UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI "M. FANNO"

DIPARTIMENTO DI SCIENZE STATISTICHE

CORSO DI LAUREA IN ECONOMIA

PROVA FINALE

"INFERENZA PER CAMPIONI NON PROBABILISTICI NELL'ERA DEI BIG DATA"

RELATORE:

PROF.SSA BISAGLIA LUISA

LAUREANDO: MARINELLI ANDREA

MATRICOLA N. 1222880

ANNO ACCADEMICO 2021 – 2022

Dichiaro di aver preso visione del "Regolamento antiplagio" approvato dal Consiglio del Dipartimento di Scienze Economiche e Aziendali e, consapevole delle conseguenze derivanti da dichiarazioni mendaci, dichiaro che il presente lavoro non è già stato sottoposto, in tutto o in parte, per il conseguimento di un titolo accademico in altre Università italiane o straniere. Dichiaro inoltre che tutte le fonti utilizzate per la realizzazione del presente lavoro, inclusi i materiali digitali, sono state correttamente citate nel corpo del testo e nella sezione 'Riferimenti bibliografici'.

I hereby declare that I have read and understood the "Anti-plagiarism rules and regulations" approved by the Council of the Department of Economics and Management and I am aware of the consequences of making false statements. I declare that this piece of work has not been previously submitted – either fully or partially – for fulfilling the requirements of an academic degree, whether in Italy or abroad. Furthermore, I declare that the references used for this work – including the digital materials – have been appropriately cited and acknowledged in the text and in the section 'References'.

Indice

1 Introduzione

Viviamo in un mondo sempre più denso di informazioni e nel quale in ogni momento vengono generate enormi quantità di dati, che stanno aprendo nuove opportunità per il mondo della ricerca. Le fonti di questi dati sono molte e varie. Esistono scienze ad alta intensità di dati come l'astronomia, la genomica, la fisica sperimentale delle particelle e l'oceanografia (Frické, 2014). Inoltre, molti dati vengono generati dal nostro comportamento e dall'interazione con computer, smartphone, transazioni elettroniche o digitali, servizi di localizzazione e così via. La società è costantemente interconnessa, il traffico di informazioni sul web ha raggiunto dimensioni mai viste prima e presenta un tasso di crescita straordinario. La recente pandemia di Covid-19 vi ha contribuito in modo sostanziale, incentivando moltissimo l'utilizzo di tecnologie digitali e del web. Secondo l'IDC (International Data Corporation, 2021) la quantità di dati creati nel mondo ha raggiunto una crescita insolitamente elevata nel 2020 arrivando a registrare 64,2 Zettabytes (1 Zb = 1012 Gb) di dati e si prevede che arrivi a toccare i 180 Zb nel 2025, che corrispondono a circa 23.200 gigabyte all'anno per ogni essere umano sulla Terra. Nell'era dei Big Data stanno dunque nascendo molte opportunità per la ricerca e per tutti gli attori del sistema economico che fanno fortemente affidamento sui dati per lo svolgimento della propria attività. Allo stesso tempo si presentano nuove sfide da affrontare: in primo luogo, una difficoltà sta nel dover gestire dati caratterizzati da alti Volumi, Velocità e Varietà. Per questo si è reso necessario l'impiego di tecniche di elaborazione dei dati automatizzate, come tecniche di *machine learning* che vengono spesso utilizzate quando si lavora con Big Data; in secondo luogo, poiché tali dati sono assimilabili a grandi campioni non probabilistici potrebbero non essere adatti per l'inferenza su una popolazione, se non vengono apportati degli aggiustamenti tali da poter giustificare la proiezione dei risultati ottenuti dal campione sull'intera popolazione. Questa seconda considerazione rappresenta il tema centrale sul quale sarà sviluppato l'elaborato. Per molti anni i campioni non probabilistici sono stati ritenuti poco affidabili per fare inferenza su una popolazione, anche se negli ultimi anni stanno progressivamente venendo rivalutati. Questo è dovuto in gran parte alla possibilità di raccogliere economicamente grandi volumi di dati. C'è chi sostiene che con così tanti dati che vengono generati, le indagini non abbiano più alcun valore. A tal proposito è possibile citare un articolo di Anderson (2008) nel Wired Magazine che è stato provocatoriamente intitolato "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", che mette in evidenza come la nascita dei Big Data abbia

sollevato molti dibattiti nei quali c'è chi sostiene che il futuro della ricerca sarà dominato da computer e algoritmi induttivi che sostituiranno progressivamente il metodo scientifico tradizionale. Rosling (2010) spiega che "Il diluvio di dati... ci sta conducendo ad una sempre maggiore comprensione della vita sulla Terra e dell'Universo... [e può] trasformare il processo di scoperta scientifica. Più dati ci sono, più scoperte possono essere fatte." D'altra parte, la quantità non va confusa con la qualità e, se certamente i Big Data rappresentano uno strumento interessante, che nasconde molte opportunità per la ricerca, non andrebbe considerato come uno strumento sostitutivo alle più tradizionali tecniche di indagine statistica, quanto piuttosto complementare (Couper, 2013). Le radici di tale dibattito risiedono nel passato della statistica. Nella storia recente della scienza, la statistica ha occupato una posizione privilegiata nell'apprendere dai dati e nel giustificare epistemicamente il ragionamento induttivo. Tuttavia, negli anni '20, i fondatori della moderna scienza statistica, per esempio Ronald A. Fisher, dichiararono che la statistica poteva studiare cause ed effetti (cioè l'inferenza causale) utilizzando i dati di esperimenti randomizzati, ma non di studi osservazionali (Raita et al., 2021). Da questo momento, il campionamento probabilistico si è affermato fino a diventare un modello di riferimento per la conduzione di indagini statistiche. Il campionamento di tipo non probabilistico è stato invece screditato in seguito ad alcuni noti fallimenti, legati soprattutto a sondaggi che non sono riusciti a prevedere correttamente i risultati elettorali, come nel caso delle elezioni statunitensi del 1936 dove venne erroneamente predetta la vittoria di Alf Landon invece che di Franklin Roosvlet (Elliot e Valliant, 2017) o nelle elezioni parlamentari britanniche del 2015 (Cowling, 2015). Tuttavia, negli ultimi anni sta crescendo sempre di più l'incentivo ad utilizzare campioni di tipo non probabilistico per due principali ragioni: da una parte, la selezione di un campione probabilistico non garantisce che le unità estratte forniscano una buona base per l'inferenza su una popolazione. Infatti, in molti tipi di sondaggi i tassi di risposta sono diminuiti drasticamente, mettendo in dubbio quanto bene questi campioni rappresentino la popolazione. Dall'altra parte è possibile ottenere grandi quantità di dati in modo rapido ed economico grazie alla diffusione delle tecnologie dell'informazione, di internet, dei social network, dell'*Internet of Things* e molto altro. Il problema si sposta quindi sul capire come rendere questi dati rappresentativi della popolazione e bisogna chiedersi se possano costituire una base solida sulla quale fare inferenza. Questo elaborato prende in esame il tema di come sia possibile utilizzare dati raccolti con tecniche di campionamento non probabilistico al fine di fare inferenza su una popolazione. Gli argomenti saranno presentati nel seguente ordine: nel capitolo

