propensione a prendere parte all'inadgine ovvero l'appartenenza al campione non probabilistico (i dettagli sul metodo di stima e sulle differenti alternative per la costruzione dei pesi vengono approfonditi nel capitolo successivo). Le tecniche di machine learning, come la classificazione e regressione ad albero(CART), bagging, o random forests [James, Witten, Hastie, and Tibshirani, 2014] rappresentano alternative alla regressione binaria per la stima delle pseudo probabilità di inclusione. L'efficacia del Propensity Score Adjustment non ha sempre trovato un riscontro empirico: in alcuni casi la riduzione delle distorsioni nelle stime si associa ad una aumento della varianza [Lee e Valliant, 2009 ; Valliant e Dever, 2011]. Per la stima di quest'ultima, gli approcci di ricampionamento come il bootstrap o il jackknife sembrano essere i più indicati [Elliot e Valliant, 2017].

La ponderazione del punteggio di propensione può essere anche utilizzata insieme alla calibrazione, inserendo i pesi correttivi ricavati dal Propensity Score Adjustment come pesi base nella calibrazione. Lee e Valliant [2009] e Ferri-Garcìa e del Mar Rueda [2018] hanno evidenziato come l'impiego combinato dei due approcci sia in grado di produrre stime più accurate rispetto all'applicazione di solo PSA o solo calibrazione.

Un altro metodo di quasi randomizzazione è lo *statistical matching*. La tecnica prevede che a ciascuna unità del campione non probabilistico venga assegnata la probabilità di inclusione di quella individuata come la corrispondente in un campione probabilistico di riferimento. La corrispondenza viene

calcolata attraverso una funzione di distanza che identifica l'accoppiata più "vicina" (come la tecnica del nearest neighbor) sulla base delle caratteristiche ausiliarie comuni [Valliant et al., 2018].

Tra gli approcci di ponderazione che si avvalgono di un campione probabilistico di riferimento si può menzionare pure l'*Entropy balancing* trattato in [Elliot e Watson, 2016]. Tale metodologia individua pesi correttivi che soddisfano una serie di vincoli di equilibrio riguardanti i momenti delle distribuzioni delle covariate nel campione e nel campione di riferimento, minimizzando al contempo la distanza tra i rispettivi pesi base. La distanza viene misurata tramite funzione di perdita (impiegando una metrica di divergenza di entropia diretta).

Gli approcci di ponderazione assumono l'ignorabilità del meccanismo di selezione nonchè presumono che le variabili ausiliarie nei dati di riferimento (censimento o indagine campionaria di probabilità) siano misurate senza errori e siano altamente correlate sia con le variabili di analisi che con la probabilità di partecipare all'indagine campionaria non probabilistica. Al fine di massimizzare tali assunzioni è necessario avere a disposizione un ampio set di covariate ovvero dati diversificati tra loro [Cornesse et al., 2020].

Le informazioni ausiliarie giocano dunque un ruolo chiave nel determinare l'efficacia della ponderazione, nel garantire stime effettivamente accurate. E può essere opportuno aggiungere ai dati demografici informazioni che differenziano il campione di convenienza dal resto della popolazione. Ad esempio, è stato mostrato che chi si autocandida nei panel online tende ad essere un consumatore precoce di nuovi prodotti e servizi: l'impiego di misure indice di tale qualià ("early adopter" variables) migliora le stime ottenute da ponderazione vincolata [DiSogra, Cobb, Chan, e Dennis, 2011; Fahimi, Barlas,

Thomas, e Buttermore 2015].

2.2 Approcci Predittivi

Gli approcci predittivi si inseriscono nel framework dei modelli di superpopolazione. Trattano la variabile d'analisi come se fosse stata generata da un modello, che viene specificato sulla base dei dati rilevati dal campione non probabilistico. A tal fine occorre avere a disposizione i valori delle variabili ausiliarie, oltre che per le unità del campione, anche per quelle della popolazione target. O almeno sono necessari i totali.

