Capitolo 7. La qualità degli indicatori: stima della varianza[1]

Qualsiasi dato statistico, indipendentemente dal modo in cui è collezionato e dal fenomeno al quale è relativo, è soggetto ad errori di vario tipo. Il fatto stesso di effettuare una indagine campionaria introduce fattori di incertezza che devono essere sempre tenuti presente per una corretta interpretazione dei risultati ottenuti. Per un corretto uso di un qualunque insieme di dati derivanti da un'indagine è necessario perciò fornire misure il più possibile accurate degli errori e, di conseguenza, della qualità dei risultati stessi.

Parte dell'informazione circa i vari tipi di errori è costituita dalla variabilità delle stime prodotte nelle elaborazioni ed è ottenibile tramite il computo degli standard error degli stimatori utilizzati insieme con quello dell'eventuale loro distorsione: queste due misure, considerate simultaneamente, costituiscono il cosiddetto *mean-squared-error* e, nonostante questo non rappresenti sempre la componente più importante dell'errore statistico, è certamente l'informazione di base e più facilmente reperibile riguardante l'errore totale. La stima della varianza totale degli stimatori risulta infatti di importanza cruciale per tutta l'inferenza statistica e permette la costruzione di intervalli di confidenza per i parametri della popolazione indagata: sebbene valutazioni più utili dell'errore campionario siano ottenibili tramite la scomposizione della varianza stessa nelle varie componenti relative ai differenti aspetti e fasi delle procedure di indagine, tale operazione risulta spesso molto complessa. L'informazione sull'errore risulta essenziale anche per la valutazione e la costruzione stessa del disegno e dell'indagine in generale. Per un qualunque problema di stima, infatti, l'entità dell'errore campionario dipende, tra gli altri fattori, dalla numerosità e dal disegno campionari adottati, i quali, a loro volta, incidono sui costi e sui tempi di indagine.

L'efficienza statistica risulta, pertanto, essere solo uno degli aspetti coinvolti in un qualunque problema di stima in quanto, nei lavori pratici di indagine, sono spesso adottate soluzioni di selezione delle unità che si discostano molto da un semplice campionamento completamente casuale. Nella trattazione di dati provenienti da tali

[1] Questo Capitolo è stato redatto in collaborazione con la Dott.ssa Silvia Innocenti.

campionamenti complessi si riscontrano numerose difficoltà nel calcolo di una corretta misura della stessa varianza poiché, introducendo strutture non iid dei dati rilevati, risulta spesso impossibile applicare le tradizionali tecniche di stima.

Lo scopo di questo Capitolo è quello di presentare alcune metodologie di costruzione degli standard error capaci di aggirare alcune di queste difficoltà. Saranno quindi affrontati, da un punto di vista teorico, alcune tecniche di costruzione della varianza degli stimatori utilizzabili in campionamenti complessi.

Nel primo paragrafo verrà presentato in generale il problema della stima della varianza in indagini complesse, come l'indagine EU-SILC.

Il secondo paragrafo è dedicato ad una ampia classe di statistiche della metodologia tradizionale – basata sulla derivazione teorica di una espressione analitica per la varianza di ciascuno stimatore, operazione spesso difficile o addirittura impossibile nel caso di molti stimatori complessi: sarà qui presentato il *metodo di linearizzazione di Taylor*, capace di riportare una qualsiasi statistica complessa ad una forma lineare. Il paragrafo tre verterà sui cosiddetti *metodi di ricampionamento* basati sulla reiterazione delle procedure di stima su un numero relativamente ampio di sub-campioni estratti a partire dal campione originale così da aggiungere variabilità artificiale alle singole stime dei parametri; tali operazioni consentono una più precisa determinazione dell'errore campionario.

Alla presentazione del software di analisi dei dati redatto in linguaggio SAS® ed utilizzato per l'applicazione della metodologia di ricampionamento *Jackknife Repeated Replication (JRR)* a data sets provenienti dall' indagini ECHP e EU-SILC, è dedicata invece una lezione in Laboratorio SAS.

7.1 Metodi di stima della varianza in indagini complesse.

Una parte fondamentale della teoria delle indagini campionarie tratta la derivazione di stimatori per la stima della varianza di statistiche utilizzati in indagini complesse. L'importanza del problema è cruciale in quanto la varianza di uno stimatore, insieme alla sua distorsione, è misura dell'incertezza e dell'accuratezza delle stime ed è quindi

necessaria per una corretta interpretazione dei risultati di una qualunque ricerca oltre che essere indicativa dell'efficienza del disegno campionario utilizzato.

Il proposito di questo capitolo è quello di presentare alcuni metodi di stima della varianza nell'ambito di indagini complesse per le quali il classico approccio per la stima degli standard error risulta poco efficace, poco preciso o alle volte addirittura inapplicabile.

Generalmente con *disegni complessi* si intendono tutti quei campionamenti che, combinando varie tecniche di selezione delle unità quali campionamento senza ripetizione o con probabilità di selezione non uniforme, stratificazione e campionamenti multistadio, introducono una struttura non IID dei dati. Inoltre, sebbene la parola complesso sia solitamente riferita allo schema campionario -ovvero alle regole secondo le quali le varie unità entrano a far parte del campione- questo non è in realtà l'unico elemento di complessità al quale il termine va riferito. Spesso infatti, i parametri stessi indagati nella popolazione e gli stimatori a ciò designati sono complicati nella forma e nella derivazione, come anche le variabili coinvolte nell'ambito di una stessa ricerca possono essere di varia natura e da riferirsi a diverse e numerose sub-popolazioni, delle quali si intende studiare la consistenza,le differenze e le relazioni reciproche. Non esistono perciò in molti casi stimatori corretti della varianza neanche per le statistiche lineari.

Di fatto, nei disegni campionari comunemente utilizzati nella pratica, soprattutto nelle indagini condotte sulla popolazione, varie tecniche campionarie sono implementate contemporaneamente ma molti di questi possono essere analizzati, come vedremo in seguito, attraverso l'utilizzo del seguente disegno multistadio: sia U la popolazione di riferimento, consistente di N distinte unità identificate dall'indice j=1,2,...N e divisa in H strati, ciascuno costituito da A_h cluster ; da ogni strato h si estraggono successivamente $a_h \geq 2$ cluster, in modo indipendente tra gli strati, mediante un determinato campionamento senza ripetizione, per evitare la selezione ripetuta di uno stesso gruppo di unità. A livello di stima della varianza, tuttavia, si considera spesso che tali cluster di primo stadio siano stati estratti con ripetizione così da evitare molte difficoltà computazionali. In tal caso l'approssimazione determina una sovrastima della varianza ma tale distorsione è di piccola entità se la frazione di campionamento a_h / A_h è

modesta. All'interno del hi-esimo cluster le n_{hi} unità elementari, con i=1,2,.. a_h , h=1,2,..H, si selezionano infine secondo una determinata procedura non specificata.
Il numero totale di unità entrate così a far parte del campione è $n = \sum\sum n_{hi}$ mentre il carattere di interesse Y -o un vettore di caratteri- rilevato sulla j-esima unità elementare nell'i-esimo cluster dello strato h è y_{hij} , j=1,2,... n_{hi} , i=1,2,.. a_h , h=1,2,..H. w_{hij} è infine il peso campionario associato all'osservazione y_{hij} , necessario alla ricalibrazione delle stime di una generica statistica. Tali pesi costituiscono essi stessi variabili casuali adattate al trattamento della post-stratificazione e delle non-risposte; gli stimatori sull'intera popolazione divengono così funzioni non-lineari delle variabili campionarie e dei pesi.
Tale disegno sarà il riferimento per i disegni multistadio considerati nel corso del capitolo.
Quando si ha a che fare con disegni di campionamento di questo tipo, i metodi convenzionali di stima della varianza sono spesso difficili da estendere a queste complesse strutture di dati oppure risultano troppo 'ingombranti' per essere implementati. Una soluzione a tali difficoltà è quella di utilizzare approcci di stima della varianza basati sulla *replicazione della stima* o i cosiddetti *metodi di ricampionamento,* che consentono di aggirare il problema mediante l'utilizzo di stime ripetute su aggregati di osservazioni estratti ripetutamente dal campione di dati disponibile o dalla popolazione stessa.
Tali metodi rispondono all'esigenza di disporre di una procedura pratica di stima degli standard error che, tenendo costantemente in considerazione la struttura complessa del disegno, risulti comunque abbastanza flessibile per essere applicata a disegni e statiche diversi e si adatti più o meno facilmente ad applicazioni anche su grandi campioni in termini di sforzi e costi implicati. E' necessario inoltre che dispongano di alcune desiderabili proprietà statistiche quali un basso mean-squared-error dello stimatore utilizzato e robustezza nei confronti dell'allontanamento pratico del disegno utilizzato in una ricerca dal modello ideale assunto di riferimento.