2 si approfondisce la differenza tra campionamento probabilistico e non probabilistico. L'analisi si concentra quindi sul campionamento non probabilistico, di cui viene proposta una classificazione e vengono discussi i principali problemi ad esso connessi. Nel capitolo 3 vengono illustrati due diversi approcci proposti come soluzione per l'utilizzo di campioni non probabilistici per fare inferenza su una popolazione finita. Il primo approccio è quello della Quasi-randomizzazione, il secondo è quello della Superpopolazione. Infine, nel capitolo 4, vengono illustrate alcune applicazioni reali di metodi di campionamento non probabilistico. Il capitolo 5 conclude l'elaborato proponendo alcune considerazioni finali sul campionamento non probabilistico, analizzandone vantaggi e svantaggi, anche alla luce dei cambiamenti in atto nell'era dei Big Data.

2 Campionamento probabilistico e non probabilistico

In questo capitolo viene approfondita la differenza tra il campionamento probabilistico e il campionamento non probabilistico, viene fornita una classificazione dei principali metodi di campionamento non probabilistico e vengono analizzati i principali problemi e rischi ad esso associati. La rilevazione campionaria è lo strumento nella prassi più utilizzato per indagare su specifiche caratteristiche di una popolazione. Questo consente di ridurre costi e tempi, analizzando solo un sottoinsieme della popolazione, per poi estendere le informazioni trovate all'intera popolazione, se si presentano le condizioni opportune. La validità dell'inferenza è determinata da quanto bene il campione rappresenta l'intera popolazione. Quando il campione è poco rappresentativo della popolazione totale, i risultati ottenuti dall'indagine possono essere distorti e quindi non validi in generale. Molta attenzione va quindi posta al modo in cui il campione è scelto, ossia alla regola di selezione delle unità che formano il campione stesso (Conti e Marella, 2012). In particolare, la regola di selezione può essere di tipo probabilistico o non probabilistico.

Nel *campionamento probabilistico* viene definito un preciso disegno campionario, ovvero chi progetta la rilevazione fissa uno schema probabilistico di selezione delle unità. Tutte le unità hanno una probabilità di essere incluse nel campione che è nota a priori e diversa da zero. Esistono diverse tecniche di campionamento probabilistico che possono essere utilizzate per estrarre un campione. Un esempio è il campionamento casuale semplice (CCS), nel quale tutte le unità hanno la medesima probabilità di essere estratte. Tuttavia, è possibile utilizzare anche tecniche più complesse che utilizzano strati, grappoli e stadi. Il vantaggio del campionamento probabilistico è rappresentato dalla possibilità di quantificare il rischio di selezionare un campione poco rappresentativo e ottenere stime distorte delle variabili d'interesse (Vehovar et al., 2016). Questo è possibile perché la selezione delle unità avviene secondo una procedura predefinita e casualizzata controllata dallo statistico. La condizione di casualità è una condizione necessaria per poter ricondurre alla popolazione, attraverso la teoria della probabilità, i risultati ottenuti dal campione con un certo grado di affidabilità (Conti e Marella, 2012).

Nel *campionamento non probabilistico* le probabilità di inclusione delle unità nel campione non sono note e non necessariamente sono diverse da zero. Il principale vantaggio del campionamento non probabilistico è che risulta ben più economico del campionamento probabilistico dal momento che consente di ridurre i costi e i tempi associati alla raccolta delle informazioni.

2.1 Campionamento non probabilistico

Una classificazione offerta dall'AAPOR (American Association for Public Opinion Research) distingue tre principali metodi di campionamento non probabilistico che includono il campionamento di convenienza, il campionamento per corrispondenza, il campionamento a rete (Baker et al., 2013).

Il *campionamento di convenienza* è una tecnica di campionamento non probabilistico in cui i soggetti vengono selezionati per la loro comoda accessibilità e la vicinanza al ricercatore. L'intervistatore potrebbe quindi scegliere di condurre le interviste in zone della città a lui più facilmente accessibili, intervistando le persone più disponibili ed evitando le zone periferiche e più scomode da raggiungere. Bisogna tenere presente che la soggettività del criterio di selezione delle unità campionarie potrebbe generare un campione poco rappresentativo della popolazione. Esempi di campionamento di convenienza sono i campioni di volontari e i sempre più diffusi opt-in web panel, ovvero campioni di volontari che vengono reclutati online quando accedono a determinati siti web.

Nel *campionamento per corrispondenza* i membri del campione sono selezionati in modo tale da riprodurre nel campione la struttura della popolazione. Lo scopo è fare in modo che il campione si distribuisca come la popolazione rispetto ad alcune variabili strutturali, legate alla variabile d'interesse (ad esempio: sesso, età, area geografica, ecc). Un esempio è il campionamento per quote dove la popolazione viene suddivisa in gruppi omogenei sulla base di variabili tipicamente demografiche. Dopo aver ricavato il peso percentuale di ogni classe, il totale delle unità nel campione viene suddiviso tra le classi in modo da rispecchiare le proporzioni esistenti nella popolazione (Conti e Marella, 2012).