Si presuppone una relazione lineare tra le y e le x ed il modello viene utilizzato per predire i valori della variabile in esame nelle unità non campionate. La stima del valore y nell'i-esima unità si può ottenere come

$$\hat{y}_i = E_m(y_i|x_i) = x_i'\beta \tag{2.2}$$

dove m identifica l'aspettativa rispetto al modello e β è il vettore dei parametri (o coefficienti) che governa la relazione tra tra Y e X, stimabile con il metodo dei minimi quadrati. L'analisi combinata dei valori osservati e dei valori predetti consente la stima dei parametri della popolazione a cui si è interessati. Il presupposto fondamentale per ottenere stime non distorte è che le unità del campione non probabilistico S e quelle di $\overline{S} = U - S$ seguano lo stesso modello.

In [Elliot e Valliant, 2017] si propone l'impiego dello stimatore di totale model-based (*model-based estimator*)

$$\hat{t}_y^{mb} = \sum_{i \in S} y_i + \sum_{i \in \bar{S}} \hat{y}_i = \sum_{i \in S} y_i + (t_{Ux} - t_{Sx})^T \hat{\beta} \tag{2.3}$$

L'equazione può essere riscritta come

$$\hat{t}_y = \sum_{i \in S} w_i y_i \tag{2.4}$$

con $w_i = 1 + (t_{Ux} - t_{sx})^T (X_S^T X_S)^{-1} x_i$. Da cui la seguente stima della media

$$\hat{\bar{Y}} = \hat{t}_y / \hat{N} \tag{2.5}$$

con $\hat{N} = \sum_{i \in S} w_i$

Per la stima della varianza Elliot e Valliant [2017] propongono diversi stimatori ma quello con un miglior riscontro empirico risulta sempre il jackknife.

Una procedura di stima analoga a quella appena descritta fà riferimento allo stimatore di totale (*model– assisted estimator*)

$$\hat{t}_y^{ma} = \sum_{i \in U} \hat{y}_i + \sum_{i \in S} (y_i - \hat{y}_i) \tag{2.6}$$

La *calibrazione assistita da modello*, invece, si caratterizza per la costruzione di pesi calibrati utilizzando il modello lineare per prevedere i valori della variabile di analisi [Wu e Sitter, 2001]. Al fine di una modellazione maggiormente parsimoniosa della variabile target Chen, Valliant ed Elliott [2018] suggeriscono di stimare β tramite LASSO (*Least Absolute Shrinkage and Selection Operator*). I pesi calibrati sono generati sulla base dei vincoli posti sulla dimensione della popolazione e sul totale dei valori previsti nella

popolazione.

Cioè

$$\hat{t}_y^{mcal} = \sum_{i \in S} y_i w_i^{cal} \tag{2.7}$$

con $\sum_{i \in S} w_i^{cal} \hat{y}_i = \sum_{i \in U} \hat{y}_i$

Nella metodo di *stima doppiamente robusta* gli approcci basati sul modello sono applicati successivamente al PSA. Valliant [2019] ha mostrato come la combinazione dei due approcci sia in grado di produrre stime più accurate rispetto all'impiego di una singola tecnica.

Anche la regressione multilivello con post– stratificazione *(MRP)* rientra negli approcci predittivi. L'adattamento di un modello di regressione multilivello offre il vantaggio di poter incorporare numerose covariate e di poter considerare interazioni gerarchiche. Dalla classificazione incrociata delle variabili ausiliarie vengono definite le celle di post-stratificazione. In ciascuna cella si procede con la stima model– based dei parametri di interesse (può essere utile un approccio Bayesiano), che vengono poi ponderati rispetto alla proporzione (deve essere nota o stimabile) di tale cella nella popolazione target [Wang et al.,2015]. Medie e proporzioni sono stimate come

$$\hat{\bar{y}} = \sum_{\gamma=1}^{G} P_\gamma \hat{\mu}_\gamma \tag{2.8}$$

con P_γ proporzione nella popolazione del poststrato $(PS)_\gamma$ e $\hat{\mu_\gamma}$ media stimata nel poststrato γ

È opportuno sottolineare come gli approcci predittivi producano correzioni risultato-specifiche, cioè ciascuna variabile d'analisi può potenzialmente seguire un modello diverso. Tuttavia l'idea è quella di identificare una forma di modello e un insieme di covariate che producano risultati ragionevoli per molte Y. Affinchè tali approcci producano stime precise è necessario che il modello sia correttamente specificato e siano inserite tutte le informazioni ausiliarie rilevanti per lo studio della variabile in esame.