Un primo semplice metodo per fornire buone approssimazioni della varianza di uno stimatore consiste comunque nell'attuare un confronto delle stime del parametro di

interesse ottenute su alcuni aggregati di osservazioni individuabili all'interno di ciascun singolo strato del un campione multi-stadio[2].

Poiché, dato un'insieme qualunque di variabili incorrelate y_j aventi un comune valore atteso

$$E[y_j] = Y \quad ,$$

la media $\bar{y}$ degli n valori y_j

$$\bar{y} = (1/n)\sum y_j$$

ha anch'essa un'aspettativa pari a Y e la sua varianza è data da:

$$Var[\bar{y}] = (1/n-1)\sum (y_j - \bar{y}),$$

molti dei problemi di stima della varianza sopra evidenziati possono essere superati semplicemente considerando l'indice j riferito non a un singolo elemento di una quantità di interesse ma ad un qualunque insieme di elementi indipendenti presenti nel campione; possiamo in tal modo considerare la quantità y_j essere una qualsiasi statistica definita su ogni insieme j, con j=1,2,… R , con l'unica riserva che le y_j siano incorrelate e con una comune aspettativa Y

Applicando tale idea ai problemi di stima riguardanti statistiche lineari in disegni complessi, è possibile eludere parte della scomposizione della varianza degli stimatori nelle sue varie componenti e considerare semplicemente le osservazioni aggregate a livello di estrazione cluster o PSU. Semplicemente assumendo che:

- La selezione campionaria dei cluster sia indipendente tra gli strati,
- Due o più estrazioni siano effettuate in ciascuno strato e
- consistano in estrazioni casuali di individui, indipendenti e con ripetizione,
- La quantità Y di interesse è lineare, ovvero valore il aggregato Y sull'intera popolazione è ottenibile come somma dei valori Y_{hij} su tutti gli elementi j, su tutte le PSU i e infine su tutti gli starti h presenti nella popolazione: $Y = \sum Y_h = \sum\sum Y_{hi} = \sum\sum\sum Y_{hij}$

[2] Il termine PSU (*primary salmpling unit*) sarà utilizzato in riferimento a tali aggregati di osservazioni; questi aggregati corrispondono ai cluster, in quanto abbiamo precedentemente modellato un campionamento nel quale l'estrazione di tali gruppi costituisce proprio l'estrazione di *primary selection*; in realtà il termine si riferisce comunque a un generico aggregato di unità di *primary selection* (PS).

è possibile semplificare il metodo classico di derivazione degli standard error in modo tale che la complessità del campionamento non compaia mai esplicitamente nei calcoli a complicare la procedura. La linearità stessa della statistica assicura, per definizione, di poter ottenere stime degli aggregati e delle loro varianze valide per l'intera popolazione tramite la semplice somma dei valori all'interno e tra gli strati. Questo rende il metodo molto flessibile e applicabile a svariati disegni campionari,cosa che rappresenta la principale ragione del suo ampio utilizzo nelle indagini. Qualche problema si riscontra invece nell'adattamento pratico del metodo ai disegni effettuati nelle indagini per quanto riguarda il fatto che, come abbiamo già evidenziato, i cluster sono ottenuti solitamente senza ripetizione e si ha inoltre a che fare con statistiche non lineari.

La quantità Y è infatti stimabile mediante la somma appropriatamente pesata dei valori Y_{hij} delle unità presenti nel campione:

$$y = \sum_h y_h = \sum_h \sum_i y_{hi} = \sum_h \sum_i \sum_j w_{hij} y_{hij}$$

dove la quantità w_{hij}, rappresentante il peso individuale dell'hij-esimo elemento del campione, può essere espressa tanto nell'unità di misura di y quanto essere riscalata tramite un arbitrario fattore costante.

Considerando tale formula di stima ai suoi vari livelli di disaggregazione si può notare come y_{hij} rappresenti la stima del totale Y_h dello strato h-esimo, mentre ciascun valore pesato y_{hij} delle PSU fornisce la stima della quantità Y_h / a_h, come anche fa la media y_h / a_h, dove a_h rappresenta il numero di PSU selezionate nello strato h.

La varianza delle singole stime di *primary selection* è stimata allora da:

$$Var[y_{hi}] = \frac{1}{a_h - 1} \sum (y_{hi} - \frac{y_h}{a_h})^2$$

e quella dei loro totale y_h :

$$Var[y_h] = (1 - f_h) \frac{a_h}{a_h - 1} \sum (y_{hi} - \frac{y_h}{a_h})^2$$

dove $(1 - f_h)$ rappresenta il fattore di correzione per popolazioni finite, con $f_h = a_h / A_h$, misurante l'effetto del mancato rimpiazzo delle unità all'ultimo stadio di campionamento considerato.

Essendo poi il campionamento indipendente tra gli strati si ha che:

$$Var[y] = Var[\sum y_h] = \sum Var[y_h] = \sum \left[(1 - f_h) \frac{a_h}{a_h - 1} \sum (y_{hi} - \frac{y_h}{a_h})^2 \right]$$

Per le statistiche lineari, a dispetto della complessità del disegno di campionamento, tale espressione della varianza risulta molto semplice, tanto da un punto di vista di derivazione teorica quanto dal lato computazionale, essendo di fatto basata solo su aggregazioni pesate di stime sulle primary selection e non richiedendo il calcolo separato delle varie componenti di varianza. Questo rende il metodo molto flessibile nell'applicazione a diversi disegni campionari.

Il limite fondamentale del metodo è riscontrabile nel fatto di non essere direttamente applicabile a statistiche non-lineari, quali ad esempio rapporti, che non possono essere semplicemente sommate tra le PSU dello strato per stimarne il totale sulla popolazione. Per tali statistiche non-lineari non esistono tipicamente stimatori corretti della varianza e solitamente questa non può neanche essere espressa esplicitamente in funzione della distribuzione campionaria delle variabili coinvolte. Approssimazioni per l'utilizzo del metodo anche un questi casi sono discusse di seguito.

7.2 Generalizzazione a statistiche non-lineari: metodo di linearizzazione di Taylor

Il metodo qui presentato prevede l'utilizzo della serie di Taylor per ottenere un'approssimazione lineare di funzioni non-lineari e una successiva stima della varianza su tale approssimazione in serie. Questa idea di stima della varianza è ricordata con molti nomi in letteratura, incluso metodo di linearizzazione, *delta method* (Kalton,1983) e *propagation of variance* (Kish, 1965).