Il *campionamento a rete* è un tipo di campionamento nel quale ai membri appartenenti alla popolazione d'interesse viene richiesto di individuare altre persone a loro collegate che appartengono alla stessa popolazione. Un esempio è il campionamento a valanga o palla di neve, molto utilizzato nelle indagini sociologiche che affrontano temi delicati come omosessualità, consumo di droga o alcool, o nelle indagini su popolazioni rare e quindi più difficile da raggiungere. In questo caso si utilizzano le reti relazionali (sociali, culturali, politiche) di un gruppo di persone inizialmente contattate per risalire velocemente ad altre con le medesime caratteristiche. Un altro esempio sono i *respondent driven sampling* (RDS), nei quali i membri sono chiamati a riferire quante altre persone appartenenti alla medesima popolazione conoscono.

2.2 Problemi dei campioni non probabilistici

L'utilizzo di campioni non probabilistici può comportare problemi di varia natura quando il fine è l'inferenza su una specifica popolazione. Il motivo risiede nel fatto che il campionatore non esercita un controllo diretto sul meccanismo di selezione delle unità campionarie e di conseguenza le stime ottenute potrebbero essere soggette a delle distorsioni.

Uno dei problemi più frequenti è la *distorsione da selezione*. In questo caso, il campione non è rappresentativo della popolazione d'interesse e quindi le stime ottenute non possono essere proiettate sull'intera popolazione. Ad esempio, quando si fa ricorso a web panel e i volontari vengono reclutati tramite l'accesso a un sito web, è necessario tenere presente che (i) non tutti hanno un accesso ad internet, (ii) non tutte le persone con accesso ad internet frequentano quel particolare sito. Di conseguenza ci sarà una parte della popolazione che ha una probabilità di essere inclusa nel campione pari a zero. Inoltre, alcune fasce della popolazione potrebbero essere sottorappresentate nel campione. Ad esempio, la probabilità di non avere accesso ad internet è molto superiore per gli individui con un reddito basso. Quindi il campione ottenuto tramite un opt-in web panel potrebbe sottorappresentare le fasce di reddito più basse.

Un secondo problema è la *non risposta*. Molti panel online prevedono una registrazione a due passaggi nella quale viene inviata un'e-mail al volontario per perfezionare la registrazione e sottoporsi al sondaggio. Tuttavia, accade spesso che il soggetto non risponda all'email e la registrazione al panel non venga ultimata. Una ricerca ha mostrato che solo il 6% delle persone porta a termine la registrazione (Alvarez et al., 2003).

Un terzo problema è quello dell'*abbandono* prima della conclusione del sondaggio. La causa potrebbe essere la perdita d'interesse, l'eccessiva lunghezza delle domande, il sovraccarico di richieste di partecipazione a diversi panel.

Infine, un altro problema molto frequente è rappresentato dagli *errori di misurazione*. Questi possono essere causati da una scorretta progettazione e implementazione del questionario tale da fornire dei dati di bassa qualità, oppure per un deliberato inserimento di risposte errate da parte del partecipante. Questo può accadere quando il sondaggio offre una ricompensa a chi lo completa. Spesso chi vi prende parte ha il solo scopo di ottenere la remunerazione e risponde alle domande in modo approssimativo o totalmente casuale.

3 Soluzioni proposte per l'utilizzo di campioni non probabilistici

Le complicazioni legate all'utilizzo di campioni non probabilistici devono essere valutate attentamente ed è necessario utilizzare adeguati metodi di aggiustamento del campione se il fine della ricerca è quello di proiettare i risultati sull'intera popolazione. Questo è possibile solo se si presentano alcune condizioni. Per procedere con l'analisi in modo ordinato si presenta un modello per l'inferenza da popolazioni finite utilizzato da Smith (1983). Il modello considera una popolazione finita U costituita da N unità, sulla quale si vuole misurare il valore della variabile casuale Y e di cui sono note a priori delle informazioni che indichiamo con X. Per l'indagine si estrae un campione e l'insieme delle unità selezionate è rappresentato dal vettore δ_s, mentre le restanti unità della popolazione si indicano con $\delta_{\bar{s}}$. La probabilità che uno specifico campione venga estratto, dati i valori delle variabili Y, X e un parametro ignoto Φ, è rappresentato dalla funzione di densità:

$$f(\delta_s|Y, X; \Phi) \tag{1}$$

che indentifica il meccanismo di selezione. La funzione di densità condizionale di Y dato X e il parametro ignoto Θ associato ad Y è:

$$f(Y|X; \Theta). \tag{2}$$

Mettendo insieme le due equazioni si ottiene la probabilità congiunta di Y e δ_s che è possibile scrivere come:

$$f(Y, \delta_s|X; \Theta, \Phi) = f(Y|X; \Theta)f(\delta_s|Y, X; \Phi) \tag{3}$$

Nel campionamento probabilistico l'inferenza si basa sull'assunzione che il ricorso alla casualizzazione nel meccanismo di selezione produca campioni rappresentativi della popolazione (Smith, 1983) e quindi si focalizza sulla distribuzione di randomizzazione $f(\delta_s|X)$. Nel campionamento non probabilistico, invece, potrebbero presentarsi distorsioni da selezione e la distribuzione di δ_s può dipendere anche dalla variabile Y e da un parametro ignoto Φ. In questo caso l'inferenza è model-based ed esistono due principali approcci proposti in Elliot e Valliant (2017). Il primo approccio è la Quasi-randomizzazione che lavora sulla probabilità di inclusione di un'unità della popolazione nel campione, quindi sviluppa un modello per $f(\delta_s|Y, X; \Phi)$. Il secondo approccio è quello della Superpopolazione che lavora sulla variabile analitica Y e sviluppa un modello per $f(Y|X; \Theta)$. Entrambi

gli approcci risultano validi e possono essere utilizzati per produrre stime di statistiche descrittive e analitiche sul fenomeno di interesse.