Capitolo 3

Propensity Score Adjustment(PSA)

Il Propensity Score Adjustment affonda le sue radici nel Propensity Scoring e nell'Inverse Probability Weighting (IPW) introdotti per analizzare gli effetti causali di un trattamento negli studi osservazionali, rendendo bilanciati e comparabili gruppo sottoposto a trattamento e gruppo di controllo [Rosenbaum e Rubin, 1983; Robins, Hernan e Brumback, 2000].

La sua applicazione è stata poi estesa alle indagini telefoniche e a quelle online. Oggi è uno degli approcci maggiormente impiegati per correggere i bias di selezione nelle analisi statistiche da campioni di convenienza.

Il campione non probabilistico S viene analizzato congiuntamente ad un campione probabilistico di riferimento S_r, rappresentativo della medesima popolazione target U. Lo scopo è stimare la propensione dei partecipanti a prendere parte all'indagine, allineando il campione non probabilistico a quello probabilistico rispetto alla distribuzione delle variabili ausiliarie. Si ritiene infatti che tale propensione dipenda dalle caratteristiche stesse dei soggetti.

3.1 Stima puntuale di Media e Totali

Sia $X = (X_1, ..., X_p)$ un vettore di variabili ausiliarie comuni, rilevate e note sia $\forall i \in S$ che $\forall i \in S_r$. I dati di S e S_r vengono concatenati e uniti in unico campione $S_c = S \cup S_r$. Quindi viene utilizzata una regressione logistica al fine di modellare la relazione tra la variabile dicotomica δ (con $\delta = 1$ se $i \in S$, $\delta = 0$ se $i \in S_r$) ed X. Ciò consente di stimare $\forall i \in S$ la "pseudo – probabilità" di inclusione π nel campione non random, detta anche punteggio di propensione.

Assumendo idealmente che le unità in U-S manchino a caso (MAR) dal campione ovvero che la mancata inclusione non sia legata alla variabile/i d'analisi Y, si ha:

$$\pi(\delta = 1|x_i, y_i; \Gamma) = \pi(\delta = 1|x_i; \Gamma) = \frac{e^{\gamma^T x_i}}{e^{\gamma^T x_i} + 1} \tag{3.1}$$

Stimato il vettore di parametri γ (massima verosimiglianza) si ricavano i punteggi di propensione π. Si procede dunque con l'inferenza utilizzando le stime $\hat{\pi}$ per costruire coefficienti w_i^{psa} ($pesi$) da inserire nei consueti (design – based) stimatori di

- media

$$\hat{\bar{Y}} = \sum_{i \in S} w_i^{psa} y_i / \sum_{i \in S} w_i^{psa} \tag{3.2}$$

- totale

$$\hat{t}_y = \sum_{i \in S} w_i^{psa} y_i \tag{3.3}$$

- proporzione (con y categoriale)

$$\hat{p}_A = \sum_{i \in S} w_i^{psa} y_i / \sum_{i \in S} w_i^{psa} \quad con \quad \begin{cases} y = 1 & i \in categoriaA \\ y = 0 & i \notin categoriaA \end{cases} \tag{3.4}$$

3.1.1 La costruzione dei pesi

In letteratura sono presenti molteplici modalità di impiego dei punteggi di propensione per la realizzazione dei pesi w_i^{psa}.

La più immediata vede il coefficiente di ponderazione corrispondere all'inverso del punteggio di propensione [Valliant, 2019]

$$w_i^{psa1} = \frac{1}{\hat{\pi}(x_i)} \tag{3.5}$$

Un'alternativa, presente in Schonlau e Couper [2017] è

$$w_i^{psa2} = \frac{1 - \hat{\pi}(x_i)}{\hat{\pi}(x_i)} \tag{3.6}$$

I punteggi di propensione possono anche essere inseriti all'interno di procedure di post-stratificazione. Il presupposto è che la stratificazione contribuisca a migliorare il bilanciamento tra il campione non probabilistico e quello di riferimento rispetto alle variabili osservate x.