Supponendo che la quantità della quale si vuole stimare la varianza sia una funzione non-lineare di S aggregati $Y_1, Y_2, \ldots, Y_s$, del tipo

$$Z = f(Y_1, Y_2, \ldots, Y_s)$$

stimata sul campione tramite una statistica z definita nella stessa forma

$$z = f(y_1, y_2, \ldots, y_s)$$

dove $y_1, y_2, \ldots, y_s$ costituiscono le stime sul campione degli aggregati $Y_1, Y_2, \ldots, Y_s$ nella popolazione, è possibile riportare il problema ad una forma simile al caso precedente

tramite un *metodo di linearizzazione* basato sull'utilizzo dell'espansione di Taylor di una qualsiasi funzione non-lineare[3].

Applicando suddetta espansione alla nostra statistica di interesse, almeno fino al primo termine, si ha che, in un intorno di Z:

$$z = Z + \sum_s (y_s - Y_s)(\delta Z / \delta Y_s)$$

e, ponendo $(\delta\delta / \delta Y_s) = D_s$,

$$z = Z + \sum (y_s - Y_s) D_s$$

dove le derivate parziali D_s sono valutate in corrispondenza di $z = Z$ e considerate quindi costanti. Un inconveniente del metodo di linearizzazione è infatti che la valutazione delle derivate parziali D_s potrebbe essere molto difficoltosa per determinati parametri, quali coefficienti di correlazione parziale o multipla; con l'aggiuntiva assunzione che tali costanti ignote siano sostituibili con una loro stima d_s sul campione è possibile però ottenere una stima della varianza tramite un'approssimazione di z non-lineare con una funzione lineare delle y_s :

$$Var[z] = Var[\sum y_s D_s] + Var[\sum y_s d_s] = Var[z_L]$$

A questo punto, per le generali proprietà di una qualunque varianza si ha che:

$$Var[z] = \sum d_s^2 \mathrm{var}[y_s] + \sum d_s d_k \mathrm{cov}[y_s y_k]$$

Notando inoltre che, per statistiche lineari, gli aggregati y_s costituiscono per definizione la somma delle stime ottenute su ciascuna PSU, ovvero $y_s = \sum_{hi} y_{s,hi}$, allora è possibile riscrivere la variabile linearizzata come:

$$z_L = \sum_s y_s d_s = \sum_s (\sum_{hi} y_{s,hi}) d_s = \sum_{hi} (\sum_s y_{s,hi} d_s) = \sum_{hi} z_{hi}$$

ovvero è possibile scrivere z_L come semplice aggregazione delle quantità z_{hi} calcolate al livello delle PSU e della quale è quindi nota la varianza:

$$Var[z_h] = Var[\sum z_{hi}] = (1 - f_h) \frac{a_h}{a_h - 1} \sum (z_{hi} - \frac{z_h}{a_h})^2$$

[3] Lo sviluppo in serie di Taylor serve appunto per approssimare una funzione localmente con un polinomio:
data una funzione f(x) (derivabile fino all'ordine n) e un punto x0 del dominio, il polinomio di Taylor di ordine n e' Pn (x)=f(x0)+f'(x0)(x-x0)+(1/2!)f''(x0)(x-x0) 2 +(1/3!)f'''(x0)(x-x0) 3 +....+(1/n!)f (n) (x0)(x-x0) n.
Sotto la condizione che la funzione abbia derivate continue fino all'ordine n, la differenza tra f(x) e Pn (x) e' trascurabile rispetto a (x-x0)n (resto della formula di Taylor secondo Peano).

Cosicché:

$$Var[z] = Var[z_L] = \sum Var[z_h] = \sum \left[(1 - f_h) \frac{a_h}{a_h - 1} \sum (z_{hi} - \frac{z_h}{a_h})^2 \right]$$

Si è giunti così ad una espressione della varianza di una statistica Z non-lineare analoga a quella ottenuta per le statistiche lineari e applicabile a qualsiasi caso non appena è possibile specificare la quantità z_{hi} al livello delle PSU. E' importante però sottolineare come tale procedimento necessiti di un'assunzione aggiuntiva sul campionamento rispetto a quelle necessarie per la stima della varianza di statistiche lineari: affinché i restanti termini dell'espansione di Taylor possano essere trascurati è necessario, infatti, che si lavori su campioni di dimensione sufficientemente ampia da consentire l'approssimazione lineare coinvolta in tale espressione della varianza.
Un problema che rimane inoltre irrisolto con l'utilizzo di questo metodo è quello di dover comunque approntare, per ogni differente statistica, una formula di derivazione dello standard error almeno al livello delle z_{hi} ; questo lo rende quindi non applicabile per gli stimatori dei quali non è possibile derivare un'espressione analitica della varianza, quali ad esempio statistiche non-funzionali come i quantili e la mediana.

Un caso di utilizzo molto frequente del metodo nei lavori di indagine è quello nel quale la statistica di interesse è rappresentata da un rapporto di due variabili, ovvero una statistica del tipo:

$$r = \frac{y}{x} = \frac{\sum_h y_h}{\sum_h x_h} = \frac{\sum_h \sum_i y_{hi}}{\sum_h \sum_i x_{hi}} = \frac{\sum_h \sum_i \sum_j y_{hij}}{\sum_h \sum_i \sum_j x_{hij}}$$

L'espressione generale della varianza per una variabile di questo tipo è:

$$Var(r) = \frac{1}{x^2}.[var(y) + r^2.var(x) - 2r.cov(x,y)]$$

dove :

$$cov(x,y) = \sum_h \left[(1 - f_h).\frac{a_h}{a_h - 1}.\sum_i \left(x_{hi} - \frac{x_h}{a_h} \right).\left(y_{hi} - \frac{y_h}{a_h} \right) \right]$$

Applicando lo sviluppo in serie di Taylor si ha quindi

$$Var[r] = r^2 [\frac{\text{var}(y)}{y^2} + \frac{\text{var}(x)}{x^2} - 2\frac{\text{cov}(y,x)}{yx}]$$

e ponendo:

$$z_{hi} = \frac{1}{x}.(y_{hi} - r.x_{hi});$$

$$z_h = \sum_i z_{hi}$$

si riporta la varianza di z, e quindi la varianza del rapporto r, alla stessa forma della varianza di una qualunque statistica lineare:

$$Var(r) = \sum_h \left[(1 - f_h).\frac{a_h}{a_h - 1}.\sum_i (z_{hi} - \frac{z_h}{a_h})^2 \right]$$

Per far si che la distorsione del rapporto sia piccola è necessario comunque assumere che le quantità x_{hi} al denominatore siano relativamente uniformi tra gli strati[4].

Per l'applicazione nella pratica dei metodi fin qui esposti è necessario comunque effettuare alcune osservazioni. Nonostante il fatto che le assunzioni di base sulla struttura del campionamento necessarie all'applicazione dei procedimenti siano solitamente soddisfatte nella maggior parte delle indagini, spesso le stesse sono soddisfatte non completamente ed è necessario confrontarsi con delle approssimazioni.
Innanzitutto è importante notare, come già accennato, che, per un migliore sfruttamento delle risorse, si effettuano in genere estrazioni senza ripetizione a tutti i livelli di campionamento così da non trovarsi di fronte al problema della duplicazione delle unità e consentire un migliore sfruttamento delle informazioni; questo comporta una sovrastima delle quantità di interesse poiché l'effetto di tale campionamento senza ripetizione è incorporato nelle formule di stima della varianza qui riportate solo per quanto riguarda il più alto stadio di campionamento, mediante l'introduzione del fattore di correzione per popolazioni finite $(1 - f_h)$. Nonostante ciò la sovrastima introdotta non è solitamente abbastanza consistente da avere grandi ripercussioni sulla qualità dei risultati.
Un altro problema frequente nell'applicazione dei metodi esposti riguarda inoltre il numero e la consistenza delle PSU: è frequente il problema di trovarsi di fronte ad un piccolo numero di cluster, ad esempio a causa dell'applicazione di un campionamento fortemente stratificato. Con questo tipo di disegno infatti, sebbene il numero totale delle PSU estratte possa essere comunque grande, alcuni strati potrebbero consistere di un piccolo numero di PSU e nel contempo contribuire sostanzialmente alla numerosità e