3.1 Quasi-randomizzazione

Nell'approccio della *quasi-randomizzazione* si stimano le probabilità di inclusione nel campione delle unità della popolazione. In questo modo è possibile eliminare la distorsione dovuta alla selezione e trattare il campione come un campione probabilistico. Nella pratica, si assume che la probabilità che un'unità venga selezionata nel campione non dipenda da Y e si produce una stima per $f(\delta_s|X;\Phi)$. Tuttavia, non esiste un modo per provare che tale condizione sia effettivamente valida. Di seguito viene riportato un esempio tratto da Elliot e Valliant (2017) per comprendere meglio come questo approccio possa essere messo in pratica. Si consideri il caso di un opt-in web panel in cui vengono reclutati dei volontari per un sondaggio tramite l'accesso a un sito web e tra questi viene estratto un campione a cui sarà effettivamente somministrato il sondaggio. La probabilità che un individuo sia selezionato dipende da diversi fattori: prima di tutto la persona deve disporre di un accesso ad internet; poi deve essere un utente del sito; inoltre deve proporsi come volontario per il sondaggio; ed infine deve essere selezionato. La probabilità $P(x_i)$ che l'i-esima unità della popolazione sia inclusa nel campione può essere espressa come:

$$P(x_i) = P(i \in I|x_i)P(i \in V|I, x_i)P(i \in s_V|V, I, x_i)P(i \in s_{VR}|s_V, V, I, x_i)$$

dove:

x_i = vettore di variabili casuali riferite all'i-esima unità della popolazione
I = insieme delle persone che dispone di un accesso ad internet
V = insieme delle persone che si propongono come volontari
s_V = insieme delle persone che vengono scelte e inserite nel campione
s_{VR} = insieme delle persone che rispondono alle domande

Queste probabilità possono essere stimate con l'aiuto di *indagini di riferimento*, ovvero con dati già disponibili nel database del ricercatore, raccolti attraverso indagini di tipo probabilistico ben eseguite e sulle quali si può fare affidamento. In questo caso sarebbe utile avere a disposizione dati sull'accesso ad internet della popolazione sulla quale si sta svolgendo l'indagine e sugli utenti che frequentano il sito web sul quale è avvenuto il reclutamento. L'approccio statistico è quello di combinare il campione di riferimento e il campione di volontari e adattare un modello per prevedere la probabilità di essere

nel campione non probabilistico (Elliot e Valliant, 2017). Affinché questo sia possibile è importante che l'indagine di riferimento utilizzi le stesse variabili x_i dell'indagine in corso sui volontari.

In alternativa è possibile utilizzare la tecnica del *sample matching* per ridurre le distorsioni da selezione. Questa consiste nel controllare che la distribuzione delle unità rispetto ad alcune variabili sia la stessa nel campione e nella popolazione. Il *campionamento per quota* è un esempio di utilizzo di questa strategia.

3.2 Superpopolazione

Nell'approccio della superpopolazione viene sviluppato un modello per la variabile Y al fine di proiettare i risultati ottenuti dal campione sull'intera popolazione. Tale approccio può essere applicato anche su campioni probabilistici e richiede che si possa ignorare il meccanismo di selezione delle unità campionarie, ovvero che $f(\delta_s|Y, X; \Phi) = f(\delta_s|X; \Phi)$. L'idea generale sulla quale si basa questo approccio è che se il campione è rappresentativo dell'intera popolazione, allora la distribuzione di Y rispetto alle variabili indipendenti X sarà la stessa sia sul campione che sull'intera popolazione. Di conseguenza stimando un modello per il campione, questo sarà valido per tutta la popolazione. Il modello è del tipo:

$$E(y_i|x_i) = x_i^T \beta \tag{4}$$

Per comprendere meglio l'utilizzo di questo approccio si consideri il caso in cui si vuole stimare il totale, t_1, della variabile Y sulla popolazione. t_1 può essere espresso come somma del totale delle y_i del campione e delle $\hat{y}_i$ predette per le unità non incluse nel campione. Per calcolare i valori predetti di $\hat{y}_i$ si utilizza quindi il modello sviluppato (4), ottenendo:

$$\hat{t}_1 = \sum_{i \in s} y_i + \sum_{i \in \bar{s}} \hat{y}_i \tag{5}$$

$$= \sum_{i \in s} y_i + \sum_{i \in \bar{s}} (t_{Ux} - t_{sx})^T \hat{\beta}, \tag{6}$$

dove t_{Ux} è la somma delle x_i delle unità dell'intera popolazione e t_{sx} è la somma delle x_i delle unità campionarie. Inoltre, quando il campione estratto è particolarmente piccolo, è possibile stimare il totale utilizzando i valori predetti $\hat{y}_i$ per tutte le unità della popolazione:

$$\hat{t}_2 = \sum_{i \in U} \hat{y}_i = t_{Ux}^T \hat{\beta} \tag{7}$$

In entrambi i casi è possibile riformulare l'equazione come media pesata dei valori y_i, ottenendo:

$$\hat{t} = \sum_{i \in s} w_i y_i \tag{8}$$

dove il peso w_i dipende solo dalle variabili esplicative X.
Tecniche utilizzate per calibrare il campione come la Post-stratificazione e il Raking rientrano in questo approccio. Il modello di regressione, in questi casi, considera anche le interazioni tra le variabili esplicative.

Nella *Post-stratificazione* la popolazione viene partizionata in H strati, ovvero in sottoinsiemi, detti appunto post-strati, sulla base di una o più variabili esplicative X, generalmente di tipo demografico (come sesso, reddito, regione di provenienza) di cui si conosce la distribuzione sull'intera popolazione. Dopo aver stimato la statistica d'interesse per ogni strato individuato, è possibile stimare $\overline{Y}$ come media ponderata dei valori predetti:

$$\hat{\overline{Y}} = \sum_{h=1}^{H} P_h \hat{\mu}_h, \tag{9}$$

dove P_h indica la proporzione dell'h-esimo strato sulla popolazione e $\hat{\mu}_h$ indica il valore predetto della statistica d'interesse nell'h-esimo strato.
Si consideri, ad esempio, il caso (tratto da Nicolini et al., 2013) di un'indagine in cui si vuole stimare il totale degli individui affetti da una malattia cronica. Si utilizzano le informazioni disponibili sulla popolazione per definire $H = 6$ post-strati, dati dall'incrocio della variabile sesso (M = maschio, F = femmina) e 3 classi di età. Il modello sviluppato utilizza delle variabili ausiliarie binarie x_{ih}, con $h = 1, 2, ..., H$, definite sull'intera popolazione e tali che

$$x_{ih} = \begin{cases} 1 & \text{se } i \in U_h \\ 0 & \text{altrimenti} \end{cases} \quad (h = 1, 2, ..., H),$$

dove $U_1, U_2, ..., U_H$ indicano i post-strati che individuano una partizione della popolazione. Dunque, il totale degli individui affetti da malattia nella popolazione può essere stimato sommando i totali di ogni post-strato aggiustati con un appropriato peso.