Lee e Valliant [2009] suggeriscono di suddividere i punteggi stimati in classi (convenzionalmente 5, in linea con Cochran [1968]), ciascuna delle quali contenente unità con uguale o molto simile punteggio di propensione. In ciascuna classe viene calcolato il seguente fattore correttivo :

$$f_{cl} = \frac{\sum_{i \in S_{r_{cl}}} d_{ri} / \sum_{i \in S_r} d_{ri}}{\sum_{i \in S_{cl}} d_i / \sum_{i \in S} d_i} \tag{3.7}$$

con d_{ri} peso base di una unità i del campione di riferimento (da disegno campionario) e d_i peso base di una unità i del campione non probabilistico (è diverso da 1 solo quando il campione viene estratto con procedura "probabi-

listica" da un più ampio panel di volontari quindi, sulla base di tale selezione, alle unità viene associato un pseudo peso base).

Ne deriva, $\forall i \in S$, il peso corretto

$$w_i^{psa3} = d_i f_{cl} \tag{3.8}$$

Da cui

$$\hat{\bar{Y}} = \frac{\sum_{cl} \sum_{i \in S} w_i^{psa3} y_i}{\sum_{cl} \sum_{i \in S} w_i^{psa3}} \tag{3.9}$$

$$\hat{t}_y = \sum_{cl} \sum_{i \in S} w_i^{psa3} y_i \tag{3.10}$$

In Dever e Valliant [2011] le stime dei punteggi di propensione sono usate per ordinare il campione congiunto S_c, che viene poi post-stratificato in 5 gruppi g di uguale dimensione. Misurato il punteggio di propensione medio di ciascun gruppo, $\bar{\pi_g}$, il suo inverso diventa il peso correttivo per ciascuna unità del gruppo:

$$\hat{t}_y = \sum_{g} \sum_{i \in S_g} \frac{d_i y_i}{\bar{\pi_g}} = \sum_{g} \sum_{i \in S_g} w_i^{psa4} y_i \tag{3.11}$$

$$\hat{\bar{Y}} = \frac{\sum_{g} \sum_{i \in S_g} w_i^{psa4} y_i}{\sum_{g} \sum_{i \in S_g} w_i^{psa4}} \tag{3.12}$$

con d_i peso base di una unità i del campione non probabilistico.

3.2 Stima della Varianza

Il metodo di stima consigliato per la varianza laddove venga applicato il Propensity Score Adjustment è il Jackknife. Le evidenze empiriche suggeriscono

come il ricorso al ricampionamento e alla replicazione dei pesi consenta stime di varianza in grado di riflettere maggiormente la variabilità legata ai pesi stimati [Valliant, 2019] .

L'applicazione del Jackknife con replicazione prevede il ricalcolo dei punteggi di propensione e dei pesi per ognuno degli n sottocampioni di S_c, generati eliminando di volta in volta l'i-esima osservazione.

Per la stima di una media (o proporzione), il relativo stimatore della varianza è

$$\hat{V}_{jk}(\hat{\bar{y}}) = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\bar{y}}_{jk(i)} - \hat{\bar{y}}_{jk})^2 \tag{3.13}$$

con $\hat{\bar{y}}_{jk(i)}$ stima della media del parametro nel campione senza l'unità i e $\hat{\bar{y}}_{jk}$ media delle stime delle medie del paramentro.

Nel capitolo successivo verranno mostrati alcuni esempi applicativi della stima di parametri mediante il Propensity Score Adjustment.

Capitolo 4

Esempi Applicativi

Il capitolo fornisce esempi applicativi dell'impiego del Propensity Score Adjustment per la stima di proporzioni da un campione non probabilistico.

Il pacchetto **NonProbEst** di *R* viene utilizzato per adattare il modello di propensione da cui ricavare i pesi correttivi, secondo le 4 diverse tecniche presentate nel capitolo precedente. I set di pesi sono poi impiegati per la stima delle proporzioni e dei totali. Ci si avvale dei dataset SampleNP, SampleP e population, contenuti nel pacchetto e creati da Rueda e Ferri-Garcìa [2018] per la simulazione con cui hanno studiato l'efficienza della combinazione di PSA e calibrazione.