[4] Ricordiamo inoltre che la distorsione della statistica è : Bias[r]= r[(var(x)/x2 - cov(x,y)/xy)]

alla variabilità campionaria. In tal caso,specialmente se siamo interessati alla stima di alcuni parametri anche per le sub-popolazioni presenti nel campione, potremmo incontrare il problema di un numero insufficiente di PSU sul quale effettuare la stima.
Quando, inoltre, un grande numero di osservazioni sono contenute in solo poche PSU è possibile costruire uno stimatore dei parametri di interesse molto preciso, ma non si riesce solitamente a fornire una buona stima della varianza; questo perché un qualunque stimatore della varianza deve, esplicitamente o implicitamente, stimare la componente di varianza che esprime la variabilità tra le PSU; la precisione di tale varianza between-PSU sarà in tal caso decisamente abbassata.
Problemi analoghi si hanno quando le PSU sono invece troppo piccole o comunque sia inappropriate per un utilizzo diretto nella formula di stima della varianza. Unità computazionali più 'adatte' alla implementazione dei metodi possono comunque essere definite mediante varie tecniche quali raggruppamenti casuali di unità all'interno degli strati o unione e combinazione di cluster tra gli strati. E' necessario infatti considerare che, in generale, le PSU sono definite a livello di campionamento con l'obiettivo di soddisfare o la minimizzazione dei costi e degli sforzi di campionamento oppure la minimizzazione della varianza between-PSU, ovvero la massimizzazione dell'eterogeneità interna delle PSU, i quali sono a loro volta obiettivi spesso in contrasto reciproco.

7.3 Stima della varianza mediante metodi di ricampionamento

Ai procedimenti appena presentati si affianca una seconda classe di metodi utilizzati in indagini complesse nella stima della varianza di stimatori, basati sul confronto tra misure ripetute di variabilità ottenute su repliche del campione, ovvero su aggregati di osservazioni ognuno dei quali riflette la struttura dell'intero campione padre, incluse eventuali stratificazioni e clusterizzazioni presenti nei vari stadi del campionamento: utilizzano procedure per la formazione di sub-campioni tali che una replica differisce dal campione originario solo nella dimensione ovvero per la numerosità di unità considerate.
Vari metodi di ricampionamento sono stati proposti negli anni più recenti, tra i quali il Jackknife,il BRR e il bootstrap hanno ricevuto maggiore attenzione, e sono in generale

tenuti in alta considerazione in quanto prescindono dalle teoriche e solitamente ignote distribuzioni degli stimatori coinvolti, ed anzi sono spesso utilizzati proprio nella stima di tali distribuzioni.

Tali metodi si basano tutti sull'utilizzazione ripetuta delle informazioni provenienti da un unico campione estratto, al fine di migliorare la precisione delle stime grazie all'aggiunta di variabilità artificiale ai dati presenti nell'unico campione osservato; differiscono tra loro unicamente per la procedura di estrazione dei sub-campioni replicati dall'unico campione padre originario.

L'idea alla base di tutti i metodi di ricampionamento utilizzati per la stima della varianza è quella di selezionare ripetutamente sotto-campioni $S_1,S_2,..S_r$ dall'intero campione originario S ottenuto mediante la strategia prescelta nel disegno di indagine, in modo da poter calcolare la statistica di interesse per ognuno di questi R sotto-campioni. E' possibile così stimare la distorsione, la varianza o addirittura l'intera distribuzione campionaria dello stimatore utilizzando appunto la variabilità delle stime ottenute sui sub-campioni.

Tali metodi traggono giustificazione formale del loro utilizzo da un principio elementare,detto *principio del plug-in* o principio *di sostituzione* secondo il quale 'sussiste ancora una buona validità statistica se alla funzione di ripartizione di una statistica sulla popolazione sostituiamo la funzione di ripartizione nel campione in quanto tali due variabili sono, al verificarsi di certe condizioni di regolarità, equivalenti in variabilità' (Efron, Thishibani e Han 92-93) ovvero hanno i momenti quarti coincidenti[5].

Il requisito di base per l'applicazione dei metodi di ricampionamento di seguito esposti è che il l'intero campione di partenza sia scomponibile in un numero di sub-campioni o repliche, aventi ciascuno lo stesso disegno e che riflettano la complessità dell'intero campione. Oltre a ciò l'ampiezza di tali sub-campioni deve essere sufficiente a consentire la produzione di stime che siano vicine o comunque confrontabili con la stima corrispondente basata sul campione-padre. Allo stesso tempo, il numero di

[5] Il principio ristretto di equivalenza in variabilità (K. Pearson) afferma propriamente che: supponendo l'esistenza della funzione f(μ1 μ2 .. μn) appartenente alla classe C3 in un intorno dei valori medi, due variabili casuali sono equivalenti in variabilità se hanno indici di simmetria e curtosi uguali; media e varianza non sono neanche prese in considerazione in quanto è sufficiente standardizzare le variabili. Molti indicatori statistici frequentemente utilizzati quali momenti, coefficienti di variazione , rapporti tra valori medi e/o varianze e molti altri rientrano in questa classe, mentre si ha difficoltà con i quantili.

repliche estraibili deve essere sufficiente affinché tale comparazione produca una approssimazione stabile della variabilità campionaria totale.

Assumendo che le repliche siano indipendenti, come nel caso trattato nei paragrafi precedenti la teoria base dei campionamenti replicati fornisce una espressione molto semplice della varianza della media tra tali repliche: posto θ parametro di interesse, sia ϑ una stima basata sull'intero campione di una qualunque statistica,non necessariamente lineare, e ϑ_r la stessa stima ottenuta sull'r-esima replica ed espressa nella stessa unità di misura; indicando con $\bar{\vartheta}_r$ la semplice media dei valori ϑ_r

$$\bar{\vartheta}_r = (1/R)\sum \vartheta_r$$

si ha che:

$$Var[\bar{\vartheta}_r] = \frac{1}{R(R-1)} \sum_r (\vartheta_r - \bar{\vartheta}_r)^2$$

In realtà,anche se la statistica di interesse è ϑ e non la media $\bar{\vartheta}_r$ delle repliche, è possibile dimostrare che nel caso di stimatori lineari le due coincidono anche se ciò non è vero per stimatori non-lineari. Cmq sia, quando ciascuna replicazione del campione è formata su molte osservazioni e riflette bene la struttura dell'intero campione-padre, le stime ϑ_r e ancor più la loro media $\bar{\vartheta}_r$ risultano molto vicine a ϑ ottenuta sull'intero campione anche per statistiche complesse non-lineari; Questo non è vero per il caso considerato nei paragrafi precedenti di stime ottenute sulle singole PSU. In questo appunto risiede la potenza di ricampionamento basato sulle replicazioni: quando applicabile, può essere utilizzato per statistiche non-lineari senza la necessità di formulare un'espressione esplicita della loro varianza per i differenti tipi di statistiche; possiamo infatti direttamente considerare:

$$Var[\vartheta] \cong Var[\bar{\vartheta}_r] = c\sum_r (\vartheta_r - \vartheta)^2$$

Oppure:

$$Var[\vartheta] \cong Var[\bar{\vartheta}_r] = c\sum_r (\vartheta_r - \bar{\vartheta}_r)^2$$

E' da notare che le due espressioni sono identiche nel caso di stimatori lineari mentre per non-lineari la prima è più conservativa. Comunque sia, in molte indagini, la differenza attesa tra le due stime $[R(\vartheta - \bar{\vartheta}_r)]$ è piccola.

Le procedure basate sul ricampionamento offrono il chiaro vantaggio di poter essere applicate anche a stimatori non esprimibili in termini analitici.