$$\hat{Y} = \sum_{h=1}^{H} \hat{Y}_h w_h,$$

dove w_h è il coefficiente di aggiustamento dell'h-esimo post-strato, tale che:

- $w_h > 1$ se l'h-esimo post-strato è sottorappresentato

- $w_h < 1$ se l'h-esimo post-strato è sovrarappresentato

La post-stratificazione è quindi una tecnica di stima che riproduce i vantaggi della stratificazione rispetto ad alcune variabili ausiliarie, senza che questa venga effettuata a livello di estrazione del campione (Nicolini et al., 2013). Questo approccio può essere esteso aggiungendo modelli di regressione gerarchica, per migliorarne l'adattamento. Infine, un'alternativa valida che può essere utilizzata è l'approccio Bayesiano all'inferenza su popolazioni finite che tratta le unità non campionarie come dati mancanti (Elliot e Valliant, 2017).

4 Applicazioni

In questo capitolo vengono presentati alcuni casi studio nei quali sono state impiegate tecniche di aggiustamento del campione non probabilistico per ridurre le distorsioni dovute alla selezione. Il primo caso, tratto da Chen (2016), rientra nell'approccio della Superpopolazione e consiste in un'applicazione del metodo LASSO (Least Absolute Shrinkage and Selection Operator) per ottenere delle stime corrette dei risultati delle elezioni statunitensi del 2014. Il secondo caso, tratto da Gazala (2018), combina la teoria statistica tradizionale con moderne tecniche di *machine learning* per sviluppare una metodologia di post-stratificazione, al fine di calibrare il campione non probabilistico. L'applicazione rientra nell'approccio della Superpopolazione e ha il fine di indagare le determinanti del livello di soddisfazione dei pazienti sulle visite mediche effettuate. Nel terzo caso esaminato, tratto da Berzofsky et al. (2009), viene presentato un approccio al disegno campionario che combina il campionamento probabilistico con il campionamento non probabilistico, in particolare il campionamento per quote.

4.1 Calibrare un campione non probabilistico con il LASSO

In questa sezione si presenta un caso, tratto da Chen (2016), di applicazione di una moderna tecnica di calibrazione del campione, utilizzata per prevedere correttamente i risultati delle elezioni statunitensi del 2014. Nell'ambito delle tecniche di calibrazione, il campione non probabilistico viene aggiustato usando dei pesi per calibrarlo, in modo che la distribuzione delle variabili demografiche sul campione aggiustato, sia la stessa della popolazione. Un esempio è la post-stratificazione presentata nel capitolo precedente. Le tecniche di calibrazione del campione necessitano di un campione di riferimento di tipo probabilistico su cui fare affidamento per aggiustare il campione non probabilistico. Nella pratica si stima un unico insieme di pesi per correggere il campione. Tuttavia, quando i campioni presentano distribuzioni sbilanciate rispetto a diverse variabili, un unico insieme di pesi può non essere sufficiente. Per aumentare la qualità delle stime è possibile utilizzare il metodo della calibrazione assistita da modello. In questo caso si entra nel contesto della Superpopolazione e si assume una relazione causale tra la variabile d'interesse Y e le variabili esplicative X tale che sia possibile stimare un modello del tipo $f(Y|X;\Theta)$. Sulla base di questo modello è possibile determinare i pesi per ottenere delle stime aggiustate delle statistiche d'interesse. Nel caso analizzato viene utilizzato come modello il LASSO (Least Absolute Shrinkage and Selection Operator). L'uso del LASSO (e in generale delle

tecniche di regolarizzazione) consente di risolvere eventuali problemi di multicollinearità, di ridurre la complessità del modello e di selezionare quindi un modello che bilanci capacità di adattamento e semplicità, includendo solo quelle variabili più influenti (Torelli, 2020). Il LASSO è anche conosciuto come metodo di regressione penalizzato, in quanto penalizza la funzione di verosimiglianza con un termine λ_n che dipende dalla complessità del modello:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i \in s_A} (y_i - x_i^T \beta)^2 + \lambda_n \sum_{j=1}^{p-1} |\beta_j| \right) \tag{10}$$

Come si può notare, se il parametro di penalizzazione $\lambda_n \approx 0$, i coefficienti tendono a quelli ottenuti con il metodo dei minimi quadrati ordinari. I passi da seguire per calibrare il campione e stimare i totali prevedono: il calcolo dei parametri β per la stima del modello di regressione $f(Y|X;\Theta)$, quindi il calcolo dei pesi w^{LASSO} per stimare i totali come $\hat{T}_y^{LASSO} = (w^{LASSO})^T y$.