4.1 I dati

SampleNP è un set di dati rilevati in 1000 unità statistiche per condurre un'ipotetica indagine sulle intenzioni di voto della popolazione.

Le variabili demografiche considerate sono il livello di istruzione conseguito (*education_primaria*,*education_secundaria*,*education_terciaria*), l'età (da 0 a 100), il genere ed il linguaggio (identifica l'essere nati nel paese del-

l'indagine simulata). Sono tutte variabili dicotomiche ad eccezione dell'età, che è una variabile numerica.

Le variabili d'analisi esprimono la preferenza nei confronti di un certo candidato (*vote_gen*, *vote_pens*, *vote_pi*).

```
'data.frame':   1000 obs. of  9 variables:
  vote_gen             : int  0 0 1 0 0 0 0 0 0 0 ...
  vote_pens            : int  1 0 0 0 0 0 0 0 1 0 ...
  vote_pir             : int  0 1 0 1 1 0 0 1 0 1 ...
  education_primaria   : int  1 0 0 1 0 1 0 0 0 0 ...
  education_secundaria: int  0 0 1 0 1 0 0 1 1 0 ...
  education_terciaria  : int  0 1 0 0 0 0 1 0 0 1 ...
  age                  : int  66 30 62 33 30 69 50 47 40 29 ...
  sex                  : int  1 1 0 0 0 1 1 1 0 1 ...
  language             : int  1 1 1 1 1 1 1 1 1 1 ..
```

Il campione è stato estratto tra le persone con accesso a Internet da una popolazione fittizia di 50.000 individui (dataset *population*). E riproduce dunque un caso di campionamento con bias di selezione (Nella tabella 4.1 si possono vedere le differenze tra campione e popolazione per quanto riguarda le proporzioni delle variabili ausiliarie).

Var.	Sample	Pop
e.primaria	0.473	0.50574
e.secondaria	0.208	0.21092
e.terziaria	0.319	0.28334
age $<$ 35	0.244	0.20448
age 35-65	0.649	0.59048
age $>$65	0.162	0.20504
M	0.498	0.48860
F	0.502	0.51140
linguaggio	0.982	0.90858

Tabella 4.1: Proporzioni

4.2 La procedura di stima

L'obiettivo è di stimare le proporzioni ed i totali delle preferenze di voto nella popolazione. La procedura di stima, a partire dall'analisi dei dati in SampleNP, prevede il calcolo delle stime dei punteggi di propensione, la creazione di un un set di pesi quindi la stima dei parametri.

4.2.1 Calcolo delle stime dei punteggi di propensione

L'applicazione del Propensity Score Adjustment necessita di un campione probabilistico di riferimento in cui sono state rilevate le medesime informazioni ausiliarie. A tal scopo viene impiegato il dataset *SampleP*, costituito dai dati raccolti in un campione casuale semplice di 500 individui (estratto dalla stessa popolazione fittizia).

```
'data.frame':   500 obs. of  5 variables:
  education_primaria  : int  1 0 1 0 0 1 1 0 0 0 ...
  education_secundaria: int  0 0 0 1 0 0 0 0 1 1 ...
  education_terciaria : int  0 1 0 0 1 0 0 1 0 0 ...
  age                 : int  35 64 55 61 35 51 53 30 49 31 ...
  sex                 : int  1 0 1 1 0 1 1 0 1 1 ...
```

Il livello di istruzione raggiunto, l'età ed il genere sono variabili comuni tra SampleP e SampleNP. Le usiamo per adattare il modello di propensione, mediante la funzione *propensities*.

```
covariates <- c("education_primaria","education_secundaria","education_terciaria",
"sex","age")
pi<- propensities(sampleNP, sampleP, covariates)
summary(pi$convenience)

Min.    1st Qu.  Median  Mean   3rd Qu.    Max.
0.4237  0.4736  0.4971  0.5032  0.5309  0.5895
```