Il maggiore limite di un approccio di questo tipo al problema di stima della varianza è quello di determinare un decisivo aumento del lavoro e dei costi: oltre a necessitare di un disegno di campionamento di difficile delineazione, in quanto deve consentire di effettuare un numero sufficiente di repliche di ampiezza 'paragonabile' all'originale, capaci di fornire ciascuna una stabile stima della varianza, prevede un elevato sforzo computazionale. Sono metodi computer-intensive che hanno visto un decisivo sviluppo e ampliamento di utilizzo solo negli ultimi anni, parallelamente allo sviluppo delle nuove tecnologie.

Esistono inoltre due distinte forme nelle quali l'idea di siffatte replicazioni può essere applicata alla stima della varianza: la più semplice è basata su repliche indipendenti e *non-overlapping*, mentre l'altra considera un ricampionamento ripetuto del campione d'origine che genera replicazioni sovrapposte e non indipendenti.

7.3.1 Stima della varianza su repliche indipendenti

La stima della varianza basata su repliche indipendenti è l'approccio più semplice al problema, almeno in linea di principio poiché prevede la semplice estrazione di sub-campioni indipendenti dal campione originario. Nella realtà tale piena indipendenza non è quasi mai ottenibile all'interno di uno stessa indagine, in quanto dovrebbe essere implementata tanto a livello di selezione campionaria, quanto a livello di raccolta e immagazzinamento dei dati come al momento dell'applicazione delle procedure di stima.

Come requisito minimo per l'applicazione del metodo si considera allora la possibilità, al momento del disegno o successiva, di dividere il campione-padre in 'più o meno indipendenti' repliche, ciascuna delle quali abbia essenzialmente lo stesso disegno del campione originario.

In un disegno multistadio questo potrebbe richiedere di dividere il campione-padre tra le repliche a livello delle PSU e, in caso di stratificazione, ciascuna replica dovrebbe essere essa stessa un sottocampione stratificato coprente tutti gli strati individuati nell'indagine. E' però necessario considerare che, per ogni data numerosità campionaria

e per ogni dato numero di PSU presenti nel campione, il numero di repliche ottenibili è inversamente proporzionale alla dimensione delle stesse; un numero troppo piccolo di repliche si risolve inoltre in una grande variabilità delle stime mentre un numero troppo alto può comportare una grande distorsione delle stesse, nella misura in cui, a causa della riduzione della dimensione dei sottocampioni-replica, non riescono questi a riflettere la piena complessità del disegno originario. In più, la necessità di dividere il campione in replicazioni indipendenti riduce la flessibilità e l'efficienza del disegno complessivo e solitamente comporta un decisivo innalzamento dei costi del lavoro sul campo.

Nonostante tali questioni debbano sempre essere tenute presenti, sussistono buone ragioni per raccomandare l'uso delle tale tecnica di ricampionamento, specialmente per indagini condotte sulle popolazioni umane, in quanto, per svariati motivi pratici e per le caratteristiche stesse di tali popolazioni, è spesso facile attuare delle buone e 'naturali' divisioni del campione, riuscendo a costruire sub-campioni al cui interno sia presente un buon numero di PSU. Diventa in questo modo possibile dividere il campione-padre in un discreto numero di replicazioni pseudo-indipendenti aventi ciascuna un ragionevole numero di PSU al proprio interno.

Nonostante, infine, non esista in realtà alcuna regola pratica che fornisca un criterio di decisione assoluto, si considera alle volte ragionevole far si che, tanto il numero di replicazioni quanto il numero di PSU al loro interno, siano all'incirca pari alla radice quadrata del numero totale di PSU disponibili.

Come estensione del metodo delle repliche indipendenti, l'approccio di stima della varianza tramite *ricampionamento ripetuto* è nato proprio in risposta al problema di avere a che fare con un numero di PSU troppo piccolo per poter costruire un numero sufficiente di replicazioni indipendenti di adeguata grandezza.

Al prezzo di decisivo incremento della complessità e del lavoro computazionale, il ricampionamento in considerazione effettua un calcolo dell' errore nel quale le stime che si vanno a comparare sono ottenute su repliche generate mediante estrazioni ripetute sul campione di origine che non assicurano però la piena indipendenza delle repliche tra loro, le quali anzi vanno a soprapporsi nelle unità considerate. Anche in questo caso ciascuna replica deve essere disegnata in modo tale da riflettere l'intera struttura del campione

Sono perciò necessarie alcune procedure aggiuntive che controllino la distorsione nella stima della varianza ottenuta mediante il confronto tra tali repliche in quanto sono meno capaci di isolare le varie componenti di varianza relative ai diversi stadi di campionamento.

Le varie procedure di ricampionamento implementabili differiscono tra loro esclusivamente nel modo nel quale le replicazioni sono generate a partire dal campione-padre e per come le rispettive stime della varianza sono determinate; tre sono le procedure generalmente più utilizzate, conosciute rispettivamente come: *Jackknife repeated replication* (JRR), *Balanced repeated replication* (BRR) e metodo *Bootstrap*.

7.3.2 Jackknife Repeated Replication (JRR)

Il metodo *Jackknife* è stato introdotto inizialmente da Quenouille (1949), come procedimento non parametrico capace di ottenere una stima della distorsione di statistiche generiche, e successivamente Tukey (1958) suggerì come la stessa procedura potesse essere utilizzata nella stima della varianza[6].

L'idea generale sottesa dal Jackknife repeated replication, per la quale tale procedura è conosciuta anche come *tecnica leave-one-out*, è quella di formare diversi sub-campioni di un generico campione estratto eliminando di volta in volta una singola osservazione o -nel caso di disegni complessi e che prevedono un'elevata numerosità- una piccola parte del campione originario, quale ad esempio una singola PSU di uno strato di un campione multi-stadio. Ciascun sub-campione misurerà così il contributo alla variabilità totale fornita da tale piccola parte eliminata dalla stima: la procedura JRR calcola infatti le stime dei parametri di interesse su ciascuna replica ed utilizza la variabilità tra tali valori per ottenere una stima della variabilità campionaria totale.

Consideriamo innanzitutto il funzionamento del metodo Jackknnife in caso di campioni di dati iid, per passare poi successivamente alla sua estensione per situazioni complesse.

[6] Il termine jackknife indica in inglese il coltello serramanico e, per estensione, il coltello degli esploratori che contiene varie lame e numerosi altri strumenti. Secondo Garhwaite *et al.* (1995), questo 'coltellino svizzero' è funzionale in situazioni di emergenza mentre risulta inutile quando si dispone degli strumenti specifici, più solidi e funzionali. Il termine è stato scelto opportunamente poiché il metodo ha un'applicazione appropriata quando non è possibile utilizzare i metodi classici.

Sia $\vartheta = t(S) = t_n(y_1, y_2, ..., y_n)$ stimatore di θ funzione dell'intero campione estratto S . Sia invece $\vartheta_{(i)} = t_n(y_1, y_2, ..., y_{i-1}, y_{i+1}, ..., y_n)$ la stessa statistica basata su (n-1) osservazioni mediante la rimozione dell'osservazione y_i , per i=1,2,...n .

Se definiamo:

$\overline{\vartheta}_n = (1/n)\sum_i \vartheta_{(i)}$

Allora:

$$\vartheta_{jk} = \vartheta - (n-1)[\overline{\vartheta}_n - \vartheta] = n\vartheta - (n-1)\overline{\vartheta}_n$$

oltre ad essere uno stimatore del parametro che riduce la distorsione della stima rispetto a ϑ , in quanto fornisce una stima $Bias[\vartheta] = (n-1)(\overline{\vartheta}_n - \vartheta)$ di $B[\vartheta] = E[\vartheta] - \theta$, può essere utilizzato per costruire uno stimatore della varianza.