Questo framework è stato utilizzato in Chen (2016) per ottenere delle stime dei risultati delle elezioni statunitensi del 2014. I sondaggi elettorali rappresentano una delle principali applicazioni delle indagini statistiche. Poiché l'arco di tempo a disposizione per raccogliere le informazioni è generalmente breve, i campioni non probabilistici costituiscono un'alternativa valida e sempre più diffusa, essendo considerevolmente più economici da raccogliere rispetto ai campioni probabilistici. Dopo aver estratto il campione è poi possibile fare affidamento su diverse tecniche per aggiustarlo e renderlo rappresentativo della popolazione. Nel caso delle elezioni statunitensi, dove i due principali partiti sono rappresentati da Democratici e Repubblicani, uno degli indicatori più utilizzati per valutare i risultati dell'indagine è lo *spread*, ovvero il differenziale tra le percentuali di voto del primo partito rispetto al secondo. In particolare, l'indagine in questione è stata condotta calcolando per ogni stato lo spread tra Democratici e Repubblicani, che può essere indicato con S_{D-R}. Il campione utilizzato per l'analisi è stato rilevato da un'indagine condotta da SurveyMonkey tramite web panel, a cui hanno risposto 85,668 persone che avrebbero votato. Come benchmark per aggiustare il campione sono stati utilizzati i dati relativi ad un'indagine telefonica di tipo probabilistico condotta da Pew Research Center (http://www.pewresearch.org) a cui hanno risposto circa 3000 persone. Nonostante il campione probabilistico fosse di dimensioni notevolmente inferiori rispetto al primo, si è rivelato molto efficiente per calibrarlo. Tuttavia, non è stato possibile includere nel campione finale quegli stati con un livello di rappresentatività nel campione probabilistico

troppo basso. Per entrambi i campioni sono state misurate variabili di tipo demografico, relative alla religione, alle opinioni politiche e al partito supportato, al fine di spiegare la decisione del voto finale. Gli stati sono stati inoltre raggruppati in 4 categorie in base ai risultati elettorali delle elezioni più recenti, per evidenziarne l'orientamento politico, che in alcuni pareva particolarmente definito e quindi un fattore capace di influenzare parzialmente gli individui nella decisione del voto, creando un effetto regionale. Complessivamente le previsioni ottenute si sono rivelate molto soddisfacenti, soprattutto se confrontare ad altre ottenute impiegando modelli alternativi al LASSO. Questa applicazione si è rivelata particolarmente interessante nel mostrare come un campione non probabilistico possa essere aggiustato e la distorsione dovuta alla selezione possa essere ridotto in modo consistente anche utilizzando come benchmark un campione probabilistico di dimensioni considerevolmente inferiori.

4.2 Post-stratificazione per correggere le distorsioni da selezione

In questa sezione viene presentata una seconda applicazione di tecniche di aggiustamento di campioni non probabilistici, tratta da Gazala (2018). L'approccio utilizzato rientra in quello della Superpopolazione ed in particolare la tecnica impiegata per calibrare il campione è quella della Post-stratificazione, di cui si è già parlato nel paragrafo 2.2. La metodologia di stratificazione è stata sviluppata integrando la teoria statistica tradizionale con tecniche di *machine learning*. La teoria statistica fornisce il framework per condurre l'analisi in modo formale, mentre il *machine learning* è impiegato per individuare schemi ricorrenti nei dati e in particolare per individuare il miglior modo di implementare la stratificazione. Il *machine learning* venne definito da uno dei suoi pionieri, Arthur Samuel, come: "la scienza che rende i computer in grado di imparare, senza essere stati esplicitamente programmati per questo". Si tratta di software basati su algoritmi matematici che simulano ragionamenti di tipo induttivo, imparando dalle informazioni. Questo avviene attraverso il riconoscimento di pattern, ovvero delle regolarità nei dati che permettono di classificare determinate situazioni e di ricondurle a specifici esiti (Musacchio et al., 2018). Il rapido sviluppo del *machine learning* è stato determinato in primo luogo dall'avvento dei Big Data: la crescente disponibilità di enormi quantità di dati ha reso necessario l'impiego di tecniche di elaborazione dei dati automatizzate. Il *machine learning* ha generato enormi impatti sociali in una vasta gamma di applicazioni come la visione artificiale, l'elaborazione del parlato, la comprensione del linguaggio naturale, le neuroscienze, la salute e

l'*Internet of Things* (Zhou et al., 2017). Una delle applicazioni più importanti riguarda il settore medico-sanitario, nel quale rientra anche il caso preso in esame. L'obiettivo dello studio condotto in Gazala (2018) è quello di generare una conoscenza riguardo a quanto i pazienti sono soddisfatti dei trattamenti e dei servizi ricevuti presso gli studi medici che frequentano. Tali informazioni rivestono un ruolo fondamentale per individuare le aree da migliorare, monitorare gli effetti del cambiamento, definire delle politiche interne e creare un'organizzazione in grado di massimizzare i benefici per i propri pazienti. In particolare, conoscere i diversi tipi di pazienti e capire quali sono le determinanti del livello di soddisfazione, può aiutare a ridurre le differenze nel livello di trattamento tra i vari gruppi di pazienti e garantire una continuità e stabilità nel livello del servizio. Oggi è possibile effettuare tali analisi ricorrendo a indagini di tipo psicometrico, che vengono utilizzare per ottenere informazioni sul grado di soddisfazione dei pazienti rispetto al trattamento ricevuto, e sono sempre più popolari nel settore medico-sanitario. Per raggiungere l'obiettivo di sviluppare un sistema sanitario equo in grado di rispondere alle esigenze dei diversi gruppi della popolazione può risultare utile l'impiego di tecniche di stratificazione. Combinando tecniche statistiche più tradizionali con il machine learning si può ottenere una metodologia di stratificazione per stimare le valutazioni e quindi le impressioni dei vari gruppi di pazienti. Per questa applicazione è stato utilizzato il dataset IPQ ("Improving Practice Questionnaire") relativo ad un'indagine condotta su larga scala che contiene le risposte di 2,546,182 pazienti riguardo le esperienze e il livello di soddisfazione percepito nelle visite mediche effettuate (Greco et al., 2003). Sui pazienti sono state misurate quattro variabili sociodemografiche: il genere (2 livelli: uomo, donna), l'età (3 livelli: giovani, adulti, anziani), se la visita è stata fatta presso il proprio medico (2 livelli: si, no), e quanti anni il paziente ha visitato l'operatore sanitario (3 livelli: meno di 5 anni, tra 5-10 anni, più di 10 anni). Dalla combinazione di queste variabili si ottengono quindi 36 strati. Ad ogni individuo sono state sottoposte 27 domande, attribuibili a tre principali aree: accesso e prenotazione, competenze interpersonali degli operatori e comunicazione con il personale. Per indagare la presenza di differenze statisticamente significative tra i sottogruppi individuati sulla base delle variabili sociodemografiche, sono state utilizzati: il test ANOVA e tecniche di *machine learning* supervisionato, come il *decision tree* e la PCA (Principal Component Analisys). In particolare, applicando le tecniche di machine learning è stato possibile stratificare i dati dell'indagine in sottopopolazioni non omogenee. La stratificazione è stata realizzata a 4 livelli, partendo dalle variabili che in base alle analisi esplorative sembravano produrre un maggior guadagno informativo, ovvero le