4.2.2 Dai valori π ai pesi

Con le stime dei π costruiamo il set di pesi correttivi che consente di bilanciare il campione non probabilistico. Individuiamo un set di pesi con ciascuna delle tecniche presentate nel capitolo 3.

```
#Valliant
Val_w=valliant_weights(pi$convenience)
Val<-summary(Val_w)

#Schonlau e Couper
sc_w<-sc_weights(pi$convenience)
sc<-summary(Sc_w)

#Dever e Valliant
vd_w = vd_weights(pi$convenience, pi$reference,g=5)
vd<-summary(vd_w)

#Lee e Valliant
lv_w=lee_weights(pi$convenience, pi$reference, g = 5)
lv<-summary(lv_w)

#confronto
pesi_con<-matrix(c(summary(Val_w),summary(sc_w),summary(vd_w),summary(lv_w)),
nrow = 4, ncol = 6, byrow = TRUE, dimnames = list(c("Valliant","Schonlau-Couper",
"Dever-Valliant","Lee -Valliant"),c("Min.","1st Qu.","Median ","Mean","3rd Qu.",
"Max.")))
```

Nella tabella 4.2 vengono riportate le diverse distribuzioni dei pesi.

Pesi	Min.	Q1	Median	Mean	Q3	Max.
Valliant	1.696	1.883	2.012	2.000	2.111	2.360
Sch-Cou	0.6963	0.8835	1.0117	0.9998	1.1115	1.3604
Dev-Val	1.780	1.916	2.016	2.000	2.100	2.225
Lee -Val	0.8037	0.8846	0.9557	1.000000	1.1250	1.2787

Tabella 4.2: Distribuzione dei pesi

4.2.3 Stima delle proporzioni e dei totali

Inseriamo i differenti set di pesi correttivi nei relativi stimatori di proporzioni e di totali.

```
#stime con pesi valliant
total_estimation(sampleNP, Val_w, c("vote_gen"), 50000)
total_estimation(sampleNP, Val_w, c("vote_pens"), 50000)
total_estimation(sampleNP, Val_w, c("vote_pir"), 50000)
mean_estimation(sampleNP, Val_w, c("vote_gen"))
mean_estimation(sampleNP, Val_w, c("vote_pens"))
mean_estimation(sampleNP, Val_w, c("vote_pir"))
```

Nella tabelle 4.3 e 4.4 vengono riportate le stime ottenute. La proporzione delle preferenze di voto è stata calcolata anche senza ponderazione.

Pesi	Gen	Pens	Pir
Valliant	4968.852	17778.04	19909.36
Scho-Cou	5137.747	18256.19	19618.66
Dev-Val	4953.108	17728.92	19951.29
Lee -Val	5125.277	17958.60	19809.38

Tabella 4.3: Stime Totali delle preferenze di voto

Pesi	Gen	Pens	Pir
No Ponder.	0.09600000	0.3460000	0.4040000
Valliant	0.09937705	0.3555607	0.3981873
Scho-Cou	0.10275494	0.3651238	0.3923731
Dev-Val	0.09906215	0.3545785	0.3990258
Lee -Val	0.10250553	0.3591720	0.3961877

Tabella 4.4: Stime Proporzioni delle preferenze di voto

4.2.4 PSA e Calibrazione

In questa ultima sezione stimiamo proporzioni e totali applicando il Propensity Scoring combinato alla calibrazione. Calcoliamo i pesi correttivi come in (3.6), quindi li inseriamo come pesi base in una procedura di calibrazione ottenendo un aggiustamento capace di "gonfiare" il campione alla popolazione.

Nelle tabelle 4.5 e 4.6 si possono osservare le stime ottenute dall'impiego combinato di PSA e calibrazione , quindi confrontarle con quelle ottenute con il solo PSA o senza nessuna correzione.