Notando infatti che ϑ_{jk} può essere altrimenti scritto come

$$\vartheta_{jk} = (1/n)\sum_i [n\vartheta - (n-1)\vartheta_{(i)}] = (1/n)\sum_i \vartheta_i^*$$

ovvero come media degli *pseudo-valori* $\vartheta_i^* = n\vartheta - (n-1)\vartheta_{(i)}$, è possibile stimare la varianza di ϑ tramite la varianza tra tali pseudo-valori ponendo le assunzioni:

- Gli pseudo-valori ϑ_i^* , per i=1,2,...n possono essere trattati come realizzazioni di variabili indipendenti ed identicamente distribuite (IID);
- $V[\vartheta_i^*] \approx V[\sqrt{n}\vartheta]$.

Sotto tali assunzioni diventa allora naturale stimate la varianza $V[\vartheta]$ dello stimatore ϑ di interesse tramite la varianza campionaria degli speudo-valori ϑ_i^* ovvero tramite:

$$Var[\vartheta_{jk}] = (1/n)Var[\vartheta_i^*] = \left(\frac{1}{n(n-1)}\right)\sum_i [\vartheta_i^* - (1/n)\sum_j \vartheta^*{}_j]^2 = \left(\frac{n-1}{n}\sum_i \vartheta_{(i)} - (1/n)\sum_j \vartheta_j\right)$$

Ad esempio, per il semplice caso di una media o una proporzione campionaria è facile mostrare che $\vartheta_{(i)} = \overline{Y}_{(i)} = (1/n-1)(n\overline{Y} - y_{(i)})$ cosicché $\overline{\vartheta}_n = \overline{Y}$ e $\vartheta_{jk} = \vartheta = \overline{Y}$. In più, in questo caso, gli pseudo-valori ϑ_i^* sono esattamente pari a $y_{(i)}$ e la loro varianza è $V[\vartheta_i^*] = V[\sqrt{n}\vartheta]$ cosicché le due assunzioni necessarie alla stima della varianza sono

valide. In questo modo lo stimatore Jackknife e l'approccio tradizionale producono lo stesso stimatore della varianza $\left[\frac{1}{n(n-1)}\sum_i (y_i - \overline{Y})^2\right]$

Da quanto appena visto è evidente come il metodo, richiedendo di ri-calcolare la statistica per n volte, implichi un notevole sforzo computazionale affrontabile solamente negli anni più recenti grazie allo sviluppo dei nuovi software.

L'estensione del procedimento al caso di campioni di dati non IID derivanti da indagini complesse, sebbene costituisca uno dei luoghi di maggiore utilizzo, non è immediata e diverse versioni ne sono state proposte.

Il primo passo è quello di decidere quale sia l'unità da eliminare in ogni replica e dividere il campione-padre in gruppi di osservazioni. Non è in generale una buona idea eliminare una unità finale, sia in quanto y_{hij} non è indipendente dalle alle altre unità presenti nello stesso cluster sia perché il numero di repliche effettuabili in tal caso risulterebbe troppo alto. Il problema diventa quindi costruire repliche in modo tale che ciascuna abbia essenzialmente lo stesso disegno campionario per assicurare che gli stimatori della varianza calcolati su ognuno di essi abbiano accettabili proprietà statistiche. In indagini a struttura complessa questo è tipicamente fatto utilizzando le osservazioni aggregate al livello delle PSU. In tal modo, se il numero di strati coinvolti è grande e contemporaneamente le PS tendono a essere piccole o variabili nella dimensione, possiamo adattare la struttura del campione dato al modello richiesto in modo efficiente e computazionalmente conveniente.

Successivamente è necessario costruire un sistema di pesi da utilizzare nelle stime dei parametri di interesse su ciascun sub-campione: poiché ciascuna replica è formata cancellando un gruppo di osservazioni presenti in uno strato del campione è necessario aumentare il peso delle restanti unità dello strato al fine di non fornire stime della varianza distorte per difetto. Normalmente il fattore c_h utilizzato a tal fine è definito come $c_h = [a_h/(a_h - 1)]$ ma si preferisce alle volte porre:

$$c_h = \frac{w_h}{w_h - w_{hi}}$$

dove

$$w_h = \sum_i w_{hi} = \sum_i \sum_j w_{hij}$$

costituisce la somma dei pesi campionari di tutte le unità j presenti nella PSU i[7]. Tale secondo modo di calcolare il fattore di correzione è preferibile in quanto conserva il peso totale del campione inalterato tra le varie repliche.

Si procede infine alla stima della varianza dei parametri di interesse: la stima di un semplice aggregato, quale ad esempio un totale, per tale (h,i)-esima replica è:

$$\vartheta_{(hi)} = \sum_{k\neq h}\sum_i\sum_j w_{kij}\vartheta_{kij} + c_h\sum_{j\neq i}\sum_j w_{hlj}\vartheta_{hlj}$$

$$= \sum_{k\neq h}\vartheta_k + c_h(\vartheta_h - \vartheta_{hi}) = \vartheta - c_h(\vartheta_{hi} - \frac{\vartheta_h}{w_h})$$

In tal modo, in caso che la statistica di interesse sia lineare e che ciascuna PSU sia eliminata una sola volta nel formare le repliche, tanto nella media $\overline{\vartheta}_{(h)}$ delle $\vartheta_{(hi)}$ sullo strato h, quanto nella media $\overline{\vartheta}$ su tutte le replicazioni, viene eliminato l'ultimo termine della formula sopra e risultano entrambe pari alla statistica ϑ stimata sul campione originario:

$$\vartheta_{(h)} = (1/w_h)\sum_i \vartheta_{(hi)} = (1/w_h)\sum_i[\vartheta - c_h(\vartheta_{hi} - \frac{\vartheta_h}{w_h})] = \vartheta - c_h[\vartheta_h - \vartheta_h] = \vartheta$$

$$\overline{\vartheta} = \frac{\sum_h\sum_i \vartheta_{(hi)}}{\sum_h w_h} = \vartheta$$

Da questo segue anche che, in modo molto semplice, l'espressione standard della varianza dello stimatore può essere scritta nelle tre forme :

$$Var[\vartheta] = [(1-f_h)c_h\sum_i(\vartheta_{hi} - \frac{\vartheta_h}{w_h})^2] = \begin{cases} \sum_h[(1-f_h)c_h\sum_i(\vartheta_{(hi)} - \vartheta_{(h)})^2] & [1] \\ \sum_h[(1-f_h)c_h\sum_i(\vartheta_{(hi)} - \overline{\vartheta})^2] & [2] \\ \sum_h[(1-f_h)c_h\sum_i(\vartheta_{(hi)} - \vartheta)^2] & [3] \end{cases}$$

Solitamente nel metodo JRR, pur potendo utilizzare queste tre espressioni indifferentemente, la formula standard della varianza viene sostituita dalla [3], in quanto è tra tutte la più conservativa[8].

[7] La correzione è stata proposta da Verma e Betti (2005).

Inoltre, sebbene tali sostituzioni siano lecite solo nel caso in cui la statistica ϑ di interesse sia lineare, queste espressioni producono una buona approssimazione anche per la stima della varianza di statistiche più complesse, anche non-lineari. Questo perché stime del tipo $\vartheta_{(hi)}$, $\overline{\vartheta}_{(h)}$ e ancor più la media $\overline{\vartheta}$, essendo basate su sotto-campioni molto vicini per struttura e dimensione al campione-padre, sono attese essere molto vicine alla stima ϑ sul campione completo anche per statistiche complesse cosicché la loro varianza fornisce una misura della varianza di ϑ altrettanto buona.