variabili in grado di massimizzare le differenze tra le sottopopolazioni. Dopodiché il test ANOVA è stato applicato per identificare sottogruppi omogenei congiungibili in un unico sottogruppo più grande. Nello specifico è stato possibile individuare 17 sottopopolazioni omogenee. A questo punto è stato utilizzato un campione probabilistico come benchmark al fine di ottenere un campione pseudo-controllato e produrre delle stime non distorte dei parametri della popolazione. Per ogni livello di stratificazione del campione è stato possibile controllare la distribuzione delle variabili indipendenti in modo che fosse la stessa del campione probabilistico, e quindi della popolazione. Questa analisi si è rivelata particolarmente interessante nel mostrare come sia possibile combinare la teoria statistica tradizionale con tecniche di machine learning per creare una metodologia di stratificazione e fare inferenza su una popolazione partendo da un ampio campione non probabilistico.

4.3 Combinare il campionamento probabilistico e non probabilistico

In questa terza applicazione tratta da Berzofsky et al. (2009) viene presentato un approccio al disegno campionario che combina il campionamento probabilistico con il campionamento non probabilistico, in particolare il campionamento per quote. Questo caso rientra quindi nell'approccio della Quasi-randomizzazione discusso nel paragrafo (3.1). Il campionamento per quote si basa sulla stratificazione della popolazione secondo alcune variabili ritenute significative per l'indagine e sulla conoscenza delle proporzioni (quote) della popolazione negli strati. A tutti gli effetti il campionamento per quote è simile al campionamento stratificato con allocazione proporzionale. Tuttavia, in quest'ultimo si dispone degli elenchi delle unità della popolazione in ciascuno strato da cui si estrae un campione probabilistico. Al contrario nel campionamento per quote non viene compilata alcuna lista. Questo riduce notevolmente i costi dell'indagine e ne spiega l'ampio utilizzo (Nicolini et al., 2013). Nel caso in questione il campione probabilistico è utilizzato per individuare delle quote e le rispettive proporzioni sulla popolazione d'interesse. Partendo dal campione viene quindi sviluppato un modello per $f(\delta_s|X;\Phi)$, ovvero viene stimata la probabilità di un'unità di essere inclusa nel campione. Successivamente viene estratto il campione non probabilistico in modo da rispettare il numero richiesto di unità in ogni quota. In questo modo si ottiene un campione parzialmente controllato (per questo quasi-randomico) e si può beneficiare delle dimensioni maggiori del campione non probabilistico per ottenere delle stime più precise e dei costi minori ad esso associati.

Questa tecnica di campionamento assistita da modello viene utilizzata in Berzofsky et al. (2009) per un'applicazione sul programma di raccolta dei dati O*NET (Occupational Information Network). Si tratta di un programma sponsorizzato dal Dipartimento del Lavoro degli Stati Uniti, che raccoglie in un database una vasta varietà di informazioni, su circa 800 diversi ambiti occupazionali, relative a tre categorie: contesto lavorativo, attività lavorative e conoscenza. La fase di estrazione del campione si divide in due fasi: nella prima, vengono individuate le strutture lavorative; nella seconda, si selezionano i lavoratori appartenenti alla categoria ricercata. Prima di procedere all'estrazione delle unità viene definita una distribuzione campionaria, sotto forma di quote, per ogni ambito occupazionale, rispetto ad alcune variabili, quali: presenza dell'occupazione nelle varie regioni, dimensione degli stabilimenti e raggruppamenti industriali nei quali è prevista quel tipo di occupazione. Si procede con l'identificazione delle strutture e quindi con la somministrazione del questionario, finché non vengono "riempite" le quote per ogni ambito occupazionale. La raccolta dei dati viene interrotta solo quando tutte le quote sono complete. A questo punto, i dati raccolti per ciascuna quota sono impiegati per ottenere le stime delle statistiche d'interesse. La stima viene effettuata seguendo l'approccio del *sample matching*, ovvero controllando la distribuzione delle unità del campione rispetto alle variabili strutturali considerate, in modo che sia la stessa della popolazione. Questa applicazione si è rivelata particolarmente interessante nel mostrare come i costi connessi ad un'indagine possano essere consistentemente ridotti utilizzando un approccio all'inferenza *model-based*, e in particolare in questo specifico caso combinando il campionamento probabilistico con il campionamento per quote, senza ottenere delle stime distorte.

5 Conclusioni

Sebbene il campionamento probabilistico rappresenti lo standard per fare inferenza su popolazioni finite, si è visto nei capitoli precedenti come sia possibile utilizzare in alternativa approcci *model-based* come la Quasi-randomizzazione e la Superpopolazione per raggiungere il medesimo obiettivo. Dopo aver presentato i due approcci e aver discusso particolari tecniche utilizzate in alcune applicazioni reali, l'elaborato si conclude riprendendo alcune considerazioni sulla validità dell'inferenza da campioni non probabilistici presentanti in Baker et al. (2013). In particolare, vengono esposte le principali conclusioni tratte dal report della task force AAPOR sull'utilizzo di metodi di campionamento non probabilistico per fare inferenza su una popolazione finita. Nei paragrafi successivi vengono discussi i pro e i contro di tali metodi, e le principali considerazioni e valutazioni proposte a riguardo.