	Gen	Pens	Pir
No Ponder.	0.09600000	0.3460000	0.4040000
PSA sc	0.10275494	0.3651238	0.3923731
PSA cal	0.09824163	0.3726149	0.3905399

Tabella 4.5: Stime Proporzioni preferenze di voto

	Gen	Pens	Pir
PSA sc	5137.747	18256.19	19618.66
PSA Cal	4912.081	18630.75	19526.99

Tabella 4.6: Stime totali preferenze di voto

Capitolo 5

Conclusioni

In letteratura si trovano molteplici proposte per l'inferenza su popolazioni finite da campioni non probabilistici. La possibilità di accedere ad elevate quantità di dati tramite i panel online o altre risorse della rete (*big data*) ha portato ad una proliferazione di trattazioni sui metodi per correggere gli stimatori migliorando la precisione delle stime. Se gli approcci di ponderazione impiegano informazioni ausiliarie per bilanciare il campione riducendo i potenziali bias di selezione, quelli predittivi le utilizzano per adattare modelli in grado di predire i valori delle variabili d'analisi delle unità non appartenenti al campione. In questo elaborato si è cercato di offrire una descrizione generale, seppur non esaustiva, dell'argomento. Ci si è poi soffermati sulla ponderazione del punteggio di propensione mostrandone un'applicazione pratica. Per concludere, nonostante il ricorso ad appositi metodi di stima, nel trarre generalizzazioni da un campione non probabilistico alla popolazione target occorre sempre tenerne presente i rischi. Non si può sapere a priori quanto un campione sia rappresentativo della popolazione target, nè quanto le stime dei parametri di interesse si discostino dai valori reali dello stesso. La corretta definizione dei modelli gioca un ruolo centrale e modelli mal spe-

cificati producono stime non accurate. Così come la mancata inclusione di informazioni ausiliarie invece fondamentali. Emerge dunque la necessità di un impianto teorico di riferimento in grado di giustificare la validità delle stime ottenute.

Bibliografia

Bowley, A. L. (1906). Address to the economic science and statistics section of the british association for the advancement of science, york, 1906. *Journal of the Royal Statistical Society*, 69(3):540–558.

Chen, J. K. T., Valliant, R. L., and Elliott, M. R. (2019). Calibrating non-probability surveys to estimated control totals using lasso, with an application to political polling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(3):657–681.

Comer, P. (2019). Sampling in the digital age. https://luc.id/blog/sampling-in-the-digital-age/.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., et al. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, 8(1):4–36.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418):376–382.

DiSogra, C., Cobb, C., Chan, E., and Dennis, J. M. (2011). Calibrating non-probability internet samples with probability samples using early adopter characteristics. In *Joint Statistical Meetings (JSM), Survey Research Methods*, pages 4501–4515.

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2):249–264.

Fahimi, M., Barlas, F. M., Thomas, R. K., and Buttermore, N. (2015). Scientific surveys based on incomplete sampling frames and high rates of nonresponse. *Survey Practice*, 8(5):1–11.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

Lee, S. and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37(3):319–343.

Martín, L. C. and Martín, M. L. C. (2020). Package 'nonprobest'. https://cran.mirror.garr.it/CRAN/web/packages/NonProbEst/NonProbEst.pdf.

Mercer, A. W., Kreuter, F., Keeter, S., and Stuart, E. A. (2017). Theory and practice in nonprobability surveys: parallels between causal inference and survey inference. *Public Opinion Quarterly*, 81(S1):250–271.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625.

Nordhausen, K. (2014). An introduction to statistical learning—with applications in r by gareth james, daniela witten, trevor hastie & robert tibshirani.

Rivers, D. (2007). Sampling for web surveys. In *Joint Statistical Meetings*, volume 4.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rueda, M. d. M. and Ferri-García, R. (2018). Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys.

Schonlau, M. and Couper, M. P. (2017). Options for conducting web surveys. *Statistical Science*, 32(2):279–292.

Valliant, R. (2020). Comparing alternatives for estimation from non-probability samples. *Journal of Survey Statistics and Methodology*, 8(2):231–263.

Valliant, R. and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40(1):105–137.

Valliant, R., Dever, J. A., and Kreuter, F. (2013). *Practical tools for designing and weighting survey samples*, volume 1. Springer.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991.

Watson, S. K. and Elliot, M. (2016). Entropy balancing: a maximum-entropy reweighting scheme to adjust for coverage error. *Quality & quantity*, 50(4):1781–1797.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453):185–193.