E' evidente come il vantaggio del ricampionamento sia quello di evitare la difficile e spesso impossibile derivazione di una formula analitica esplicita della varianza delle statistiche considerate: pur fornendo solitamente risultati molto simili a quelli ottenibili con l'approccio tradizionale[9], gli stimatori Jackknife sostituiscono tale derivazione teorica con un maggiore sforzo computazionale e, sebbene non sia teoricamente applicabile a statiche funzioni di statistiche ordinali -quali ad esempio i quantili- , studi empirici dimostrano come esso rimanga metodo efficace, in caso di campionamenti complessi in cui si considerano PSU formate da gruppi consistenti di osservazioni, anche per stimatori di questo tipo.

7.3.3 Balanced Repeated Replication (BRR)

Il BRR è stato proposto per la prima volta da McCarthy nel 1969 come tecnica di ricampionamento da applicare a campioni stratificati che prevedessero esattamente $a_h = 2$ cluster per strato estratti con ripetizione e solo successivamente è stato adattato al caso di un generico numero a_h di cluster presenti in ciascuno strato. Poiché infatti il metodo prevede l'utilizzazione di esattamente metà campione nella formazione di ciascuna replica mediante l'esclusione di uno dei due gruppi presenti in ciascun strato, se alcuni strati sono formati da più di due cluster è necessario o raggruppare quest'ultimi

[8] E' dimostrato che questi tre stimatori Jackknife sono asintoticamente equivalenti fino al secondo ordine: Vjk[1]/Vjk [2]= Vjk[1]/Vjk [3]= Vjk[2]/Vjk [3]=1+ Op(n-2).

[9] In termini di comportamento asintotico, il Jackknife e il metodo tradizionale di linearizzazione sono nella maggior parte dei casi asintoticamente equivalenti fino al primo ordine : Vjk/VL=1+Op(n-1) e Vjk/VL=1+Op(n-2) se ah=2 per ogni h.

in due *super-PSU* o dividere gli strati stessi in più piccoli strati artificiali così da riportare la situazione a quella originariamente proposta.

Una volta che tali divisioni artificiali del campione sono state effettuate possiamo illustrare il metodo considerando il caso di esattamente due PSU estratte in ciascun strato formando le repliche mediante l'eliminazione di una PSU per ciascuno strato.

L'insieme delle 2^H repliche così ottenibili può essere definito tramite una matrice di dimensione $R * H$ nella quale l' (r,h)-esimo elemento $\varepsilon_{rh} = \pm 1$ indica quale delle due metà dello strato è stata selezionata nell' (r,h)-esima replica, con $r = 1,2,..,R$, $h = 1,2,..,H$, $\Sigma_r\ \varepsilon_{rh} = 0$ per ogni h e $\sum \varepsilon_{rh}\varepsilon_{rk} = 0$ per ogni h≠k - ovvero le colonne della matrice sono ortogonali. Esiste una seria difficoltà pratica nel calcolare le stime su ciascuna replica pure quando si ha a che fare con un piccolo numero di strati. Nell'utilizzo effettivo del metodo si considera perciò solamente un sottoinsieme delle 2^H possibili repliche – e delle 2^{H-1} possibili repliche distinte –, formato dal numero minimo di sub-campioni necessari alla cattura delle informazioni e rappresentato dalle $H < R' < H + 3$ repliche ottenute scegliendo H colonne da una matrice di Hadamard[10] $R * R$ ad esclusione della riga rappresentata da tutti elementi uguali a 1[11].

Sia allora $\vartheta^{(r)}$ lo stimatore di un parametro θ analogo a $\vartheta = t(S)$ lineare ma calcolato sulla r-esima metà-campione, si ha che:

$$\vartheta^{(r)} = \sum_h (1/2) \sum_i (\varepsilon_{rh} + 1) \sum_j w_{hij} \vartheta_{hij}$$

per $r = 1,2,..,R'$. Definendo allora:

$$\overline{\vartheta}^{(r)} = (1/R') \sum_r \vartheta^{(r)}$$

la varianza di ϑ può essere scritta come:

$$Var[\vartheta] = Var[\vartheta^{(r)}] = (1/R') \sum_r (\vartheta^{(r)} - \overline{\vartheta}^{(r)})^2$$

---

[10] Matrice quadrata le cui entrate sono $+1$ o -1 e le cui righe sono mutuamente ortogonali del tipo di quella R*H appena costruita.

[11] Per l'applicazione del metodo è infatti chiaramente desiderabile trovare un BRR con R più piccolo possibile; nel caso di ah>2 è tuttavia molto più difficile trovare un numero di repliche fattibile rispetto a quando ah=2 e il più piccolo R possibile potrebbe essere molto più grande di H+4. Nel caso di ah=p>2 per ogni strato con p numero primo o potenza di primo un BRR può essere implementato attraverso campioni bilanciati ottenuti selezionando un solo cluster per ciascuno strato ma questo è solo uno tra i tanti metodi proposti.(vd Gupta and Nigam 1987; Wu 1991; Sitter 1993).

Nel caso qui riportato, ovvero quello in cui lo stimatore è lineare, tale stimatore BRR della varianza corrisponde allo stimatore Jackknife a causa della natura bilanciata delle repliche, mentre sono necessarie alcune approssimazioni per l'estensione al caso di statistiche non-lineari in quanto ne le stime calcolate su ciascuna replica $\vartheta^{(r)}$ ne la loro media $\overline{\vartheta}^{(r)}$ risultano uguali alla stima ϑ ottenuta sul campione originale; si preferisce allora utilizzare i seguenti stimatori della varianza - anch'essi analoghi agli stimatori JRR in caso di linearità- che risultano di maggiore precisione:

$$Var[\vartheta] = Var[\vartheta_{BRR-(r)}] = (1/R')\sum_r (\vartheta^{(r)} - \vartheta)^2$$

$$Var[\vartheta] = Var[\vartheta_{\text{BRR-D}}] = (1/4R')\sum_r (\vartheta^{(r)} - \vartheta^{(r)c})^2$$

con $\vartheta^{(r)c}$ stimatore basato sulle repliche contenenti la metà-campione complementare a quella utilizzata in $\vartheta^{(r)}$. E' da notare come, mentre la prima espressione semplicemente sostituisce alla media delle repliche il valore dello stimatore sull'intero campione, la seconda, considerando $\vartheta^{(r)c}$ implica uno sforzo computazionale esattamente doppio alle altre due, essendo infatti basata su $2R'$ stime ripetute. Infine, poiché sempre con lo stesso numero 2R' di repliche è possibile misurare la variabilità di entrambe le statistiche $\vartheta^{(r)}$ e $\vartheta^{(r)}$ rispetto alle proprie medie, un'ulteriore e migliore stima della varianza di ϑ è ottenibile mediante:

$$Var[\vartheta] = (1/2R')\sum_r [(\vartheta^{(r)} - \vartheta\ )^2 + (\vartheta^{(r)c} - \vartheta)^2] = (1/2)Var[\vartheta_{BRR-(r)}]Var[\vartheta_{BRR-(r)c}]$$

la quale però, pur essendo la più precisa, è preferibile solo se il lavoro computazionale non risulta di per se troppo impegnativo.

Tali stimatori BRR della varianza risultano essere stimatori consistenti nonostante forniscano risultati meno vicini a quelli ottenibili con il metodo tradizionale di linearizzazione rispetto agli stimatori JRR; a differenza di questi ultimi però non si riscontrano problemi teorici nell'applicazione ai quantili.

I maggiori limiti del metodo risiedono invece nelle restrizioni che pone alla struttura del campione oltre che nell'elevato sforzo computazionale del quale necessita. Soluzioni ad entrambi i problemi che sono state proposte nel corso degli anni rispondono, da una parte, alla necessità di allargare lo spettro di utilizzo del BRR a campioni per i quali non si verifica la condizione di avere ($a_h = 2$) cluster per ogni strato, dall'altra all'esigenza di diminuire i lavori di calcolo coinvolti.