Una prima considerazione da fare è che se per il campionamento probabilistico è possibile individuare un'unica base teorica e delle specifiche proprietà statistiche, questo non è possibile per il campionamento non probabilistico, che comprende una varietà di metodologie distinte. Di conseguenza anche le prestazioni e le proprietà di tali metodi variano considerevolmente. Spetta al ricercatore saper individuare l'alternativa migliore in base allo scopo e al tipo di ricerca che sta effettuando. In ogni applicazione è fondamentale comprendere la variabilità delle caratteristiche studiate all'interno della popolazione obiettivo e le caratteristiche tecniche dei metodi utilizzati per valutare la validità dell'inferenza. Uno degli errori commessi di frequente negli studi che analizzano il comportamento umano è che si assume più omogeneità di quanto non sia realmente. Questo può comportare distorsioni nelle stime. Un altro problema è la distorsione dovuta alla selezione, che si presenta quando il campione non è rappresentativo della popolazione, di cui si è già detto nel paragrafo 2.2. Ad esempio, quando il processo di estrazione del campione non è sotto il controllo del ricercatore, come nel caso di un campionamento di convenienza, è necessario apportare degli aggiustamenti al fine di ottenere informazioni sulla popolazione. Un modo per ridurre i rischi associati al fare inferenza partendo da campioni non probabilistici è quello di esercitare un controllo sul campione e utilizzare delle variabili ausiliarie adatte ad aggiustare il campione per renderlo rappresentativo della popolazione. Al fine di valutare la bontà delle stime è inoltre importante che la metodologia utilizzata sia definita chiaramente, con precise assunzioni che sottostanno al modello. Come si è detto gli approcci *model-based* risultano un ottimo strumento che riserva un grande potenziale per

fare inferenza. Questi approcci stanno diventando sempre più popolari soprattutto nei panel condotti online, in virtù dell'economicità con cui si può accedere a grandi quantità di informazioni nell'era dei Big Data. Tuttavia bisogna anche considerare che il "risparmio" connesso alla facilità di accesso a grandi volumi di dati è compensato dalla necessità di disporre delle conoscenze adeguate per saper individuare e implementare il metodo che più si addice al caso specifico. Per condurre l'analisi in modo opportuno è bene tenere presente che "il livello di rigore impiegato dovrebbe essere commisurato all'importanza e alle esigenze del contesto applicativo e decisionale" (National Research Council, 2012).

Riferimenti bibliografici

Alvarez, R., Sherman, R. & Van Beselaere, C. (2003). Subject acquisition for web-based surveys. *Political Analysis*, *11*, 23–43. https://doi.org/https://doi.org/10.1093/pan/11.1.23

Anderson, C. (2008). *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. https://www.wired.com/2008/06/pb-theory/

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. & Tourangeau, R. (2013). *Report of the AAPOR Task Force on Non-probability Sampling*.

Berzofsky, M., Williams, R. & Biemer, P. (2009). Combining Probability and Non-Probability Sampling Methods: Model-Aided Sampling and the O*NET Data Collection Program. *Survey Practice*, *2*(6). https://doi.org/10.29115/SP-2009-0028

Chen, K. T. (2016). *Using LASSO to Calibrate Non-probability Samples using Probability Samples*. https://hdl.handle.net/2027.42/120734

Conti, P. & Marella, D. (2012). *Campionamento da popolazioni finite*. Springer.

Couper, M. (2013). Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods*, *7*(3), 145–156. https://doi.org/https://doi.org/10.18148/srm/2013.v7i3.5751

Cowling, D. (2015). *Election 2015: How the opinion polls got it wrong*. Recuperato luglio 29, 2022, da https://www.bbc.com/news/uk-politics-32751993

Elliot, M. & Valliant, R. (2017). Inference for Nonprobability Samples. *Statistical Science*, *32*(2), 249–264. https://doi.org/https://doi.org/10.1214/16-STS598

Frické, M. (2014). Big Data and Its Epistemology. *Journal Of The Association For Information Science And Technology*, *66*(4), 651–661. https://doi.org/https://doi.org/10.1002/asi.23212

Gazala, A. (2018). *Evidence-based stratification methodology for non-probabilistic sampling surveys*. http://hdl.handle.net/10292/11784

Greco, M., Powell, R., K, S. & Carter, M. (2003). The Improving Practice Questionnaire (IPQ): A practical tool for general practices seeking patient views. *Education for Primary Care*, *14*, 440–448. https://www.researchgate.net/publication/233572339

International Data Corporation. (2021). *Data Creation and Replication Will Grow at a Faster Rate than Installed Storage Capacity, According to the IDC Global DataSphe-*

re and StorageSphere Forecasts. https://www.idc.com/getdoc.jsp?containerId=prUS47560321

Musacchio, N., Guaita, G., Ozzello, A., Pellegrini, M., Ponzani, P., Zilich, R. & De Micheli, A. (2018). Intelligenza Artificiale e Big Data in ambito medico: prospettive, opportunità, criticità. *The Journal of AMD*, *21*(3), 204–218. https://www.jamd.it/wp-content/uploads/2018/11/2018_03_03.pdf

National Research Council. (2012). *Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification*. The National Academies Press. https://doi.org/10.17226/13395

Nicolini, G., Marasini, D., Montanari, G. E., Pratesi, M., Ranalli, M. G. & Rocco, E. (2013). *Metodi di stima in presenza di errori non campionari*. Springer.

Raita, Y., Camargo, C. A., Liang, L. & Hasegawa, K. (2021). Big Data, Data Science, and Causal Inference: A Primer for Clinicians. *Frontiers in Medicine*, *8*. https://doi.org/https://doi.org/10.3389/fmed.2021.678047

Rosling, H. (2010). *The Joy of Stats*. https://www.gapminder.org/videos/the-joy-of-stats/

Smith, T. (1983). On the validity of inferences from non-random samples. *Journal of the Royal Statistical Society*, *146*(4), 394–403. https://doi.org/https://doi.org/10.2307/2981454

Torelli, N. (2020). *Un esempio di LASSO e regressione ridge in R*. https://moodle2.units.it/pluginfile.php/364129/mod_resource/content/2/esempio-lasso.pdf

Vehovar, V., Toepoel, V. & Steinmetz, S. (2016). Non-probability sampling. Sage.

Zhou, L., Pan, S., Wang, J. & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, *237*, 350–361. https://doi.org/https://doi.org/10.1016/j.neucom.2017.01.026