Per quanto riguarda il primo aspetto una semplice ma inefficiente soluzione adottabile è quella di divedere gli a_h strati presenti nello strato h in due gruppi contenenti rispettivamente $(a_h / 2)$ e $[a_h - (a_h / 2)]$ cluster e applicare il BRR su tali cluster artificiali. Tale metodo, chiamato *grouped* BRR (GBRR), fornisce per H alto uno stimatore della varianza consistente ma la sua efficienza risulta decisamente ridotta quando il numero di strati è molto minore al numero $a_h = \sum a_h$ di cluster totali previsti dal campionamento.

Un'altra tecnica adottata per lo stesso scopo è quella di dividere casualmente gli a_h cluster di ciascuno strato in $a_h / 2$ gruppi e dopo trattare tali $a / 2$ gruppi costruiti come nuovi strati aventi 2 cluster ciascuno. Lo stimatore della varianza ottenuto con questa seconda approssimazione risulta consistente all'aumentare di *a* tuttavia la sua applicazione può non essere semplice quando *a* risulta molto più grande del numero di 'veri' strati H (può infatti risultare molto difficile costruire una matrice di Hadamard con un numero molto alto di colonne).

Altri metodi di raggruppamento di cluster e/o strati sono spesso utilizzati anche come soluzione al secondo problema di avere un numero troppo elevato di repliche sulle quali calcolare ripetutamente gli stimatori. Nella stessa direzione si muovono infine le versioni 'parzialmente bilanciate' del BRR ovvero che estendono il metodo a ricampionamenti basati su repliche non esattamente formate da metà campione.

E' interessante infine menzionare il *metodo di Fay* che costituisce anch'esso una versione modificata del BRR: mentre nel BRR viene assegnato peso zero a metà campione e all'altra metà viene raddoppiato, questo secondo metodo consiste nell'assegnare un peso ρ, $0 < \rho < 1$, a una metà del campione e un peso $2 - \rho$ all'altra.

7.3.4 Bootstrap

Il *metodo bootstrap* è stato proposto da Efron nel 1979 come evoluzione del metodo Jackknife per il caso di campioni aventi struttura di dati IID ed è stato esteso da Rao e Wu (1988) a disegni multistadio stratificati.[12] In pochi anni, questa procedura ha avuto

[12] Il nome Bootstrap, letteralmente "stringhe" o "lacci delle scarpe", deriva, per dichiarazione dello stesso autore, dal romanzo del diciottesimo secolo "Adventures of Baron Munchausen" di Rudolph Erich Raspe ed in particolare dall'espressione inglese che ritroviamo nel romanzo "to pull oneself up by one's

una evoluzione così rapida e accompagnata da una serie di approfondimenti da parte dello stesso autore e dei suoi colleghi, che è stata resa ad oggi la tecnica di ricampionamento più diffusa e nota[13].

Il metodo proposto si presenta in stretta relazione con gli altri ricampionamenti ed è possibile dimostrare come gli stimatori della varianza costruiti mediante il Jacknife possano addirittura essere pensati come espansione lineare per approssimare il bootstrap nel caso in cui la statistica di riferimento sia sufficientemente liscia[14]. Questo non implica comunque che l'applicazione del metodo risulti sempre più efficiente degli altri ricampionamenti, o che produca stimatori 'migliori'. Nell'ambito di stima della varianza stimatori JRR e BRR richiedono uno sforzo computazionale decisamente minore in quanto sono basati su un numero più piccolo di repliche e il bootstrap necessita inoltre di modifiche sostanziali per una sua corretta applicazione a disegni complessi -si dimostra in sostanza meno robusto al variare delle condizioni di campionamento. Tali modifiche sono state approntate in varie versioni del bootstrap che risultano però applicabili, di volta in volta, solo a situazioni ben specifiche mentre l'argomento deficia di una trattazione teorica generale che ne consenta l'applicazione a svariati disegni complessi. Sotto un altro punto di vista il metodo qui presentato risulta comunque di più ampio utilizzo per affrontare altri problemi di inferenza statistica.

La particolarità delle tecniche bootstrap risiedono nel modo di costruire le repliche a partire dal campione di dati ottenuto nell'indagine: considerando tale campione-padre come nuova popolazione artificiale sulla quale implementare nuovamente il disegno campionario e utilizzando estrazioni delle unità con ripetizione, il metodo consente di costruire repliche indipendenti aventi la stessa dimensione del campione originario oltre che rifletterne la struttura. In questo modo il numero di repliche ottenibili risulta essere molto maggiore rispetto agli altri metodi: nel caso di campioni di dati IID, ad esempio, è

---

bootstrap", letteralmente "tirarsi su attaccandosi ai lacci delle proprie scarpe". Tale espressione evidenzia, in modo scherzoso, il fatto paradossale che l'unico campione disponibile serve per generarne molti altri.

[13] In realtà il Bootstrap non è di per se tecnica necessariamente di ricampionamento: solo quando la funzione di ripartizione empirica dello stimatore non è una funzione esplicita dei parametri e della funzione di ripartizione delle variabili allora si utilizzano metodi di stima con ricampionamento.

[14] Il Bootstrap fornisce stime della varianza non consistenti in caso che la distribuzione abbia code molto pesanti e/o in presenza di valori anomali: rispetto al Jackknife costruisce stimatori consistenti per una più ristretta classe di statistiche perché molto sensibile a osservazioni estreme ed in tali casi si distrugge la logica stessa del bootstrap.

quello di tutte le possibili permutazioni degli n elementi presenti nel campione originario (n!) anziché le n ottenibili tramite jackknife.

Sia $\vartheta = t(S) = t_n(y_1, y_2, ..., y_n)$ stimatore di θ funzione dell'intero campione estratto S. Nel caso di campioni di dati IID per costruire le repliche bootstrap, si considera S essere una nuova popolazione U* di riferimento e si estraggono R campioni con ripetizione di numerosità n da tale popolazione artificiale. Calcolando la statistica ϑ su ciascuna replica si così una successione di R realizzazioni $\vartheta_r^* = t_n(y_1^*, y_2^*, ..., y_n^*)$, dello stimatore bootstrap $\vartheta^* = t(S^*) = t_n(Y_1^*, Y_2^*, ..., Y_n^*)$. Per R elevato -non si considerano infatti solitamente tutte le n! possibili ripetizioni- tale stimatore fornisce buone approssimazioni per le misure sullo stimatore ϑ sul campione-padre, la realizzazione del quale sul campione S svolge per ϑ^* lo stesso ruolo assunto da θ in U. Ponendo allora:

$$\overline{\vartheta}^* = (1/R)\sum_r \vartheta_r^*$$

è possibile, come negli altri metodi di ricampionamento, stimare la varianza $V[\vartheta]$ di ϑ con

$$Var[\vartheta] = Var[\vartheta_{boot}] = Var[\vartheta^*] = (1/R-1)\sum_r (\vartheta_r^* - \overline{\vartheta}^*)^2 .^{15}$$

Per l'estensione del metodo al caso di campioni complessi un primo -ingenuo- modo di procedere consiste nella semplice applicazione del bootstrap al livello dell'estrazione dei cluster in ogni singolo strato h. Attuando tale estrazione con ripetizione in modo indipendente tra gli strati si ha che una espressione per la varianza dello stimatore ϑ è:

$$Var[\vartheta] = Var[\vartheta_{boot}] = \sum_h (1 - f_h) V[\vartheta_{boot,h}] .$$

Spesso tale stimatore non gode però di buone proprietà ed è, ad esempio, inconsistente per molte statistiche. Sarebbe possibile descrivere molte varianti del metodo che possono migliorare le proprietà degli stimatori bootstrap ma, come già accennato, esse risultano molto specifiche e adatte alla trattazione di casi particolari.

[15] Una stima della distorsione è invece Bias[ϑ]= $\overline{\vartheta}^*$ - ϑ .